Thèse
présentée par

Plínio Almeida Barbosa

pour obtenir le titre de
DOCTEUR DE L'INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE
(Spécialité : Signal-Image-Parole)

Caractérisation et génération automatique de la structuration rythmique du français

Écrire c'est tenter de savoir ce qu'on écrirait si on écrivait – on ne le sait qu'après – avant, c'est la question la plus dangereuse que l'on puisse se poser. Mais c'est la plus courante aussi. *Marguerite* DURAS, Écrire

La voix ne peut porter la langue et les lèvres qui lui donnent des ailes. Seule, elle affronte l'air. *Khalil* GIBRAN, Le Prophète

À Rose,

sine qua nihil esse posset.

Remerciements

Eu vim de infinitos caminhos,
e os meus sonhos choveram lúcido pranto
pelo chão.

Quando é que frutifica, nos caminhos infinitos,
essa vida, que era tão viva, tão fecunda,
porque vinha de um coração?

E os que vierem depois, pelos caminhos infinitos,
do pranto que caiu dos meus olhos passados,
que experiência, ou consolo, ou prêmio alcançarão?

Je suis venu d'infinis chemins,
et mes songes ont plu de lucides pleurs
sur le plancher.

Quand fructifie-t-elle, dans les chemins infinis,
cette vie, qui était si vivante, si féconde,
parce que venant d'un cœur ?

Et ceux qui viendront après, par les chemins infinis,
des pleurs qui sont tombés de mes yeux passés,
quelle expérience, ou consolation, ou prix conquerront-ils ?

Cecília MEIRELES, Herança

À Gérard Bailly, mon Directeur de Thèse. Depuis le début tu m'a appelé *partenaire* : dans nos échanges tu as toujours été à l'écoute de mes idées, sans manquer de me donner de nombreuses suggestions, qui m'ont guidé dans les méandres labyrinthiques de la prosodie (plusieurs fois tu as cru à mon travail plus que moi-même). Travailler avec toi m'a instruit et m'a stimulé par l'exemple de ton enthousiasme et le sérieux que tu mets dans tout ce que tu entreprends. Cette formation m'a beaucoup aidé dans mes rapports avec les chercheurs de différentes disciplines. Pour des gens comme toi, *merci* est un mot presque dérisoire. Les traces de ma plus profonde reconnaissance seront longtemps gravées dans ces pages : ceci me réjouis.

À Jean-Marc Dolmazon, Directeur de l'ICP. Je te remercie de m'avoir accueilli dans ce laboratoire exceptionnel avec ton sourire habituel. Si je me suis plu tellement ici, je le doit en partie à toi, par la qualité de ta direction. Un grand merci d'avoir approuvé les missions qui m'ont conduit à Bruxelles, à Bourges, à Paris et à Aix-en-Provence, cette formation m'a été précieuse. Un encore grand merci d'avoir voulu lire un travail fortement empreint de linguistique en acceptant de faire partie du jury.

À Jacqueline Vaissière, Rapporteur. Je vous remercie de votre intérêt si visible pour notre travail en l'enrichissant avec vos suggestions et vos observations si opportunes, fruit de votre grande expérience scientifique. Un grand merci d'avoir accepté de faire partie du jury et de m'avoir encouragé et félicité pour mon travail.

À Daniel Hirst, le plus anglais des nos chercheurs français. Je te remercie de ta disponibilité, de ta sympathie légendaire, d'avoir accepté de faire partie de ce jury dans la qualité de Rapporteur. En tant que tel, tu as su lire notre travail avec beaucoup d'intérêt et de sérieux. Ton intérêt remonte loin, puisque tu nous as donné une copie du précieux travail de Dogil et as essayé de valider sur tes propres corpus quelques-uns de nos résultats. À toi un grand et sincère merci.

À Michel Contini, le plus Sarde de nos chercheurs. Je vous remercie de votre gentillesse, de votre intérêt envers les langues (passion qu'apparemment, j'ose dire, nous partageons) et d'avoir accepté de faire partie de ce jury en enrichissant notre travail avec votre expérience linguistique.

À celle qui a quitté mère, frères, neveux, de vrais amis, travail, pays aimé et la possibilité d'entendre partout le son de notre belle langue pour vouloir traverser tout un océan pour me suivre, la gratitude devient presque une dévotion. Rose, tu m'as soutenu dans mes tristesses et tu ne t'es pas découragée pendant mes angoisses et énervements. Tu t'es intéressée à mon travail en l'écoutant avec attention : que de fois tu as voulu connaître ce que de nouveau je présentais dans un article. Tu es même restée cinq jours seule, quand j'étais à Bourges, et deux jours, quand j'étais à Aix-en-Provence (je sais combien cette solitude forcée t'a coûté). Sans toi ce travail n'aurait vraiment pas vu le jour. Rose, *Deus te abençoe para sempre !*

À Lídio et Cynthia, mes parents. Je vous remercie pour le don de la vie. Ce travail est aussi fruit de l'arbre né des semences de votre douleur : combien la *saudade* a pesé dans vos cœurs !

À Fábio, mon frère. Je te remercie de ton amour fraternel toujours si spontané, même s'il ne s'est concrétisé qu'en une seule lettre !

À mes beaux-parents, José Antônio et Adolphina, pour le don de la vie de Rose. Je vous remercie de m'avoir accueilli dans cette famille si sympathique : j'ai beaucoup reçu. Un merci spécial à mes petites nièces, Francine et Talita : *tio Primus (sic)* ne vous a pas oubliées !

À mes oncles et tantes d'Outre-Rhin, Cyanne et Roland, Clyzélia et Horst. Je vous remercie spécialement, mes tantes, par votre joie débordante et votre préoccupation constante.

À Rafael Laboissière, mon compatriote. Je te remercie d'avoir beaucoup fait pour que je vienne ici et surtout, d'avoir choisi mon Directeur de thèse. Merci aussi pour ton aide dans l'Informatique, particulièrement avec ce programme LaTeX.

À Rudolph Sock, le plus créole de nos chercheurs anglophones. Je te remercie pour les étiquettes posées au corpus *Bureau*, pour ton sourire contagieux et ta bonne humeur constante, pour la correction de l'anglais de l'article publié aux États-Unis et pour ta disponibilité : *Thanks a lot, Rudo !*

À Frédéric Bimbot, notre locuteur. Je te remercie de nous avoir prêté ta voix en acceptant pour cela de venir de Paris. Merci encore pour les fonctions d'émergence.

À nos auditeurs, pour nous avoir prêté leurs oreilles. Je vous remercie de votre patience, en jugeant avec attention des paires de phrases parfois si semblables ! Soyez individuellement remerciés : Jean-Yves Antoine, Laurent Aublet-Cuvelier, Gérard Bailly, Denis Beautemps, Christian Bulfone, Bertrand Caillaud, Olivier Delemar, Chantal Jacquet, Anne-Sophie Laréal, Hélène Lœvenbruck, Nino Medves, Yann Morlec, Mariette Parbelle, Jérôme Piard, Michel Piquemal, Emmanuel Reynier, Clément Rossignol, Jean-Luc Schwartz, Emmanuel Scoton, Stéphane Tourret, Laurent Varin, Marie-Laure Zanello, Jérôme Zeiliger.

À Jean-Luc Schwartz. Je te remercie de ta disponibilité et de ta lucidité scientifique. Tu as toujours du temps pour les autres, malgré tes innombrables engagements. Je suis convaincu que ta carrière musicale sera aussi grandiose que ta carrière scientifique, parce que tu mets ton âme en tout ce que tu fais.

À Véronique Aubergé. Je te remercie de l'intérêt continuel démontré envers mon travail, même si je ne t'ai pas mise fréquemment au courant. Merci de ta sympathie et de ton enthousiasme scientifique.

À Geneviève Caelen-Haumont. Je te remercie (permets-moi de te tutoyer) de ton sourire et de ton accueil chaleureux à mon travail, sachant toujours m'écouter, me renseigner, me fournir du matériel bibliographique, malgré tes nombreuses occupations.

À Nick Campbell. Je te remercie pour ton modèle de prédiction de la durée, qui nous a inspiré, pour ta sympathie, pour les exemplaires de ta thèse et de tes articles.

À Bernd Pompino-Marschall. Pour ta bonhomie, pour le rapport annuel de l'Institut für Phonetik und Sprachliche Kommunikation de l'Université de Munique avec ton article sur les *P-centers*.

À Piet Mertens. Je te remercie de ta sympathie belge, de ta disponibilité et de ton intérêt pour mon travail.

À Sophie Scott. Pour ton intérêt de vouloir échanger quelques idées sur le *P-center* par *e-mail* et de mettre ton travail à disposition avec tant de professionnalisme.

À tous les chercheurs avec qui j'ai pu discuter aux cours de mes années en France et qui m'ont beaucoup apporté.

À Nadine Bioud, Joëlle Roques, Nino Medves, Monique Revil-Signorat, Danuta Emonet, Christian Bulfone, Jean-Claude Caërou et Dalila, notre personnel administratif et technique. Je vous remercie de votre travail important et essentiel en donnant le support nécessaire au bon déroulement de mon travail.

À tout le personnel INPG qui a travaillé pour cette thèse. Je remercie de façon particulière le personnel du service Scolarité, spécialement M[me] Christiane Vacher, pour sa gentillesse exquise, et le personnel du service Reprographie de l'ENSERG.

À Pierre Escudier, notre futur Directeur. Je te remercie d'avoir toujours rendu nos pauses-café plus drôles et humaines.

À Jérôme Zeiliger, le plus biologique de nos Cogniticiens. Je te remercie de nos « causeries » amicales, toujours pleines de ton humanité si attentive aux autres. Tu me manqueras.

À Jean-Yves Antoine, le plus artiste de nos chercheurs. Je te remercie de ton amitié, d'avoir rendu mon séjour plus agréable, par nos conversations pendant lesquelles nous avons pu aller au-delà des frontières de la parole.

À Olivier Delemar. Ta disponibilité aux autres est une source éternelle d'admiration. Ta bonté, camouflée derrière ton esprit moqueur, me sera un souvenir inoubliable.

À Bertrand Caillaud et Michel Piquemal. Je vous remercie d'avoir tant de fois, malgré votre travail de thèse, avec votre gentillesse habituelle, répondu à mes questions en Informatique.

À Hélène Lœvenbruck. Je te remercie pour ton aide avec le programme *xrbp*.

À Jordi et Ester Robert-Ribes. Je vous remercie pour nos dîners amicaux et de nous avoir emmené pour voir les cieux catalans.

À Yann Morlec, qui prendra littéralement ma place. Je te remercie d'avoir rendu notre bureau plus agréable avec ta personne si sympathique et d'avoir pris notre flambeau par l'étude de la fréquence fondamentale.

Aux compagnons de route : Christophe Savariaux, Jordi Robert-Ribes, Jordi Camps i Boixadera, Claudine Cabrera, Olivier Delemar, Bertrand Caillaud, Jean-Yves Antoine, Yen Pham Thi et Nadine Ozkan, avec qui j'ai commencé mon travail de thèse : bonne chance dans vos soutenances, qui ne tarderont pas.

À tous les collègues du laboratoire et tous ceux qui ont passé par lui. Je vous remercie d'avoir rendu mon travail plus joyeux par la bonne ambiance de cet institut.

À la langue française. Je te remercie pour ta beauté aux multiples facettes et l'exemple de ton art dans tant de belles œuvres ! Merci surtout à toi, Baudelaire.

À la langue portugaise, chère langue maternelle. Je te remercie pour ta rare beauté, que j'ai appris à apprécier particulièrement, et pour l'art tendre et inimitable de ta littérature. Un merci surtout à vous, Cecília Meireles et Florbela Espanca.

Au Poète Éternel. Tu me fais le don continuel de Ta vie et m'as toujours cumulé d'espérance en rendant mes fardeaux plus légers : « Venez à moi et je vous soulagerai. » Combien cette phrase est vraie !

À tous ceux que d'une façon ou d'une autre ont pu contribuer au succès de ce travail, particulièrement à ceux que mon ingrate mémoire aurait pu oublier.

Avertissement

Cette thèse comporte un important travail bibliographique. Or, la plupart des articles que j'ai eu l'occasion de lire sont en anglais et c'est pourquoi j'ai été confronté à des termes utilisés fréquemment par les anglophones et pour lesquels il est difficile de donner une traduction précise.

Quelques-uns de ces termes n'ont pas été traduits dans un soucis de précision et afin d'éviter une part d'ambiguïté.

Le terme de *stress* est associé à des notions de force articulatoire et de sonie. Il n'a pas de correspondant direct en français, langue qui ne met pas en relief les mêmes paramètres que l'anglais . Le terme d'*accent* n'a pas été utilisé pour le traduire. Cependant, le terme de *stressed syllable*, en tant qu'il indique la mise en relief d'une syllabe dans la phrase, a été traduit par *syllabe accentuée*. Le mot *glide* est évoqué dans ce travail pour faire référence à un son complexe comprenant un changement de timbre en cours d'énonciation, comme dans le cas des diphtongues.

La dichotomie entre langues *stressed-timed* et *syllable-timed* n'est pas traduite systématiquement. Pour la deuxième notion, on a utilisé parfois le terme de *langue à organisation temporelle syllabique*. Les prédécesseurs de ces notions sont les rythmes du type *machine-gun* et *morse-code* dont les traductions littérales sont *mitraillette* et *code Morse*. Les termes de *trailer-timed* et *leader-timed* font référence à la position de l'accent dans un groupe accentuel. Le *trailer* est la *roulotte*, qui se place derrière, le *leader* est celui qui commande, qui se place devant. *Stress-based language* désigne une langue qui possède la notion de *stress*.

Les termes de *beat* et *tapping* sont aussi difficiles à traduire en peu de mots. Le premier est lié à une notion de frappe, mais faisant référence au repère lui-même et non à l'acte de frapper. Le deuxième fait référence à une séquence de frappes successives.

À *vowel onset* j'ai donné une demi-traduction, *onset vocalique*, car cette notion me semble plus associée à un point précis que celle d'*établissement*. Le mot *offset* n'a pas été traduit et se réfère aux notions de relâchement, finalisation. Il peut correspondre à la fin du voisement vocalique, par exemple (*vocalic voiced offset*). Le *cluster* est un ensemble de consonnes prononcées sans qu'il y ait une jointure entre elles, comme dans *bleu* ou dans *très*.

Quoique la traduction de *tampon* est proposée pour *buffer*, j'ai préféré maintenir le dernier, car il est le seul effectivement utilisé par les informaticiens français. Dans ce travail *click* est un petit bruit sec, un son pur de courte durée. Un *parser* est un algorithme qui parcourt la chaîne phrastique en détectant des événements singuliers et définis d'avance.

Les notions de *top-down* ou de *bottom-up* sont courantes pour faire référence, respectivement, à des modèles descendants et ascendants.

Le terme de $z-score$ a aussi été maintenu inchangé dû à son utilisation fréquente dans la littérature et l'absence, à notre connaissance, d'un équivalent français susceptible de le remplacer. Certains auteurs utilisent le terme de *durée normalisée*.

Il y a évidemment le terme directeur de notre thèse, le *Perceptual-center* (et ses dérivés, *groupe inter-Perceptual-center* et *intervalle inter-Perceptual-center*), pour lesquels je n'ai pas proposé de traduction.

Toutes les citations, en anglais ou en allemand, ont été traduites par mes soins et j'espère ne pas avoir rudement modifié la pensée originale des auteurs.

Quant au titre de *Sommaire*, donné à quelques sections, il a été utilisé dans son acception de *bref résumé, abrégé des idées principales*.

Introduction

> Cinq aveugles voulaient savoir à quoi ressemble un éléphant. Le cornac leur permit de toucher l'animal pour s'en faire une idée. « L'éléphant ressemble à un gros serpent » dit l'aveugle qui avait touché la trompe. « Non, il est semblable à un chasse-mouche » protesta celui qui tâtait l'oreille. « Allons donc, c'est un pilier ! » déclara celui qui tâtait la jambe. Celui qui tenait la queue affirma : « Pas du tout, c'est une corde ! » Et celui qui palpait une défense se mit à rire : « Êtes-vous sots ! Un éléphant, cela ressemble à un os ! »...

Contes populaires de l'Inde, *L'éléphant et les aveugles*,
traduction en russe, Moskva, 1956

Ce passage illustre bien, à notre avis, ce que le présent travail veut faire : donner des éléments pour que nous puissions mieux comprendre ce qu'est l'Éléphant rythmique.

Pour réaliser cette tâche, il semble pertinent de l'envisager sous une optique : le fait que la parole soit un **acte**.

Dans tout acte humain, un double mécanisme est en jeu. D'abord celui de l'action, renfermant l'ensemble des activités nécessaires à transformer l'intention de réaliser quelque chose en acte proprement dit. Cet ensemble comprend les activités cognitives, les commandes neuro-motrices et les déplacements des articulateurs pertinents. Le deuxième mécanisme est celui de la perception : tout acte est perçu, d'abord par l'acteur lui-même et, ensuite, par un deuxième personnage, l'individu percevant.

S'il s'agit de parole, l'acteur est un locuteur et, celui qui perçoit, un auditeur (celui-ci reçoit en fait un message simultanément acoustique et visuel). Les deux mécanismes, **produire** et **percevoir** sont, cependant, intrinsèquement liés.

La Perception comprise dans le cadre de la production

Ce lien est largement exploité par Viviani & Stucchi (1992), qui posent l'hypothèse que le processus de sélection perceptive est contraint ou guidé par des patrons moteurs, c'est-à-dire, par la connaissance, implicite et de procédure, que le système nerveux central a des mouvements qu'il est capable de produire. Les liens perception-action sont au centre du débat de la perception de parole (Shepard 1984) et, plus récemment, cette approche a trouvé un écho très dur chez les tenants de la perception motrice où les véritables objets de la perception sont les gestes articulatoires intentionnels (Liberman & Studdert-Kennedy 1978; Liberman & Mattingly 1985).

Cette connaissance a priori que l'on a du système de production est illustrée par l'expérience de Viviani & Stucchi, qui consistait à faire voir à des sujets le dessin dynamique d'un cercle par un point lumineux. Toutefois, le profil de vitesse de l'acte de dessiner ce cercle était le même que celui que l'on obtient en dessinant une ellipse (*ibid.*, p. 606). Le résultat étonnant est que les sujets perçoivent une ellipse en train de se dessiner et non un cercle. La connaissance implicite du mécanisme de production a fait basculer le conflit entre la cinématique du stimulus et sa géométrie vers l'information donnée par le mouvement :

> ... notre pouvoir de discrimination prend sa base dans la connaissance implicite du mécanisme de fonctionnement du système moteur.
>
> ibid.*, p. 621*

Nous ne voyons aucune raison de penser que ce lien entre production et perception ne soit pas aussi valable pour la parole. En ce qui concerne la coarticulation (un phonème qui influe sur un autre), ce lien est prôné par Whalen (1990, p. 33) :

> La coarticulation est réalisée dans le but de rendre l'acte d'énonciation plus simple et/ou réhausser l'efficacité du signal de parole pour l'auditeur. Quand les phonèmes s'influencent les uns les autres, la cohésion du son est augmentée.

L'Organisation gestuelle des articulateurs

Selon Whalen, la coarticulation ne se construit pas comme la conséquence d'un enchaînement des divers articulateurs lors de l'acte de phonation mais elle est planifiée d'avance.

Cette planification est illustrée par une série d'expériences réalisées avec des occlusives bilabiales dans des contextes vocaliques. Quand les consonnes à prononcer sont connues avant le début de l'énonciation, un phénomène de compensation s'opère entre les durées de la voyelle et de la consonne : lorsque la voyelle est suivie de /p/, sa durée est plus courte que lorsque elle est suivie de /b/ — le rapport entre les durées des consonnes est inverse. Un tel effet ne se vérifie pas si les consonnes ne sont connues qu'une fois l'énonciation commencée.

L'organisation temporelle des voyelles et consonnes au cours de l'énonciation est une conséquence de l'orchestration d'un ensemble de gestes articulatoires. Edwards et collègues (1991) suggèrent qui il y a deux stratégies articulatoires différentes qui peuvent influer sur la durée d'une syllabe :

- **la raideur** propre aux gestes d'ouverture et de fermeture. Cette spécification est d'ordre intragestuel ;
- **l'organisation temporelle intergestuelle**, c'est-à-dire, l'organisation temporelle relative entre deux gestes consécutifs.

La première stratégie a été liée, d'après les résultats des expériences entreprises par les auteurs, à la variation du débit et, la deuxième stratégie, au mécanisme d'accentuation. Les auteurs concluent que « cette forte évidence pour deux stratégies articulatoires différentes soutient l'idée que la durée seule n'est pas une mesure très informative sur les phénomènes liés à la durée en parole (Edwards & Beckman 1988). » (*ibid.*, p. 380) Cet aspect est illustré par une étude récente (Lœvenbruck & Perrier 1993) sur la réduction vocalique, où les auteurs proposent que le contrôle prosodique des signaux de parole peut se faire à partir de deux paramètres, la durée de la commande motrice et la raideur des articulateurs concernés dans la production. Ces éléments deviennent prépondérants lorsque le débit augmente. Les modèles suggèrent donc que le geste articulatoire est contrôlé par deux mécanismes agissant sur le même système dynamique : un contrôle des cibles (points d'équilibre) et, en parallèle, un contrôle prosodique qui agirait sur la durée de l'acte et sa précision. Il convient maintenant de préciser un peu plus le sens du qualificatif *prosodique*, par le biais du nom qui l'a généré : *prosodie*.

La Prosodie vue sous son aspect articulatoire

Une étude récente (de Jong, Beckman & Edwards 1993) propose une définition de la prosodie comme étant

> the organizational framework that measures off speech into countable units within which various items are highlighted.

> *le cadre organisationnel qui segmente le signal de parole en unités entières dans lesquelles plusieurs éléments sont mis en relief.*

La prosodie est donc comprise comme mécanisme organisateur associé aux gestes articulatoires produits à but de communication. Avant de l'étudier sous un autre angle (chapitre suivant), on constate que l'une de ses composantes, le rythme, participe doublement à l'organisation temporelle des gestes articulatoires.

Le rythme, plus que toute autre composante de la parole (Fraisse 1974; Allen 1975), est fortement contraint par les négociations entre production et perception. En effet, le rythme se structure à plusieurs niveaux : d'abord au niveau **microrythmique**, par l'organisation temporelle des articulateurs responsables de la production des sons de parole : il doit permettre d'organiser des gestes articulés suffisamment clairs pour rendre possible le décodage des sons (Lindblom 1990). Ensuite, aux niveaux **macrorythmiques**, comprenant des unités de taille égale et supérieure à celle de la syllabe, où il est un véhicule essentiel de la structure accentuelle.

La référence abstraite du rythme a été appelée *régulateur* par Wenk & Wioland (1982, p. 204). Elle se charge de la manifestation physique des niveaux rythmiques cités ci-dessus. Il est clair qu'au niveau microrythmique le régulateur peut avoir une fonction linguistique plus précise dans le cas où le système phonologique exploite des traits distinctifs de longueur (voyelles longues et brèves).

La composante macrorythmique doit aussi être planifiée d'avance, par un compromis entre les mécanismes de production et perception, tout en servant un but de communication, comme il est suggéré par Whalen (*ibid.*, p. 33) pour la coarticulation : « les locuteurs structurent ainsi leurs énoncés vraisemblablement parce que ceci sert à un but de communication. »

Les Hypothèses de travail

Les Repères biologiques du rythme

Pour que cette structuration complexe ait lieu, des repères temporels internes sont nécessaires. Ces horloges, ou cette **horloge interne** (Leiner, Leiner & Dow 1991), seraient le point d'ancrage pour la production et perception du niveau rythmique de la parole. La production du rythme se réalise ainsi sous le pilotage d'un *pacemaker* (cadenceur). Toute déviation par rapport à ce cycle quasi-périodique serait perçue comme porteuse d'information.

L'auditeur, qui connaît implicitement la nature de ce pilotage, doit avoir des points de repère de cette rythmicité.

Les Repères perceptifs du rythme

Nous avons exploité le paradigme du *Perceptual-center*, notion développée par Marcus et collègues (Marcus 1976; Morton, Marcus & Frankish 1976), comme étant ce point de repère pour la perception de l'**isochronisme**, c'est-à-dire, la tendance à percevoir les repères comme intervenant à des intervalles de temps réguliers. Quoique son inventeur ait nié la possibilité d'une localisation absolue de ce repère sur le signal acoustique (cf. section 3.4), Pompino-Marschall (cf. section 3.4.2) a rendu cette notion opérationnelle en proposant un modèle psychoacoustique qui calcule une position absolue pour le *P-center*.

Les résultats de l'analyse de corpus de parole naturelle à la lumière de ces paradigmes méthodologiques ont été exploités dans notre travail. Dans le cadre de la prédiction automatique de la durée, nous proposons un module de génération qui tient compte de la composante macrorythmique de la parole. Le paradigme de l'apprentissage automatique, permettant la mémorisation, à partir d'exemples d'élocution, des patrons accentuels propres à la structure rythmique de la langue étudiée, sera présenté.

Ce qui sera abordé dans notre travail

Dans le premier chapitre, nous présenterons le cadre général dans lequel se trouve inséré ce travail, à savoir, celui de la caractérisation et de la génération automatique de la prosodie.

Le chapitre 2 sera dédié à la présentation des modèles de prédiction de la durée couramment utilisés dans la littérature. Nous examinerons les divers choix adoptés, selon que l'unité choisie pour la génération est le phonème ou des unités de niveau supérieur.

Le chapitre 3 développe le thème de la composante rythmique sous les aspects de production et de perception. La présentation du débat sur l'isochronisme sera reliée à la notion de *Perceptual-center*, ce qui nous permettra de proposer une nouvelle unité de programmation rythmique à la place de la syllabe, le GIPC.

Les raisons par lesquelles le GIPC est choisi pour l'obtention des durées au niveau microrythmique et pour caractériser la structuration macrorythmique de la phrase sont présentées au chapitre 4. Un test de perception permet de confirmer les liens entre les mécanismes de production et de perception pour le rythme.

Le modèle combiné de génération de la durée est présenté au chapitre 5. Ce modèle intègre l'émergence de la pause au plan de la substance de l'expression (cf. premier chapitre) selon l'hypothèse d'une horloge interne, point d'ancrage de la production rythmique.

Les conclusions et perspectives permettront de faire un bilan de ce travail tout en proposant les développements naturels qui pourront suivre.

Chapitre 1

La Prosodie

> En apprenant la prosodie d'une langue, on entre plus intimement dans l'esprit de la nation qui la parle que par quelque genre d'étude que ce puisse être.
>
> *M^me de* STAËL, *De l'Allemagne*

Le linguiste danois Hjelmslev (1968) a essayé de clarifier les distinctions saussuriennes entre forme et substance, expression et contenu. Dans un schéma fonctionnel de la langue il affecte le domaine de l'**expression** aux sons que la langue choisit pour transmettre la signification, et celui du **contenu** à la façon dont la langue présente la signification (Ducrot & Todorov 1972, p. 36). Chacun de ces domaines possède une **forme** et une **substance**.

La forme du contenu est du ressort de la morphologie et de la syntaxe tandis que sa substance est du ressort de la pragmatique et de la sémantique. La forme de l'expression est décrite par la phonologie, tandis que sa substance, formée par les sons de parole, est étudiée par la phonétique (Di Cristo 1981). On divise classiquement la phonologie en **phonématique**, traitant de l'étude des unités discrètes et distinctives représentées par les phonèmes, et en **prosodie**.

Le mot *prosodie* est originaire du grec $\pi\rho o\sigma\omega\delta \acute{\iota}\alpha$, qui signifie « accent, quantité, dans la prononciation ». À l'origine employé dans la poésie, où il est identifié aux caractéristiques quantitatives et mélodiques en versification, le terme de prosodie a pris un sens spécialisé dans la linguistique moderne (Ducrot & Todorov 1972, p. 228).

Martinet (1991, p. 83) définit la prosodie de façon négative[1] comme recouvrant

> tous les faits de parole qui n'entrent pas dans le cadre phonématique, c'est-à-dire ceux qui échappent, d'une façon ou d'une autre, à la deuxième articulation[2].

1. Le rôle de la prosodie chez Martinet est marginal du point de vue linguistique en n'ayant qu'une fonction de suppléance à l'enchaînement phonématique (*ibid.*, p. 101)

2. La parole est caractérisée classiquement par une double articulation. La première articulation con-

Les caractères prosodiques sont susceptibles de varier d'une façon continue et sont associés aux facteurs les plus divers : d'abord linguistiques, comme l'accentuation, l'intonation (terme abordé plus loin) et le rythme et ensuite, para- et extralinguistiques, comme l'état d'esprit du locuteur, les caractéristiques régionales, l'âge et le sexe (Di Cristo 1985). Ces caractéristiques sont étudiées par le biais d'un ensemble de paramètres prosodiques mesurables.

1.1 La Substance acoustique de la prosodie

> La métrique grecque et latine reposait sur l'étude de la durée, de la hauteur et éventuellement de l'intensité, qui font aujourd'hui l'objet des études linguistiques de prosodie.
>
> *(Ducrot & Todorov 1972, p. 228)*

Il semble ainsi qu'une tradition s'est établie autour de l'étude de trois paramètres principaux, liés à la production de la parole (Calliope 1989) :

- **la fréquence fondamentale**, corrélat acoustique de la fréquence laryngienne, fréquence de vibration des cordes vocales. Celle-ci est essentiellement la conséquence des variations de la pression trans-glottique et de la tension des cordes vocales ;
- **la durée**, en tant que mesure du temps, est inhérente à tout processus physique. Elle sera étudiée avec plus de détail dans ce chapitre ;
- **l'intensité**, corrélat acoustique de la pression sous-glottique et de l'ouverture du conduit vocal. Ce paramètre est moins étudié, peut-être dû à sa forte dépendance de la nature intrinsèque des sons.

Nous avons ébauché précédemment le lien étroit qui existe entre production et perception de parole. L'ensemble des phénomènes prosodiques perçus est étudié classiquement sous les formes de **hauteur**, de **durée perçue** et de **sonie**.

Il faut éviter de courir le risque d'associer respectivement les paramètres physiques décrits plus haut aux corrélats perceptifs. Dans toute perception de la hauteur, la durée et l'intensité jouent un rôle qui ne doit pas être négligé (Botte, Canévet, Demany & Sorin 1989) ; dans la perception de la durée, la fréquence fondamentale[3] et l'intensité collaborent (Fraisse 1974) ; on sait que la sonie d'un son pur varie en fonction de la

cerne le choix des unités significatives minimales, les **monèmes**, et, la deuxième, concerne le choix des unités distinctives minimales, les **phonèmes**.

3. Une syllabe est perçue plus courte si elle possède un ton statique (Lehiste 1978).

fréquence (Fletcher & Munson 1933; Botte, Canévet, Demany & Sorin 1989) et de la durée (Munson 1947).

Les notions d'**intonation** et de **rythme** sont associées respectivement aux variations à long terme de hauteur et de durée perçue. Il est clair que, selon ce qui a été discuté précédemment (cf. Introduction) pour la prosodie, des caractéristiques d'organisation de la substance mélodique ou rythmique entrent en jeu dans ces deux notions.

Les paramètres étudiés plus haut décrivent la phénoménologie prosodique au niveau de sa substance. Ceci ne constitue qu'une face de la monnaie d'échange d'information entre locuteur et auditeur.

1.2 La Prosodie en tant que signe

> Interlanguage resemblances of sound and meaning are so far-reaching and so persistent in the face of the relative arbitrariness and unintelligibility of the words and structures of a language that one does not "know", that there must be a common base somewhere. The present volume carries on with this assumption and undertakes to show that even when it interacts with such highly conventionalized areas as morphology and syntax, intonation manages to do what it does by continuing to be what it is, primarily a symptom of how we fell about what we say, or how we feel *when* we say.
>
> *Les ressemblances de* son *et de* sens *entre les langues vont si loin et sont si persistentes face à l'arbitraire et à la non-intelligibilité des mots et des structures d'une langue que l'on ne connaît pas qu'il doit avoir une base commune quelque part. Le présent volume poursuit cette hypothèse et s'engage à montrer que, même quand elle interagit avec des domaines hautement conventionalisés comme la morphologie et la syntaxe,* l'intonation parvient à faire ce qu'elle fait tout en continuant à être ce qu'elle est, *essentiellement un indice de ce qui on pense de ce que nous disons ou comment on se sent quand nous disons*[4]. *Dwight* BOLINGER *(1989), p. 1*

Ce passage introduit l'ouvrage de Bolinger sur l'intonation et ses usages.

4. Les passages soulignées dans notre traduction sont réalisées par nos soins. Ce qui a été souligné par l'auteur est maintenu dans le texte original.

Bolinger étudie l'intonation dans son aspect de substance de l'expression, c'est-à-dire, comprenant les variations de hauteur attachées à des unités au niveau phrastique (Ducrot & Todorov 1972) ou les variations à long terme de la fréquence fondamentale (Calliope 1989)[5]. Pour Bolinger, elle interagit avec la composante linguistique sans se dénaturer, c'est-à-dire, sans cesser de véhiculer des informations extralinguistiques sur le locuteur (âge, sexe, niveau social, état d'âme) et sur la situation du discours. Pour Bolinger, l'intonation et des domaines comme la morphologie et la syntaxe sont indépendants mais interagissent au niveau pragmatique (*ibid.*, p. 67).

Pour Rossi et collègues (1981, p. 14), l'intonation est un signe linguistique qui associe une forme de l'expression à une forme ou substance du contenu. Ils la conçoivent comme

> une forme **discontinue**, constituée d'unités discrètes, coextensives aux phonèmes et commutables en un même point de la chaîne, ces unités, toujours significatives s'organisant pour former des contours ou syntagmes intonatifs dont le domaine est la phrase ou ses constituants (*ibid.*, p. 14).

Les caractères *discret et coextensif au phonème*, qui lui sont attribués ci-dessus, contredisent la définition de Martinet ou de celle du terme *suprasegmental*, créé par Hockett (1942) et critiqué par Rossi et collègues (1981, p. 9-15).

Bolinger a évoqué plus haut et a étudié l'intonation dans son double aspect de son et de sens ou, en d'autres termes, de **signifiant** et de **signifié**. Pour cela, l'intonation peut être liée à la notion de **signe**. Nous pouvons appliquer ses réflexions sur l'intonation au domaine plus large de la prosodie sans prendre le risque de compromettre l'une ou l'autre.

« Le signe, notion de base de toute science du langage, est en raison de cette importance l'une de plus difficiles à définir », disent Ducrot & Todorov (1972, p. 131). Ils proposent cependant (*ibid.*, p. 132) que le signe est une unité qui (1) peut devenir sensible, et (2) pour un groupe défini d'usagers, marque un manque en elle-même. La première notion s'appelle signifiant et, la deuxième, signifié. La relation qu'ils entretiennent entre eux s'appelle la **signification**.

Le signifiant est la face qui peut éventuellement être perçue par un individu appartenant à la communauté où le signe a été institué. Il est par exemple la représentation graphique d'un mot ou son image acoustique (ou acoustico-visuelle). Le signifié, défini comme manque ou absence, n'existe pas sans le signifiant et a son existence liée à un

5. Cet aspect qui ne concerne que les variations de la hauteur ou de la fréquence fondamentale, sera appelé le sens restreint, par opposition au sens large où intonation se confond avec prosodie.

système de relations ; il a été traditionnellement associé à la notion de concept et est un objet d'étude de la sémantique[6].

Martinet attribue la double face introduite par la notion de signe à la prosodie :

> Un trait prosodique, comme la montée de la courbe mélodique qui fait d'*il pleut ?* une question, combine un signifiant, la montée de la courbe, et un signifié, celui qu'on reconnaît en français au monème *est-ce que*. Ces signes jouent, dans les communications humaines, un rôle qui n'est pas négligeable.
>
> *André* MARTINET *(1991, p. 101)*

La phrase interrogative *il pleut ?* et l'assertive *il pleut.* ne se différencient que grâce à la prosodie. Son rôle dans l'intelligibilité du discours n'est pas sans importance. Bolinger relate que dans une étude de prose et poésie, présentée à deux auditeurs de deux façons, une avec intonation normale et l'autre, monotone, il a été trouvé que la monotonicité produit une perte d'intelligibilité (Glasgow 1952). Dans la parole synthétique, l'intonation représente aussi une partie importante dans l'intelligibilité globale et dans le naturel du signal synthétisé (Olive & Nakatani 1974; Nooteboom, Brokx & de Rooi 1976; Benoît 1990).

En reliant ces propos à la notion d'acte discutée dans l'introduction, on peut dire que la prosodie associe un mouvement, le signifiant, à un but de communication, dont la signification peut être mise en relation avec un signifié.

Notre propos ne consiste pas à bâtir une théorie de la signification des mouvements prosodiques. Comme nous le verrons par la suite (cf. section 1.4), la nature et la fonction des objets prosodiques de la perception sont fort discutées. Peu d'expériences de perception catégorielle (Kohler 1987; Kohler 1991) ont été entreprises de manière à faire émerger, comme il a été démontré en phonématique (Vallée 1994; Lindblom 1986), l'ensemble des unités minimales de la composante prosodique. De plus, le rapport de ces éventuelles unités à la structuration linguistique sont eux aussi fortement discutés. Notre propos est, au contraire, d'offrir le moyen d'apprendre automatiquement les correspondances entre le signifiant prosodique (dans notre cas signifiant macrorythmique) et un certain nombre d'éléments de signification linguistique, à l'intérieur d'un système de synthèse.

Les choix des unités qui supportent la signification – signifiant et signifié – utilisées dans le cadre de cette thèse seront explicités : la structuration du signifiant, que cette

6. On assiste actuellement à un basculement dans la théorie du signe, représenté par ce que l'on appelle la primauté du signifiant, en indiquant que quelque chose fonctionne comme signifiant jusque dans le signifié. Pour cela, voir Derrida (1968) et, Rifaterre (1979). Ce dernier exploite la construction de la signifiance d'un texte à partir du réseau de relations déclenché par la disposition des signifiants.

thèse cherche à démontrer, et le signifié, représenté par les marques relationnelles (cf. section 1.4.1 et annexe C).

1.3 Le Système de synthèse de la parole

La tâche d'un modèle de génération n'est pas mince. Il réalise un passage au niveau de l'expression, de la forme vers sa substance. Ici se situe toute la difficulté, car la forme appartient au monde mental, donc symbolique, et celui de la substance, au monde physique.

Ce passage crucial a été appelé **le fossé de Smolensky**, en hommage à celui qui continue à réaliser sa recherche dans le domaine du connexionisme — dont cette thèse est une application concrète —, qu'il propose comme la composante sub-symbolique des opérations cognitives (Smolensky 1992).

À partir d'une description symbolique, éventuellement paramétrique, de la tâche linguistique, issue d'une série de traitements symboliques appliqués au texte, le modèle de génération permet de générer les paramètres prosodiques permettant de faciliter l'« écoute» de cette tâche linguistique.

Un système de synthèse (figure 1.1) est le cadre dans lequel ce genre d'opérations sera effectué.

À partir du texte écrit, un ensemble de traitements symboliques est réalisé, de façon à transformer la chaîne graphémique dans une chaîne phonétique. Ces traitements incluent la transcription orthographique-phonétique, la syllabation et l'analyse morpho-syntaxique. Cette étape est appelée aussi de **transfert** (cf. annexe C).

Le texte phonétique étiqueté doit être suffisamment complet pour permettre les générations segmentale (la substance phonétique dépourvue de son actualisation prosodique) et prosodique (l'actualisation de paramètres prosodiques présents dans la substance phonétique selon le nouveau texte que l'on synthétise). Le signal acoustique est obtenu avec le concours de ces deux processus de génération. Une **méthode de synthèse** doit être choisie pour effectuer cette opération. Un ensemble de paramètres de contrôle est ensuite utilisé par une **technique de synthèse** appropriée qui, par l'intermède d'un **synthétiseur de parole**, délivre le signal de parole synthétique. Les méthodes et techniques de synthèse les plus connues sont présentées dans l'annexe D.

La génération automatique de la prosodie peut se valoir des mêmes méthodes utilisées pour la génération segmentale. Dans ce cadre, plusieurs travaux ont essayé de « capturer» le phénomène prosodique en vue d'un traitement automatique ultérieur. La recherche s'est concentrée surtout dans le domaine de l'intonation.

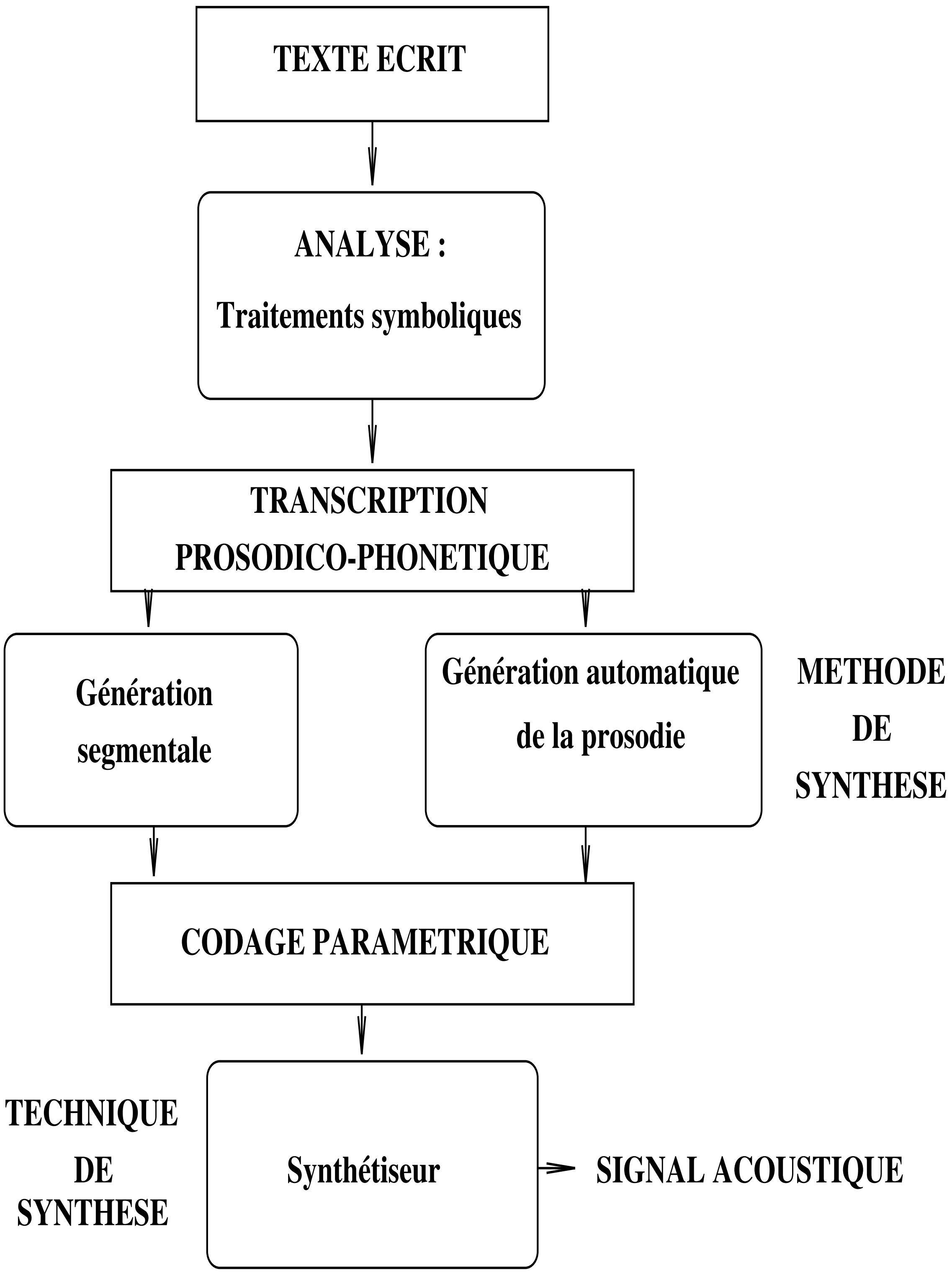

FIG. 1.1 - *Schéma d'un système de synthèse à partir du texte.*

1.4 Caractérisation et génération de l'intonation

Ladd (1988) propose de diviser les études axées sur la description de la fréquence fondamentale (F0) en deux traditions. Ces deux façons de caractériser l'intonation constituent deux points de vue à part entière d'analyse du phénomène intonatif. Ils trouvent dans les cadres fournis par les méthodes de synthèse, d'excellents moyens pour valider leur voie de recherche.

1.4.1 Les Formes, contours ou prototypes

Cette tradition modélise la F0 par l'imposition de formes ou de pentes couvrant des composants phrastiques de taille variable, dans le but de générer la complexité des contours réels de F0. Il s'agit d'une recherche à fort a priori de la façon par laquelle les informations linguistiques sont véhiculées.

Cette approche peut être reliée à la notion de forme, telle comme elle est vue par la *Gestalttheorie* (cf. section 4.4.1).

Les travaux qui décrivent la F0 de cette façon sont représentés, entre autres, par le modèle de Fujisaki et collègues (Fujisaki & Sudo 1972; Fujisaki & Kawai 1988) et ceux de Cooper & Sorensen (1981), Gårding (1983) et Vaissière (1983). Ces auteurs décrivent d'abord une forme globale de la courbe de fréquence fondamentale dont l'extension est celle d'un syntagme ou de la phrase[7]. Ensuite, des mouvements locaux de F0 représentent les variations aux frontières prosodiques ou au voisinage des syllabes accentuées.

Aubergé (1991) exploite la notion de **rendez-vous**: l'intonation serait une structure indépendante de la syntaxe qui présente des points de rendez-vous de synchronisation avec la structure syntaxique pour réaliser un but de communication.

La notion de *rendez-vous* est plus ou moins présente dans ce que nous pouvons appeler l'École française, représentée par la proposition d'un ensemble de marques syntactico-prosodiques ou marques relationnelles, issues de l'analyse linguistique de la phrase (cf. annexe C). C'est au voisinage d'une telle marque que les niveaux intonatif et linguistique se synchroniseraient. À cette École nous pouvons associer Emerard (1977), Caelen-Haumont (1978), Martin (1982), Di Cristo (1985), Bailly (1989), Aubergé (1991) ou Vaissière (1971), pour ce qui concerne la génération des pauses.

Les formes délimitées dans le continuum intonatif peuvent être stockées et on comprend pourquoi cette approche se prête bien à une méthode de synthèse par unités

7. La diminution progressive de la F0 attestée dans les phrases assertives est connue par le nom de **ligne de déclinaison**

stockées[8] (cf. section D.1.2). Dans l'utilisation de l'approche *contours* avec ce type de méthode de synthèse, seul fait exception dans l'École française, Vaissière (1971), qui a contribué à la synthèse par règles pour ce qui est de la mélodie.

La deuxième tradition, au contraire, cherche à faire émerger du continuum acoustique, des symboles puis, cherche à aligner ces symboles avec des structures linguistiques.

1.4.2 Le Point de vue phonologique

Les tenants de cette ligne de recherche (Crystal 1969; O'Connor & Arnold 1973; Ladd 1983; Liberman & Pierrehumbert 1984) supposent qu'une ligne mélodique peut être décrite par un enchaînement d'unités phonologiques. Dans beaucoup de modèles développés, il y a une spécification séquentielle de symboles (les tons) associée aux niveaux intonatifs au long de la phrase : les transitions entre symboles n'ont aucune importance perceptive, seules comptent les cibles.

Chez Pierrehumbert (1980, 1981), par exemple, l'évolution des contours de fréquence fondamentale est générée par des valeurs cibles qui sont reliées par des fonctions de transition (du type parabolique, par exemple). Ces valeurs-cible peuvent appartenir à deux catégories, *haut* (H : *high*) et *bas* (L : *low*). On voit bien l'aspect phonologiquement dirigé de ce modèle. D'autres représentants de cette école sont Hirst (1977) et Witten (1977). Les symboles utilisés peuvent cependant décrire un véritable mouvement, comme les termes *rise* (montée) et *fall* (descente) adoptés par O'Connor & Arnold (1973), ou les mouvements codés par une suite de deux symboles, chez Mertens (1993).

Dans le but de ne considérer que la partie de la substance prosodique qui soit réaliste du point de vue de la perception, certains travaux proposent la stylisation de contours élémentaires à partir de tests de pertinence perceptive (Willems, Collier & Hart 1988; Cohen, Collier & Hart 1982; Beaugendre & Lacheret-Dujour 1993), pour étudier a posteriori les relations que ces contours stylisés entretiennent avec les niveaux linguistiques. La difficulté essentielle du travail de Beaugendre & Lacheret-Dujour réside dans le passage des mouvements issus de l'analyse interactive qu'ils obtiennent à partir de la stylisation – « égalité perceptive» – à des mouvements dits standards, qui agissent comme de véritables signes, par « équivalence perceptive» , où l'expert joue un rôle déterminant. Seule une intuition géniale permet de savoir si cette dernière simplification conserve l'intégralité des signifiés linguistiques et paralinguistiques.

8. Aubergé (1991) présente une méthode de génération de l'intonation par concaténation de contours préalablement stockés selon des critères linguistiques. Les formes stockées par Aubergé sont issues d'une analyse statistique guidée par des clés linguistiques du type nombre de syllabes du constituant et fonction grammaticale. Ces contours fonctionnent comme de véritables prototypes d'une classe.

L'approche phonologique, en proposant un système de symboles, se prête bien à être intégrée dans un système de synthèses par règles (cf. section D.1.1).

L'apprentissage automatique résoud le conflit entre ces deux traditions en proposant un système qui permet d'apprendre les relations entre la structure phonétique du signifiant avec celle du signifié.

1.4.3 L'Apprentissage automatique : un associateur

Cette méthode de synthèse (cf. section D.1.3), qui exploite la capacité de généralisation des réseaux neuromimétiques (cf. annexe E et section 5.5), est représentée, par exemple, par les travaux de Scordilis & Gowdy (1989), Sagisaka (1990) et Traber (1992).

Les réseaux neuromimétiques permettent d'associer description linguistique et substance prosodique. Si un sous-ensemble de la caractérisation linguistique est non pertinent, les poids [9] qui lui correspondent s'affaiblissent. Si la caractérisation est insuffisante, les prédictions seront mauvaises. Une solution possible serait l'élaboration préalable d'une analyse de correspondances ou de facteurs pertinents pour le phénomène étudié (cf. section 2.1.5).

Dans la façon par laquelle le niveau linguistique est représenté à l'entrée du réseau, on peut rattacher ces systèmes à la première ou à la deuxième tradition décrites dans cette section. Le travail de Traber (cf. section 5.5.4), par exemple, se sert principalement d'une description du niveau accentuel du mot, de la nature des phonèmes de chaque syllabe et de la valeur intrinsèque de la fréquence fondamentale (*high* ou *low*) associée à la syllabe. Pour cela on peut rattacher son approche de la tradition phonologique (cf. van Santen 1994 et section 1.4.2).

Nous avons conçu notre étude dans le cadre de la synthèse par systèmes dynamiques. Notre travail propose un modèle de génération pour la durée en utilisant une méthodologie d'analyse ascendante sur des corpora de parole naturelle pour faire émerger des contours. Dans ce système, le modèle de transfert basé sur le modèle de marquage de Bailly (cf. annexe C), a servi d'entrée de référence, dans une démarche descendante. La prédiction de la durée est réalisée avec le souci de conserver l'information macrorythmique. La description linguistique à l'entrée du réseau neuromimétique est dépouillée : les marques relationnelles sont accentuables, certes, mais n'imposent pas, de prime abord, un type d'accent particulier car les marques sont effaçables. Les marques fonctionnent comme

9. Ces poids représentent la force ou le degré de pertinence d'un facteur linguistique dans la manifestation du phènomène présenté en sortie du réseau

un rendez-vous possible entre les niveaux linguistique et macrorythmique. De plus, la structure macrorythmique a été étudiée dans le cadre de la notion de forme globale et par là, nous nous rattachons à la tradition gestaltiste (section 1.4.1).

Il a été dit précédemment que les paramètres prosodiques classiques jouent un rôle prépondérant dans la construction du rythme, où *les facteurs temporels sont premiers* (Fraisse 1974, p. 84). Le temps, composant de tout processus physique doit être mesuré.

1.5 La Durée

> Sans une mémoire élémentaire qui relie les deux instants l'un à l'autre, il n'y aura que l'un ou l'autre des deux, un instant unique par conséquent, pas d'avant et d'après, pas de succession, pas de temps.
>
> *Henri* BERGSON*, Durée et simultanéité*

Des trois paramètres classiques utilisés dans la recherche en prosodie, la durée est le plus difficile à préciser, car elle n'est pas directement associable à un corrélat biologique du système phonatoire. La fréquence fondamentale est le corrélat acoustique de la fréquence laryngienne, due à la mise en vibration des cordes vocales. L'intensité est liée à la pression sous-glottique qui s'exerce par l'air qui vient des poumons sur le larynx et à l'ouverture du conduit vocal contrôlée par les lèvres et le velum. Ces deux paramètres évoluent continûment lors de l'élocution de chaque individu ainsi que le temps mais, pour que l'on puisse parler de durée, il faut fixer au moins deux instants, liés à des **événements singuliers**, entre lesquels le temps s'écoule.

Le temps se fait mesurable quand il se spatialise en laissant la trace d'un mouvement. C'est seulement quand ce mouvement existe, lequel peut être divisé, décomposé en unités plus petites, que l'on peut alors parler de *durée*. Mais on ne conçoit pas celle-ci sans être directement associée à une quelconque activité cognitive de la part de l'individu percevant. On ratifie ainsi ce que Bergson (1968) avait affirmé à ce propos :

> (...) on ne peut concevoir un temps sans se le représenter perçu et vécu. Durée implique donc conscience ; et nous mettons de la conscience au fond des choses par cela même que nous leur attribuons un temps qui dure.

On parlera de **durée observée**, quand il s'agira de la mesure objective du temps, exigeant pour cela d'être exprimée en secondes ou dans une autre unité pertinente. Cette durée observée est associée au mécanisme de production de la parole. En ce qui concerne la composante perceptive, il sera toujours question de **durée perçue**, au sens expliqué ci-dessus.

Même quand il s'agit de durée observée, la tâche de mesurer n'est pas pour autant simplifiée puisque il faut choisir une paire d'événements représentative du phénomène observé dans une constellation d'événements détectables dans le signal de parole.

1.6 La Segmentation du signal de parole

L'étude exhaustive d'Abry et collègues (1985) retient une dizaine d'événements sur le signal acoustique susceptibles de fonctionner comme frontières pour la délimitation de la durée. Quelques-uns de ces événements sont le début et la fin de voisement, début et fin de voisement vocalique, début et fin de frication consonantique. L'adoption d'une durée comprendra donc un choix d'événements pertinents.

Mais l'oscillogramme du signal de parole ne fournit qu'une partie de l'information nécessaire à une segmentation du signal de bonne qualité et susceptible d'être automatisée. Le domaine fréquentiel fournit la complémentarité indispensable à la localisation des frontières acoustiques. Une part d'incertitude est cependant toujours présente dû aux phénomènes de coarticulation — l'influence d'un phonème sur un autre — discutés précédemment. Même la segmentation manuelle a sa part d'incertitude due à des décisions hautement subjectives et à la fatigue inhérente à ce processus qui consomme beaucoup de temps (Leung & Zue 1984). La segmentation automatique diminue le temps d'un tel processus en éliminant de surcroît les différences inter- et intrasujet d'une segmentation manuelle. La segmentation proposée par Leung & Zue a des erreurs inférieures à 10 ms en comparaison à une segmentation manuelle soignée. Des erreurs de quelques dizaines de millisecondes dans un modèle de prédiction de durée se trouvent dans la plage des erreurs humaines de segmentation. Cependant ces erreurs, quoique de faible valeur, peuvent être gênants si l'on ne respecte pas le niveau de structuration macrorythmique de l'énoncé (cf. section 5.7). La complexité de la tâche de segmentation est une justification du nombre de techniques proposées pour segmenter le signal acoustique.

Dans la figure 1.2, nous avons illustré en (a) l'approche triviale ou la séquence idéale de segments identifiés directement aux phonèmes. La figure 1.2 (b), suggère des segments de taille infra- ou supraphonémique selon le degré d'homogénéité du continuum acoustique (Vigouroux & Caelen 1985).

L'approche schématisée en (c) illustre des informations provenant de canaux différents (temporel et fréquentiel, par exemple) qui fonctionnent en complémentarité pour la délimitation des frontières (Abry, Autesserre, Barrera, Benoît, Boë, Caelen, Caelen-

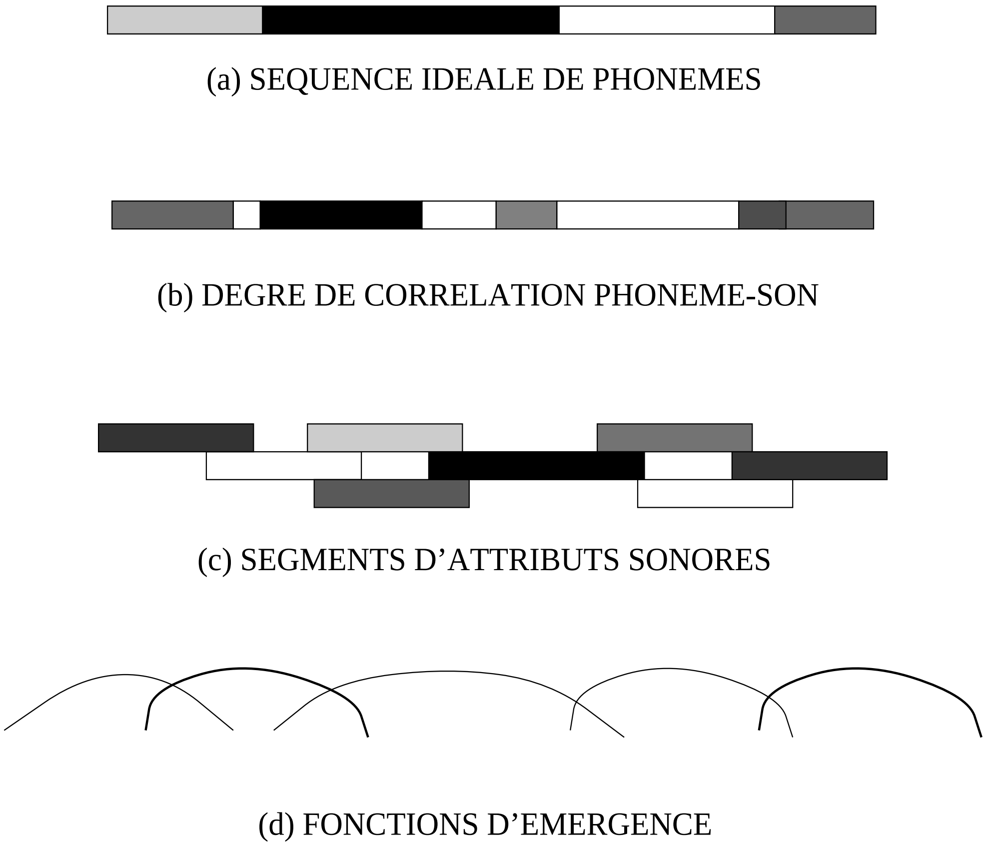

FIG. 1.2 - *Segmentation du signal de parole. En (a), la séquence phonémique idéale ; en (b), les phones homogènes ; en (c), l'approche d'Abry et collègues (1985) ; en (d), les fonctions d'émergence. D'après (Fant 1973), mais les figures ont été attribuées aux notions expliquées dans le texte.*

Haumont, Rossi, Sock & Vigouroux 1985) des réalisations phonémiques. Des connaissances articulatoires peuvent aussi donner des renseignements sur les segments ou ils peuvent être identifiés selon l'évolution d'un paramètre articulatoire (Browman & Goldstein 1985).

Dans la figure 1.2 (d), nous avons représenté la technique de la *Décomposition Temporelle* (Atal 1983; Bimbot 1988) (cf. section 4.1) qui établit des fonctions d'émergence représentant les points où les phonèmes se réalisent pleinement et leur influence sur les phonèmes voisins. Le résultat d'une telle démarche est une suite de réalisations phonémiques ou infraphonémiques[10], comme dans (b), mais les frontières ne sont délimitées que par le point où l'influence d'un phonème A sur son voisin B égale l'influence de ce voisin B sur le phonème A.

10. Dans notre travail, les fonctions d'émergence attachées à un même percept (phonème) sont regroupées.

Le signal de parole est destiné à être perçu ainsi, il ne doit pas seulement être segmenté mais aussi étiqueté : il faut identifier le segment délimité par les événements choisis au moyen d'un symbole, l'étiquette. Cette phase d'étiquetage est nécessaire quoi que soit la technique de segmentation utilisée. Pour cette phase la connaissance d'un expert est nécessaire.

Les corpora que nous avons utilisés ont été segmentés par un processus semi-automatique utilisant un algorithme de *Décomposition Temporelle* (Bailly, Barbe & Wang 1992). Les durées qui sont ainsi déterminées représentent la séquence de réalisations phonémiques dont les frontières ont été interprétées selon la technique des fonctions d'émergence (cf. section 4.1).

On entrevoit que la durée est fortement dépendante de caractéristiques de nature segmentale et que ce paramètre prosodique suggère une notion plus restreinte de la prosodie. C'est à ce niveau que l'on peut parler de *microprosodie*.

La microprosodie concerne des variations prosodiques qui se manifestent au niveau phonémique, comme la conséquence acoustique de l'abaissement de la pression sous-glottique, entraînant un abaissement de la fréquence fondamentale lors de la production des occlusives voisées. Par le truchement de la durée, on peut parler ainsi de micro- et macrorythme (cf. Introduction et les termes de micro- et macrodurée chez Vaissière 1971, p. 133-136). Cette nature segmentale de la durée a influencée fortement les modèles de prédiction que l'on retrouve dans la littérature.

1.7 Discussion

Nous avons réalisé un travail dont les objectifs principaux sont :

- quant à l'aspect segmental, nous ne cherchons pas à récupérer la trace d'un événement acoustique ou articulatoire particulier sur le signal de parole, comme c'est le cas chez Keller et collègues (1994), qui vont chercher à modéliser les organisations micro- et macrotemporelles à partir de leur lien au système articulatoire. Notre démarche se propose à prédire une durée segmentale qui s'identifie avec la réalisation d'un phonème particulier, dont les frontières sont délimitées par une suite de fonctions d'émergence. Les éventuelles durées de parties infraphonémiques (phénomènes de frication avant ou après certaines voyelles ou semi-voyelles, notamment /i/, /y/ et /j/) permettent de rétablir la durée d'une réalisation complexe d'un phonème unique ;

- quant aux niveaux organisationnels, nous rappelons que notre modèle de génération vise à obtenir des *contours de durée*, caractérisant la composante macrorythmique de la parole et ceci le plus indépendamment possible des constituants segmentaux ;
- choix a priori d'une représentation de la tâche linguistique à « faire écouter » par la prosodie : marques relationnelles.

La présentation détaillée des modèles de prédiction de la durée qui va suivre culminera justement dans les modèles qui proposent de préserver la composante macrorythmique de la parole.

Chapitre 2

Modèles de prédiction de la durée

Ce chapitre examine de près quelques modèles pour la prédiction de la durée segmentale. Nous avons pris la décision de les classer selon la taille de l'unité utilisée au départ pour la prédiction de la durée, puisque ce découpage nous semble plus conforme à la problématique soulevée par Denis Klatt (1987, p. 760) :

> *In fact, one of the unsolved problems in the development of rule systems for speech timing is the size of the unit (segment, onset/rhyme, syllable, word) best employed to capture various timing phenomena.*
>
> *En fait, l'un des problèmes non résolus dans le développement de systèmes de règles pour l'organisation temporelle de la parole est la meilleure taille de l'unité (réalisation phonémique, attaque/rime, syllabe, mot) employée pour rendre compte de plusieurs phénomènes temporels.*

Nous considérerons tout d'abord les modèles utilisant le phonème comme unité de base pour l'obtention de la durée phonémique, ensuite les modèles qui privilégient l'aspect de structuration rythmique de l'énoncé : ceux qui ont proposé le pied comme unité de départ pour la prédiction ou qui utilisent la notion de *stress* avant de prédire la durée segmentale et ceux qui partent d'une unité au même niveau que la syllabe, comme le modèle de Campbell. Celui-ci sera examiné avec plus de détail par son intérêt particulier utilisant, à la fois, la syllabe comme unité de programmation rythmique et un modèle de génération qui possède de bonnes capacités de généralisation. Le modèle que nous avons déjà eu l'occasion de présenter à diverses reprises, hérite de cette philosophie de génération hybride proposée par Campbell.

2.1 Modèles fondés sur le phonème

2.1.1 Le Modèle de Klatt ou le modèle de référence

L'auteur a consacré toute sa vie au développement d'un système de synthèse à partir du texte. Ce qui est présenté ici ce n'est que la partie concernant le modèle de prédiction de la durée qui, multiplicatif dans ces premières versions (1976), a évolué vers un modèle additif et multiplicatif.

De par son importante contribution à l'évolution des systèmes de synthèse de la parole (Stevens 1991), le système de Klatt sert de cadre théorique à bon nombre de développements postérieurs en prédiction de la durée basée sur le phonème : les modèles qui n'utilisent pas une version modifiée de son modèle, comparent la performance de leur système à la performance de celui de Klatt (cf. section 2.1.5).

Cette section décrit la version résumée du modèle de prédiction de la durée segmentale de l'anglais, publiée en 1987. Elle présente peu de modifications par rapport au modèle de 1979.

Les principes fondamentaux de ce modèle sont :

- à chaque phonème on associe une durée intrinsèque spécifiée comme une de ces propriétés distinctives (il cite House 1961) ;
- chaque règle essaie d'effectuer un certain pourcentage de modification à la durée de chaque réalisation ;
- les réalisations ne peuvent pas être comprimées en deçà de la durée minimum.

La raison par laquelle Klatt ne définit pas cette durée intrinsèque explicitement semble être parce qu'elle doit être trouvée par essai et erreur de façon à réaliser un ajustement entre durée observée et prédite (Klatt 1979, p. 294, noté par Campbell 1992, p. 15).

La durée phonémique est obtenue par :

$$Dur = MinDur + \frac{(InhDur - MinDur) \times PRNCT}{100} \tag{2.1}$$

où Dur est la durée de la réalisation, $InhDur$, la durée intrinsèque, $MinDur$, la durée minimum calculée en fonction de $InhDur$ (en général $MinDur = 0,45 \times InhDur$) pour chacun des phonèmes et pour les réalisations inaccentuées. Pour la réalisation accentuée la valeur de la durée minimum correspondante est fixée à $2 \times MinDur$ (valeur correspondante à la durée de la réalisation inaccentuée). $PRNCT$ est le pourcentage de rétrécissement déterminé de façon cyclique par application des règles. Chez Klatt, la durée intrinsèque n'a d'autre sens qu'en tant que point de départ pour l'application des règles. En fait, la valeur finale du paramètre $PRNCT$ est obtenue en multipliant cumulativement les facteurs qui influent sur la durée segmentale tels que le contexte phonétique et l'environnement syntaxique.

Il est important de noter que pour l'application des règles contextuelles, il faut commencer par l'attribution des éventuelles pauses silencieuses dès le départ [1], de façon non intégrée au mécanisme général des règles (cette règle n'est pas appliquée à l'équation 2.1, mais comme facteur contextuel des règles précédentes).

En ayant adopté un modèle basé sur le phonème, Klatt affirme que les influences rythmiques et sémantiques peuvent être intégrées a posteriori (1975). Il conclut en fait (Klatt 1987, p. 761):

> Other authors postulate rules related to rhythm and isochronous principles (Lehiste 1977). Neither of these kind of rules is incorporated explicitly in the Klatt system, but partial isochrony is achieved through rules that shorten unstressed syllables and consonant clusters (Carlson, Granström & Klatt 1979).
>
> *D'autre auteurs postulent des règles basées sur le rythme et des principes d'isochronisme. Aucun de ces types de règles n'est incorporé explicitement dans le système de Klatt mais, un isochronisme partiel est atteint à partir des règles qui rétrécissent les syllabes inaccentuées et les consonnes dans les clusters.*

2.1.2 Les Modèles d'O'Shaughnessy

O'Shaughnessy a développé deux modèles différents pour la prédiction des durées phonémiques : un modèle multiplicatif et un autre, additif.

Le modèle multiplicatif (O'Shaughnessy 1981) a été réalisé pour le français québécois dans le cadre de la synthèse par règles en analysant 285 mots monosyllabiques en position accentuée (tous les mots ont été suivis d'un silence de 200 à 300 ms). Pour étudier les facteurs qui ont une influence sur la durée phonémique, il ne considère que les effets qui introduisent des variations dépassant 10 % de changement pour la durée de base, les facteurs introduisant des variations en dessous de 10 % ayant été attribués à la variabilité intralocuteur.

Le modèle propose de générer la durée phonémique à partir d'une durée de base. Cette durée de base est, pour la voyelle dans une syllabe CVC, de 90 ms (consonne non

1. « *Règle 1*. INSERTION DE LA PAUSE : insérer une pause brève (200 ms) avant chaque syntagme principal à l'intérieur de la phrase et à d'autres frontières délimitées orthographiquement par une virgule. (voir Goldman-Eisler (1968), Cooper, Paccia & Lapointe (1978)) » (Klatt 1987, p. 761)

occlusive suivie de voyelle haute suivie d'occlusive non voisée) à partir de laquelle des pourcentages de modifications sont appliqués selon certains facteurs influant sur les durées vocaliques. Pour la consonne, la durée de base dépend de sa nature (occlusives labiales, 120 ms, d'autres occlusives, 101 ms, fricatives non voisées, 176 ms, voisées : 115 ms (/v/), 128 ms (/z/), 150 ms (/ʒ/), nasales, 113 ms, liquides : 101 ms (/l/), 113 ms (/ʀ/), *glides*, 143 ms).

Les facteurs intervenant dans le changement de la durée vocalique sont la hauteur de la voyelle, le trait de nasalité et le voisement et le mode d'articulation de la consonne qui la suit [2]. Pour les consonnes, le trait de voisement, le mode d'articulation et le voisement et mode d'articulation des consonnes adjacentes à l'intérieur des *clusters*. Dans les syllabes ouvertes ($Durée = 163\ ms$), p. ex., les durées vocaliques ne sont pas influencées par la hauteur ou nasalité de la voyelle. La consonne du coda a été systématiquement plus allongée que la consonne initiale dans les syllabes CVC : ne pas oublier que les monosyllabes sont toujours en position prépausale. Certaines caractéristiques syllabiques dans les expériences où le mot reçoit de plus en plus de suffixes (« mal », « malheureux », « malheureuse », « malheureusement ») vont dans le sens d'une conservation de la durée du mot : « rajouter une syllabe en fin de mot polysyllabique réduit la durée de chacune de ses syllabes. Le nombre total de syllabes a peu d'effet sur la durée du mot. Quand plus de syllabes sont rajoutées en fin de mot les syllabes non finales se rétrécissent légèrement. » O'Shaughnessy démontre ainsi une tendance des syllabes à présenter une durée égale à l'exception de celles précédant une pause.

Dans ce modèle, l'influence du débit, de l'accentuation et de la position dans la phrase sur la durée ne sont pas pris en compte, ni même la génération de la pause silencieuse.

Dans le modèle additif (O'Shaughnessy 1984) pour le français de l'hexagone (Bretagne) il s'agit d'une étude sur 29 locuteurs lisant un paragraphe de 111 mots à un seul débit. Le modèle part d'un ensemble de durées de base sur lesquelles des durées sont ajoutées ou retranchées, en fonction de facteurs du type de ceux déjà mentionnés dans le modèle multiplicatif et, en fonction de facteurs dus à la position dans la phrase et à la frontière de pauses mineures et majeures, celles-ci attribuées au départ ! Les 33 règles sont obtenues de façon à maximiser la ressemblance entre durées observées et prédites.

Le modèle a été capable de spécifier les durées de 281 phonèmes avec un écart-type d'à peine 9 ms pour l'erreur. Les voyelles prépausales ont été prédites avec un écart-type de 15 ms. Ce résultat moins performant est attribué à la grande variabilité de la durée vocalique dans une telle position. Dans une étude avec un nombre aussi important de sujets, des phénomènes fort différents en frontière prosodique peuvent se produire tels

2. Ceci est une constatation capitale sur l'importance du groupe VC.

l'allongement final (associé ou non à la pause silencieuse), ce qui justifie la diminution de la performance du système.

O'Shaughnessy est bien conscient que d'autres facteurs comme le débit, le *stress* et l'environnement syntaxique affectent la durée phonémique mais ils ne sont pas considérés dans son modèle.

2.1.3 Le Modèle de Bartkova & Sorin

Ce modèle de prédiction pour le français est un modèle multiplicatif appliqué à un système de synthèse par diphones. Les auteurs se proposent à obtenir la durée phonémique à partir d'une durée intrinsèque — mesurée sur un corpus de logatomes du type CVC et VCV dans des phrases porteuses enregistrées par vingt locuteurs dont quatre, statistiquement représentatifs de l'ensemble, sont utilisés pour le calcul des moyennes pour l'obtention de la durée intrinsèque — et d'un ensemble de coefficients dont les valeurs sont données en fonction du contexte (allongement en fin de mot, allongement vocalique dans certains contextes consonantiques, rétrécissement des consonnes dans les *clusters* entre autres) et d'informations grammaticales et syntactico-prosodiques (voir Di Cristo 1985 pour une analyse des variations intrinsèques et co-intrinsèques de la durée vocalique). Le corpus à partir duquel les durées sont prédites est un texte lu de la taille d'un paragraphe (approx. 1 min de lecture).

Les pauses sont générées au départ, avant l'application des règles de modification sur les durées phonémiques.

Les durées vocalique et consonantique sont données par :

$$Durée\,voyelle = DI.V_i.m_c \tag{2.2}$$

$$Durée\,consonne = DI.C_{ij} \tag{2.3}$$

où DI est la durée intrinsèque du phonème concerné (voir tableau 2.1), m_c (au nombre de 6) les coefficients co-intrinsèques, et l'ensemble V_i et C_{ij} sont les coefficients respectivement des voyelles et des consonnes dépendant de contexte. Les coefficients V_i sont au nombre de 11 et les coefficients C_{ij} sont au nombre de 9 pour chacune des 7 classes de consonnes du modèle (au total on a 80 coefficients à estimer !).

Ces coefficients sont censés être indépendants du locuteur avec certains d'entre eux (les coefficients co-intrinsèques) représentant l'influence des consonnes post-vocaliques sur les durées des voyelles et spécifiques au locuteur. Ils sont déterminés pour assurer le meilleur ajustement entre les durées prédites et les durées observées (1987, p. 250-251).

TAB. 2.1 - *Durées intrinsèques en millisecondes mesurées sur des logatomes. Moyenne sur 4 locuteurs couvrant des débits de 4,2 à 6,1 syll/s (d'après Bartkova & Sorin 1987, p. 247).*

a	177	ɛ	175	e	180	i	170	œ	185	ø	186
y	167	ɔ	170	o	186	u	170	j	150	w	144
ɥ	150	ɑ̃	200	ɔ̃	200	ɛ̃	200				
p	193	t	210	k	210	b	174	d	167	g	163
f	239	s	254	ʃ	250	v	150	z	163	ʒ	173
m	167	n	157			ʀ	132	l	134		

Pour les durées prédites des voyelles la variation maximum est de 30 % et pour les consonnes, un maximum de 20 % par rapport aux durées observées.

2.1.4 Le Modèle de Riley

L'approche de Riley (1992) est essentiellement statistique : il modèle la durée phonémique en appliquant des arbres de décision statistiquement générés. Son modèle établit une hiérarchie des facteurs influant sur la durée phonémique : contexte phonétique (3 phonèmes à gauche et 3 à droite), *stress* (aucun, primaire et secondaire), position dans le mot et position dans le syntagme.

Chaque phonème est caractérisé par un ensemble de 4 attributs : mode et lieu d'articulation consonantiques et mode et lieu d'articulation vocaliques. Si le phonème en question est une voyelle, p. ex., les deux premiers attributs (mode et lieu d'articulation consonantiques) sont vides. À chaque attribut on associe un ensemble de 12 valeurs décrivant le phonème.

Un ensemble de 1500 phrases a été utilisé pour la prédiction des durées de l'anglais. Cet ensemble a nécessité un arbre d'approximativement 250 nœuds pour les prédictions de durée dont les erreurs ont présenté un écart-type de 23 ms. Les erreurs ont été plus importantes pour les phonèmes qui étaient moins nombreux dans le corpus d'apprentissage. Les durées phonémiques en position prépausale ont été mal prédites lorsque qu'il ne s'agissait pas de la fin de la phrase (1992, p. 270).

2.1.5 Le Modèle de van Santen

Jan van Santen (1992) propose un modèle basé sur l'analyse ordonnée de données (*ordinal data analysis*) : une méthodologie pour l'analyse d'interactions complexes entre facteurs prépondérants pour la détermination de la durée segmentale.

Citant Klatt (1987), van Santen donne comme exemple sept facteurs qui influent de façon pertinente sur la durée phonémique, à savoir, *stress*, emphase, phonèmes précédents, phonèmes suivants, position dans le syntagme, position au sein du mot et nature du phonème. À chacun de ces facteurs on peut associer un ensemble d'indices ({primaire, secondaire, inaccentué}, pour le facteur *stress*).

Dans une présentation complète et récente de son modèle (1994), van Santen se réclame d'une non appartenance aux méthodes traditionnelles de prédiction de la durée[3].

La raison de cette prise de position est la constatation d'un certain nombre de défis à relever dans le processus de conversion texte-parole. Ces défis sont liés essentiellement à deux propriétés.

Problèmes liés à la généralisation

La première vient d'une observation concernant des matrices d'indices creuses. Si l'on construisait une grille de référence (*lookup table*) en utilisant tous les indices associés aux facteurs donnés plus haut, il y aurait un nombre très élevé (de l'ordre de plusieurs milliers) de cases vides, parce que peu de combinaisons de facteurs sont possibles, l'espace linguistiquement pertinent étant une petite fraction de l'espace d'indices. Quoiqu'il soit difficile de fournir des données pour l'apprentissage couvrant certaines cases fort rares (ce qui est impraticable même en utilisant de très grosses bases de données), le système doit cependant prévoir ces cas exceptionnels.

Problèmes liés à l'interaction entre facteurs

La deuxième propriété vient du fait que les facteurs interagissent, dans un sens quantitatif, de telle sorte que l'ampleur d'un effet n'est pas constante mais, affectée par d'autres facteurs.

L'auteur argumente que la première propriété pose une série de problèmes aux méthodes statistiques (voir effectivement les modèles de Bartkova & Sorin et d'O'Shaughnessy dans ce chapitre) et que la deuxième , n'est pas correctement prise en compte dans les modèles de durée actuels.

C'est pour cela qu'il propose une analyse du type ordonnée de façon à déterminer quel est l'ordre de facteurs à être considéré dans la caractérisation de la durée phonémique. L'auteur constate, par l'étude approfondie de la deuxième propriété, que les interactions

3. L'auteur fait référence aux méthodes statistiques comme les *Arbres de classification et régression* (Breiman, Friedman, Olshen & Stone 1984) ou les réseaux neuromimétiques et les modèles de durée comme celui de Klatt (1976)

entre facteurs ont un comportement régulier : l'ampleur de l'effet d'un facteur (mesuré soit en millisecondes soit en pourcentage) est modulée par d'autres facteurs. Il y a des facteurs qui produisent toujours le même effet lorsque qu'ils sont combinés à d'autres. Ces facteurs sont appelés ordonnés (*ordered factors*). Un exemple de facteur de ce type est le *stress* : il est difficile d'imaginer deux conditions en anglais différant seulement par le *stress* syllabique où la durée de la voyelle soit plus longue en condition inaccentuée qu'en position accentuée. Les effets de ce type de facteur peuvent être amplifiés mais ils ne peuvent jamais être inversés. Ils peuvent avoir la propriété d'indépendance de facteur simple (*single-factor independance*), si cet effet se produit indépendemment d'autres facteurs ou, celle d'indépendance jointe (*joint independance*), si l'effet concerné se réalise par paire de facteur ou par une association d'un nombre plus élevé.

D'autres facteurs ne fonctionnent de la manière décrite plus haut qu'à l'intérieur d'une catégorie comme c'est le cas, cité par l'auteur, des voyelles et des consonnes intervocaliques. C'est pour cela qu'une distinction entre classes de phonèmes est nécessaire avant l'application du modèle de prédiction (cette distinction permet d'établir des ensembles homogènes à l'intérieur desquels les effets d'interaction entre facteurs peuvent être ordonnés). Les classes homogènes sont constituées par les voyelles, les consonnes intervocaliques et les consonnes dans les *clusters*. Pour chacune de ces catégories il y a un ensemble de paramètres du modèle de prédiction à estimer.

Le modèle que propose van Santen est constitué d'une *somme de produits*. Il est introduit au moyen d'une comparaison avec le modèle de Klatt, « what is arguably the best known duration model. »(van Santen 1994, p. 102) (« celui que l'on peut soutenir être le meilleur modèle connu pour la durée. »).

En reprenant la formule 2.1 du modèle de Klatt, il la réécrit de façon à préciser l'ensemble de facteurs intervenant sur la durée phonémique :

$$Dur(P, \mathbf{f}) = k_{f_1} \times \ldots \times k_{f_n} \times (D_{inh,P} - D_{min,P}) + D_{min,P} \tag{2.4}$$

Dans cette équation il est explicite que la durée phonémique Dur est une fonction du phonème P et de la série de paramètres k_{f_i} (le paramètre composé $PRNCT$ de Klatt) associés au vecteur de facteurs $\mathbf{f}$ ($F_1 \ldots F_{n+1}$). Les valeurs $D_{inh,P}$ et $D_{min,P}$ sont respectivement la durée intrinsèque et la durée minimum du modèle de Klatt. On voit bien qu'il s'agit d'un modèle de somme de produits dont le modèle de van Santen est la généralisation :

$$Dur(\mathbf{f}) = \sum_{i \in T} \prod_{j \in I_i} S_{i,j}(f_j) \tag{2.5}$$

où l'indice j du terme $S_{i,j}$ (appelé facteur d'échelle) fait référence au facteur F_j et l'indice i est le nombre de termes de la somme. Pour mieux illustrer cette généralisation il réécrit l'équation 2.4 du modèle de Klatt comme :

$$Dur(\mathbf{f}) = S_{1,1}(f_1) \times S_{1,2}(f_2) \times \ldots \times S_{1,n+1}(f_{n+1}) + S_{2,n+1}(f_{n+1}) \tag{2.6}$$

En comparant avec l'équation 2.4 : $S_{1,n+1}(f_{n+1}) = (D_{inh,P} - D_{min,P})$, $S_{2,n+1}(f_{n+1}) = D_{min,P}$, $S_{1,j} = k_{f_j}$, $T = \{1, 2\}$, $I_1 = \{1, \ldots, n+1\}$, $I_2 = \{n+1\}$.

Pour estimer les paramètres, van Santen a eu recours à une base de données de 2 162 phrases courtes, lues par un seul locuteur américain. Les phrases ont fourni 41 588 réalisations phonémiques couvrant 5 073 types de vecteur d'indices. Toute la base a été segmentée manuellement !

Pour l'estimation des paramètres, trois classes ont été formées (voir plus haut). L'estimation est réalisée de façon à reproduire au mieux les durées observées. Pour cela, dans le cas des voyelles, 32 paramètres ont dû être estimés, pour un ensemble de huit facteurs linguistiquement pertinents : nature de la voyelle, accent mélodique, *stress* syllabique, consonne postvocalique et prévocalique, position dans le mot et dans la phrase.

Pour les consonnes intervocaliques cinq facteurs, dont deux joints – nature de la consonne et position dans le mot –, se sont avérés pertinents : nature de la consonne, niveau de *stress* des voyelles environnantes, position dans le mot et dans le syntagme et statut accentuel du mot (accentué, désaccentué ou clitique). L'estimation de 196 paramètres a dû être effectuée. Pour les consonnes dans des *clusters* des équations multiplicatives simples ont été utilisées car la quantité de données n'a pas été suffisante pour utiliser des sommes de produits.

L'une des façons d'évaluer le modèle a été de calculer la corrélation entre les durées prédites et cellles observées dans un texte lu par le même locuteur à trois débits différents. Les résultats sont donnés dans le tableau 2.2.

TAB. 2.2 - *Corrélations entre durées prédites par le système (dans le débit par défaut) et les durées observées dans trois débits (d'après van Santen 1994, p. 123).*

	prédiction	rapide	normal	lent
prédiction	1,000	0,874	0,884	0,882
rapide	0,874	1,000	0,928	0,889
normal	0,884	0,928	1,000	0,912
lent	0,882	0,889	0,912	1,000

Un test perceptif a montré que les sujets préfèrent nettement ce nouveau modèle de van Santen à celui contenant un ensemble de règles semblable à celles du modèle de Klatt (quoiqu'il soit un modèle plus simple, dit l'auteur).

2.2 Modèles fondés sur le pied

Dans ces modèles, la nature rythmique de la parole est considérée comme point de départ pour la prédiction de la durée. La durée au niveau segmental est déterminée en fonction de celles des niveaux supérieurs.

2.2.1 Le Modèle de Witten

Dans son modèle de prédiction de la durée (Witten 1977) pour l'anglais, Witten essaie de mettre au point un système qui intègre les caractéristiques de la dynamique de la parole tels que le tempo, le phénomène pausal, le rythme, le contour de la fondamentale, la sonie déterminés selon ce qui a été écrit sur le sujet par Abercrombie (1967). Son point de départ a été l'« imitation» du rythme de l'anglais parlé à partir de la notion de pied.

Pour intégrer cette rythmicité, l'auteur localise tout d'abord les *onsets* vocaliques à partir de la durée du pied, sans insister sur l'isochronisme de ces repères (cette notion sera abordée plus longuement dans le chapitre 3). Dans son système, le pied a une durée de départ standard de 480 ms et les durées des syllabes du pied doivent s'ajuster initialement dans ce cadre temporel. Si, en se faisant, la durée d'au moins une syllabe est inférieure à une valeur minimum, la contrainte d'isochronisme est relaxée et le pied est allongé. Il n'y a pas en principe de durée maximum pour la syllabe, mais des limites sont imposés à partir des limites des phonèmes à s'allonger.

Le débit de parole est déterminé par la durée du pied. Si ce temps est court, d'autres pieds seront forcés à avoir une durée plus longue que le standard et le signal sera « moins isochrone» Le débit rapide est obtenu en imposant les durées minima (140 ms) et les débits plus lents dans les limites de la capacité des phonèmes à s'allonger ou à se rétrécir. Les pieds sont du type (a) syllabe courte suivie de syllabe longue, (b) syllabe longue suivie de syllabe courte et deux syllabes de longueur égale. Il y a quatre types de syllabes quant à la longueur et trois quant à la position dans la phrase.

Il est intéressant de noter que les durées syllabiques sont délimitées par les points de *beat* de la syllabe (il cite Allen 1972) que Witten appelle « tapping points» en les plaçant comme hypothèse de travail à l'*onset* de la voyelle accentuée ou avant un *glide* si la syllabe a deux ou trois consonnes initiales (cf. choix de l'*onset* vocalique pour le *P-center*, au chapitre 3).

Pour lui, la façon dont on doit distribuer la durée syllabique entre les phonèmes qui la composent est inconclusive et il propose un choix de théories se basant sur la notion de durée intrinsèque ou extrinsèque : dans le premier cas la durée est déterminée par la nature des phonèmes qui la composent tandis que dans le second par la propre durée de la syllabe (il cite Lawrence 1974).

2.2.2 Les Modèles de Kohler et de Coker et collègues

Kohler(1986) propose un modèle *top-down* basé sur le pied, en allemand. La durée syllabique est une fonction linéaire de la durée du pied et celle des réalisations phonémiques, une fonction linéaire de la durée de la syllabe (deux modèles différents pour syllabe accentuée et inaccentuée). Les paramètres du modèle sont obtenus par analyse statistique au moyen de coefficients de régression.

Le modèle basé sur le *stress* de Coker, Umeda & Browman pour prédire la durée dans le cadre de la synthèse articulatoire, commence en insérant des pauses ou pseudo-pauses dans la phrase et en attribuant le patron de *stress* de chaque mot avant de calculer les durées segmentales. Il s'agit somme toute d'un modèle de prédiction basé sur le phonème et son contexte.

2.3 Modèles fondés sur des unités apparentées à la syllabe

2.3.1 Le Modèle de Monnin & Grosjean

François Grosjean et ses collègues (Monnin & Grosjean 1993; Grosjean & Dommergues 1983; Grosjean & Lane 1979) ont fait des études approfondies sur la réalité psychologique des structures de surface des phrases, ce qui les a conduits à poser l'existence des structures de performance. Les travaux de cette équipe de psycholinguistes ne visent pas, en principe, à la prédiction de la durée mais plutôt celle de la structure prosodique de l'énoncé. Toutefois, ces algorithmes se prêtent bien à la prédiction de la durée de l'unité $V + P$ (voyelle + pause) comme suggèrent les travaux de Gee & Grosjean (1981) et ceux de Keller et ses collègues (1993).

Les structures de performance ne correspondent pas à la structure de surface classique (décrivant les relations syntaxiques entre les constituants de la phrase) mais plutôt aux structures prosodiques, représentant les rapports effectifs entre les constituants phrastiques qui émergent lors de l'énonciation. Ce concept de structure prosodique se rattache à des théories linguistiques (Selkirk 1980; Liberman & Prince 1977) qui se proposent d'explorer plus à fond les mécanismes de production de la parole. Ces théories suggèrent que la phrase possède une organisation hiérarchique sous la forme d'arbre métrique à un niveau supérieur au niveau phonématique. Dans cet arbre, on trouve à chaque niveau (syllabe, mot, syntagme) une relation « faible» (inaccentuée) ou « forte» (accentuée) permettant l'établissement de règles qui expliquent le regroupement de certains éléments de l'arbre à un certain niveau (selon leur relation faible/fort et leur fonction dans la phrase) pour constituer les éléments du niveau supérieur.

Les structures prosodiques peuvent être obtenues par le biais de certaines tâches expérimentales et pour différents types de langue. Elles présentent les propriétés de constituer des unités de base d'à peu près la même longueur[4], de hiérarchie et de symétrie. La hiérarchie établit que les unités de base se regroupent au sein d'unités plus grandes qui, à leur tour, se regroupent pour former des unités encore plus grandes. La symétrie établit que la pause principale se trouve aux environs du milieu de la phrase (par rapport au nombre de mots) ; chacun de deux segments situés de part et d'autre de cette coupure est scindé à son tour en deux parties plus ou moins égales.

Pour pouvoir faire émerger ces structures de performance une tâche expérimentale (décrite dans Grosjean & Lane 1979) a consisté à demander aux sujets de lire des phrases à cinq vitesses différentes. Ensuite, les pauses silencieuses des sujets sont mesurées. Des moyennes inter-sujets sont calculées et exprimées en fonction du pourcentage du temps total de pause silencieuse de chaque sujet. Ce procédé a permis d'effectuer des représentations hiérarchiques des phrases du corpus. Les auteurs démontrent que les structures de performance ne sont pas le résultat de contraintes de respiration (la même tâche effectuée sans reprise de souffle a produit sensiblement les mêmes résultats).

Dans une étude plus récente (Monnin & Grosjean 1993) les auteurs ont pris huit sujets lisant neuf phrases à deux débits. Ils décident cette fois de mesurer la durée de la voyelle à laquelle on rajoute la durée du silence adjacent qui peut éventuellement la suivre[5]. Le phénomène pausal, disent-ils citant Duez (1987) est indissociable de l'allongement final. Cette nouvelle unité présente l'avantage de fournir une mesure entre chaque mot de la phrase indépendamment de la présence ou non de pause silencieuse. La hiérarchie des composants de la phrase est obtenue de la même manière que celle décrite précédemment.

4. Cette propriété va dans le sens du principe universel d'eurythmie exploré par Dell (1984, p. 90)

5. On verra que cette unité qui commence à l'*onset* vocalique et finit à la fin du silence va dans le sens de l'unité de programmation rythmique proposée dans notre thèse, à savoir le groupe inter-*Perceptual-center*.

L'exemple ci-dessous (figure 2.1) permettra à la fois d'illustrer la différence entre structure de surface et de performance et la façon dont la structure de performance est obtenue à partir des pourcentages des unités $V + P$ par rapport à la somme totale de ces durées dans la phrase.

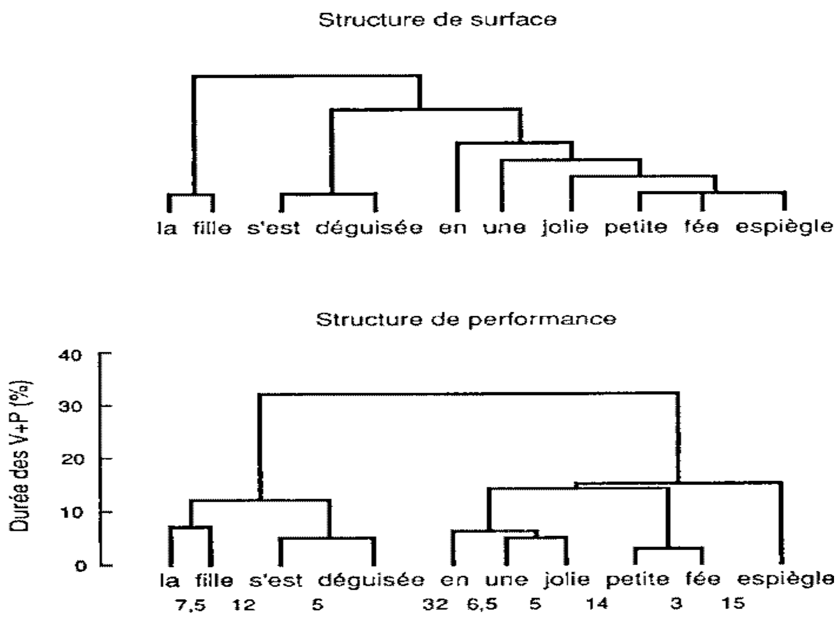

FIG. 2.1 - *Structures de surface et structures de performance, d'après (Monnin & Grosjean 1993, p. 15). La structure de performance est élaborée à partir des durées des $V+P$ converties en pourcentage (de la somme totale de durées $V + P$ de la phrase).*

Les auteurs testent ensuite quatre algorithmes qui visent à la prédiction automatique des structures de performance à partir des relations syntaxiques entre les composants phrastiques et de principes phonotactiques comme l'eurythmie (les auteurs citent Bailly 1989 et Pasdeloup 1990).

En conclusion, Grosjean et ses collègues garantissent que le fait de dégager les structures de performance permet l'amélioration de la qualité de la synthèse à partir du texte. En psycholinguistique, la compréhension de ces structures aura des conséquences sur l'élaboration des modèles de production et de perception de la parole (Monnin & Grosjean 1993, p. 27).

2.3.2 Le Modèle de Campbell

Campbell (1992) définit son modèle de prédiction de la durée segmentale pour l'anglais britannique comme un modèle combiné : il sépare le contrôle du temps au niveau plus élevé de la syllabe (décrivant d'abord la structure prosodique et rythmique de la parole) de l'attribution de la durée des réalisations phonémiques (située à un niveau inférieur) selon le cadre temporel trouvé précédemment au niveau de la syllabe. L'une des justificatives de Campbell pour son modèle est puisée dans les travaux de Kozhevnikov & Chistovich (1965) et Collier (1992, p. 206). Les premiers ont vérifié qu'en répétant la prononciation de la même phrase, au même débit, la variance en durée d'une réalisation phonémique était compensée dans les réalisations adjacentes (cf. Whalen, dans l'Introduction). Chistovich conclut alors que l'organisation temporelle des articulateurs est organisée, au moins en partie, à des niveaux supérieurs au phonème. Collier avance la même conclusion en disant que, si tel est le cas, la longueur des réalisations phonémiques n'est pas critique tant que la somme de leurs durées ait la valeur requise pour l'unité de plus haut niveau.

Le modèle conçu par Nick Campbell fonctionne en deux étapes. La première consiste dans l'obtention de la durée syllabique par apprentissage automatique en utilisant un réseau neuromimétique du type perceptron multicouches (voir annexe E pour explication du terme). Dans une seconde étape, cette durée est distribuée entre chacun des composants de la syllabe en utilisant un modèle statistique que nous avons appelé modèle de répartition.

Prédiction de la durée de la syllabe : apprentissage automatique

Une lecture de la nouvelle « Through the Tunnel » écrite par Doris Lessing a été enregistrée d'une émission à la *BBC Radio 4* et utilisée comme corpus. Ce sont 20 minutes de lecture contenant 3 959 syllabes.

Campbell utilise un perceptron à trois couches pour l'apprentissage du réseau. La première couche a six unités d'entrée variant entre 0 et 1, la deuxième a trois unités cachées et la sortie a une unité. L'entrée est une description de la syllabe en six facteurs : nombre de phonèmes, nature du noyau vocalique (voyelles réduites, voyelles relaxées, voyelles tendues ou consonnes syllabiques et diphtongues ou triphtongues), position dans le groupe tonal, type de pied (employé pour décrire le contexte rythmique de la syllabe), *stress* et classe du mot contenant la syllabe (mots clitiques ou non clitiques). À la sortie, la durée syllabique est exprimée en termes de son logarithme naturel.

Pour l'apprentissage l'algorithme classique de rétropropagation (McClelland & Rumelhart 1988) est utilisé. Il est basé sur la méthode du gradient calculé à partir de l'erreur quadratique globale entre la sortie désirée et celle obtenue en propageant les activités de la première couche au long du réseau.

Prédiction de la durée segmentale : modèle de répartition

La base de données ATR obtenue par l'enregistrement de la lecture par quatre locuteurs (deux hommes et deux femmes) d'anglais britannique de 200 phrases phonétiquement équilibrées forme le deuxième corpus de Campbell. Il l'utilise pour étudier le mécanisme de distribution (*accomodation*, en anglais) des durées des réalisations phonémiques à l'intérieur du cadre fourni par la durée syllabique. Ceci consiste à spécifier une valeur simple d'allongement associée à une syllabe qui produit les durées appropriées pour chacune des réalisations des phonèmes composant cette même syllabe. Cette valeur représente le degré d'étirement appliqué à l'élasticité (*elasticity*, capacité de s'allonger ou de se rétrécir) des phonèmes. Si on associe les durées des réalisations phonémiques à une fonction du type gaussienne, cette valeur unique d'allongement syllabique représenterait la déviation en unités d'écart-type au-dessus ou au-dessous de la moyenne.

Le principe d'élasticité dans sa version la plus forte propose que tous les phonèmes d'une syllabe donnée ont un même facteur d'allongement z (le $z-score$) : la durée de chacune des réalisations est déterminée selon les paramètres des distributions statistiques spécifiques à chaque phonème.

Campbell a constaté que les distributions des durées des réalisations de chaque phonème a une forme très proche de la gaussienne quand on extrait le logarithme de ces durées exprimées en millisecondes. Dans ce cas, on peut dire que la distribution de chaque phonème est complètement caractérisée par une moyenne et un écart-type. La moyenne peut être mise en relation avec une durée inhérente au son concerné (le temps qui s'écoule en moyenne pour la réalisation de ce phonème pour un débit donné) et l'écart-type avec sa capacité de s'allonger ou de se rétrécir (les consonnes fricatives, par exemple, sont plus allongeables que les occlusives). Pour illustrer ce mécanisme, nous allons procéder à un calcul de $z-scores$ à partir de données exprimées en millisecondes.

Admettons que l'on a deux syllabes, /as/ et /aʀ/, possédant la même durée, mettons 300 ms. À partir du tableau 4.2, section 4.2, on obtient pour le /a/ : $\mu_a = 4,69$ et $\sigma_a = 0,31$, pour le /s/ : $\mu_s = 5,07$ et $\sigma_s = 0,23$ et pour le /ʀ/ : $\mu_R = 4,49$ et $\sigma_R = 0,31$. Calculons par un algorithme récursif le $z-score$ de la première syllabe en imposant $\exp(4,69 + z_1.0,31) + \exp(5,07 + z_1.0,23) = 300$ et celui de la deuxième en imposant $\exp(4,69 + z_2.0,31) + \exp(4,49 + z_2.0,31) = 300$. Pour la première syllabe $z_1 = 0,45$ et pour la deuxième, $z_2 = 1,35$. Avec z_1 et z_2, on peut calculer les valeurs des durées segmentales dans la première syllabe : $D(a) = 125\ ms$, $D(s) = 175\ ms$ et dans la deuxième, $D(a) = 165\ ms$, $D(\text{ʀ}) = 135\ ms$. Il est intéressant de noter que le /a/ qui précède le /ʀ/ est plus long que celui qui précède le /s/, caractéristique connue des méthodes fondées sur les coefficients co-intrinsèques, comme celui de Bartkova & Sorin[6], décrit dans la section 2.1.3. Pour tester l'hypothèse d'élasticité dans le cadre syllabique, Campbell a établi une comparaison entre les moyennes des $z-scores$ catégorisées par rapport à la position de chaque phonème au sein de la syllabe.

Tout d'abord, il a calculé la moyenne ($\mu_{phonème}$) et l'écart-type ($\sigma_{phonème}$) obtenus à partir des durées absolues des réalisations de chaque phonème du corpus ATR : chaque phonème est caractérisé ainsi par une paire de paramètres statistiques ($\mu_{phonème}$, $\sigma_{phonème}$). Pour chaque réalisation phonémique du corpus ATR son $z-score$ réel est calculé par :

$$z_{réalisation} = \frac{(durée\ observée_{réalisation} - \mu_{phonème})}{écart-type_{phonème}} \tag{2.7}$$

De cette façon la distribution de ces $z-scores$ réels sera normalisée (moyenne = 0 et écart-type = 1).

Ensuite, Campbell a étiqueté les phonèmes et les syllabes. Toutes les voyelles et consonnes syllabiques ont été étiquetées *peak* (noyau). Pour les mots monosyllabiques les consonnes précédant le noyau et après une frontière de mot ont été étiquetées *onset* (attaque) et celles suivant le noyau, avant une frontière de mot, *coda* (coda). La même chose a été faite pour les mots plurisyllabiques mais les consonnes potentiellement ambisyllabiques de l'anglais ont été étiquetées *medial*. Les syllabes ont été groupées en quatre classes : courtes (dont la moyenne entre les écarts-type des ces phonèmes est inférieure à -1), normales (dont la moyenne entre les écarts-type des ces phonèmes se trouve entre -1 et 1), longues (dont la moyenne entre les écarts-type des ces phonèmes est supérieure à 1) et finales (dernière syllabe de chaque phrase). Il a calculé les moyennes et écarts-type des $z-scores$ des réalisations pour chaque sous-groupe (attaque, noyau, coda et *medial*). Ses résultats sont montrés dans le tableau 2.3.

En ayant réalisé les tests de différences pour les résultats ci-dessus, Campbell conclut que, dans le groupe des syllabes courtes, la compression est similaire pour tous ses éléments. La même conclusion a pu être avancée pour l'allongement des sous-groupes *attaque* et *noyau* dans le groupe des syllabes longues et pour l'allongement des sous-groupes *noyau* et *coda* dans le groupe des syllabes finales. Ainsi, dans une expansion ou compression, ce n'est pas la voyelle qui reçoit la plupart des changements (ce qu'une analyse avec des durées absolues montrerait), mais un facteur agissant sur toute la syllabe, allongeant la partie initiale des syllabes accentuées et la partie finale des syllabes en fin de phrase et rétrécissant les syllabes courtes uniformément.

Campbell développe son modèle de prédiction en admettant l'hypothèse forte : les

6. Noter le rapport entre les durées du /a/ devant /ʀ/ et du /a/ devant /s/, de 1,41 dans (Bartkova & Sorin 1987) et le rapport de 1,32 dans les données présentées ici.

TAB. 2.3 - *Allongement dans les syllabes longues, courtes et finales, d'après Campbell 1992, p. 84*

	longues			*courtes*			*finales*		
type phon.	moy.	éc.-t.	n	moy.	éc.-t.	n	moy.	éc.-t.	n
attaque	1,56	0,93	102	-1,22	0,56	99	0,24	0,98	149
noyau	1,47	1,16	187	-1,22	0,51	170	1,09	1,25	245
coda	1,03	1,08	87	-1,12	0,38	37	1,14	1,21	242
medial	1,48	0,92	63	-1,26	0,55	37	0,48	0,96	83

$z-scores$ sont les mêmes au long de la syllabe. Le modèle calcule alors le $z-score$ associé à chaque syllabe en résolvant l'équation ci-dessous.

$$Durée(syllabe) = \sum_{i=1}^{n} \exp(\mu_i + z\sigma_i) \tag{2.8}$$

La somme concerne les éléments phonémiques de la syllabe, z est le $z-score$ associé à cette syllabe et la paire (μ_i, σ_i) contient respectivement la moyenne et l'écart-type associés au phonème i et obtenus à partir des logarithmes des durées des réalisations (exprimées en millisecondes) de ce phonème dans le corpus. À la réalisation du phonème i dans la syllabe en question sera attribuée la durée $\exp(\mu_i + z\sigma_i)$. Les durées sont exprimées en fonction logarithmique parce que l'histogramme obtenu est plus proche de la fonction normale que si les durées étaient exprimées en millisecondes. Les estimateurs naturels de la moyenne et de l'écart-type[7] sont donc de bons estimateurs pour la distribution des durées des réalisations phonémiques exprimées en logarithme. Campbell a essayé de caractériser la distribution des durées phonémiques par une fonction Gamma à deux paramètres en trouvant un ajustement légèrement meilleur que celui obtenu avec la gaussienne appliquée à l'ensemble des durées des réalisations de chaque phonème. C'est pour cela qu'il reprend la fonction gaussienne, vu que les résultats de la prédiction ne diffèrent pas suffisamment pour justifier de l'adoption d'une fonction dont les paramètres ont un coût calculatoire plus grand.

Il sera obligé par la suite de modifier le premier modèle afin de distinguer les valeurs du $z - score$ selon qu'ils sont associés à des réalisations phonémiques en position de *stress* ou avant frontière prosodique. Campbell montre que l'allongement des réalisations phonémiques en position de *onset* sont plus prononcés dans une syllabe proéminente par

7. Les estimateurs naturels pour l'ensemble des n données y_j sont calculés par $\mu = \sum_{j=1}^{n} \frac{y_j}{n}$, pour la moyenne et $\sigma = \sum_{j=1}^{n} \sqrt{\frac{(y_j-\mu)^2}{n-1}}$, pour l'écart-type.

rapport au coda, tandis que la rime est plus allongée lorsque la syllabe se trouve dans une position avant frontière. Ceci, dit-il, n'invalide pas son modèle de répartition car le $z-score$ doit être manipulé de façon à s'allonger ou se rétrécir selon sa position dans la syllabe. Campbell utilise ces caractéristiques pour réaliser un algorithme de détection automatique de frontières prosodiques (1993).

En français, l'accent primaire a une fonction démarcative dans l'énoncé, de par sa position finale dans le mot. Étant donné que l'accent dans le mot isolé peut être effacé au cours de l'énonciation pour constituer des groupes accentuels ou mots prosodiques, on dit aussi que le français n'a pas d'accent lexical. La syllabe en frontière de mot prosodique est toujours accentuable et la syllabe accentuée détermine presque toujours une frontière prosodique (cf. l'accent initial chez Pasdeloup 1987). Il n'y aurait pas de raison d'appliquer un modèle de répartition au français en faisant une différence entre phénomènes liés à l'accentuation et à l'établissement de frontières. En outre, le choix d'une unité de programmation rythmique où les consonnes de l'attaque ne font pas partie de leur syllabe d'origine, évitera de faire une modification au modèle de répartition en sa version première, à savoir, que tous les éléments au sein de l'unité sont soumis à un même et simple facteur d'allongement.

2.4 Discussion

Les auteurs sont d'accord pour dire que le nombre de facteurs phonémiques et contextuels interagissant pour la détermination de la durée au niveau phonémique est important et ces facteurs ne sont pas clairement compris. Ceci est corroboré par la complexité des systèmes décrits plus haut.

Il est important de jeter un regard sur le commentaire de van Santen sur les modèles fondés sur des unités phonémiques :

> Another avenue is the role of units other than phonetic segments in timing control. This issue is quite subtle, and must be phrased as: is there evidence that, once we have taken care of all standard durational factors (e.g., the eight factors we used for vowel duration), there is still something left to be explained by higher order control processes that pay attention to macroscopic rhythms and isochronies? Although such control processes are fascinating from a theoretical point of view, macroscopic temporal structure may be much more a perceptual than a speech production phenomenon. To decide these issues,

very precise conceptual and mathematical models are needed.

> *Un autre point à soulever est le rôle d'unités autres que les phonémiques dans le contrôle temporel. Cette question est tout à fait subtile et doit être paraphrasée ainsi : Y a-t-il une évidence qu'une fois que l'on n'ait traité de tous les facteurs standards de la durée (les huit facteurs utilisés pour la durée vocalique, par exemple), il reste encore quelque chose à être expliqué par des processus de contrôle d'ordre supérieur qui rendent compte de rythmes macroscopiques et de l'isochronisme ? Quoique tels processus de contrôle soient fascinants d'un point de vue théorique, la structure temporelle macroscopique peut être beaucoup plus un phénomène perceptif qu'un phénomène de production de la parole. Pour se décider sur ces questions, des modèles conceptuels et mathématiques très précis sont nécessaires.*

Sans négliger l'immense travail accompli par van Santen pour la réalisation de son modèle de somme de produits à partir du niveau phonémique, il nous semble cependant que cet effort humain et calculatoire aurait pu être allégé si le modèle de durée partait des niveaux macrorythmiques. Il laisse vraisemblablement à l'aspect perceptif un rôle secondaire. Or nous pensons qu'en tant que mécanisme de production d'un système biologique, production et perception collaborent l'un avec l'autre : nous parlons pour être écoutés (cf. introduction). La prise en compte des niveaux macrorythmiques permet l'intégration de la composante perceptive (en considérant les événements qui entrainent des erreurs qui ont une conséquence perceptive) : il nous paraît ainsi que les problèmes que van Santen soulève dans les méthodes de prédiction statistiques et dans les modèles actuels de durée pourraient être résolus en tenant compte des niveaux supérieurs de structuration rythmique, sans le besoin de « modèles conceptuels et mathématiques très précis.» La question à se poser est celle qui concerne la précision de l'estimation de la durée : comme pour la fréquence fondamentale, toute erreur de prédiction n'a pas la même valeur perceptive (cf. section 5.7).

On pourra noter facilement que, dans tous les modèles de prédiction présentés, la pause est traitée toujours à part, avant toute application d'un processus de calcul des durées phonémiques.

L'émergence de la pause, étudiée de façon analytique par Grosjean et collègues, est un phénomène du mécanisme de production, intrinsèquement lié au ralentissement progressif de l'élocution. Ce mécanisme doit être reproduit, si l'on souhaite réaliser des modèles de prédiction biologiquement réalistes.

Chapitre 3

La Composante rythmique

> Un petit coup au carreau, comme si quelque chose l'avait heurté, suivi d'une ample chute légère comme de grains de sable qu'on eût laissés tomber d'une fenêtre au-dessus, puis la chute s'étendant, se réglant, adoptant un rythme, devenant fluide, sonore, musicale, innombrable, universelle : c'était la pluie.
>
> *Marcel* PROUST, *Du côté de chez Swann*

3.1 Du rythme

Qui mieux que les écrivains de qualité pour susciter notre sensibilité à cet aspect de la langue qu'est le rythme ? Comment ne pas percevoir dans ce paragraphe anthologique — même pour celui qui ne comprend pas le français — la description presque tangible d'une averse ? L'espacement des virgules guide la lecture, à l'ampleur des propositions au début se substitue la succession des mots qui caractérisent la pluie : la fluidité, la sonorité, la musicalité mais aussi l'*adoption d'un rythme*.

On pourrait être tenté de dire qu'il n'y a rien d'étonnant dans l'obtention de cet effet puisque l'activité rythmique se trouve partout dans l'expérience quotidienne et est même la propriété fondamentale de la nature vivante (Fraisse 1974). Mais qu'est-ce le rythme ?

Nous avons choisi de ne pas donner une définition précise vu l'état presque embryonnaire des études rythmiques — surtout en parole —, mais on évoquera certaines idées qui nous guideront par la suite.

Le mot *rythme* vient du latin *rhythmus*, lui-même emprunté au grec *ῥυθμός*. D'abord le mot eut le sens de « forme » dans la philosophie ionienne (surtout chez Leucippe et Démocrite), selon Benveniste (1951), pour être innové après par Platon, qui réduisit le sens rythme-forme des Anciens à la forme des mouvements humains : « C'est l'*ordre* dans le *mouvement*. » (*Lois, 665a*), écrit-il. Ainsi, selon Fraisse (1974), « le concept de rythme ne viendrait pas de quelque expérience de la nature mais bien de l'organisation du mouvement humain.»

Quoique dans le rythme les facteurs temporels soient les plus importants, cette notion de régularité est rendue manifeste non seulement au moyen de la composante temporelle du signal de parole mais aussi par les autres paramètres prosodiques. Ces éléments interagissent les uns avec les autres, parfois de façon fort imbriquée, en donnant comme effet une sensation perceptive de longueur.

Dans son dictionnaire de linguistique et phonétique, Crystal (1985) écrit :

> Une application du sens général de ce terme en phonologie se réfère à la régularité perçue d'unités proéminentes en parole. Ces régularités peuvent être établies en termes de patrons de syllabes accentuées ou non accentuées, longueur syllabique (longue ou courte) ou ton (haut ou bas) ou une combinaison quelconque de ces variables.

Le Petit Robert, par contre, ne le définit que par rapport à la durée. Toutefois sa définition présente quelques notions (en italiques) que nous ont semblées pertinentes :

> Distribution d'une durée en une suite d'*intervalles réguliers*, rendue possible par le retour d'un *repère* et douée d'une *fonction et d'un caractère psychologiques et esthétiques*.

Pour ce qui est de l'esthétique (l'aspect psychologique, considéré ici dans son sens commun, joue un rôle dans tout jugement de caractère esthétique et ne sera pas abordé dans ce travail), l'extrait proustien en est un exemple très illustratif. Nous n'y reviendrons plus puisque ce sujet est étranger à la nature de ce travail. Nous renvoyons le lecteur pour cela à des auteurs comme Grammont (1965) pour les aspects rythmiques en versification française et Rifaterre (1979) pour l'étude des procédés en production littéraire en vue de susciter des effets esthétiques chez le lecteur.

On mettra la notion platonicienne d'*ordre* à côté de celle d'*intervalles réguliers*, évitant ainsi son caractère par trop général et ambigu. Cet aspect de régularité est évoqué dans ce passage où Woodrow (1951, p. 1232) essaye une définition pour le rythme :

> *By rhythm, in the psychological sense, is meant the perception of a series of stimuli as a series of groups. The successive groups are ordinarily of similar pattern and experienced as repetitive.* Each group is perceived as a whole *and therefore has a length lying within the psychological present.*

> *Par rythme au sens psychologique, on entend la perception d'une série de stimuli comme une série de groupes. Les groupes successifs sont ordinairement de patron semblable et ressenti comme répétitif. Chaque groupe est perçu comme un tout possédant ainsi une longueur qui s'étend dans le présent psychologique.*

Le présent psychologique (assimilable à la notion plus récente de mémoire à court terme, quand on explore ce présent par une méthode d'évocation) cité plus haut est une « caractéristique fondamentale de la perception globale du successif et, sans perception globale, il n'y a pas de structure rythmique » (Fraisse 1974, p. 75) : rappelons pour cela le passage précédent de Woodrow : « Chaque groupe est perçu comme un tout. »

Cet aspect de totalité permet de mettre le groupement rythmique en relation avec la notion d'une forme selon la *Gestalttheorie* (gestaltisme) dans le sens où les stimuli ne sont pas perçus indépendamment, ils interagissent entre eux favorisant ainsi la perception d'une globalité, la *Fugengestalt* (forme jointée). Fraisse (1974, p. 74) explique que dans certaines conditions de succession, les stimulations sont perçues comme groupées et que la répétition de ces groupes donne naissance à la perception du rythme. Même si la séquence des stimuli est parfaitement régulière, nous pouvons les percevoir comme étant groupés par deux ou trois. On parle alors de *rythmisation subjective*, « parce qu'aucun facteur lié à la suite objective des stimulations ne détermine le groupement. » Dans ce type de rythmisation, deux modifications subjectives apparaissent peu à peu : l'intervalle entre deux groupements successifs apparaît plus long que les intervalles entre les éléments du groupe (qui se trouvent dans la condition d'égalité physique). On peut parler dans ce cas de pause subjective. En plus, le premier et parfois le dernier élément apparaît plus accentué que les autres (ce phénomène sera rappelé plus loin par l'étude de Wenk & Wioland sur la caractéristique rythmique du français). On parlera de *rythmisation objective* si un intervalle périodique physiquement plus long que les autres crée une pause ou si un élément sur deux ou trois par exemple est accentué. Ces deux groupements ont une caractéristique liée à la perception.

Ce phénomène d'assimilation d'éléments et distinction des ensembles joue un rôle important pour qu'une forme émerge, pour qu'elle soit perçue en tant que telle. Mais pour qu'il y ait perception de groupement, l'intervalle entre les stimuli n'est pas quelconque. Fraisse en donne comme valeurs extrêmes les durées de 15/20 cs et 150/200 cs pour des stimuli liés à l'activité motrice humaine (Fraisse 1974) (ceci est confirmé par Allen (1975) : « la nature précise des groupements rythmiques dans les langues à organisation syllabique mérite beaucoup d'attention. Toutefois, le débit de succession de syllabes et groupes se trouve essentiellement compris dans la région des comportements moteurs autres que la parole.»).

3.2 Le Rythme dans la typologie des langues

> (...), it is likely that the nature of rhythm will not be long in emerging: until such time, however, any decision as to its basis in English seems premature.
>
> *(...), il est probable que la nature du rythme ne mettra pas longtemps à émerger : cependant, jusque là, toute décision quant à sa base pour l'anglais semble prématurée.* *David* CRYSTAL*, 1969*

Allen introduit son article par la phrase : « No one doubts that spoken language has rhythm.» (Allen 1975) (*Personne ne doute que la parole ait du rythme.*). Mais, selon lui, comment il est incorporé dans une théorie de la langue est beaucoup moins clair. Avant de vouloir apporter une première réponse à cette lacune, il convient de jeter un regard sur certains aspects de la structuration rythmique en parole.

Comment la notion de rythme est-elle liée au mécanisme de production ? Pour Abercrombie (1967), ce mécanisme de phonation s'établit par une série de mouvements périodiques constituant le processus de production syllabique. Celui-ci est accompagné à certains intervalles par des renforcements (*stress mechanism*) qui produisent les syllabes accentuées. Des langues différentes coordonneraient ces deux activités différemment (l'alternance entre syllabes accentuées et non accentuées) en produisant ce qui est propre au rythme de chacune d'entre elles. Ainsi, faisant appel aux travaux de Pike (1945), il affirme que dans l'état actuel des connaissances (1967), les langues du monde peuvent être divisées de façon exclusive en deux types fondamentaux de rythme : les langues dites « stressed-timed» et les langues « syllable-timed». Cette nomenclature (ou au moins la constatation d'un tel phénomène) remonte assez loin comme expliqué par Abercrombie (1967) :

> *The existence of the two basically different kinds of speech rhythm was pointed out by Arthur Lloyd James in "Speech Signals in Telephony" (1940, p. 25): he called them 'machine-gun rhythm' and 'morse-code rhythm'. The more apt terms 'syllable-timed' and 'stress-timed' were coined by K. L. Pike and first used by him in "The Intonation of American English", p. 35. Many writers since the 18th century have pointed out that in English stressed syllabes tend to be isochronous.*

> *L'existence de deux types de rythme en parole, différents dans le fond, a été signalée par Arthur Lloyd James dans « Speech Signals in Telephony » (1940, p. 25) : il les appela « machine-gun rhythm» et « morse-code rhythm» . Les termes plus précis « syllable-timed» et « stress-timed» ont été fixés par K. L. Pike et utilisés par lui-même dans « The Intonation of American English» , p. 35. Beaucoup d'auteurs depuis le 18e siècle ont signalé qu'en anglais, les syllabes lexicalement accentuées avaient la tendance à l'isochronisme.*

Dauer (1983) cite plusieurs linguistes et phonéticiens qui se servent de la dichotomie entre langues *stress-timed* et langues *syllable-timed* en essayant de la rendre opérationnelle par la recherche de l'isochronisme dans les langues qu'ils ont étudiées. On peut citer à titre d'exemple Roy Major.

Major (1981) étudie le portugais du Brésil et écrit que cette langue a une tendance au *stress-timing* : « (1) les durées entre les syllabes accentuées ne sont pas directement proportionnelles au nombre de syllabes entre elles ; (2) beaucoup de différences dans les durées entre les syllabes accentuées ne sont pas perceptibles ; (3) la durée syllabique est inversement proportionnelle au nombre de syllabes du mot ; (4) dans le langage courant les syllabes inaccentuées sont effacées, ce que a l'effet d'égaliser le nombre de syllabes à l'intérieur du groupe accentuel ; (5) les phénomènes de réduction de la durée aident le *stress-timing*. Ces processus de réduction augmentent avec le passage du niveau de langue de soutenu à courant. Comme le niveau courant est le plus évolué (*progressive*) et est fréquemment l'indicateur de la direction du changement de la langue, les données suggèrent que le portugais est en train de changer de *syllabe-timing* à *stress-timing*.» Un tel processus dans un état synchronique revient à dire que les deux types d'organisation temporelle peuvent être présents dans une langue donnée. Ceci vient corroborer ce que Crystal (1985, p. 48) écrit à propos des conclusions de Pike sur l'organisation temporelle de l'anglais dans son *The Intonation of American English* :

> *He gives the name* stress-timed *to this kind of rhythm ["rhythm units tend to follow one another in such a way that the lapse of time between the beginning of their prominent syllables is somewhat uniform" (Pike 1945, p. 34)], but points out that English does not restrict itself to this kind of rhythm, but may occasionally display other rhythmic patterns, for example* syllable-timed.

> *Il donne le nom de stress-timed à ce type de rythme [dont « les unités rythmiques ont la tendance à se suivre de façon à ce que le laps de temps entre le début de leurs syllabes proéminentes est dans un certain sens uniforme» (Pike 1945, p. 34)], mais signale que l'anglais ne se restreint pas à*

ce type de rythme, mais peut occasionnellement montrer d'autres patrons rythmiques, par exemple, le syllable-timed.

Il n'est pas étonnant que Dauer remette cette dichotomie en cause en refusant de diviser les langues dans deux catégories aussi strictes et préférant dire qu'une langue est plus ou moins *stress-based* selon le rôle que le *stress* joue dans cette langue, surtout par rapport à la structure syllabique, au phénomène de réduction vocalique et à l'accent lexical. Toujours selon lui, l'anglais, où le *stress* a d'importantes conséquences dans ces trois domaines, peut être vraiment qualifié de *stress-timed*. Il y aurait ainsi un continuum à partir d'une langue très peu liée au *stress* (il y met le japonais mais ceci est contesté par Fujimura, dans une communication personnelle à Dauer) jusqu'à l'anglais. L'isosyllabisme a été fortement contesté par Wenk & Wioland (1982, p. 203), qui ont préféré opposer le terme de « trailer-timed », pour le français, à « leader-timed », pour l'anglais. Ils justifient leur choix en se référant à Fraisse (1974), entre autres (cf. section 3.1) : « Dans les études de rythmisation objective, la position perçue de l'élément accentué varie selon la nature physique de l'accent. Si l'accent est produit par une augmentation dans la sonie ou la hauteur, l'élément accentué est perçu à la tête du groupe dans lequel il se trouve[1]. Cependant, si l'élément accentué est tout simplement plus long que les autres, il sera perçu comme apparaissant à la fin de son groupement. » [2] À Allen (1975) d'affirmer encore : « Le rythme perçu sera différent pour chaque langue. Cependant, selon la nature des séquences sonores et leurs rapports réciproques, les langues avec accent tonique (par exemple, anglais et allemand) auront les groupements rythmiques avec les syllabes accentuées en tête ; les langues avec accent fondé sur la durée (par exemple, le français) auront la syllabe accentuée en queue ; d'autres aspects de la structure du groupement rythmique vont dépendre de règles syntaxiques et phonologiques spécifiques à la langue. »

Ce qui a été dit serait suffisant pour justifier de la nomenclature utilisée par les

1. Dans Allen (1975, p. 78) : « Fry (1955) a montré que notre indice le plus important pour le *stress* est la hauteur, avec durée et sonie moins importantes, dans cet ordre. Cependant ces trois paramètres sont fortement corrélés les uns aux autres tels qu'ils apparaissent en anglais : les syllabes qui ont plus de hauteur sont normalement plus longues et prononcées avec plus d'intensité. *Nous ne devrions pas être surpris alors, que les syllabes accentuées semblent mener le groupe dans lequel elles se trouvent, exception faite en fin de syntagme, où elles ont une durée plus longue.* » (c'est nous qui soulignons)

2. Toujours dans Allen (1975, p. 78) : « En français le patron rythmique est différent. Les syllabes successives dans un groupe de sens sont semblables en durée, hauteur et sonie, avec l'exception de la dernière, laquelle est en moyenne plus haute, légèrement moins intense et beaucoup plus longue en durée (Delattre 1966). Quoique la syllabe plus élevée en hauteur aurait une tendance à conduire le groupe rythmique, la grande augmentation en durée et la légère diminution en intensité coopèrent avec le groupement sémantique pour donner un accent final et alors une séquence de pieds « montants » pour le rythme du français parlé. »

auteurs en opposant *leader-timed* à *trailer-timed* , mais ils développent aussi un concept qui nous semble pertinent pour relier les facteurs qui entrent en jeu pour la réalisation des groupements rythmiques et les aspects perceptifs du rythme lui-même. L'unité rythmique abstraite qui détermine les limites des groupements rythmiques est appelé *régulateur*. L'importance du choix de ce terme va dans le sens de l'évocation de la notion de régularité, tout en permettant de se passer de la confusion apportée par les diverses connotations du mot rythme. Au niveau phonétique, le régulateur se réalise par l'accent. Ainsi, le français est réglé à la fin de groupe tandis que l'anglais l'est en début de groupe. D'autres caractéristiques phonétiques différencient ces deux types de rythme, par exemple, syllabes accentuées avec intensité diminuée et syllabes non accentuées tendues pour la langue *trailer-timed* par opposition à l'augmentation de l'intensité et syllabes non accentuées relaxées dans la langue *leader-timed*.

Les auteurs se gardent en principe d'étendre ces notions à d'autres langues, en évitant ainsi ce qui reviendrait à un simple changement de nom dans l'ancienne dichotomie.

3.3 L'Isochronisme en parole

Selon Abercrombie, dans une langue où le rythme possède une organisation temporelle syllabique, les syllabes sont produites à des intervalles réguliers : elles sont isochrones. Dans le cas d'une langue où le rythme possède une organisation temporelle structurée par le *stress*, ce sont les syllabes accentuées qui sont isochrones. Ainsi le français, le Telugu et le Yoruba seraient des langues avec une organisation temporelle syllabique tandis que l'anglais, le russe et l'arabe seraient des langues avec une organisation temporelle à *stress* (Abercrombie 1967, p. 97).

Ilse Lehiste (1977) rappelle cette définition de l'isochronisme :

> *The term "isochrony" refers to the phenomenon that in a stress-timed language, such as English, stressed syllables follow each other at approximately equal time intervals (Pike 1945)*[3].
>
> *Le terme d'« isochronisme » se réfère au phénomène que, dans une langue stress-timed comme l'anglais, les syllabes accentuées se succèdent à des intervalles approximativement égaux (Pike 1945).*

3. On voit bien que la littérature anglo-saxonne limite la notion d'isochronisme à la succession approximativement isochrone des *syllabes accentuées*. Dorénavant nous allons utiliser le terme d'isochronisme dans ce sens restreint et réserver le terme d'*isosyllabisme* à la notion de succession approximativement isochrone des *syllabes* (et accentuées et non accentuées).

Elle fait une révision détaillée de la littérature dans le domaine. Ces études vont d'abord considérer le problème du point de vue de la production, pour ensuite l'analyser sous l'angle de la perception.

3.3.1 De l'isochronisme en production

Lehiste résume les conclusions de la recherche pour un isochronisme en production : Shen & Peterson (1962) rejettent cette notion ; O'Connor (1965) la met toujours en doute (ces observations sur les travaux de ces deux auteurs étaient déjà avancées en 1969 par David Crystal) ; la notion d'isochronisme parfait est finalement abandonnée, comme il est résumé par les résultats de Lea (1974) : « le temps moyen entre deux *stresses* qui se suivent semble croître presque linéairement avec le nombre de syllabes du pied. »

Ceci va dans le sens des études récentes de Fant &Kruckenberg (1989) montrant que la durée de l'intervalle entre deux *stresses* consécutifs est une fonction linéaire du nombre de phonèmes. Plus récemment Fant (1991) suggère que l'isochronisme peut être présent dans des mots plurisyllabiques isolés ou dans les phrases courtes de laboratoire. Pour lui ce phénomène est toutefois présent en poésie lue (Nord, Kruckenberg & Fant 1990), où un effet stylistique est recherché. Nooteboom (1991, p. 230) se sert des résultats de Fant et ses collaborateurs pour nier toute forme d'isochronisme en production : « There is no tendency towards isochrony in speech production.» (*Il n'y a pas de tendance à l'isochronisme en production de parole.*)

Wenk & Wioland (1982), en ce qui concerne l'isosyllabisme en français, sont catégoriques en affirmant que cette notion va à l'encontre du mécanisme d'accentuation en français qui consiste justement à signaler l'accent entre autres par une augmentation importante de la durée en fin de syntagme[4].

Lehiste va prendre une position plus nuancée en affirmant d'après Classe (1939) que l'isochronisme peut exister, mais sous des critères favorables comme la proximité du nombre de syllabes, de leur structure phonétique et de la structure grammaticale des groupes concernés.

Il y a encore, toujours d'après Lehiste, quelques facteurs qui contraindraient la durée des réalisations phonémiques dans la direction de l'isochronisme. Les réalisations phonémiques non finales dans les mots de plus d'une syllabe sont d'autant plus courtes en durée qu'elles sont éloignées de la fin du mot. Au contraire, les réalisations phonémiques

4. Il est cependant étonnant de vérifier dans nos données la distribution à faible écart-type de la syllabe inaccentuée (cf. section 5.3). Nous pensons avec Lehiste que la fonction de la déviation par rapport à la durée de cette syllabe ou d'un groupe au même niveau est de signaler l'accent. Cette déviation aurait ainsi une fonction linguistique.

en fin de mot sont plus longues en durée. Ceci va dans le sens de la préservation de la durée totale des mots dans la phrase de façon à ce que, en moyenne, elles soient semblables en durée. Un autre exemple qui va dans la même direction est le rétrécissement des consonnes dans les *clusters*.

Ces quelques exemples confirment une tendance à produire les séquences sonores en parole de façon isochrone.

3.3.2 De l'isochronisme en perception

Lehiste décrit une expérience dans laquelle 17 phrases avec quatre pieds chacune sont dupliquées sous la forme d'enveloppes de bruit (à la place du signal de parole) séparés par des *clicks*. Trente auditeurs ont décidé dans les deux cas (parole et bruit) laquelle des quatre unités de la phrase était plus longue ou plus courte que les autres. Les auditeurs ont eu beaucoup de difficulté dans le cas de la parole pour prendre une telle décision, la tâche ayant été beaucoup plus simple avec la séquence de bruit. Lehiste conclut donc que la difficulté des auditeurs peut être due au fait que les séquences doivent paraître pour eux de durées égales. L'isochronisme serait alors un phénomène lié à la perception et en plus limité à la parole (ce qui justifie la facilité de la tâche pour le bruit). Elle affirme encore que l'auditeur doit probablement imposer une structure rythmique dans les séquences d'intervalles inter-*stress* (*cf.* la rythmisation subjective de Fraisse, dans la section 3.1). Ceci est corroboré par Allen (1975), en ajoutant qu'il y a une sous-estimation de la durée des intervalles de temps longs et une surestimation des intervalles courts : on écouterait des séquences ayant une durée sensiblement égale comme étant plus proches qu'elles ne le sont en réalité. Ces deux caractéristiques contribuent à la perception de rythme régulier dans les langues à accent du type *stress*.

Elle incorpore cette notion dans une théorie de la langue en affirmant que, comme l'auditeur s'attend à ce qu'il y ait de l'isochronisme, il serait possible de manipuler la durée à des buts linguistiques.

En ayant manipulé systématiquement la durée de l'intervalle[5] entre deux syllabes accentuées consécutives (dans des phrases obtenues par analyse-synthèse avec fréquence fondamentale monotone), de façon à étudier les paramètres dont les auditeurs se servent pour désambiguïser une paire de phrases, Lehiste, Oliver & Streeter (1976) démontrent qu'ils réussissent à le faire lorsque cet intervalle atteint une certaine valeur de durée, c'est-à-dire, l'augmentation de la durée de l'intervalle inter-*stress* est suffisante à elle seule pour signaler la présence d'une frontière syntaxique, mais précisément parce que

5. Dans un même intervalle les rapports entre les durées des réalisations phonémiques sont toujours maintenus constants.

cette augmentation est une déviation d'un patron attendu.

Peut-on parler la même chose pour le français ? Une modification dans la rythmicité de la phrase est-elle perçue comme signe linguistique ou comme un manque de naturel ? On essayera d'apporter une réponse à cela dans le chapitre 4.

Mais sur quoi l'auditeur s'appuie-il pour percevoir des stimuli en tant que des séquences isochrones ? Y a-t-il des points d'ancrage, des repères, des événements singuliers sur lesquels il « calculerait » en quelque sorte le maintien ou les déviations par rapport à la régularité de base ?

3.4 Le « Perceptual-center »

Après des expériences de synchronisation entre le battement d'un doigt sur une surface et la production simultanée d'une séquence vocale, Fraisse suggère l'adoption de la *jonction consonne/voyelle* comme point de repère pour la perception d'une certaine régularité en parole (Fraisse 1974).

George Allen réalise le même type d'expérimentation (Allen 1972) et conclut que, « au moins pour l'anglais, l'*onset* du noyau vocalique d'une syllabe accentuée est un focus en parole, c'est-à-dire, les locuteurs natifs de l'anglais, quand ils parlent ou écoutent de façon rythmique, placent les temps frappés (*downbeats*) de leurs rythmes près du voisinage des *onsets* des voyelles accentuées. » Et juste après : « Quoique beaucoup de langues n'aient pas des accents du type *stress* au même degré qu'en anglais, toutes ont un certain type d'accent. Nous pouvons alors généraliser à partir du *stress-beat* en anglais vers l'*onset* du noyau vocalique de n'importe quelle syllabe comme étant un *beat* rythmique potentiel en parole. » Une étude récente réalisée pour le franco-provençal montre une situation semblable pour le synchronisme entre un geste de percussion (la fabrication artisanale d'un sifflet) et la parole (Berthier, Abry & Lalouache 1991). Ils montrent qu'il y a une relation temporelle entre la percussion et le début de la voyelle qui le suit (de 0 à 100 ms), mais jamais avec la fin des voyelles précédant cette percussion.

Quand la séquence de référence est aussi de la parole, l'établissement d'un repère sur le signal est plus difficile. Dans ces cas, produire des stimuli à des intervalles réguliers ne consiste pas tout simplement à avoir les *onsets* à des intervalles réguliers. C'est la conclusion principale de Morton, Marcus & Frankish (1976) dans l'article qui fixe le début d'un nouveau paradigme en perception, le *Perceptual-center*[6] :

> *We wish to introduce a new term into the language of perception, in particular, speech perception.*

6. Ce concept est développé par Marcus à deux reprises dans des textes non publiés (1975, 1976).

Nous souhaitons introduire un nouveau terme dans le langage de perception, tout particulièrement en perception de la parole.

3.4.1 Les Travaux de Marcus et ses collègues

La définition de ce *Perceptual-center* (P-center) est liée en principe au mot :

> *The* P-center *of a word corresponds to its psychological moment of occurrence. This definition is not completely satisfactory for polysyllabic words but will serve as a first approximation.*
>
> *Le P-center d'un mot correspond à son moment psychologique d'occurrence. Cette définition n'est pas complètement satisfaisante pour des mots plurisyllabiques mais servira comme première approximation.*

Et à propos de leur motivation :

> *(...) we were forced to ask ourselves what it was that was regular in a rhythmic list. To simplify our discussions we defined this as the P-center of each item. This act of reification completed, we began to ask questions about the concept itself.*
>
> *(...) nous avons été forcés de nous demander qu'est-ce qui était régulier dans une liste rythmique. Pour simplifier nos discussions nous définîmes cela comme le P-center de chaque item. Cet acte de réification terminé, nous commençâmes à nous poser des questions sur le concept lui-même.*

On peut dire que l'émergence de ce concept est le fruit des recherches sur l'isochronisme en parole.

Pour estimer la localisation du *P-center* sur le signal, Morton, Marcus & Frankish ont élaboré un protocole expérimental qui consiste dans la présentation d'une paire de sons alternés (des chiffres de *un* à *neuf*) contrôlés par ordinateur. Un des sons se répète à des intervalles fixes et l'autre se répète à des intervalles déterminés par l'auditeur en ajustant un bouton jusqu'à que la paire de sons soit perçue comme se succédant à des intervalles égaux. Après la présentation de toutes les paires, une procédure de calcul par la technique des moindres carrés est utilisée pour déterminer la position des *P-centers* de chaque son par rapport aux autres sons présentés. Il s'agit alors d'un alignement *relatif* des *P-centers* que, selon les auteurs, « montre peu de variabilité individuelle, contrairement aux techniques de Rapp (1971) et Allen (1972) basées sur des positions *absolues*. »

Les auteurs écrivent que le locuteur peut produire ainsi des séquences de mots que l'auditeur percevra comme étant régulières. Ceci impliquerait que l'organisation temporelle des articulateurs du locuteur obéit aux mêmes règles que sa perception (ou celle d'autrui). Dans ce sens, Fowler (1979) suggère que le P-center correspond à l'organisation temporelle de gestes articulatoires, codés de façon complexe dans le signal de parole (la *perception directe*). Les sujets pourraient juger la régularité perceptive, selon Fowler, sur la base de l'organisation temporelle des *onsets* des gestes articulatoires.

Mais pour Marcus le P-center ne semble correspondre

- ni à un simple événement articulatoire : « Même si on admet qu'il est possible de déterminer le moment de l'*onset* articulatoire à partir du signal acoustique, celui-ci est concerné comme un tout dans la perception de l'organisation temporelle et c'est comme une totalité qu'il doit être considéré dans cette détermination » ;
- ni à un simple événement acoustique (qui serait clairement identifiable, p.ex., l'*onset* vocalique ou le maximum d'intensité du signal) : « (...)les P-centers sont déterminés non par un simple événement acoustique mais par des événements distribués sur un intervalle considérable de temps au long du stimulus. » (Marcus 1975, p. 252).

De plus, selon lui, basé sur les observations des changements d'énergie produits par les explosions des occlusives comme le /t/, les P-centers semblent être invariants par rapport à des changements considérables dans la distribution d'énergie et déterminés de façon critique par le patron temporel du stimulus. C'est pour cette raison qu'il propose un modèle acoustique — qui ne tient compte que du facteur temps — pour la localisation du P-center dans des monosyllabes (du type C_1VC_2, où C_1 et C_2 peuvent être des consonnes, des *clusters* ou absents et V peut être une voyelle ou une diphtongue). Ce modèle est exprimé par :

$$P = \alpha dur(C_1) + \beta dur(VC_2) + k, \tag{3.1}$$

où P est la localisation du P-center par rapport à l'*onset* du stimulus, α et β sont les paramètres du modèle, $\mathrm{dur}(C_1)$ et $\mathrm{dur}(VC_2)$ sont respectivement la durée des éléments précédant la voyelle et celle de la rime et k est une constante arbitraire représentant le fait que la position du P-center est relative par rapport à d'autres stimuli.

Dans ses expériences, Marcus a établi que $(\alpha; \beta) = (0,65; 0,25)$ en affirmant que si ces valeurs-là doivent être déterminées pour chaque ensemble de stimuli, le modèle garde son caractère descriptif. Toutefois ces valeurs ont été utilisées par lui-même (Marcus 1976) et Eling, Marschall & van Galen (1980) pour le hollandais, avec succès.

Il est intéressant de noter que l'une des conclusions de Marcus (1975, p. 256) d'après ses résultats est que, avant de parler de *syllable-* ou *stress-timing* en parole continue, « on a besoin d'être clair à propos de quels sont les événements sur lesquels on mesure l'organisation temporelle» et, remettant en cause les recherches sur l'isochronisme : « Nous devrions être avertis que beaucoup des données, qui ont été utilisées pour démontrer de la présence ou de l'absence d'isochronisme, ont maintenant besoin d'être soigneusement réexaminées. »

Pompino-Marschall (1989) a élaboré une technique semblable à celle de Marcus, mais cette fois pour une estimation de la localisation absolue des *P-centers* au moyen d'expériences de perception pour des séquences de syllabes du type CV ou VC. Ses intéressants résultats et sa technique seront sommairement présentés ci-dessous.

3.4.2 Les Travaux de Pompino-Marschall

Pompino-Marschall a voulu tester l'influence de la durée de l'attaque et de la rime (voyelle et coda) dans la localisation du P-center. Le plan de l'expérience est constitué par des séquences de cinq syllabes synthétisées (S), présentées aux sujets au moyen d'un casque en alternance avec un ton (C) à un tempo prédéterminé (de l'ordre de la seconde). Les durées de la consonne et de la voyelle peuvent être variées indépendamment et ceci est réalisé par des pas de 40 ms. Ces stimuli quasi-naturels sont présentés sous forme de séquences du type C-S-C-S-C ou S-C-S-C-S. Les sujets devaient ajuster l'alignement temporel de la séquence de clicks de façon à percevoir chacune de deux types (C-S-C-S-C ou S-C-S-C-S) de séquence comme isochrone en tournant un potentiomètre. La moitié de la durée entre deux instants consécutifs de clicks (mesurés par rapport à l'*onset* acoustique de la syllabe) a été prise comme un indicateur de la position du P-center (figure 3.1).

Dans la première expérience il teste le modèle de Marcus (1976, 1975) concernant l'effet additif de la durée de la consonne et de la voyelle dans la détermination du P-center avec des syllabes /ma/. On voit clairement dans la figure 3.2.b l'effet de la durée consonantique, qui déplace linéairement la localisation du *P-center* par rapport à l'*onset* de la syllabe. Légèrement biaisée par la durée croissante de la voyelle /a/, la localisation du P-center ne s'écarte pas beaucoup de l'*onset* vocalique (pente continue au milieu de la figure). Cette expérience a été répétée avec des stimuli qui n'étaient pas de la parole de façon à tester les modèles basés uniquement sur l'enveloppe d'amplitude du signal (Howell 1988; Schütte 1978). Selon ces deux modèles cités, les prévisions basées sur l'enveloppe de signal auraient dû donner la même localisation pour le P-center. Dans le cas des stimuli

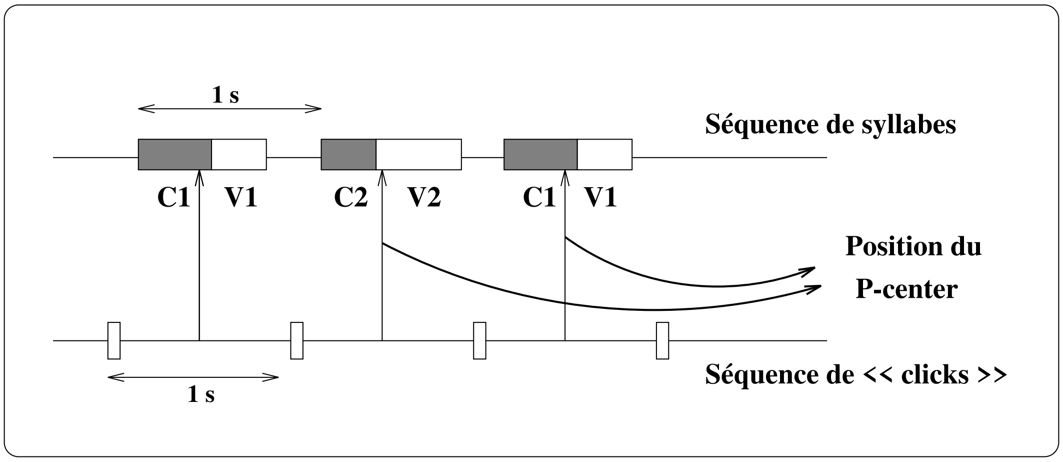

FIG. 3.1 - *Paradigme de Pompino-Marschall pour l'estimation de la position du* P-Center

non parole, Pompino-Marschall se demande pourquoi les résultats de la localisation du P-center sont alors différents dans ce cas (voir les signes « + » dans la figure 3.2.b. Si l'hypothèse de la dépendance unique par rapport à l'enveloppe d'amplitude s'avérait être vrai, chaque signe « + » aurait dû coïncider avec un signe correspondant des stimuli quasi-naturels).

Dans une autre expérience, en voulant vérifier si l'effet P-center était déterminé uniquement par les durées et non affecté par la catégorie phonétique des éléments de la syllabe, il utilise cette fois-ci la syllabe /ʃi/. Les conclusions sont semblables à celle de l'expérience précedente (voir figure 3.2.c). On y voit clairement la différence de la position du P-center pour les stimuli qui n'étaient pas de la parole. Le P-center est encore plus nettement proche de l'*onset* vocalique (représenté par la pente continue) que dans l'expérience précédente.

Une syllabe du type VC (/am/) est testée dans la troisième expérience. En comparant la figure 3.2.d avec 3.2.b et 3.2.c, on voit que la durée croissante de la voyelle affecte moins la localisation du P-center que ne le fait la durée croissante de la consonne. Ici l'*onset* vocalique coïncide avec celui de la syllabe (dans la ligne horizontale $LocalisationPC/onsetS = 0$). La différence d'effet pour chaque durée de consonne va à l'encontre d'un modèle basé uniquement sur la durée de la rime prise comme un tout, comme celui de Marcus présenté plus haut ou celui de Cooper, Whalen & Fowler (1988).

Tout en affirmant que le modèle de Marcus défini par l'équation 3.1 offre des caractéristiques intéressantes, puisque il donne plus de poids pour la durée des consonnes en position d'attaque pour la détermination du P-center, l'auteur expose les facteurs suivants qui vont à l'encontre de ce modèle de l'effet P-center : (a) les influences des durées des réalisations phonémiques dans la localisation du P-center ne sont pas indépendantes

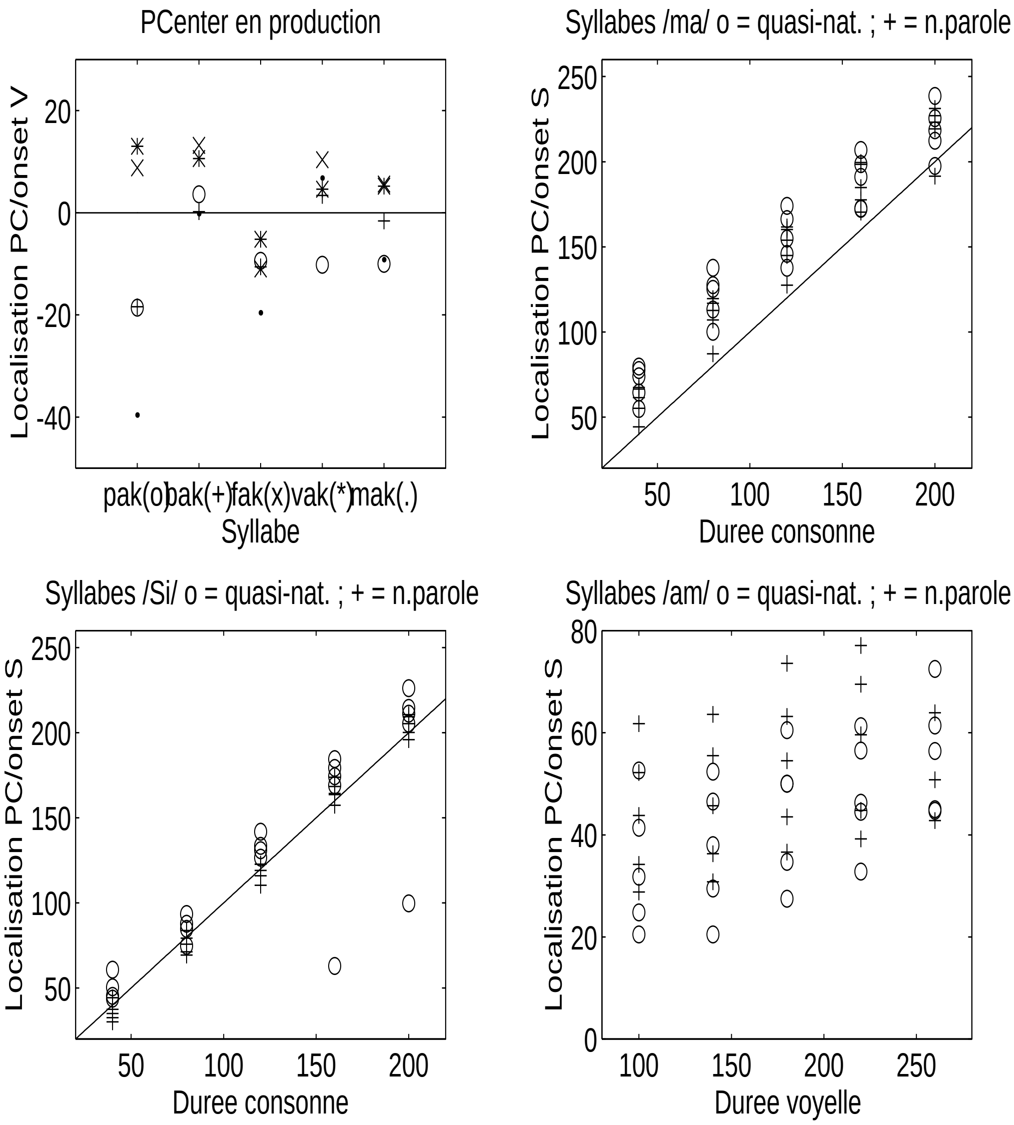

FIG. 3.2 - **En production.** *Dans la figure en haut à gauche (a) il s'agit de produire des séquences en* beat *avec un métronome de 60 pulsations/min (un ton de 3 kHz de 5 ms). Il s'agit de moyennes pour deux sujets masculins.* **En perception.** *En haut à droite (b) il s'agit d'intercaler un métronome (C : un ton de 1 kHz pendant 5 ms, 60 pulsations/min) avec une séquence de syllabes (S) de période d'une seconde de façon à percevoir l'alternance comme étant isochrone. Il y a deux séquences (C-S et S-C), 4 ajustements et 5 sessions pour deux sujets. La syllabe /ma/ a été modifiée avec un synthétiseur Klatt, consonne et voyelle indépendamment. Le matériel marqué « n. parole» est un signal carré de 100 Hz qui possède la même enveloppe de pression sonore que les syllabes. En bas à gauche (c) la même chose a été faite avec la syllabe /ʃi/ : la partie « n. parole» a été réalisée en ajustant l'amplitude de la fricative et l'enveloppe de la voyelle selon leurs correspondants synthétisés. Pour la syllabe /am/, en bas à droite (d), même opération que pour la syllabe /ma/.*

les unes des autres ; (b) la rime n'affecte pas son P-center comme un tout ; (c) il y a de fortes non linéarités dans la détermination de cet effet et (d) les effets ne sont pas spécifiques à la parole mais ne sont pas indépendants du spectre des stimuli testés.

Ces résultats le conduisent à adopter une explication fondée sur des paramètres psychoacoustiques. Tout comme Marcus, il s'oppose aux affirmations de Fowler et de ses collègues sur le fait que « les auditeurs extraient de l'information du signal acoustique spécifiée par l'organisation temporelle des articulateurs. » (Fowler, Whalen & Cooper 1988, p. 94). Pour lui, le rôle de la psychoacoustique pour guider la récupération du geste articulatoire à partir du signal de parole est sous-estimé dans la plupart des considérations en perception de parole.

Pompino-Marschall propose (1989) un modèle psychoacoustique fondé sur l'information présente dans des bandes fréquentielles qui représentent la résolution fréquentielle de l'oreille humaine. La première version de son modèle réussit à prédire de façon très précise la localisation des P-centers calculée avec son paradigme expérimental. Une deuxième version a permis d'améliorer la concordance du modèle avec les résultats trouvés avec la syllabe du type VC. Pour cela il a considéré cette fois non seulement des événements du type *onset* mais aussi des événements partiels d'*offset* et ainsi tenir compte de la totalité de la forme sonore de la syllabe, avec le calcul d'une sorte de centre de gravité syllabique.

Dans ses études en production de parole, les sujets devaient prononcer des séquences de monosyllabes identiques ou en alternance avec d'autres types de monosyllabes en *beat* avec un métronome généré par ordinateur. Ce qu'il a mesuré comme un indicateur de la position du P-center a été la position du début du métronome par rapport à l'*onset* vocalique (Pompino-Marschall 1991). Les résultats se trouvent dans la figure 3.2.a. On voit encore une fois comment, malgré la diversité des syllabes, l'*onset* vocalique semble être une espèce d'attracteur pour le P-center.

Une autre expérience de Pompino-Marschall (1991, p. 110) avec des séquences de syllabes /mam/ suggère que les sujets semblent s'appuyer sur l'intervalle entre deux P-centers consécutifs pour réaliser un jugement de changement de tempo. Ce résultat corrobore lui aussi l'importance du P-center du point de vue perceptif.

3.4.3 Les Travaux de Sophie Scott

Le travail de Sophie Scott (1993) propose une évaluation des théories actuelles sur la détermination de la localisation du P-center. Elle conteste les modèles uniquement basés sur la distribution d'énergie du signal (Howell 1988) ou les modèles basés sur une théorie de la perception directe (Fowler 1979; Tuller & Fowler 1981). L'effet de l'augmentation en intensité à des différents points au long du signal a été étudié et les résultats montrent que les changements à l'*offset* d'un signal n'affectent pas la localisation du P-center. Continuant la recherche des paramètres qui pourraient avoir une contribution à l'effet P-center, Scott analyse les stimuli en sortie d'un banc de filtres modélisant l'action de la cochlée. Les résultats ont montré que l'instant du maximum de la dérivée de l'intensité d'un sous-ensemble des filtres était corrélé avec la localisation du P-center : « (...) le taux de variation des changements dans la bande fréquentielle de 195 à 1638 Hz détermine de façon importante la localisation du P-center.» Elle suggère ainsi l'adoption d'un modèle fondé sur la dérivée de premier ordre de l'intensité du signal. Vu que normalement c'est aux environs de l'*onset* vocalique qui se produit un changement significatif de l'intensité du signal, elle est d'accord pour dire que la localisation du P-center se trouve dans son voisinage (communication personnelle).

3.5 Sommaire

Nous avons donné un éclairage sur l'importance de la composante rythmique dans le mécanisme de production et perception de la parole : la notion d'isochronisme y est directement liée.

L'intuition d'auteurs comme Witten ou Campbell nous semble fondamentale : un modèle de prédiction de la durée qui se veut intégrer les contraintes du mécanisme de production de la parole a besoin de considérer dès le départ la composante rythmique. On verra que cette prise de position a simplifié la génération automatique (on relègue le calcul de la durée phonémique à un ajustement dans un cadre macrorythmique) et a permis d'envisager l'émergence de la pause de façon concomitante au propre mécanisme d'obtention des durées au niveau segmental.

Nous ne proposons pas dans notre travail un nouveau modèle pour la localisation du P-center, tel n'est pas notre intérêt. Notre problématique est de savoir si la syllabe est ou non la meilleure unité pour la programmation rythmique, de façon à pouvoir obtenir selon un modèle de répartition, comme celui proposé par Campbell (voir chapitre précédent), les durées des réalisations phonémiques. Pour ce faire, nous retiendrons du paradigme P-center une localisation acoustique remarquable (Howell & Scott 1992): même s'il s'agit de parole continue (les expériences effectuées ne concernent que des séquences de monosyllabes), le P-center sera placé en principe à l'*onset* vocalique. Toutefois, suivant notre interprétation des résultats de Scott, si un silence s'introduit dans la séquence (s'il s'agit de ceux dus à la nature des réalisations phonémiques elles-mêmes, comme les occlusives non voisées, ou de ceux dus au phénomène de pause), le P-center sera placé à la fin de celui-ci, en début de la partie sonore du signal, où le changement de l'intensité du signal commence à se faire sentir[7].

Deux P-centers consécutifs établissent ainsi un intervalle de temps que l'on appelera l'*intervalle inter-P-center (IIPC)*. Chaque IIPC contient un nombre défini de réalisations de phonèmes qui forment le *groupe inter-P-center (GIPC)*.

Dans le chapitre suivant il sera question de la candidature du GIPC comme unité de programmation rythmique.

7. Sophie Scott semble un peu sceptique concernant cette conclusion. Pour elle, moins de sonorité une consonne présente, plus le P-center sera retardé. En n'ayant pas travaillé jusqu'à présent en parole continue, elle préfère attendre les résultats des travaux qui la considéreront (communication personnelle).

Chapitre 4

L'Unité de programmation rythmique

Dans ce chapitre, une série d'analyses nous a permis de choisir le GIPC comme unité de programmation rythmique en français. Il convient d'éclaircir ce que nous comprenons par ce terme. Il s'agit de l'unité rythmique minimale, puisque elle semble constituer une unité sur les plans de la production et de la perception (comme on a vu précédemment pour l'organisation temporelle donnée par les *P-centers*). De plus, nous verrons que l'élément inaccentué d'un tel groupe semble constituer la valeur de référence dans l'organisation du rythme (voir Pasdeloup (1987), pour la syllabe). Le terme de programmation est maintenu dans le sens où l'unité minimale permet la structuration des groupements de niveau supérieur et où elle est planifiée d'avance.

4.1 Les Corpus

Les trois corpus utilisés dans notre travail ont été obtenus dans trois sessions d'enregistrement différentes en chambre sourde. Dans les trois sessions, notre locuteur Frédéric Bimbot a lu un ensemble de phrases isolées présentées à l'écran.

Le corpus *Logatomes* est composé de mots dépourvus de sens (logatomes) insérés dans une phrase porteuse (« C'est *logatome* ça ?»). Un ensemble de 2 012 logatomes a été enregistré à débit normal avec notre locuteur. Ce corpus a été construit pour que chaque phonème puisse avoir tous les contextes gauches et droits possibles en français, de façon à estimer un modèle de répartition.

Le corpus *Bureau* se compose d'un ensemble de phrases lues originalement à trois débits obtenus par des instructions du type « Parle plus lentement !» , « Parle plus rapidement !» . Une analyse ultérieure a montré en effet que seulement deux débits effectivement distincts ont pu être obtenus. Les deux débits retenus ont été baptisés *rapide* et *normal*. Il s'agit d'un corpus contenant 88 phrases enregistrées qui a été conçu de manière à mettre le locuteur en situation de délivrance d'information, en vue d'une étude plus globale sur les phénomènes prosodiques : nous avons voulu l'exploiter pour l'étude spécifique de la durée. Les phrases de ce corpus, possédant entre 4 et 33 syllabes, se trouvent dans la section A.1. Il convient de préciser que ce corpus présente la répétition voulue de mots et syntagmes (« Rappelez Monsieur Dupont », « à son bureau », « à quinze heures dix », par exemple) afin d'étudier les structures rythmiques dans plusieurs phrases, dans plusieurs positions dans la phrase, toute chose étant égale par ailleurs (cf. Article présenté à La Baume-lès-Aix). Le paramètre de taille des mots nous a semblé judicieux : au-delà de trois syllabes, le mot est susceptible de recevoir une marque d'emphase, que nous voulions, dans la mesure du possible, éviter. Ce qui nous intéresse dans une première étape est de révéler une structuration macrorythmique dans une situation de lecture la plus « neutre » possible (où le lien à la structure syntaxique sera le plus fort).

Les phrases du troisième corpus *PCenter* se composent uniquement de séquences de syllabes CV de façon à faciliter la localisation du phénomène *P-center* et étudier le problème de l'émergence de la pause. Les phrases se trouvent dans la section A.2. La distinction effective entre les cinq débits (*très rapide*, *rapide*, *normal*, *lent* et *très lent*) a été assurée en posant au locuteur un ensemble de questions en synthèse générées à cinq débits[1]. À chaque question synthétisée présentée par un casque correspondait une phrase qui apparaissait à l'écran et était lue par le locuteur, qui essayait de reproduire un débit proche de celui de la question posée.

Les corpus ont été étiquetés manuellement au centre de leurs réalisations phonémiques (un exemple d'étiquetage est donné en annexe B). Celles du troisième corpus (environ 4 000 étiquettes) ont été posées par nos soins. Les frontières ont été déterminées automatiquement par un algorithme de *Décomposition Temporelle* comme décrit pour la première fois par Atal (1983) et appliquée à la segmentation par Bimbot (1988). La performance de cet algorithme est très correcte, toutefois, les éventuelles erreurs ont été corrigées manuellement. Celles-ci ont été retrouvées dans la grande majorité des cas lorsque des points isolés dans l'histogramme des erreurs de prédiction de notre modèle sont apparus.

Les durées observées dans les phrases naturelles et considérées lors des analyses qui

1. Ces phrases ont été obtenues par une application *grosso modo* du modèle de prédiction par méthode de répartition – en multipliant les moyennes et écarts-type caractéristiques de chaque phonème par un facteur unique α (Wightman, Shattuck-Hufnagel, Ostendorf & Price 1992) (voir section 4.6) – inséré dans le système de synthèse de l'*Institut de la Communication Parlée*.

vont suivre concernent donc les intervalles de temps compris entre deux frontières. Les erreurs de prédiction sont calculées entre ces durées observées obtenues par étiquetage semi-automatique et les durées délivrées par le modèle.

4.2 Caractérisation des phonèmes du locuteur

On a calculé à partir du corpus *Logatomes* les moyennes et écarts-type des durées log-transformées des réalisations phonémiques[2]. Le tableau 4.1 montre les moyennes et écarts-type obtenus pour les phonèmes du français de notre locuteur et calculés à partir de la distribution des durées exprimées en millisecondes. Le tableau 4.2 présente des moyennes et écarts-type obtenus à partir de la distribution des logarithmes naturels de chaque durée. Le premier tableau nous permet d'avoir une idée plus précise des caractéristiques d'allongement liées à chaque phonème. À chaque phonème, on peut donc associer une paire ($\mu_{phonème}$,$\sigma_{phonème}$) contenant la moyenne et l'écart-type de ses réalisations dans le corpus *Logatomes*.

TAB. 4.1 - *Moyennes (et écarts-type) des durées, exprimées en millisecondes, des réalisations des phonèmes français de notre locuteur. Remarquez que les voyelles les plus fermées sont d'une façon générale moins longues (sauf pour les contrastes de timbres intermédiaires), les occlusives et les fricatives sourdes sont plus longues que leur contre-partie voisée, les liquides ont une durée moyenne semblable, entre les voyelles, les nasales sont les plus longues.*

a	115(38)	ɛ	109(32)	e	118(41)	i	101(33)	œ	107(36)	ø	116(29)
y	105(30)	ɔ	110(32)	o	125(34)	u	108(32)	j	94(27)	w	95(25)
ɥ	92(20)	ɑ̃	136(42)	ɔ̃	138(42)	ɛ̃	134(39)				
p	125(38)	t	128(39)	k	114(35)	b	92(22)	d	94(25)	g	87(23)
f	143(32)	s	163(36)	ʃ	144(26)	v	106(30)	z	116(25)	ʒ	104(26)
m	104(24)	n	102(24)	ɲ	110(26)	ʀ	94(30)	l	93(27)		

4.3 Le Choix de l'unité de programmation rythmique

Un modèle de répartition a déjà été décrit au chapitre 2, section 2.3.2. Rappelons qu'il s'agissait d'associer à une syllabe donnée, un facteur d'allongement ou $z-score$, calculé à partir de sa durée et des paires ($\mu_{phonème}$,$\sigma_{phonème}$) qui caractérisent les phonèmes qui

2. $durée_{log-transformée} = \ln(durée_{exprimée\ en\ millisecondes})$

TAB. 4.2 - *Moyennes (et écarts-type) des logarithmes naturels des durées des réalisations des phonèmes français de notre locuteur.*

a	4,69(0,31)	ɛ	4,65(0,27)	e	4,72(0,32)	i	4,57(0,32)	œ	4,62(0,31)	ø	4,72(0,25)
y	4,62(0,28)	ɔ	4,67(0,28)	o	4,79(0,26)	u	4,64(0,29)	j	4,50(0,30)	w	4,52(0,25)
ɥ	4,50(0,23)	ɑ̃	4,87(0,29)	ɔ̃	4,89(0,30)	ɛ̃	4,86(0,27)				
p	4,79(0,30)	t	4,81(0,31)	k	4,69(0,32)	b	4,49(0,24)	d	4,49(0,24)	g	4,43(0,24)
f	4,94(0,24)	s	5,07(0,23)	ʃ	4,95(0,19)	v	4,63(0,29)	z	4,73(0,21)	ʒ	4,63(0,22)
m	4,62(0,23)	n	4,60(0,23)	ɲ	4,67(0,24)	ʀ	4,49(0,31)	l	4,48(0,26)		

la composent. Cependant certaines différences par rapport à l'algorithme de Campbell doivent être mises en relief.

Le groupe concerné dans notre travail n'est pas la syllabe mais le GIPC (voir chapitre précédent pour sa définition). De par sa propre définition, ce groupe peut contenir au long d'une phrase des éléments appartenant à une autre syllabe. Par exemple, à partir de la suite de réalisations phonémiques /ʀaplemøsjødypɔ̃asɔ̃byʀo/ (« Rappelez Monsieur Dupont à son bureau.»), les GIPC sont, successivement : /apl/, /em/, /øsj/, /ød/, /yp/, /ɔ̃/, /as/, /ɔ̃b/, /yʀ/. Le dernier GIPC débuterait avec le phonème /o/ mais on ne peut déterminer sa fin puisque il est en fin de phrase (il n'a pas un *onset* vocalique à droite pour que l'on puisse établir sa deuxième frontière . Dans des phrases isolées, le dernier GIPC s'arrête toujours au dernier *onset* vocalique). Le fait de ne pas prendre la rime du dernier GIPC de la phrase ne pose pas de problème particulier pour la génération automatique si l'on prend en compte la grande variabilité de celle-là en production, comme on verra dans la section 4.5.

L'équation pour déterminer le $z-score$ associé au GIPC courant est donnée dans notre cas par :

$$Durée(GIPC) = \sum_{i=1}^{n} \exp(\mu_i + z\sigma_i) \tag{4.1}$$

avec la paire (μ_i,σ_i) correspondant au phonème i et prise dans le tableau 4.2. La somme concerne dans notre cas tous les phonèmes du GIPC correspondant. Les $z-scores$ ne seront pas modifiés selon la position du phonème dans le groupe pour les raisons déjà évoquées précédemment.

4.3.1 Test de l'hypothèse d'élasticité

Un test de l'hypothèse d'élasticité a été réalisé sur le corpus *Bureau* puisque il possède une richesse plus grande de phrases à un même débit que celle du corpus *PCenter* (plus de signifiance pour le test statistique de différence entre les distributions) ainsi que la présence de réalisations dans les trois catégories testées par Campbell : *attaque*, *noyau* et *coda*.

Analyse par comparaison entre moyennes

De façon à tester la pertinence de l'hypothèse d'allongement uniforme au sein du GIPC (hypothèse forte d'élasticité), nous avons tout d'abord refait l'analyse à partir des moyennes par catégorie phonémique déjà réalisée par Campbell et présentée dans la section 2.3.2 au chapitre 2. Nous avons jugé plus pertinent de faire une telle analyse sur le propre corpus dont on se propose de générer automatiquement les durées phonémiques.

Tous les phonèmes du corpus *Bureau* ont été divisés en trois catégories. Pour chaque syllabe, tous les phonèmes précédant la voyelle (semi-voyelles incluses) ont été marqués *attaque*[3], la voyelle a été marquée *noyau* et les consonnes la suivant, *coda*. Le $z-score$ réel de chaque réalisation phonémique est calculé à partir de l'équation

$$durée\,réalisation = \exp(\mu_{phonème} + z\sigma_{phonème}) \tag{4.2}$$

En isolant la valeur de z pour la réalisation, on obtient :

$$z_{réalisation} = \frac{(\ln(durée\ réalisation) - \mu_{phonème})}{\sigma_{phonème}} \tag{4.3}$$

Les syllabes ont été catégorisées selon leur distance par rapport à la distribution des $z-scores$ du corpus. La distribution des $z-scores$ réels ayant moyenne μ_z et écart-type σ_z (pour le corpus rapide, $\mu_z = -0,42$ et $\sigma_z = 0,02$ et pour le normal, $\mu_z = 0,27$ et $\sigma_z = 1,16$), les syllabes ont été catégorisées en longues (quand la moyenne pondérée entre les $z-scores$ des ses composants est supérieure à $\mu_z + \sigma_z$), courtes (quand cette moyenne est inférieure à $\mu_z - \sigma_z$), normales (quand cette moyenne se trouve entre $\mu_z - \sigma_z$ et $\mu_z + \sigma_z$) et finales (pour la dernière syllabe de chaque phrase).

On a calculé ensuite les paramètres statistiques des distributions des *z–scores* réels (équation 4.3) par catégorie phonémique et pour chaque type de syllabe. Les résultats se trouvent dans les tableaux 4.3 et 4.4 pour chaque débit pour les syllabes longues, courtes et finales. Dans les deux débits les syllabes d'allongement normal sont de loin les plus nombreuses. Pour ce cas, les trois catégories phonémiques ont des moyennes statistiquement égales avec $p < 10^{-4}$.

3. Cette étiquette désigne ainsi les consonnes et semi-voyelles appartenant à l'attaque. Deux consonnes appartenant à l'attaque seront étiquetées, chacune, par le mot *attaque*. Même observation pour l'étiquette *coda*.

TAB. 4.3 - *Z−scores moyens pour les syllabes longues, courtes et finales dans le débit* **normal** *par type de réalisation phonémique.*

	longues			*courtes*			*finales*		
type phon.	moy.	éc.-t.	n	moy.	éc.-t.	n	moy.	éc.-t.	n
attaque	1,63	1,15	66	-0,84	1,23	67	-0,24	1,06	89
noyau	2,01	1,20	71	-1,46	0,97	92	0,11	1,34	88
coda	0,76	0,79	23	-1,03	0,68	3	2,87	0,99	24

TAB. 4.4 - *Z−scores moyens pour les syllabes longues, courtes et finales dans le débit* **rapide** *par type de réalisation phonémique.*

	longues			*courtes*			*finales*		
type phon.	moy.	éc.-t.	n	moy.	éc.-t.	n	moy.	éc.-t.	n
attaque	-0,20	0,24	79	-0,69	0,04	70	-0,46	0,12	87
noyau	0,00	0,27	90	-0,68	0,09	43	-0,12	0,24	89
coda	-0,18	0,24	6	-0,71	0,13	3	0,20	0,36	25

Discussion

Pour le débit normal, le *coda* est distinct de l'*attaque* et du *noyau* avec $p < 10^{-4}$ dans les syllabes longues. Ceci se reproduit dans les syllabes finales où la catégorie *coda* est nettement plus allongée que les deux autres, qui ne peuvent pas être distinguées statistiquement. Dans les syllabes courtes, la situation est plus compliquée que dans les résultats de Campbell puisque dans la catégorie *attaque* le $z-score$ moyen est distinct des ceux des catégories *noyau* et *coda* ($p < 10^{-3}$).

Pour le débit rapide, le *coda* et le *noyau* sont distincts de l'*attaque* avec $p < 10^{-3}$ seulement dans les syllabes finales (ceci rejoint les résultats de Campbell). Dans les autres types de syllabe, les trois catégories semblent constituer un groupe homogène.

Certaines caractéristiques comme l'allongement important du coda peuvent être dues à la spécificité du corpus que possède une majorité de /s/ dans cette position. Cependant, cet effet d'allongement est intéressant à noter dans les deux débits. Dans le débit rapide, p. ex., la rime (*noyau* + *coda*) pourrait constituer une seule unité soumise à l'allongement en final de phrase. Une chose intéressante à noter est que le débit rapide semble plus homogène (voir syllabes courtes ou longues) que le débit normal, attestant la diminution des effets d'allongement dus à l'accentuation.

À notre avis, la méthode utilisée par Campbell — calculant la moyenne entre toutes les réalisations phonémiques d'une catégorie et pour un type de syllabe donné — n'est pas la meilleure méthode pour vérifier la validité de l'hypothèse d'élasticité uniforme au sein de la syllabe puisque il fait intervenir l'allongement de réalisations appartenant à des syllabes différentes. On peut très bien avoir, par exemple, des moyennes semblables entre *attaque* et *noyau*, même en ayant tous les allongements en position *attaque* décorrélés des allongements en position *noyau* dans une même syllabe. De plus, les différences entre les moyennes n'étant pas toujours significatives vu le nombre de valeurs pour le calcul du test de signifiance (surtout en position de coda), il faut avoir recours à une analyse plus fine sur les données. Ainsi, nous avons été conduit à utiliser la corrélation entre les z–*scores* des réalisations phonémiques adjacentes au long de chaque phrase. Les résultats de cette analyse nous ont permis de choisir le GIPC et rejeter la syllabe comme unité de programmation rythmique.

Analyse par corrélation

Ces corrélations ont été calculées par paire de classe du type (*catégorie du phonème courant, catégorie du phonème suivant*), entre les z–*scores* réels associés aux réalisations appartenant à chacune des catégories de la paire. Les catégories sont étiquetées au sein de chaque syllabe de la même façon que dans l'analyse par moyenne : *attaque* (a), *noyau* (n) ou *coda* (c). Ainsi, dans la séquence /akɛ̃zœʀdisʒødi/ (À quinze heures dix jeudi) les réalisations sont respectivement catégorisées comme /nanancancanan/ (la consonnne de la liaison est difficile à étiqueter : nous avons choisir de la considérer comme formant l'attaque d'une nouvelle syllabe) et on peut former les classes *na*, *an*, *nc* et *ca*. Dans la classe *na*, p. ex., se trouvent réunis tous les points sur un espace bidimensionnel définis en abscisse par le $z-score$ d'une réalisation catégorisée n et en ordonnée par le $z-score$ de la réalisation adjacente qui la suit et catégorisée a.

La corrélation calculée à l'intérieur de chaque classe permet d'établir si les $z-scores$ de cette classe sont corrélés ou non entre eux. Il convient de noter que, pour chaque paire de valeurs d'une classe, il est possible de dire si les réalisations appartiennent à la syllabe ou à un autre groupe. Les classes possibles sont *aa*, *an*, *nc*, *cc*, *na*, *ca*, *nn* et *cn*. La classe *ac* est impossible, puisque entre une consonne du type *attaque* et une consonne du type *coda* il y a au moins une voyelle. Dans les classes *na*, *ca*, *nn* et *cn*, il y a une frontière syllabique, mais il n'y en a pas dans *aa*, *an*, *nc* ou *cc*.

Les classes *na*, *ca*, *aa*, *nc* et *cc* regroupent des phonèmes appartenant au même GIPC tandis que dans les classes *aa*, *an*, *nc* et *cc* les phonèmes appartiennent à une même syllabe. Les paires qui n'appartiennent qu'à un seul type de groupe sont *na*, *ca* (GIPC) et *an* (syllabe). La corrélation entre ces classes sera décisive pour choisir entre syllabe ou GIPC.

Les résultats des corrélations pour les deux débits sont présentés ci-dessous.

TAB. 4.5 - *Corrélation par méthode de Pearson entre $z-scores$ réels adjacents par paire (catégorie réalisation courante,catégorie réalisation suivante) pour le* **débit rapide**. *Les probabilités indicant la signifiance des résultats sont données entre parenthèses (pour $\alpha = 5\%$). Les classes qui n'appartiennent qu'à un seul groupe (syllabe ou GIPC) ont été mis en caractères gras.*

	réalisation suivante		
réal. cour.	attaque	noyau	coda
attaque	0,48($p \leq 10^{-4}$)	**-0,02**($p < 0,5$)	-
noyau	**0,33**($p \leq 10^{-4}$)	0,20($p < 0,3$)	0,43($p \leq 10^{-4}$)
coda	**-0,13**($p < 0,3$)	-0,04($p < 0,9$)	-

TAB. 4.6 - *Corrélation par méthode de Pearson entre $z-scores$ réels adjacents par paire (catégorie réalisation courante,catégorie réalisation suivante) pour le* **débit normal**. *Les probabilités indicant la signifiance des résultats sont données entre parenthèses (pour $\alpha = 5\%$). Les classes qui n'appartiennent qu'à un seul groupe (syllabe ou GIPC) ont été mis en caractères gras. La catégorie* **coda/noyau** *n'est pas représentée ici parce qu'après tous les* codas, *il y avait un silence qui appartenait à une quatrième catégorie de réalisation.*

	réalisation suivante		
réal. cour.	attaque	noyau	coda
attaque	-0,18($p < 0,05$)	**-0,10**($p < 0,01$)	-
noyau	**0,37**($p \leq 10^{-4}$)	0,19($p < 0,7$)	-0,10($p < 0,4$)
coda	**0,02**($p < 1$)	-	-

Discussion

Dans les deux débits de ce corpus il n'y a pas de classe *cc*. Si l'on considère l'ensemble des classes qui prônent le GIPC (*na*, *ca*, *aa*, *nc* et *cc*), on obtient des corrélations de 0,15 pour le débit normal et 0,26 pour le rapide, les deux avec $p \leq 10^{-4}$. Pour la syllabe (*aa*, *an*, *nc* et *cc*) les corrélations sont de -0,12 (débit normal) et 0,00 (débit rapide), $p \leq 10^{-4}$.

En regardant les tableaux 4.5 et 4.6, on voit bien que les $z-scores$ des réalisations vocaliques (catégorisées n) sont corrélés positivement avec les $z-scores$ des réalisations consonantiques qui les suivent appartenant à une autre syllabe (catégorie a) : $0,33$ dans le débit rapide et $0,37$ dans le débit normal, ces deux valeurs étant statistiquement significatives. Au sein d'une même syllabe, la corrélation entre les $z-scores$ des réalisations adjacentes (catégories a et n) est négative, quoique possédant une signifiance moins importante. La classe *ca* n'est pas suffisamment significative pour que l'on puisse faire une analyse correcte. Il est intéressant de noter que, dans le débit rapide, les réalisations de l'attaque sont corrélées positivement $(0,48, p \leq 10^{-4})$ de façon significative, ce qui démontre la cohérence du groupe des consonnes en début de syllabe ou en fin de GIPC (rappelons que la voyelle est corrélée à la consonne que la suit).

Si l'on résume ces résultats à l'aide du schéma donné dans la figure (4.1), l'hypothèse d'élasticité sera plus valable au sein d'un groupe délimité entre les débuts consécutifs de deux voyelles. Considérant une analyse en absence de silence interposé (le débit rapide n'en possède pas), ce groupe est le GIPC (la syllabe ne peut pas servir puisque la corrélation entre les réalisations consonne et voyelle consécutives est négative).

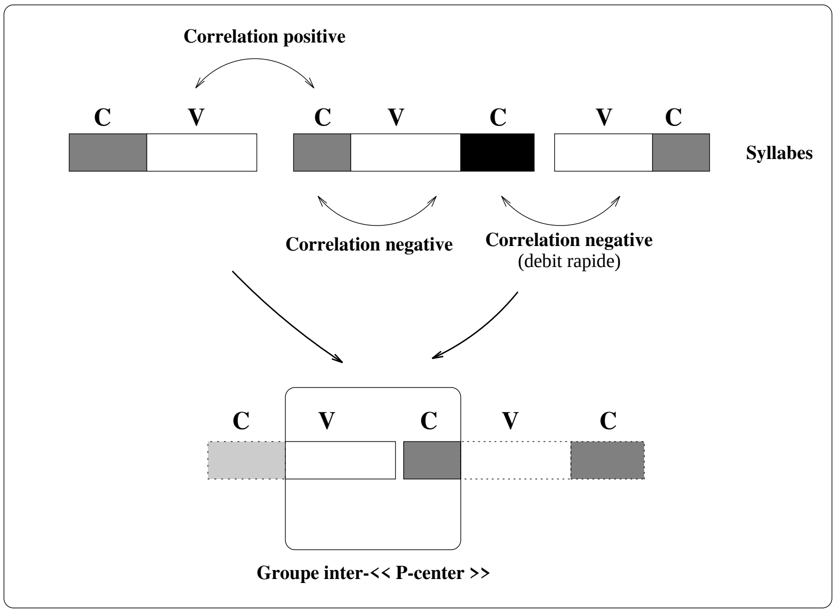

FIG. 4.1 - *Émergence du GIPC, le meilleur groupe avec lequel le modèle de répartition sera plus efficace. Ce schéma résume les résultats des analyses de corrélation effectuées.*

D'autres résultats vont aussi prôner le choix du GIPC comme unité de programmation rythmique.

4.3.2 Corrélation entre débits

Ce test a été réalisé avec les durées observées du corpus *PCenter* dans les cinq débits. Il s'agit d'étudier la corrélation entre les durées observées de deux types de groupe, la syllabe et le GIPC. Entre chaque paire de débits dont la corrélation est calculée, seuls les groupes qui ont la même position dans la phrase et formés des mêmes phonèmes sont concernés. Quand le GIPC à un débit donné possède un phonème de moins par rapport au GIPC correspondant dans un autre débit, ce groupe n'est pas pris en compte dans le calcul de la corrélation. Tous les groupes ont été étiquetés *accentué* ou *inaccentué*. Un groupe accentué correspond à une frontière syntaxique, à laquelle une marque est associée. Le groupe au débit *très lent* qui lui correspond (même phrase, même position) possède nécessairement une pause silencieuse associé à la réalisation de l'accent. Les groupes délimités par deux accents successifs est appelé groupe accentuel ou mot prosodique (Martin 1982).

Dans le tableau 4.7, les corrélations concernent le GIPC. Dans le tableau 4.8, elles concernent la syllabe.

Il sort des tableaux précédents que les corrélations entre durées pour les groupes accentué et inaccentué sont plus élevées si le GIPC est pris comme unité de programmation rythmique.

TAB. 4.7 - *Corrélation entre débits en considérant le GIPC comme unité de programmation rythmique. Corrélations entre débits adjacents en caractères gras.*

	Accentué				*Inaccentué*			
débit	très lent	lent	moyen	rapide	très lent	lent	moyen	rapide
lent	**0.76**	1.00	0.84	0.82	**0.86**	1.00	0.90	0.84
moyen	0.70	**0.84**	1.00	0.84	0.83	**0.90**	1.00	0.92
rapide	0.76	0.82	**0.84**	1.00	0.72	0.84	**0.92**	1.00
très rapide	0.72	0.61	0.71	**0.81**	0.43	0.59	0.73	**0.89**

TAB. 4.8 - *Corrélation entre débits en considérant la syllabe comme unité de programmation rythmique. Corrélations entre débits adjacents en caractères gras.*

	Accentué				*Inaccentué*			
débit	très lent	lent	moyen	rapide	très lent	lent	moyen	rapide
lent	**0.63**	1.00	0.80	0.77	**0.5**2	1.00	0.72	0.64
moyen	0.72	**0.80**	1.00	0.86	0.47	**0.7**2	1.00	0.80
rapide	0.63	0.77	**0.86**	1.00	0.44	0.64	**0.80**	1.00
très rapide	0.53	0.72	0.75	**0.75**	0.43	0.66	0.73	**0.87**

Ceci est d'autant plus vrai pour les groupes inaccentués (comparez les diagonales des groupes inaccentués dans les tableaux précédents), justement où les effets de différence de comportement pour l'allongement de l'attaque et de la rime sont moins importants au sein de la syllabe, par rapport aux mêmes effets en position accentuée. Ceci vient confirmer la cohérence plus grande entre les réalisations au sein du GIPC. Les corrélations pour les groupes accentués (comparez les diagonales des groupes accentués dans les tableaux précédents) sont équivalentes, à l'exception de celle entre débits *lent* et *très lent*. Ceci est probablement dû à la séparation entre attaque et rime pour la syllabe : quand la phrase ralentit, l'effet s'intensifie et fait baisser la corrélation.

Le phénomène pausal explique les corrélations supérieures du GIPC inaccentué par rapport à l'accentué. Ce résultat préconise le choix de la durée du GIPC inaccentué comme période d'une horloge interne qui assurerait le maintient d'un cycle isochrone qui ne serait brisé que par l'accent (avec d'éventuels phénomènes d'alternance de durée courte et longue que l'on trouve dans la littérature pour la syllabe [4]). En indiquant que quelque chose est en train de se passer, par une atteinte à l'isochronisme, l'accent assurerait une fonction linguistique.

Les signifiances plus basses entre les débits *très lent* et *rapide* ou *très rapide* ou entre les débits *lent* et *très rapide* ne nous autorisent pas à expliquer les valeurs inférieures des corrélations pour les groupes accentué versus inaccentué dans le GIPC ou la syllabe.

D'autres résultats puisés dans la littérature viennent renforcer notre choix du GIPC comme unité de structuration rythmique.

Campbell (1993) étudie le comportement de l'attaque et de la rime en frontière prosodique (voir section 2.3.2). Les durées observées des réalisations des phonèmes de l'attaque sont sur-prédites (la durée prédite est très supérieure en moyenne à la durée observée. Ce n'est pas le cas de la rime dont la durée est bien prédite en moyenne) par rapport à ceux de la rime, suggérant que l'hypothèse forte d'élasticité n'est pas valable : la rime s'allonge de façon plus importante que l'attaque. Il est donc clair qu'un groupe cohérent, notamment en position de frontière syntaxique, doit commencer à l'*onset* vocalique.

Le choix de la syllabe par Campbell a été justifié par les travaux de Kozhevnikov & Chistovich (1965) or, selon Allen (1973, p. 234), « leur modèle statistique n'a pas été suffisamment protégé des erreurs de mesure (Ohala 1970) et ainsi leur revendication que la syllabe CV soit une unité temporelle en russe doit être vue comme non prouvée. Des études similaires en anglais souffrent du même problème (Lehiste 1970; Allen 1970). »

4. Ces alternances ont été constatées avec une approche syllabique — on peut même la voir en utilisant des durées normalisées, comme dans la figure 4.5 — et la plupart du temps par l'usage de la durée intrinsèque des syllabes. On peut voir ces alternances, par exemple, chez Pasdeloup 1992.

Les résultats de plusieurs autres auteurs concernant la rime en tant qu'unité au sein de laquelle l'allongement a lieu confirme notre choix (Klatt 1975; Vaissière 1983; Duez 1987; Edwards & Beckman 1988; Crystal & House 1990; Wightman, Shattuck-Hufnagel, Ostendorf & Price 1992). Dans la contribution de Vaissière, où elle résume les recherches effectuées en quelques domaines prosodiques comme la variation de la courbe de fréquence fondamentale, l'intensité, la pause et l'allongement final, ce dernier semble être porté principalement par la voyelle de la syllabe en final de syntagme (1983, p. 62). Wightman et ses collègues ont étudié le degré d'allongement des réalisations phonémiques en voisinage de frontière prosodique et montrent qu'il se restreint à la rime de la syllabe précédant la frontière et que cette unité est la plus appropriée pour décrire ce phénomène d'allongement.

On a vu dans le chapitre 3 que Pompino-Marschall (1991, p. 110) suggère que la durée du GIPC est la période de référence pour la perception du tempo spontané.

Les résultats déjà montrés au chapitre 3 sur les *beats* en parole (Allen 1975; Fraisse 1974) favorisent aussi la durée du GIPC comme référence du rythme.

Hirst (1993, p. 34) affirme en outre : « Most of the arguments in favour of the syllable as a unit are really arguments in favour of syllabicity rather than syllabic constituents (...)» (La plupart des arguments en faveur de la syllabe comme une unité sont en fait des arguments en faveur de syllabicité plutôt qu'en faveur de constituants syllabiques (...)). La notion de GIPC garde en soi cet aspect de syllabicité. Dans une communication personnelle[5], Hirst avait calculé la variance pour les durées de la syllabe et du groupe VC dans un texte d'une durée d'environ une minute (« l'atome ») et a trouvé une valeur plus petite pour la variance des unités de programmation commençant à l'*onset* de la voyelle. Nous avons fait la même chose entre la durée syllabique et la durée du GIPC pour le corpus *PCenter*. Les résultats sont présentés ci-dessous, figure 4.2.

On voit bien que, aussi dans notre cas, l'écart-type (189 ms, pour le GIPC contre 220 ms, pour la syllabe) et la fraction $écart - type/moyenne$ (environ 0,6, pour le GIPC contre 0,7, pour la syllabe) sont plus petits pour le GIPC. Ceci est un indicateur de la plus grande cohérence de ce groupe rythmique. La conclusion est la même si l'on regarde chaque débit individuellement. Par contre, pour les positions inaccentuées, tous débits confondus, on obtient des écarts-type de 112 ms pour le GIPC contre 78 ms pour la syllabe (fractions $écart - type/moyenne$ de respectivement 0,5 et 0,4 pour les groupes

5. « Avez-vous essayé simplement de comparer la variance brute des syllabes d'un texte avec la variance des groupes IPC ? Je l'ai fait sur les 3/4 du texte « l'atome» sur lequel on a travaillé et effectivement la variance est nettement moins élevée pour les inter-Pivots que pour les syllabes — normal si les pivots sont les points de synchronisation. » Message électronique adressé à Barbosa et Bailly le 1er décembre 1992. En ce qui concerne les pivots il fait référence au travail de Dogil dont il sera question plus loin.

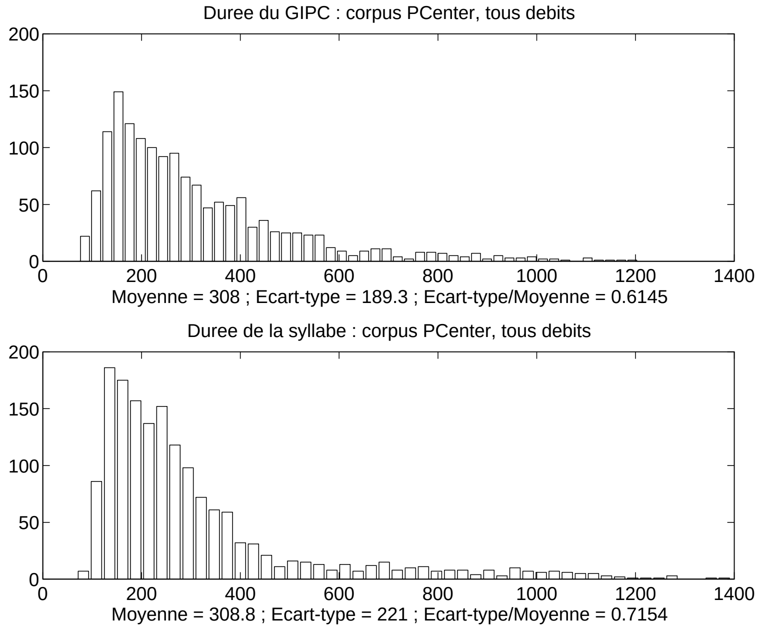

FIG. 4.2 - *Histogrammes des durées du GIPC (en haut) et de la syllabe (en bas) pour le corpus* PCenter*, tous débits confondus. Le silence est pris en compte, que ce soit pour la syllabe ou pour le GIPC.*

GIPC et syllabe). Ceci veut dire que la différence plus importante entre la syllabe et le GIPC se trouve en position accentuée, où la différence d'allongement entre attaque et rime est bien connu, comme déjà discuté plus haut.

Les études sur le décodage de l'information par le mécanisme de perception viennent apporter un renfort pour le choix d'un groupe délimité par les transitions CV.

4.3.3 L'Apport du décodage de la parole

Quelques résultats de la perception de parole véhiculent le même message : la transition CV est un lieu privilégié pour le décodage de l'information spectrale.

L'étude réalisée par Stevens & Blumstein (1978) concernant l'identification du lieu d'articulation des consonnes occlusives voisées (/b/, /d/ et /g/) en contexte vocalique (/i/, /ɑ/ et /u/) suggère que ces consonnes peuvent être identifiées sur la seule base de l'information spectrale au voisinage du relâchement de la consonne, indépendamment de la nature de la voyelle. Ceci implique que le système auditif possède un mécanisme de détection de la présence d'un *onset* d'énergie brusque dont on trouve la trace dans la région spectrale supra-citée.

Dans sa thèse, Morris (1992) corrobore les résultats de Stevens & Blumstein en affirmant que « les pics d'information pour distinguer entre les trois plosives voisées [b,d,g] (qui ont des lieux d'articulation différents) se produisent autour des positions de fermeture de la consonne et de déclenchement de l'explosion. »

Les données sur la perception des voyelles isolées suggèrent que leur identification est basée sur l'extraction d'événements sur le signal acoustique associés aux maxima spectraux (Chistovich 1980). Il s'agit d'une intégration temporelle de l'information distribuée dans des points privilégiés, comme atteste l'existence de neurones qui ne répondent qu'à des événements du type *onset* (Chistovich & Ogorodnikova 1982; Creutzfeldt, Hellweg & Schreiner 1980).

La théorie du *pivot* de Grzegorz Dogil (1988, 1987) propose la transition CV comme lieu privilégié pour le décodage de l'information segmentale. Il met en avant la structure CV comme syllabe prototypique en présentant des arguments tels que sa présence dans toutes les langues (Maddieson 1983), son acquisition avant toute autre dans le processus d'apprentissage de la langue (Allen 1981; Locke 1983) ou sa préservation dans des cas sérieux d'aphasie (Dogil 1984).

Ce que Dogil utilisera en réalité pour réaliser son *parser* c'est ce qu'il appelle la *représentation diphonique* de la syllabe CV, une représentation spectrale à partir de la transition consonne-voyelle (leur frontière acoustique, pour reprendre le terme de l'auteur : 1988, p. 15). C'est la région autour de l'*onset* de la voyelle qui possède le plus d'information (Morris 1992, pour la transition consonne occlusive - voyelle) : le *parser* saute de transition CV en transition CV. Ceci n'est rien d'autre qu'une version simplifiée de la séquence des *P-centers*. Ceci est corroboré par Dogil lui-même qui cite les travaux de Marcus sur le *P-center*. D'autres arguments de Dogil pour la syllabe CV ne sont rien d'autre que des arguments pro-*P-center* : « quand les indices de lieu d'articulation sont différents aux transitions VC et CV, les auditeurs ont la tendance à suivre les indices de la transition CV. » ou « les locuteurs tentent de créer des articulations plus bien définies, plus précises temporellement à la proximité de la transition CV et non de la transition VC. » (Dogil 1987, p. 361).

Des différences entre la caractérisation de la structure rythmique des phrases en français en utilisant soit la syllabe soit le GIPC seront encore présentées dans la section suivante, mais les résultats que nous avons présenté précédemment et ces quelques exemples puisés dans la littérature nous ont convaincu d'adopter le GIPC comme unité prosodique minimum pour la caractérisation de la structuration rythmique en français.

4.4 Étude de la structuration rythmique des corpus

Wandelt oben des Lichtes Fluren
Göttlich unter Göttern die Gestalt

Marche parmi les champs lumineux
Divine entre les déesses, ô forme
SCHILLER, *Das Ideal und das Leben (Strophe III), Gedankenlyrik*

Cette section est dédiée à l'analyse du corpus *Bureau* à partir du calcul des *z–scores* des GIPC successifs de chaque phrase. Ces analyses nous ont permis de dégager certains aspects de la structure accentuelle de la phrase lue en français.

Il convient d'exposer avant tout quelques aspects du gestaltisme qui nous seront utiles par la suite.

4.4.1 De la « Gestalttheorie»

Le gestaltisme ou théorie de la forme (*Gestalttheorie*) est un système psychologique qui a connu une grande popularité depuis ses origines. L'une des raisons de son succès est due au fait que cette école a énoncé et diffusé des concepts assez clairs à première vue, n'exigeant sur le plan du raisonnement qu'un minimum d'information technique (Thinès 1990, p. 434 c).

Les théoriciens du gestaltisme ont utilisé des exemples dans le domaine de la perception visuelle : les principales figures sur lesquelles ils fondaient leur analyse appartiennent en majorité au domaine des illusions optico-géométriques (*ibid.* p. 435 a). Les principes fondamentaux de cette pensée sont basés sur les notions développées dans les travaux de psychologues expérimentaux comme Wilhelm Wundt et Franz Brentano et de l'école phénoménologique fondée par le mathématicien philosophe Edmund Husserl.

Étudiant la notion d'ensemble, ce dernier a contribué à l'élaboration du concept de forme. « Envisageant l'intervention de processus symboliques dans la perception des ensembles numériques, Husserl remarque que ceux-ci s'imposent dès que le dénombrement devient impossible. Cette limite est atteinte à partir d'une dizaine d'objets. Au-delà, c'est l'*unité figurale de la totalité*[6] qui est perçue. Or celle-ci n'est pas réductible à la somme de ses parties constitutives, lesquelles fusionnent de manière à donner naissance au caractère figural.» (*ibid.* p. 437 a)

Pour la *Gestalttheorie*, la forme est perçue immédiatement. Les lois de cette structuration perceptive sont basées sur la différenciation entre *fond* et *forme*. En voici quelques-unes (*ibid.* p. 437 c) :

– la forme est nettement distincte du fond ;

6. C'est nous qui soulignons.

- la forme est close et structurée. C'est à elle que le contour semble appartenir ;
- son émergence dépend des caractères objectifs de structuration (relations géométriques, relations de contraste, etc) ;
- l'ensemble détermine les caractéristiques phénoménales des parties et réciproquement.

Thinès continue : « En résumé, la partie du champ [visuel] qui est vue comme forme est celle qui est délimitée phénoménalement par un contour précis et *retient l'attention.*» Nous avons souligné la notion d'attention, qui nous semble importante, signalant un mécanisme qui ne se construit pas seulement à partir des données elles-mêmes, comme pourrait penser un empiriste, mais avec le concours des processus cognitifs de l'individu percevant.

Concernant le domaine auditif, le psychologue Katz écrit : « Le son est immédiatement présent et s'offre à l'auditeur. Mais on ne peut pas en dire autant des relations entre les notes ; celles-ci doivent être recherchées. De toute évidence, la forme sonore ne saurait être identique à la compréhension de la position relative de ses notes constitutives. » (Katz 1950) Il donne un exemple très illustratif de cette notion de forme appréhendée immédiatement : les lettres P et D ne sont jamais perçues dans la lettre R bien qu'elles soient incluses comme constituants partiels.

Nous concluons avec Thinès (*ibid.* p. 438 a) : « (...) les formes qui se dégagent de façon élective dans un contexte particulier constituent un résultat **adaptatif** optimal. On parlera à leur sujet de « bonnes formes » ou de « formes prégnantes » . La prégnance est considérée par les gestaltistes comme l'expression des capacités auto-régulatrices de l'organisme ; elle intervient, selon eux, dans tous les processus qui régissent les relations entre l'organisme et le milieu.»

Il nous paraît fondamental de conclure que l'unité de la forme émergente (appelée *Fugengestalt* par Fraisse, comme on a vu au chapitre 3) est obtenue avec le concours de ses parties constitutives par le truchement des processus cognitifs de l'individu déclenchés ou assurés par un mécanisme attentionnel. Cette forme privilégiée, la « forme prégnante» , est structurée selon des lois observables démontrant que **les stimuli individuels ne sont pas indépendants les uns des autres dans le mécanisme de perception mais contribuent de façon imbriquée à la perception de la globalité**. Le paradigme gestaltiste est encore vivant et fructueux, comme démontrent les développements de la Théorie des prototypes, dans le domaine de la Sémantique, et la notion de processus holistiques, utilisée dans les Sciences cognitives.

La Sémantique prototypique propose que les exemplaires d'un concept ne sont pas équivalents : certains sont plus représentatifs de la catégorie que d'autres (Rosch 1973). Ces exemplaires qui sont plus typiques peuvent servir de **prototype** de la classe. Les prototypes sont les exemplaires les plus semblables aux exemplaires de la classe et les moins semblables aux exemplaires des autres classes. Dans ce sens on voit clairement l'émergence d'une forme (associée au sémème du prototype) qui se dégage du fond (associée au sémème des autres classes).

Gardner (1974) distingue les processus perceptifs analytiques, traitant des dimensions physiques séparables, des processus holistiques, analysant de manière globale et synthétique les représentations. L'analyse holistique est de ce fait directement liée à la notion de forme introduite par le gestaltisme. Elle présente l'avantage d'être sensiblement plus rapide (cf. Cooper 1982, pour la vision et section 5.4.1, au chapitre 5, pour la durée). Cette notion rencontre une plausibilité neurophysiologique dans le travail de Zaidel (1984).

En ce qui concerne les lois de la structuration perceptive, nous verrons comment elles peuvent être mises en relation avec les patrons rythmiques que l'analyse par les $z\!-scores$ a pu dégager.

4.4.2 Caractérisation de la structure rythmique du corpus

En résolvant par récursion l'équation donnée par la formule 4.1 du modèle de répartition, il est possible de calculer la séquence des $z-scores$ associés aux GIPC de chaque phrase du corpus. Chaque $z-score$, en tant que durée normalisée, est une mesure de l'allongement global du GIPC correspondant et ainsi, on peut étudier les patrons rythmiques révélés par cette procédure.

Les Patrons accentuels révélés par les $z - scores$

Si l'on représente par un graphique avec, en abscisse, la séquence des GIPC et, en ordonnée, le $z - score$ associé à chaque GIPC, on obtient des patrons du type présenté dans les figures 4.3 et 4.4.

Si l'on regarde les patrons d'allongement au long de chaque groupe terminé par une marque prosodique (position accentuable) on vérifie que :

- les $z - scores$ évoluent fréquemment de façon monotone et croissante ;
- les maxima locaux de cette évolution coïncident toujours avec une marque prosodique ;

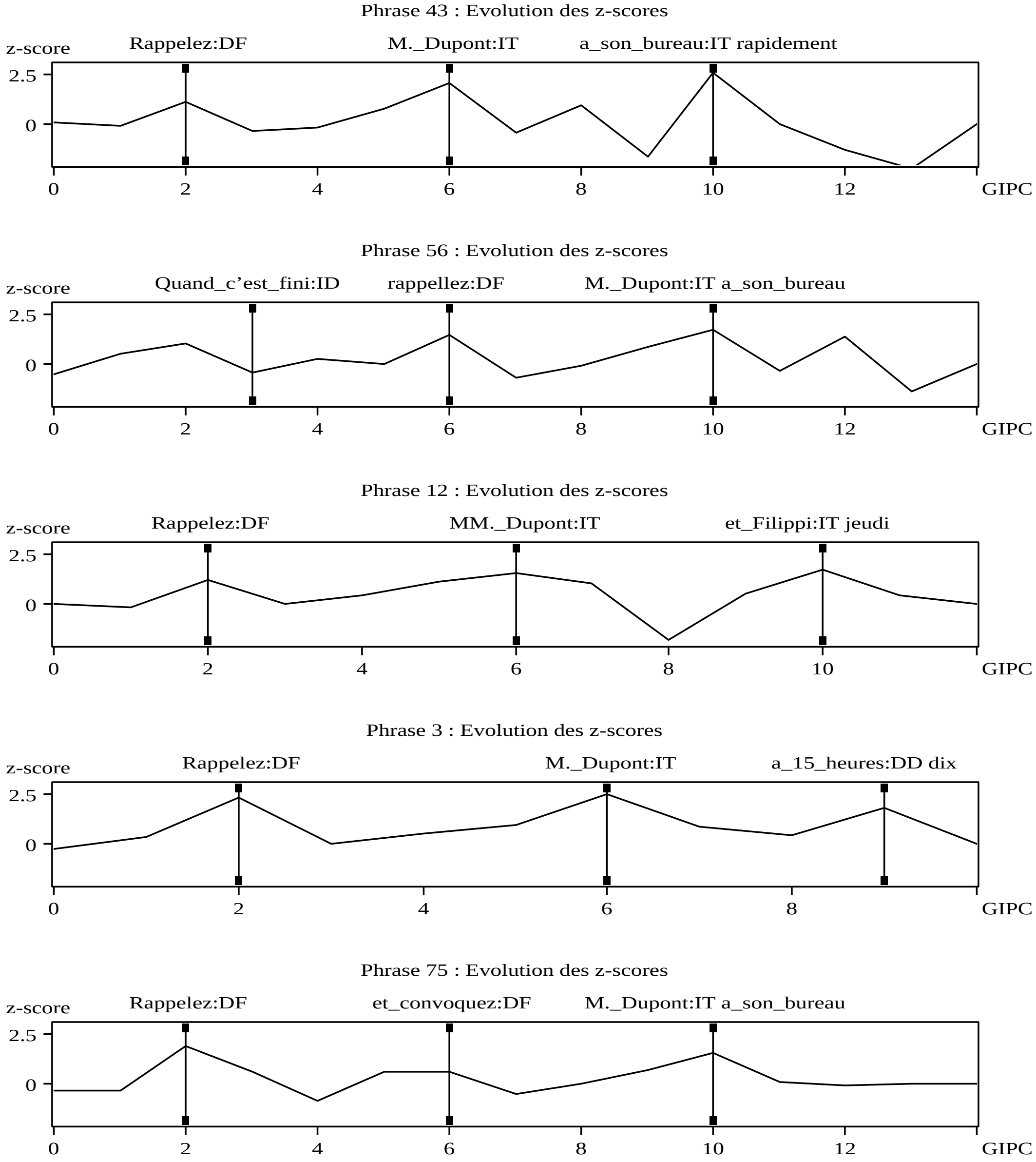

FIG. 4.3 - *Évolution des $z - scores$ dans les phrases 43, 56, 12, 3 et 75 du corpus* Bureau *(voir section A.1 pour le texte), débit normal. Dans ces courbes, les valeurs discrètes des $z - scores$ sont reliées de façon à révéler plus clairement les caractéristiques rythmiques des phrases. Les silences n'ont pas été considérés pour le calcul du $z - score$. Chaque position accentuée a été indiquée par une étiquette indiquant la syllabe concernée dans l'accentuation et la marque prosodique qui lui est associée. Dans les derniers groupes accentuels, le $z - score$ a été mis artificiellement à* 0 *(rappeler que les GIPC commençant à la dernière voyelle ne sont pas pris en compte).*

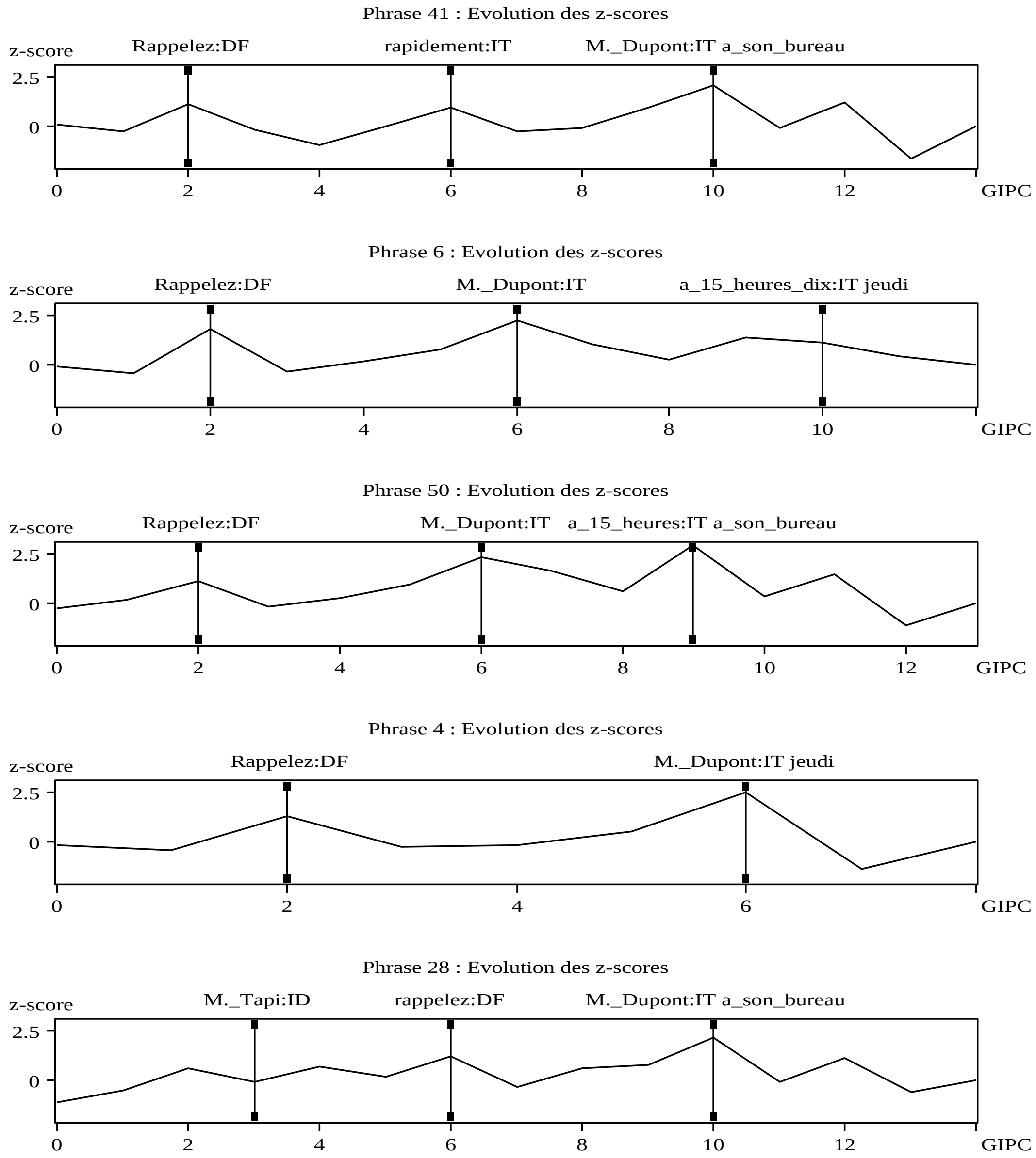

FIG. 4.4 - *Évolution des $z-scores$ dans les phrases 41, 6, 50, 4 et 28 du corpus* Bureau *(voir section A.1 pour le texte), débit normal. Dans ces courbes, les valeurs discrètes des $z-scores$ sont reliées de façon à révéler plus clairement les caractéristiques rythmiques des phrases. Les silences n'ont pas été considérés pour le calcul du $z-score$. Chaque position accentuée a été indiquée par une étiquette indiquant la syllabe concernée dans l'accentuation et la marque prosodique qui lui est associée. Dans les derniers groupes accentuels, le $z-score$ a été mis artificiellement à* 0 *(rappeler que les GIPC commençant à la dernière voyelle ne sont pas pris en compte).*

– après atteindre un maximum, il y a un « reset » et les $z - scores$ repartent d'une position neutre ;

– l'allongement est plus significatif au dernier GIPC, celui qui porte l'accent.

Il est remarquable de constater que les maxima des $z - scores$ suivis du « reset » correspondent toujours avec une marque prosodique correspondant à une position accentuée. L'allongement de la durée est un indice privilégié de l'accent de phrase. La caractérisation par $z - scores$ permet ainsi de délimiter les groupes accentuels (ou mot prosodique : Martin 1982, Llorca 1984).

Cette caractérisation a été effectuée sur tout le corpus *Bureau* et pour le corpus *PCenter*. Pour le premier, 43 % des patrons accentuels sont monotones et croissants et pour le deuxième, 80 % le sont. Pour le calcul de ce pourcentage et pour le relevé des caractéristiques des patrons accentuels, le dernier syntagme de chaque phrase (entre la dernière marque prosodique et la fin de la phrase) n'a pas été pris en compte, car on n'a pas d'information sur le $z - score$ d'un GIPC commençant à l'*onset* de la dernière voyelle de la phrase. Ceci serait possible si les phrases étaient enchaînées (dans un paragraphe, par exemple). Dans ce cas, le point final serait considéré comme une marque prosodique à part entière (indépendance forte).

L'augmentation de la durée au long du groupe accentuel a déjà été constatée par exemple par Vaissière (1980), Llorca (1984), Touati (1987) et Pasdeloup (1992). Les données de Pasdeloup montrent bien une oscillation au début de groupe accentuel, caractéristique d'une approche tenant compte des durées intrinsèques des syllabes ou d'un repère rythmique mal choisi. On peut dire avec Allen (1975, p. 84) que « l'une des fonctions de cette structuration rythmique serait d'organiser l'information apportée par les éléments de la phrase dans un paquet cohérent, permettant ainsi l'efficacité de la communication. »

Cette augmentation continuelle de la durée est aussi décrite par Pasdeloup (1992), qui la fait émerger par moyennage, et va dans le sens de la proposition de Llorca (1984) d'élargir le domaine de l'accent à plusieurs syllabes (*ibid.*, p. 77), selon une suggestion de Garde (1968).

Les patrons non monotones signalent une influence de l'information segmentale dont la caractérisation par un unique $z - score$ par GIPC n'arrive pas à se dégager complètement (on a vu dans le chapitre précédent comment le $z - score$ tenait compte d'effets de coarticulation très connus). Dans la séquence /efilipi/ (et Filippi), par exemple, le GIPC /ef/ a un $z - score$ plus élevé que celui de /il/ ou de /ip/, ce qui va à l'encontre des aspects relevés plus haut. Ceci vient du fait que les réalisations /e/ et /f/ sont très allongées par rapport à celles de /i/, /l/ et /p/. L'écart est encore plus important si l'on avait considéré la syllabe comme unité rythmique (voir contour 8 de la figure 4.5).

Les patrons non monotones peuvent aussi être signe d'une emphase ou d'un accent secondaire du groupe accentuel comme on peut constater dans le groupe /asɔ̃byʀo/ (à son bureau) de la phrase *43*, dans la figure 4.3. L'emphase est fréquente sur un possessif et la place d'un accent secondaire sur l'antépénultième unité du groupe accentuel est une position prévue par les règles rythmiques de Pasdeloup (1987).

Revenons aux groupes accentuels à évolution de $z-score$ monotone ascendante. Cet aspect peut être mis en relation avec les caractéristiques de l'émergence de la forme dans le gestaltisme :

- les phénomènes d'accentuation et de « reset» permettent de délimiter une « forme» structurée, celle-ci étant monotone et croissante ;
- l'émergence de cette forme dépend du contraste entre accent et « reset» ;
- la « forme» peut être associée aux positions successives des *onsets* vocaliques, qui sont les points singuliers de la chaîne vocale, les points d'ancrage de la composante macrorythmique. Le « fond» peut être associé aux phénomènes acoustiques compris entre les *onsets* vocaliques, contrôlées par des variables de la composante microrythmique.

Cette façon de considérer les séquences sonores comme formes privilégiées n'est pas nouvelle, comme le démontrent les travaux de quelques chercheurs (cf. Fraisse 1974 et Llorca 1984, qui cite Guillaume 1979).

Toutefois notre approche permet de dégager la structuration rythmique de façon très simple par l'intermédiaire des « contours de durée» [7] définis par la séquence des $z-scores$.

Il est possible de comparer par cette caractérisation le comportement de la syllabe avec celui du GIPC. Si l'on regarde de près la figure 4.5, on voit que les contours de durée associés à la syllabe présentent des oscillations dues au problème, déjà soulevé, de considérer des réalisations phonémiques qui ne s'allongent pas de la même manière au sein d'un même groupe. La caractérisation par le GIPC semble révéler un allongement continu au long du groupe accentuel, à partir de son début jusqu'à la réalisation de l'accent. Ce résultat confirme la robustesse du GIPC comme unité de programmation rythmique : si le *P-center* est un point d'ancrage d'une horloge interne (cf. section 5.2), il est naturel de penser que l'information pertinente pour le décodage serait délimitée entre deux *P-centers* consécutifs. Le *P-center* serait un point de contrôle qui assurerait la régularité

7. On utilisera le terme de *contour* dorénavant pour se référer à la séquence de valeurs discrètes reliées des $z-scores$ au sein du groupe accentuel.

du cycle de production (et un point privilégié pour le décodage réalisé par le système auditif. Cf. section 4.3.3 plus haut). Le peu de variance montré par les GIPC inaccentués confirme aussi la robustesse d'un groupe compris entre deux *P-centers* consécutifs (cf. section 5.2.1).

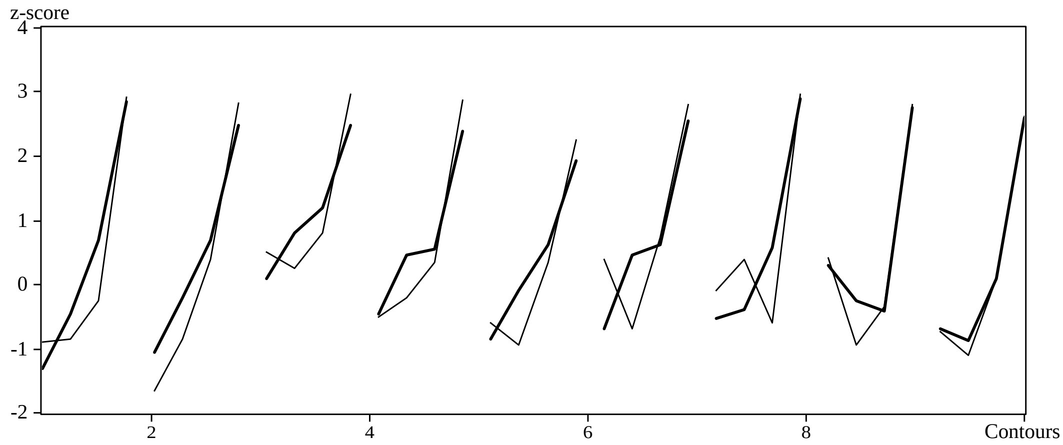

FIG. 4.5 - *Superposition de l'évolution des z−scores pour deux types de groupe : la syllabe (ligne fine) et le GIPC (ligne épaisse). Les z−scores sont des valeurs discrètes mais les points correspondants ont été reliés pour des raisons de visibilité. Notez que le troisième, le cinquième, le sixième et le septième contours sont monotones croissants pour le GIPC et non pour la syllabe. Dans les autres contours, l'évolution des z−scores pour le GIPC est moins brusque. Tous les contours sélectionnés ont 4 syllabes et finissent avec la marque prosodique ID. Notez les problèmes d'oscillation dus au repère syllabique dans les contours 3, 6 et 7 par rapport à la monotonicité des contours obtenus en choisissant le GIPC.*

Une étude de la forme des contours de durée mettant en relation la marque qui les domine et les phonèmes qui sont concernés dans le groupe accentuel a révélé une dépendance résiduelle de l'approche *z−score* à la composante segmentale. Chaque valeur de *z−score* garde en soi une information double : des allongements (ou rétrécissements) de nature microrythmique dus, par exemple, à la coarticulation, comme a été exemplifié dans la section 2.3.2, et des allongements (ou rétrécissements) de nature macrorythmique dus, par exemple, au mécanime d'accentuation. Ce travail a été présenté dans l'article « Prédiction de la durée segmentale : le paradigme du groupe inter-*P-center* » qui est reproduit au final de ce volume, dans lequel une approche par lexique de contours de durées est discutée. Cette étude visait, entre autres, à savoir si l'on pouvait alléger l'entrée du réseau neuromimétique utilisé en génération (et qui sera étudié dans le chapitre suivant) de l'information concernant la nature des phonèmes constitutifs de la phrase. Ceci s'est montré infructueux, démontrant la dépendance de la durée – même normalisée – de la nature du phonème, sinon de l'influence de son contexte.

Un système de production a des liens étroits avec le système perceptif et il était important de tester si le patron accentuel dégagé était nécessaire à la perception de l'accent. Ceci a été étudié par un test de perception expliqué dans la section 4.7.

Si l'on considère uniquement les z–*scores* associés à une marque prosodique, on est capable de dégager la hiérarchie des marques relationnelles de Bailly (cf. annexe C).

La Hiérarchie des marques prosodiques

En calculant la moyenne et l'écart-type des z–*scores* associés au GIPC accentué (systématiquement correspondant à une marque prosodique), on obtient une hiérarchie d'accentuabilité des marques. Ceci a été réalisé dans le corpus *Bureau* pour le débit normal car on pouvait avoir de surcroît le nombre de pauses silencieuses (Grosjean & Dommergues 1983) associées à ces marques, ce qui est un contrôle des résultats présentés dans les tableaux 4.9 et 4.10 et illustrés par la figure 4.6.

TAB. 4.9 - *Moyennes (et écarts-type) des z–scores par marque prosodique (associée à la frontière prosodique) pour le corpus normal. Le pourcentage de marques suivies de pauses silencieuses est aussi donné (avec leur nombre d'occurrences sur le nombre de marques). Pour la signification des marques voir annexe C.*

	Marque				
	ID	IT	DD	DF	DG
z–*scores*	4.4 (0.9)	3.6 (1.1)	1.7 (0.9)	1.3 (0.9)	0.4(0.4)
pauses (%)	96 (26/27)	92 (106/115)	28 (23/81)	19 (17/91)	-

TAB. 4.10 - *Test de différences entre les marques prosodiques selon distribution des z – scores qui leur sont associés. Le symbole* **dl** *est le nombre de degrés de liberté.*

Paire comparée	t-stat	dl	$p <$
ID et IT	3,57	45	0,001
IT et DD	12,3	162	10^{-4}
DD et DF	2,5	147	0,02
DF et DG	8,7	116	10^{-4}

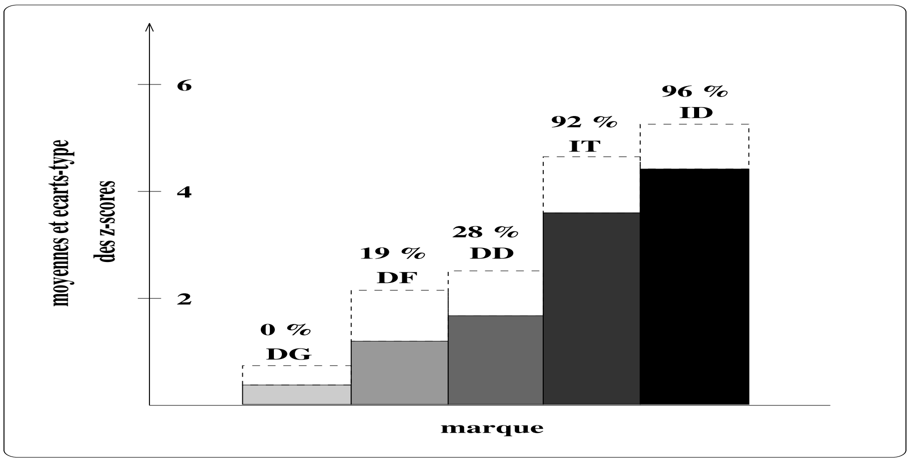

FIG. 4.6 - *Hiérarchie des marques relationnelles de Bailly selon force des $z - scores$ moyens à la frontière prosodique correspondante. Les rectangles au sommet des histogrammes correspondent à l'écart-type et les pourcentages sur les barres verticales correspondent au pourcentage de pauses silencieuses associé à chaque marque. D'après tableau 4.9. Il n'y a pas de marque* **GF** *dans le corpus.*

On voit dans la figure 4.6 que l'ordre de « force » des marques est semblable à celui défini par le système de marques de Bailly (ID, GF, IT, DG, DF, DD, en sens décroissant de force), dans l'annexe C. Étant donné que les différences entre DD et DF sont peu significatives, la seule marque qui semble ne pas suivre l'ordre des marques de Bailly est la marque DG. En fait, Bailly a classé sa « force » selon le lien étroit entretenu par les deux prosodèmes concernés. Dans le cas de la marque DG, les prosodèmes sont difficilement séparables, et la plupart du temps forment un seul groupe accentuel. C'est pour cette raison que le degré d'accentuabilité de cette marque est en fait inférieur à celui des marques DF ou DD. Si l'on souhaite ordonner les marques de Bailly selon ce degré d'accentuabilité, il faut prendre la hiérarchie donnée dans la figure ci-dessus.

La hiérarchie des marques établit une hiérarchisation des accents qui leur sont associés, selon l'observation de Dell (1984, p. 77)[8] : « En français, chaque mot porte un accent dont la force est proportionnelle à l'importance de la coupe syntaxique qui suit ce mot.» ou celle donnée par Vaissière (1983, p. 61) en affirmant d'après les travaux dans le domaine que « si l'allongement relatif de différents éléments est perçu, elle contribue à la récupération de la structure syntaxique par l'auditeur : plus grand est l'allongement, plus importante est la frontière syntaxique. »

Si l'on considère que les marques DF et DD (et éventuellement ID et IT) ne sont pas séparables, statistiquement, on établit par ce fait trois (ou quatre) catégories d'accen-

8. Il cite dans ce passage Kellenberger (1932) et Martin (1973)

tuabilité. Ce résultat peut être comparé avec celui de Wightman et ses collègues (1992) qui, ayant utilisé la même procédure, soutiennent l'existence d'au moins quatre catégories d'allongement en frontière prosodique pour l'anglais. Ceci est corroboré par les études de Ladd & Campbell (1991).

On vérifie de plus que le $z-score$ pour la dernière voyelle de la phrase n'est pas systématiquement allongé, pouvant prendre une gamme de valeurs assez large.

4.5 La Variabilité de la dernière voyelle des phrases

Examinons les figures 4.7 et 4.8 : deux aspects peuvent être comparés entre les *z–scores* réels pour chaque débit et les $z-scores$ des dernières voyelles : d'abord, la forme des distributions, montrant la distribution plus étendue pour les voyelles finales et, ensuite, la valeur de la fraction $écart - type/moyenne$ qui est une mesure de l'extension des d'un histogramme : cette valeur est plus élevée pour l'histogramme correspondant aux $z-scores$ des dernières voyelles. Ceci pourrait suggérer que l'oreille serait moins sensible aux allongements en fin de phrase. Le travail de Nishinuma & Santi (1992) confirme cette hypothèse.

Ces auteurs démontrent que l'atténuation progressive de l'intensité au long de la phrase déclarative joue un rôle significatif et dégrade le seuil de discrimination de durée : « (...) lorsque l'intensité décroît, la détection de la durée totale ne s'effectue pas aussi bien qu'avec un son d'intensité stable et, de ce fait, la comparaison des sons en durée devient moins performante.» (*ibid.* p. 157). Selon les auteurs, Klatt & Cooper (1975) avait trouvé le même phénomène pour les consonnes fricatives en fin de phrase. Pour la génération, la rime finale de chaque phrase peut être générée de manière triviale en affectant la valeur de la moyenne puisée au tableau 4.2, corrigée pour le débit correspondant, comme on verra par la suite.

4.6 De la compensation des effets dus au débit

Wightman et ses collègues (1992) ont aussi étudié le problème que pose la représentation des durées phonémiques en termes de durée intrinsèque : en anglais, une réalisation contenant une voyelle longue sera plus longue qu'une réalisation contenant une voyelle courte dû aux différences entre les durées intrinsèques des réalisations ; « ce serait une erreur de dire que l'une d'elles était allongée. » (*ibid.*, p. 1711). Les auteurs proposent, pour pallier à ce problème, une durée normalisée qui n'est rien d'autre que le $z-score$ réel de la réalisation phonémique. La moyenne et l'écart-type caractéristique de chaque

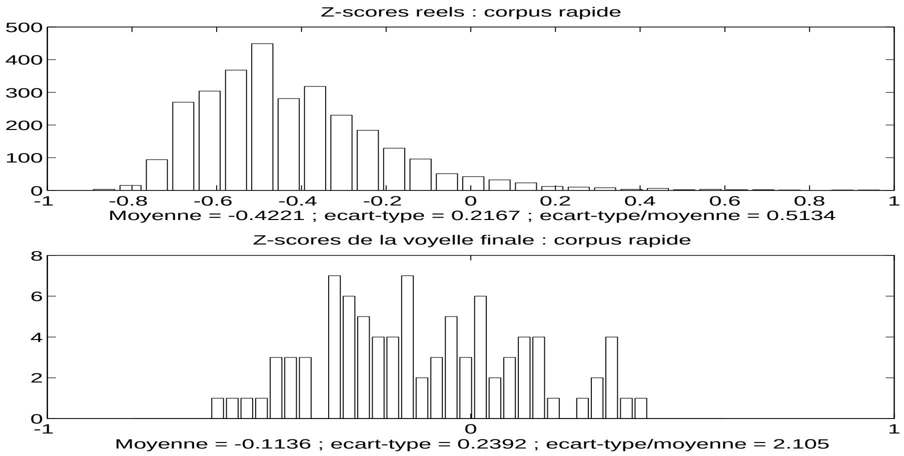

FIG. 4.7 - *Histogramme des $z-scores$ réels du débit rapide (en haut) et histogramme des $z-scores$ de la dernière voyelle des phrases du même corpus (en bas). Les variables m et s sont respectivement la moyenne et l'écart-type des distributions. La valeur s/m est calculé en absolu et mesure l'extension des distributions respectives.*

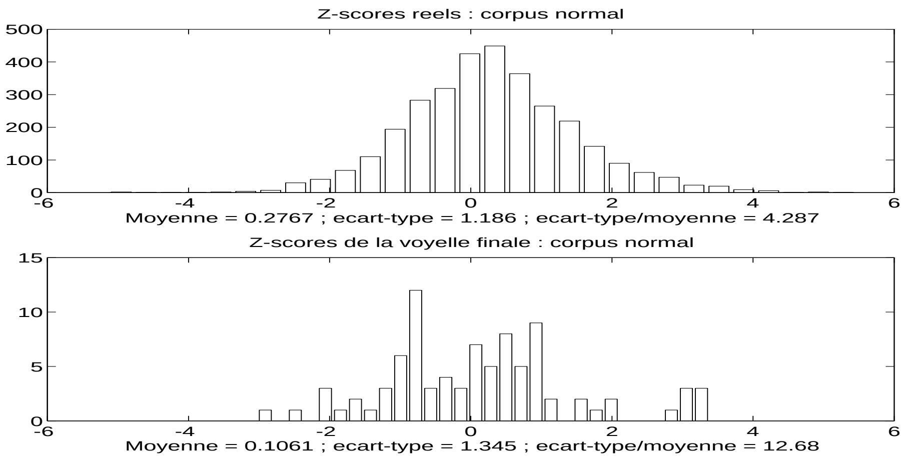

FIG. 4.8 - *Histogramme des $z-scores$ réels du débit normal (en haut) et histogramme des $z-scores$ de la dernière voyelle des phrases du même corpus (en bas). Les variables m et s sont respectivement la moyenne et l'écart-type des distributions. La valeur s/m est calculé en absolu et mesure l'extension des distributions respectives.*

phonème (notez les valeurs des $z - scores$ à partir des logatomes au tableau 4.1) sont estimés pour chaque locuteur à partir d'un ensemble de phrases lues, exactement comme déjà décrit par Campbell et expliqué dans le chapitre 2.

Ce qui est proposé de façon originale par les auteurs est une compensation des effets dus au débit de parole pour que la durée normalisée ne manifeste que l'allongement dû à la frontière syntagmatique. Ils suggèrent, d'après un résultat puisé chez Crystal & House (1988), que l'influence du débit sur la moyenne et l'écart-type de chaque phonème peut être décrite par un simple facteur α tel que :

$$\hat{\mu_p} = \alpha\mu_p; \hat{\sigma_p} = \alpha\sigma_p \tag{4.4}$$

Les auteurs affirment que cette caractérisation est cohérente avec une description des durées phonémiques en termes d'une fonction gamma.

Le paramètre α présenté plus haut peut être estimé par la formule (*ibid.*, p. 1712) :

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\frac{d_i}{\mu_{p_i}} \tag{4.5}$$

Dans cette équation, $\hat{\alpha}$ est l'estimateur de α, d_i est la durée de la réalisation phonémique i et μ_{p_i} est la moyenne entre les durées associées au phonème p_i.

Si l'estimation est faite sur le même corpus sur lequel la moyenne μ_p est calculée, la valeur de $\hat{\alpha}$ est 1. De façon alternative, plus la valeur de $\hat{\alpha}$ sera proche de 1, plus le débit du corpus sera proche de celui sur lequel la moyenne μ_p a été calculée.

Les durées phonémiques normalisées sont calculées chez Wightman et collègues pour chaque corpus en utilisant les valeurs corrigées des moyennes et écarts-type de phonèmes, $\hat{\mu_p}$ et $\hat{\sigma_p}$.

Nous avons vérifié si la transformation donnée par la formule 4.4 était possible dans notre cas. Pour cela, nous avons calculé les moyennes et écarts-type de chaque phonème pour chacun des débits du corpus *Bureau*. On obtient les paires ($\mu_p\, débit\, normal$, $\mu_p\, débit\, rapide$) et ($\sigma_p\, débit\, normal$, $\sigma_p\, débit\, rapide$) pour chaque phonème. En analysant les valeurs $\mu_p\, débit\, normal$, $\mu_p\, débit\, rapide$, $\sigma_p\, débit\, normal$ et $\sigma_p\, débit\, rapide$, on peut savoir si l'adoption d'un seul facteur α pour transformer les moyennes et les écarts-type en millisecondes entre deux débits est ou non possible. Les résultats se trouvent dans la figure 4.9.

En regardant les distributions, on constate qu'il est possible d'adopter une transformation simple entre les moyennes des débits normal et rapide mais que tel n'est pas le cas entre les écarts-type : la distribution n'est pas repartie avec un faible écart-type autour de la même valeur trouvée pour la distribution des α de la courbe d'en haut. Le

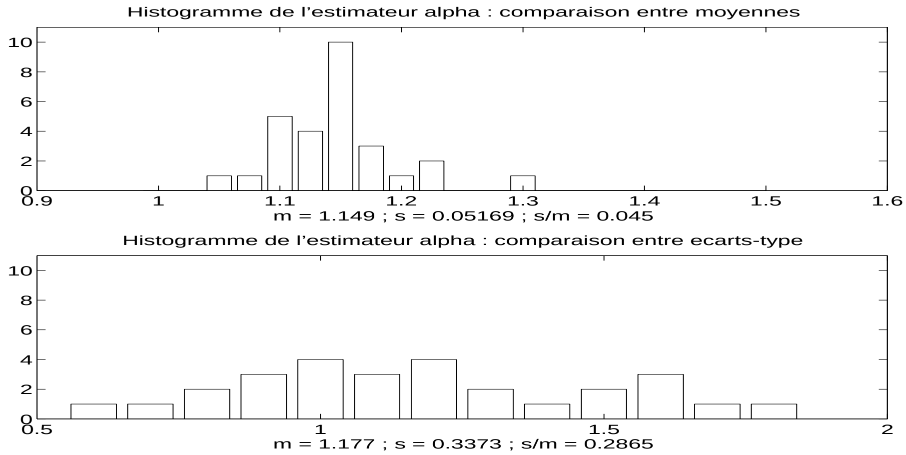

FIG. 4.9 - *Histogramme des facteurs* α *du corpus* Bureau. *En haut, les valeurs de la fraction* $\mu_p\, débit\, normal/\mu_p\, débit\, rapide$ *et en bas celles de la fraction* $\sigma_p\, débit\, normal/\sigma_p\, débit\, rapide$. *Les variables* m *et* s *sont respectivement la moyenne et l'écart-type des distributions. La valeur* s/m *est calculé en absolu et mesure l'extension des distributions respectives.*

même test a été effectué avec les moyennes et écarts-type log-transformés et la conclusion est la même. On a dû abandonner la compensation due au débit : le calcul du $z-score$ dans notre approche est réalisé toujours avec la même paire ($\mu_{phonème}$,$\sigma_{phonème}$) puisée dans le tableau 4.2 et ceci de façon globale, pour le GIPC. On verra que, même dans des débits très différents (corpus *PCenter*), on obtiendra toujours des patrons accentuels avec les caractéristiques décrites dans la section 4.4.2. L'important pour la caractérisation de l'allongement est le rapport entre les $z-scores$ au long de chaque phrase. Dans notre approche, la valeur absolue de chaque $z-score$ est révélatrice non seulement de l'allongement mais aussi du débit de la phrase (surtout les $z-scores$ en début de groupe accentuel).

La valeur du $z-score$ de chaque GIPC étant définie par des contraintes de niveau supérieur, aucune durée phonémique a reçu une valeur en dessous des durées minima réalisées pour chaque catégorie phonémique.

L'estimateur $\hat{\alpha}$ est toutefois une bonne méthode pour la comparaison entre débits. On présente ici les valeurs trouvées par l'application de la formule 4.5 en prenant les valeurs de la paire(μ_p, σ_p) dans le tableau 4.1 du corpus *Logatomes*. Considérant la fraction d_i/μ_{p_i}, sa moyenne arithmétique est l'estimateur $\hat{\alpha}$ et son écart-type, une mesure de la dispersion autour de la valeur de l'estimateur. Les résultats se trouvent dans le tableau 4.11.

On voit dans le tableau 4.11 que le débit *normal* est le plus proche du corpus *Logatomes* ($\hat{\alpha}$ proche de 1). Un test de signifiance permet de constater que les débits *normal*

TAB. 4.11 - *Moyennes (μ), médianes (m) et écarts-type (σ) des facteurs $\hat{\alpha}$ calculés pour chaque débit du corpus* Bureau

débit	*normal*			*rapide*		
	μ	m	σ	μ	m	σ
$\hat{\alpha}$	1,15	1,05	0,47	0,63	0,58	0,23

et *rapide* sont effectivement distincts, si l'on considère les facteurs α comme base pour la distinction. Dans le chapitre suivant on proposera une mesure pour le débit fondée sur le IIPC (intervalle inter-*P-center*) dont nous testerons la capacité à discriminer les styles d'élocution..

L'estimateur $\hat{\alpha}$ permet en outre de générer la durée des premières consonnes (avant le premier *onset* vocalique) et les éléments de la dernière rime de la phrase en multipliant l'exponentielle de la moyenne puisée dans le tableau 4.2 par le facteur $\hat{\alpha}$ correspondant au débit souhaité.

4.7 La Pertinence perceptuelle des patrons rythmiques

> Nous ne connaissons a priori des choses que ce que nous y mettons nous-mêmes.
>
> *Immanuel* KANT, *La Critique de la raison pure*

Un test de préférence a été conçu de façon à montrer la pertinence perceptuelle de l'allongement progressif au long du groupe accentuel en production (Viviani & Stucchi 1992).

Dix paires de phrases ont été générées avec un système d'analyse-synthèse de haute qualité (Moulines 1992; Bailly, Barbe & Wang 1992). La configuration A a été obtenue par le calcul des $z-scores$ successifs de tous les GIPC dans une phrase. Dans la configuration B, seulement les $z-scores$ associés au dernier GIPC du groupe accentuel (correspondant à une marque) ont été maintenus égaux aux valeurs originales tandis que les autres ont été mis à 0 (voir figure 4.10). Le dernier groupe accentuel n'a pas été modifié puisque on ne peut rien dire quant à la monotonicité de certains groupes accentuels en fin de phrase (voir section 4.4).

Les durées phonémiques dans les configurations A et B ont été calculées par l'application de l'algorithme de répartition. Signalons que dans A les durées phonémiques sont différentes des durées correspondantes dans la phrase naturelle car le $z-score$ du GIPC représente un moyennage entre les $z-scores$ réels de chaque réalisation (on a déjà vu que l'hypothèse d'élasticité uniforme ne correspond pas à la réalité). Toutefois la localisation des *onsets* vocaliques au long de la phrase est la même pour les phrases naturelles et celles du type A puisque les durées des GIPC sont les mêmes. Les pauses silencieuses et les autres paramètres prosodiques n'ont pas été changés. Dans le type B la succession des *onsets* vocaliques n'est pas la même que celle de la phrase naturelle puisque les durées phonémiques correspondant aux GIPC inaccentués est obtenue en faisant $z-score = 0$.

Un exemple de phrase avec les $z-scores$ calculés en configuration A et la phrase correspondante en configuration B à être testée est montré dans la figure 4.10.

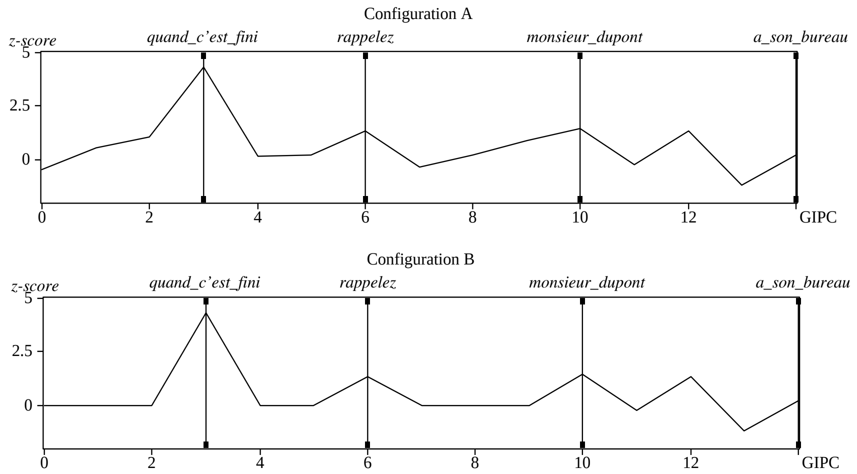

FIG. 4.10 - *Exemple de configurations A et B pour la phrase « Quand c'est fini, rappelez Monsieur Dupont à son bureau.» Les valeurs discrètes du z−score ont été liées pour une meilleure visualisation des « mouvements» d'allongement. Les groupes accentuels ont été indiqués avec des étiquettes verticales placées en position d'accent . Le dernier z−score n'est normalement pas calculé. La valeur qui apparaît ici est celle du z−score réel de la réalisation du phonème (/o/).*

Onze sujets travaillant à l'*Institut de la Communication Parlée* mais pas dans le domaine de la synthèse ont participé à cette expérience. Chaque session a duré entre 10 et 15 minutes. Dans un test ABBA, chaque sujet a été instruit de sélectionner la phrase de la paire que lui semblait la plus naturelle[9]. Chaque sujet pouvait répéter l'écoute une fois. Si en définitive il ne voulait pas faire un choix (soit parce que les phrases lui ont paru semblables ou que ni l'une ni l'autre ne lui a semblé naturelle), il pouvait répondre avec une touche prévue à cet effet.

9. Cette question est relativement bien définie dans ce contexte car la qualité de l'analyse-synthèse la justifie.

4.7.1 Résultats

Tous les sujets confondus, les scores obtenus sont de 65% pour A, 20% pour B et 15% d'indécision. Entre les réponses A et B, le résultat est alors de 77% de préférence pour la configuration A, tous sujets confondus. Les scores individuels sont donnés dans la figure 4.11 et les scores pour chaque phrase, tous sujets confondus, dans la figure 4.12.

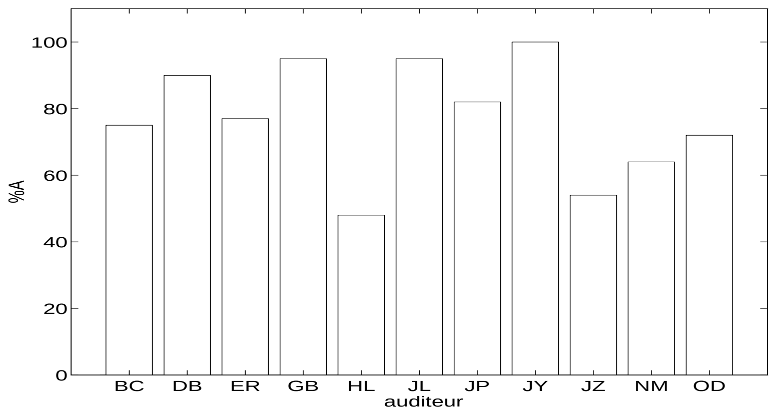

FIG. 4.11 - *Moyenne (entre les deux ordres de présentation) pour la sélection de la configuration A par sujet.*

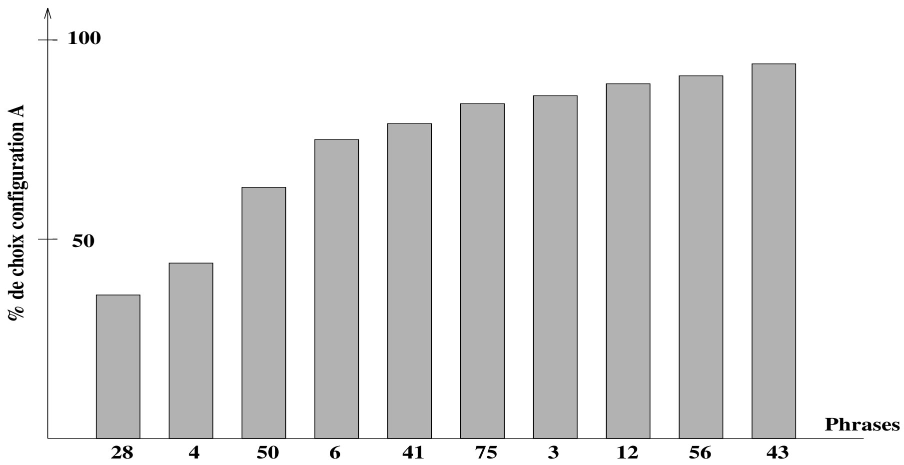

FIG. 4.12 - *Pourcentage de choix de la configuration A par phrase du corpus Bureau (voir texte dans la section A.1), tous sujets confondus.*

Il n'y a pas eu d'effet d'ordre de présentation. Tous les auditeurs étaient d'accord sur la difficulté de la tâche. Quelques commentaires nous ont paru révélateurs du phénomène perceptif en jeu[10] :

Au cours de ce test, j'ai eu l'impression de basculer de critère pour choisir

10. C'est nous qui avons souligné les termes qui apparaissent en italique

> entre une phrase et une autre. Au début j'ai choisi celles qui étaient plus *constantes* au niveau *rythme*. Après, j'ai cru avoir la tendance de préférer celles qui étaient plus rapides quoique plus brusques. *JL*
>
> J'ai choisi celles qui étaient les plus *constantes* au niveau *rythme*. *OD*
>
> Pour moi, les deux étaient artificielles. Pour mes choix, j'ai rejeté les phrases dont le *rythme* n'allait pas. *DB*
>
> Tu es sûr que des fois *c'était pas la même phrase* ? *JY*

Il est intéressant de noter le vocabulaire utilisé dans les commentaires des sujets : les mots *constant* et *rythme* reviennent fréquemment [11], ce qui conforte le fait que les sujets étaient sensibles au facteur qui nous voulions tester, à savoir la composante rythmique. Aucune remarque sur la nature des sons n'a pas été émise. La notion de *constance* est peut-être un indice d'isosyllabicité attendue. La remarque du sujet *JY* est très intéressante dans la mesure où c'est lui qui a préféré à 100 % les phrases du type A. Ceci indique la difficulté de la tâche demandée. C'est pour cette raison qu'un petit nombre de phrases a été choisi pour le test, limitant les effets dus à la fatigue.

4.7.2 Discussion

Le taux de patrons monotones et ascendants des 10 phrases utilisées dans le test de perception – qui sont celles qui se trouvent dans les figures 4.3 et 4.4 – est de 66 % (ceci ne tient pas compte de groupes accentuels finals). Il n'est cependant pas facile d'expliquer pourquoi les sujets ont jugé une phrase plus naturelle que l'autre dans une paire : ceci concerne l'analyse de plusieurs groupes accentuels et des capacités générales de perception de chaque sujet.

Toutefois, si l'on compare les résultats de la performance des sujets par phrase avec l'évolution des $z\!-\!scores$ de ces mêmes phrases (respectivement figure 4.12 et figures 4.3 et 4.4), il est possible d'expliquer les bonnes performances des phrases 56, 41 et 6, notamment pour des groupes comme « Monsieur Dupont », qui sont ascendants et monotones en opposition au patron binaire de la configuration B. La performance de la phrase 43 est difficile à expliquer du point de vue de notre hypothèse de préférence pour la structure ascendante, car le choix a pu concerner la préférence de structures du type « à son bureau ». Mais si tel est le cas, il ne serait pas anormal de considérer un tel patron accentuel (accent secondaire sur le groupe /ɔ̃b/) comme étant plus naturel que celui de

11. Les commentaires des autres sujets n'ont pas été reproduits afin d'éviter une répétition excessive.

la configuration B. Dans la phrase 4, les patrons accentuels des configurations A et B sont très proches puisque l'allongement au long des patrons accentuels est moins marqué : ceci justifie son score proche de 50 %. Le score de la phrase 28 est curieux. Serait-il possible que le patron du premier groupe accentuel (« Monsieur Tapi ») de la configuration B (quoique binaire, elle présente un allongement final au dernier GIPC) a été préféré au patron exceptionnellement non monotone du groupe accentuel correspondant dans la configuration A ?

Étant donné que la majorité des jugements entre les configurations A et B était un choix entre un patron ascendant monotone contre un patron binaire et que les commentaires des sujets après le test semblent indiquer un jugement plutôt global sur la phrase, on peut affirmer que la préférence générale pour les patrons rythmiques monotoniquement croissants prônent la nécessité de ce phénomène de production pour la perception de l'accent. Le manque d'allongement dans les GIPC le précédant semble brusque. L'hypothèse d'une horloge interne, qui sera abordée dans le chapitre suivant peut expliquer le comportement perceptif des sujets : les GIPC rétrécis sont des indices d'une accélération locale non attendue. La littérature semble corroborer cette conclusion en liant le phénomène à l'isosyllabicité (Duez 1987; Lehiste 1977) : l'allongement progressif en production serait nécessaire à la perception de l'isosyllabicité.

Il est clair que le contrôle rythmique par $z-score$, mis en exergue ici, n'a aucun fondement biologique et ne vise à aucune théorie physiologique.

4.8 Sommaire

Les analyses effectuées sur le corpus *Bureau* nous ont conduit à adopter, dans le cadre de la génération automatique, une unité minimum de programmation rythmique, le GIPC, plutôt que la syllabe.

La plus grande cohérence entre les allongements des réalisations phonémiques qui constituent le GIPC fait de la durée de ce groupe un cadre privilégié dans les limites duquel une durée sera affectée à chacune des réalisations qu'il contient.

La caractérisation des IIPC successifs en termes d'une durée normalisée ($z-score$), a permis de révéler une structuration rythmique de l'énoncé en une suite de groupes accentuels cohérents. Dans ces groupes, l'accent est réalisé avec un allongement important de la durée du dernier GIPC, mais cet allongement est progressif tout au long du groupe accentuel.

On a montré, en production de parole, une hiérarchisation des frontières syntaxiques en au moins trois catégories d'allongement.

Un test de perception a montré que les auditeurs ont une préférence pour les patrons accentuels trouvés en production : leur durée normalisée monotoniquement croissante est une condition nécessaire à la perception du naturel dans les phrases lues. L'allongement progressif semble être être perçu comme relativement isochrone.

Si l'on essaie de mettre en relation la composante macrorythmique que nous avons étudiée avec sa fonction de communication, on remarque que les principales fonctions de la prosodie, à savoir, celles d'organisation du signal acoustique émis par un locuteur donné en un message cohérent et structuré, d'anticipation et de segmentation (Dubois, Giacomo, Guespin, Marcellesi, Marcellesi & Mével 1994) sont confirmées par la caractérisation de la composante rythmique du corpus par des patrons accentuels révélés par les $z-scores$:

- la notion de structure mise en relation avec le gestaltisme abordée dans ce chapitre assurerait la première fonction : organisation du signal acoustique en un message cohérent ;
- la monotonicité croissante de ces patrons (mais perçus comme relativement isochrones) assurerait une fonction d'anticipation ;
- la réalisation de l'accent et le phénomène de « reset » qui suit cette réalisation, serait en relation avec la troisième fonction : segmentation de l'énoncé en unités plus petites dans le cadre de la phrase.

Chapitre 5

Le Modèle dynamique de génération de la pause

Dans ce chapitre, les caractéristiques rythmiques du corpus *PCenter* sont présentées. Essentiellement, les mêmes aspects abordés, dans le chapitre précédent, pour le corpus *Bureau* ont été retrouvés dans ce corpus spécialement conçu pour l'étude du *Perceptual-center*. Ceci est présenté dans la première section.

Ensuite il sera question de l'étude du rattachement de la structure rythmique de la parole à une structure biologique sous-jacente, à travers l'hypothèse d'une horloge interne. Celle-ci va nous permettre de proposer une mesure du débit de chaque phrase, obtenue en termes du IIPC moyen. Quelques aspects concernant le changement de la structuration rythmique avec le débit sont discutés à partir de cette caractérisation.

Le chapitre se termine avec la présentation d'un modèle combiné de génération de la durée phonémique qui tient compte de l'émergence de la pause. Ce modèle intègre l'apport de l'apprentissage automatique à un modèle statistique de répartition.

5.1 Caractéristiques rythmiques du corpus *PCenter*

On a voulu vérifier comment se présente ce nouveau corpus en termes de corrélation entre $z-scores$ réels des réalisations phonémiques, de l'évolution des $z-scores$ associés aux GIPC et de la hiérarchisation des marques prosodiques.

5.1.1 L'Unité de programmation rythmique : analyse par corrélation

Ce corpus ne présente pas de réalisation du type coda, étant donné qu'il a été conçu de façon à rendre la localisation du P-center plus facile, par le choix de séquences du type CV. On a calculé les $z-scores$ réels de chaque réalisation en les classant comme *attaque* et *noyau* à l'intérieur de chaque syllabe CV. La voyelle a été classée *noyau* et la consonne qui la précède, *attaque*. Les corrélations ont été calculées selon les procédures expliquées dans le chapitre précédent et les résultats sont présentés ci-dessous.

TAB. 5.1 - *Corrélation par méthode de Pearson entre $z-scores$ réels adjacents par paire (catégorie réalisation courante,catégorie réalisation suivante) pour le corpus* PCenter. *Entre les catégories* noyau *et* attaque, *il n'y a pas de silence.*

débit	(*attaque*,*noyau*)	(*noyau*,*attaque*)
très lent	0,38	0,27
lent	0,42	0,15
normal	0,43	0,15
rapide	0,36	0,28
très rapide	0,22	0,35

Il est surprenant que dans un corpus contenant uniquement des séquences CV, les corrélations entre réalisations consécutives étiquetées (*noyau*,*attaque*) soient plus basses (à l'exception du débit *très rapide*) que celles correspondantes à la séquence (*attaque*,*noyau*) : tel n'a pas été le cas dans le corpus *Bureau* (cf. tableaux 4.5 et 4.6). Dans les débits *très lent* et *rapide* les deux unités rythmiques, syllabe et GIPC, seraient encore comparables mais tel n'est pas le cas dans les débits *lent* et *normal*.

En regardant de près les $z-scores$ des réalisations phonémiques, on constate que les plus grandes causes des basses corrélations ci-dessus sont dues principalement au groupe /ɔm/ dans « sommet» et au groupe /ɔ̃k/, dans « mont Karimo» : les consonnes (respectivement /m/ et /k/) sont très allongées par rapport aux voyelles (respectivement /ɔ/ et /ɔ̃/). En réécoutant les phrases on s'aperçoit que le mot « sommet» semble être emphatisé et que le /m/ pourrait être une consonne géminée, de par sa longueur excessive. En ce qui concerne le groupe « mont Karimo» les mêmes observations sont valables tout en rajoutant la difficulté d'attribuer une durée à la réalisation /k/. En enlevant les GIPC correspondant à ces cas problématiques du calcul de la corrélation, sa valeur s'élève à 42 %

pour les débits *lent* et *normal* pour la paire (*noyau*,*attaque*) et la paire (*attaque*,*noyau*) pour ces débits continue à avoir 42 % de corrélation. Malgré ces difficultés, le GIPC a été toujours maintenu comme unité privilégiée pour le calcul des durées phonémiques à la place de la syllabe. On verra en fait que, quoique le choix du GIPC ou de la syllabe comme cadre pour la génération phonémique entraîne des erreurs de prédictions qui restent semblables dans le cas de la durée des sons, le GIPC est un meilleur choix quand il s'agit de la génération de la pause silencieuse ou des réalisations phonémiques dans les deux débits plus rapides (cf. section 5.4.3).

Pour des raisons dues à d'éventuelles erreurs de segmentation du corpus ou d'éventuelles erreurs dans les valeurs du tableau contenant les statistiques des phonèmes du locuteur, l'analyse par corrélation peut ne pas être une mesure très sûre pour tirer des conclusions sur la nature de l'unité de programmation rythmique. L'analyse concernant la corrélation entre débits deux à deux, pour les durées des syllabes ou des GIPC, a déjà été présentée dans la section 4.3.2 pour ce corpus : quelques raisons pour lesquelles le GIPC nous semble être la bonne unité de programmation rythmique, selon ce critère, ont été discutées à cette occasion.

5.1.2 Les Patrons rythmiques du corpus *PCenter*

Le même type de graphique réalisé dans la section 4.4.2 pour le corpus *Bureau* a été refait pour cet autre corpus : chaque $z - score$ est calculé à partir de la durée des GIPC au long de chaque phrase. L'évolution de ces valeurs au long de la phrase est illustrée dans la figure 5.1.

On confirme la même tendance : 80 % des groupes accentuels sont monotones et ascendants en ce qui concerne l'évolution des $z - scores$. Les caractéristiques d'accent coïncidant avec une marque prosodique et de « reset » après l'accent sont retrouvées encore pour ce corpus.

On peut observer, en outre, un effet de restructuration rythmique intéressant. Regardons le groupe « Filez les sous» (4e à 7e GIPC) dans la figure 5.1 : le 5e GIPC, apparemment accentué dans les débits *très lent* et *lent* (délimitant ainsi deux groupes : « Filez» et « les sous»), semble se désaccentuer dans les débits *normal*, *rapide* et *très rapide* (formant ainsi un seul groupe : « Filez les sous »). La tendance semble être la même pour « baisser les mains» (11e à 13e GIPC) mais dans ce cas, le dernier GIPC calculable est le groupe /em/, inaccentué.

Cet effet sera étudié plus longuement dans la section 5.3, lorsqu'on caractérisera les GIPC en termes de nombre de périodes d'horloge interne.

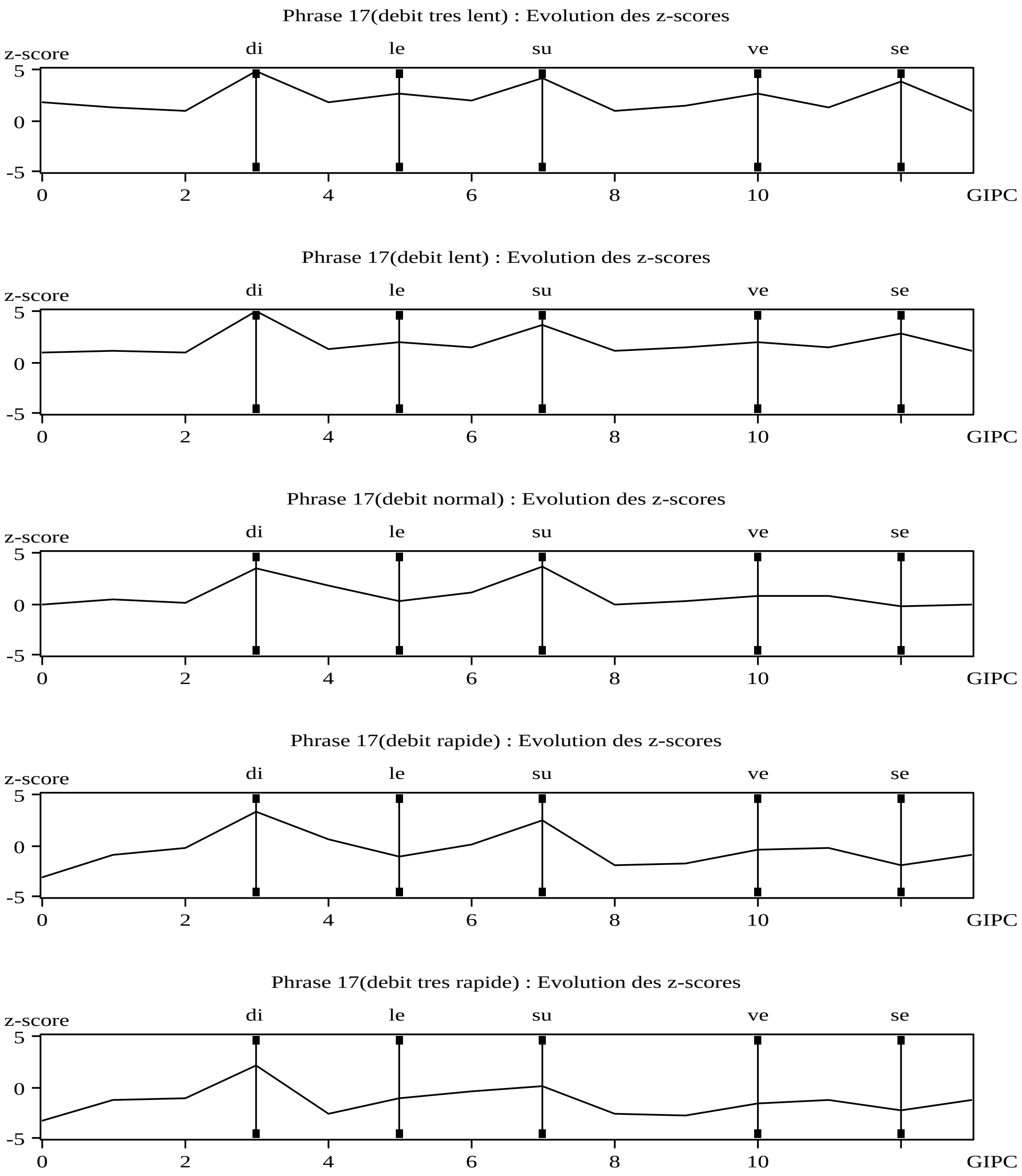

FIG. 5.1 - *Évolution des $z-scores$ dans la phrase 17 du corpus* PCenter *:* Tu m'avais dit : « Filez les sous, vous pouvez baisser les mains.» *Dans ces courbes, les valeurs discrètes des $z-scores$ sont reliées de façon à révéler plus clairement les caractéristiques rythmiques des phrases. En position accentuée, le silence n'a pas été considéré pour le calcul du $z-score$. Chaque position accentuée a été indiquée par une étiquette indiquant la syllabe concernée dans l'accentuation et la marque prosodique qui lui est associée. Les débits concernés sont, de haut en bas, très lent, lent, normal, rapide et très rapide.*

5.1.3 La Hiérarchie des marques prosodiques

La hiérarchie des marques prosodiques a été présentée ci-dessous selon des procédés décrits dans le chapitre précédent. La difficulté supplémentaire est un problème de signifiance statistique, vue la quantité inférieure de marques par débit, étant donné la longueur et le nombre de phrases du corpus *PCenter*. On présente en outre (tableau 5.2), la durée moyenne et l'écart-type des GIPC en position de marque prosodique.

TAB. 5.2 - *Moyennes (et écarts-type) de la durée et du $z-score$ associés aux groupes en frontière prosodique. Les résultats sont présentés par débit pour le corpus* PCenter. *On peut établir une hiérarchie à partir de ces valeurs mais deux marques voisines ne sont pas statistiquement différentes. On peut établir en fait de trois à quatre groupes.*

	Durée (ms) et $z-score$				
Marque	tr. lent	lent	normal	rapide	tr. rapide
ID	825(194)	752(141)	597(77)	507(106)	283(72)
	3.9(0.5)	*4.0(0.5)*	*3.4(0.4)*	*2.6(0.7)*	*0.7(0.9)*
GF	607(101)	518(97)	383(71)	247(38)	169(17)
	3.6(0.4)	*3.5(0.8)*	*2.2(0.7)*	*0.2(0.5)*	*-1.1(0.4)*
IT	654(141)	614(172)	371(131)	388(192)	177(44)
	2.9(0.4)	*2.8(0.6)*	*1.8(1.6)*	*1.8(2.0)*	*-0.9(1.0)*
DD	700(275)	464(146)	363(84)	239(74)	166(24)
	3.2(0.8)	*2.7(1.0)*	*1.6(0.8)*	*0.0(0.7)*	*-1.2(0.4)*
DF	561(207)	395(114)	270(76)	175(34)	134(19)
	2.8(1.1)	*2.1(1.1)*	*0.7(0.9)*	*-0.6(0.6)*	*-1.6(0.5)*
DG	417(104)	396(181)	306(37)	239(0.8)	188(11)
	1.7(1.1)	*1.6(0.2)*	*0.5(0.5)*	*-0.6(0.6)*	*-1.6(0.3)*

Si l'on regarde l'ordre défini par les valeurs moyennes des $z-scores$ associés à chaque marque, on constate que, pour les trois débits plus lents la hiérarchie pourrait être, en ordre décroissant de force, ID, GF, IT, DD, DF et DG, avec trois catégories distinctes parce qu'il y a absence de signifiance pour les marques adjacentes deux à deux (ceci est dû au nombre réduit de données). Pour les deux autres débits, la hiérarchie pourrait être, en ordre décroissant de force, ID, IT, GF, DD, DF/DG, même observation quant au nombre de catégories : trois au minimum. Cette autre hiérarchie n'est pas différente, en termes statistiques, de la première. Il y aurait plus de différences entre les hiérarchies des marques pour chaque débit si la durée intrinsèque moyenne du GIPC directement dominé par une marque prosodique, était prise comme élément pour constituer la hiérarchie.

Cette hiérarchie des marques, si l'on excepte l'effet de l'inclusion de la marque GF, est essentiellement la même trouvée pour le corpus *Bureau*.

5.1.4 Les Facteurs α associés aux débits du corpus *PCenter*

Calculons l'estimation $\hat{\alpha}$ des facteurs α, selon les procédés de la section 4.6. Les résultats se trouvent dans le tableau 5.3.

TAB. 5.3 - *Moyennes, écarts-type) et médianes des facteurs $\hat{\alpha}$ calculés pour chaque débit du corpus* PCenter.

débit	$\hat{\alpha}$		
	Moyenne	Écart-type	Médiane
très lent	1.51	0.75	1.29
lent	1.42	0.67	1.23
normal	1.19	0.53	1.04
rapide	0.87	0.40	0.77
très rapide	0.69	0.24	0.65

Les valeurs de $\hat{\alpha}$ correspondant à chaque paire de débits sont statistiquement distincts. Ceci est un indice de distinction effective entre les débits. L'analyse des histogrammes des IIPC pour chaque débit confirmera cette affirmation (cf. section 5.2.1). Nous pouvons constater en outre que le débit *normal* est le plus proche de celui du corpus *Logatomes* (valeur de $\hat{\alpha}$ proche de 1).

N'oublions pas que ces valeurs peuvent être utilisées pour obtenir la durée des réalisations en début et en fin de phrase qui n'entrent pas dans le cadre du GIPC.

5.2 L'Hypothèse de l'horloge interne

Étant donné que la parole est une mécanisme de production d'un système biologique, il est naturel que ce mécanisme soit assujetti aux contraintes de nature physiologique et cognitive (contraintes propres à l'état psychologique actuel du sujet et contraintes physiques de nature neuronale). En ce qui concerne la structure neuronale sous-jacente, il est nécessaire de réfléchir sur la question des repères physiologiques de la temporalité.

Turvey et collègues (1990) ont étudié la coordination absolue des mouvements rythmiques du corps humain, notamment la marche. Ils proposent un modèle représentatif de l'acte de locomotion basé sur deux fonctions : celle de fonction de maintien de la temporalité (*timekeeping fonction*), exécutée par des cellules centrales ou des populations de cellules centrales qui produisent un *signal périodique*, et celle de fonction motrice (*motor fonction*) qui, réalisée par des populations de cellules centrales qui prennent comme repère ce signal périodique, transmet des impulsions aux muscles.

Ce *signal périodique* semble être la référence pour le contrôle du temps lors de l'acte d'énonciation comme est suggéré par Allen quand il commente le contrôle de l'organisation temporelle des articulateurs (1975, p. 84) :

> Avant le début de l'énonciation les attributs phonétiques (incluant les valeurs de temps) pour tout le syntagme sont calculés et stockés comme une chaîne ordonnée dans un *buffer* de sortie. (...) Le syntagme est alors transformé en commandes neuromotrices par le programme de contrôle moteur, qui contrôle l'organisation temporelle des articulateurs, à travers la comparaison des valeurs stockées des temps des réalisations phonémiques avec une horloge interne (Allen 1973).

Les résultats des études de coordination entre le *tapping* et un stimulus périodique externe tendent à renforcer l'hypothèse de l'horloge interne (Fraisse 1974; Allen 1975). Dans l'étude réalisée par Andras Semjen et collègues (1992), l'existence de cette horloge est même une nécessité, « d'après les performances que ce genre d'activité exhibe» , disent les auteurs. Ils présupposent que la période de cette horloge peut être modifiée selon les besoins de la tâche à exécuter. L'existence de cette horloge interne est une évidence aussi pour Mates (1992), qui présente des sujets réalisant le même type de tâche.

Pour Semjen, cependant, s'il existe au moins une horloge interne nécessaire au contrôle de ces activités périodiques, le nombre d'horloges internes concernés peut très bien être supérieur à 1.

Un travail sur les capacités cognitives et les fonctions langagières du système cérébro-cérébellaire humain semble indiquer que, d'après les résultats obtenus par l'étude de lésions sur la capacité langagière, la localisation de la fonction de maintien de périodocité pourrait être le cervelet. En fait, cet organe transmet des signaux, à travers le thalamus ventro-latéral, aux aires motrices et aux aires responsables de fonctions langagières et cognitives, telles que l'aire de Broca (Leiner, Leiner & Dow 1991). La fonction du cervelet dans l'organisation temporelle de la parole est aussi démontrée dans le travail de Bell-Berti et collègues (1990).

Constatant que, sous certaines conditions expérimentales, les cellules du thalamus génèrent des réponses oscillatoires à des fréquences comprises entre 6 et 10 Hz, Llinás

(1989) affirme que ces cellules peuvent se synchroniser par couplage et générer une modulation de phase pour les neurones moteurs. La fonction principale de cette entrée serait la synchronisation de l'activité de groupes de muscles dans le corps afin de générer des réponses motrices organisées (*ibid.*, p. 13).

Lame (1990) propose une horloge interne qui trouve son ancrage sur la succession des *P-centers* :

> In a speech production task requiring the synchronization of monosyllables with a metronome, it was found that, while relative timing between syllables conforms to the P-center description, the alignment of a syllable with the metronome is influenced by the length of the preceding syllable's rhyme. Such behavior is **consistent with the operation of an internal clock**[1], used for perception and production of rhythmic intervals, which is phase shifted by acoustic events.
>
> *Dans une tâche de production de parole exigeant la synchronisation entre des monosyllabes et un métronome, il a été trouvé que, tandis que l'organisation temporelle relative entre les syllabes est conforme au paradigme P-center, l'alignement d'une syllabe avec le métronome est influencé par la longueur de la rime de la syllabe précédente. Un tel comportement est compatible avec le fonctionnement d'une horloge interne, utilisée pour la perception et la production d'intervalles rythmiques, dont la phase est décalée par des événements acoustiques.*

Le mouvement de la mâchoire est depuis longtemps décrit comme périodique (Abercrombie 1967). Nous pensons qu'une horloge interne, générée par une structure neuronale mais non nécessairement spécifique aux mécanismes de production de la parole, peut fonctionner comme repère de périodicité ou régularité (Mates 1992) lors de l'énonciation. Un repère neuronal expliquerait cette tendance universelle à l'isochronisme, même en production (cf. section 3.3.1). Toute déviation entre cette horloge interne (événement proximal) et la période du cycle moteur (événement distal), serait perçue comme ayant une fonction de communication (cf. observation de Lehiste dans la section 3.3.1), comme tel est le cas pour le mécanisme accentuel. Cette déviation du mécanisme quasi-isochrone de production par la réalisation de l'accent est inhérente à la notion d'accentuation et est l'une des raisons pour laquelle Wenk & Wioland (1982) rejettent la notion d'isosyllabisme pour le français.

[1] C'est nous qui soulignons.

L'existence de cette horloge interne est postulée dans notre travail. Même s'il s'avère un jour, quoique nous ne soyons pas du tout convaincus, que son existence physiologique soit scientifiquement niée, ceci ne nuit en rien notre travail car cette hypothèse a cependant un rôle méthodologique très fructueux.

Nous avons pensé à caractériser chaque phrase du corpus *PCenter* par une mesure indirecte de cette horloge à partir de la durée du GIPC, appartenant au cycle de production motrice (au plan distal). Cette durée a été appelée plus haut l'intervalle inter-*P-center* (IIPC).

5.2.1 Caractérisation du débit de la phrase par le calcul de l'IIPC

En ce qui concerne le patron d'accentuation au sein du mot prosodique, on constate que :

– au sein du groupe accentuel, le dernier élément est le plus allongé (l'augmentation de la durée a une forme quasi-exponentielle, si l'on regarde la figure 4.5). Ceci a été montré par Pasdeloup (1992). Elle conclut que le ralentissement au long du groupe accentuel s'accentue dans la syllabe pénultième (par rapport à la position de l'accent) pour atteindre un maximum dans la syllabe accentuée ;

– l'accent semble véhiculer une fonction linguistique au moyen d'une déviation — qui se manifeste déjà à partir de la syllabe juste avant la syllabe accentuée — par rapport à une activité quasi-isochrone.

Ceci nous a conduit à choisir la moyenne, pour chaque phrase, des IIPC non correspondant à une marque prosodique, comme une mesure de la période de l'horloge interne. Cette mesure sera appelée PH (pour *période d'horloge*). Le résultat pour toutes les phrases du corpus *PCenter* est donné dans le tableau 5.4 et dans la figure 5.2.

Une analyse statistique des différences révèle que les cinq débits ainsi caractérisés sont distincts (entre *très lent* et *lent*, $p < 0,01$; entre *lent* et *normal*, $p < 10^{-8}$; entre *normal* et *rapide*, $p < 10^{-12}$ et entre *rapide* et *très rapide*, $p < 10^{-12}$, avec $intervalle\, de\, confiance =$ 5 %).

Si l'on divise la durée de chaque GIPC au long des phrases par la PH correspondant à la même phrase, on obtient une évolution de la durée en termes d'unités de PH. Les patrons de durée ainsi obtenus ont pu alors être étudiés.

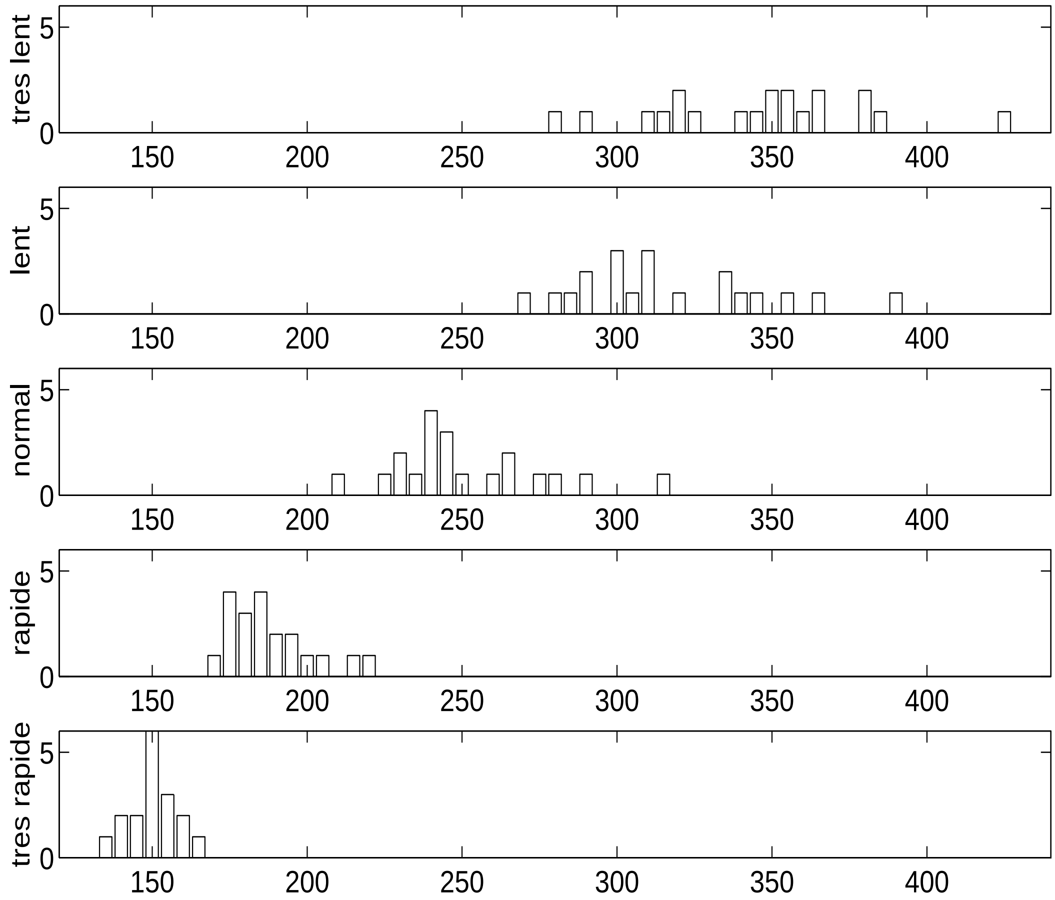

FIG. 5.2 - *Période de l'horloge interne, à partir de la moyenne entre les durées des GIPC non correspondant à une marque prosodique. Les cinq débits du corpus* PCenter *sont montrés.*

5.3 Le Patron accentuel révélé par l'approche horloge interne

On a voulu vérifier certaines caractéristiques de la structuration rythmique au long de la phrase et leur relation avec le débit. Ces caractéristiques concernent le statut du GIPC quant à l'accentuabilité et la restructuration rythmique lorsque l'énonciation se fait plus rapidement.

Regardons, par exemple, sous la forme d'un graphique illustrant la moyenne et l'écart-type de l'évolution d'un contour spécifique d'IIPC en termes d'unités de PH. Le résultat est donné dans les figures qui suivent. Les phrases concernées sont celles dont le premier syntagme est « Tôt dans la matinée» (phrase 1 à 5) et les positions 1 à 6, en abscisse, correspondent aux GIPC successifs. Dans tous les cas, le silence a été considéré comme faisant partie de la durée totale du GIPC, selon la définition donnée pour la constitution du GIPC.

TAB. 5.4 - *Durée d'horloge moyenne, extrema des histogrammes, nombre de pauses silencieuses et pause minimum relevée par débit. Il y a 285 GIPC par débit.*

débit	période d'horloge moyenne et écart-type (ms)	plage de variation (ms)	nombre de silences	silence minimum (ms)
très lent	345(35)	281-423	96	75
lent	317(32)	270-367	64	48
normal	252(24)	212-314	32	57
rapide	187(13)	173-218	26	51
très rapide	150(7)	137-164	4	59

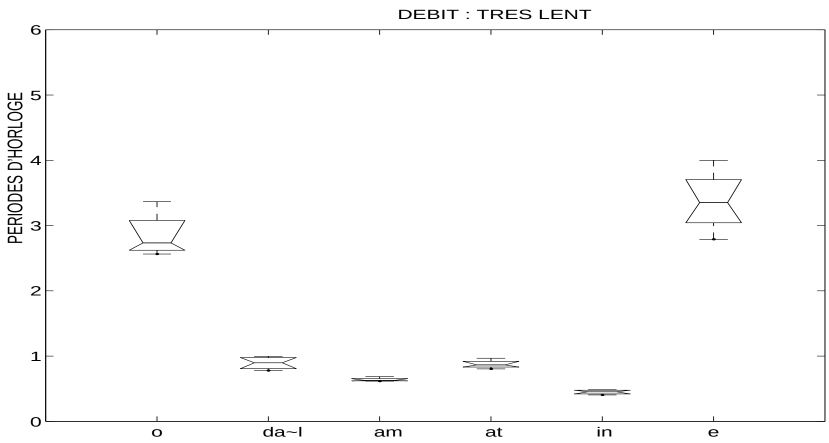

FIG. 5.3 - *Évolution du nombre de PH dans le corpus* PCenter*, débit* très lent*. Il s'agit du syntagme « Tôt dans la matinée», des phrases 1 à 5. Dans les 5 phrases de ce débit il y a un silence après « Tôt» et un autre après « matinée». Chaque bloc de la figure représente plusieurs données statistiques. L'étendue des valeurs est indiquée par les traits horizontaux les plus extrêmes, la médiane par le trait horizontal à la rencontre des deux trapèzes, la hauteur de chaque trapèze, en dessous et en dessus du trait représentatif de la médiane, indique l'étendue de valeurs respectivement inférieures et supérieures à la médiane, comprenant la plage entre 25 et 75 % de tous les valeurs préalablement ordonnées.*

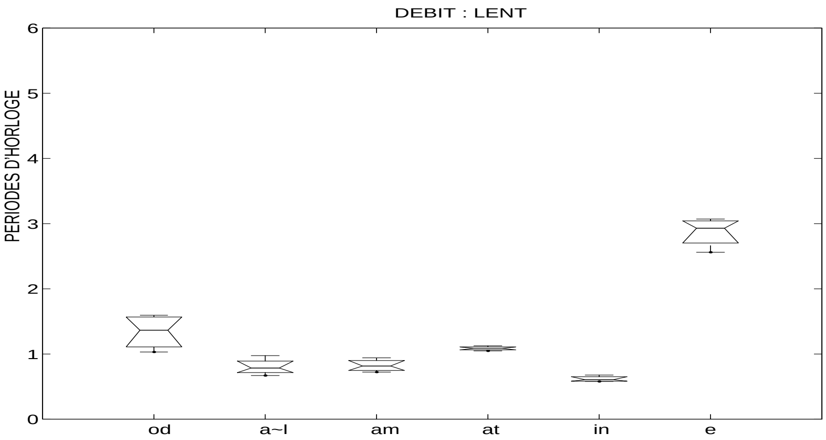

FIG. 5.4 - *Évolution du nombre de PH dans le corpus* PCenter, *débit* lent. *Il s'agit du syntagme « Tôt dans la matinée», des phrases 1 à 5. Dans les 5 phrases de ce débit il y a un silence après « matinée».*

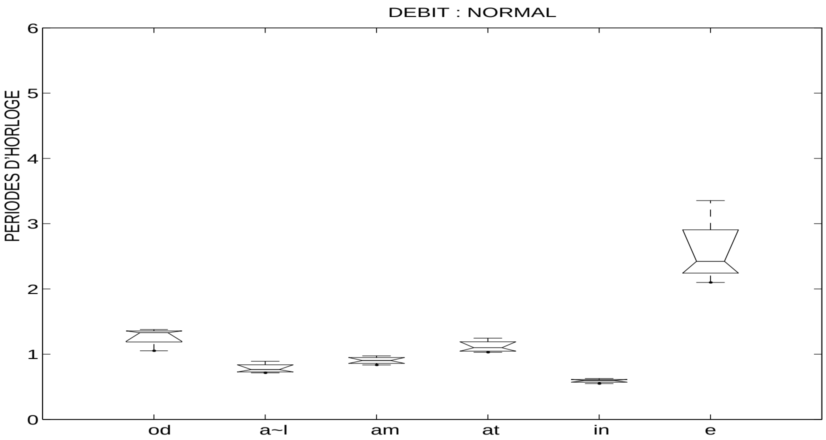

FIG. 5.5 - *Évolution du nombre de PH dans le corpus* PCenter, *débit* normal. *Il s'agit du syntagme « Tôt dans la matinée», des phrases 1 à 5. Dans les 5 phrases de ce débit il y a un silence après « matinée».*

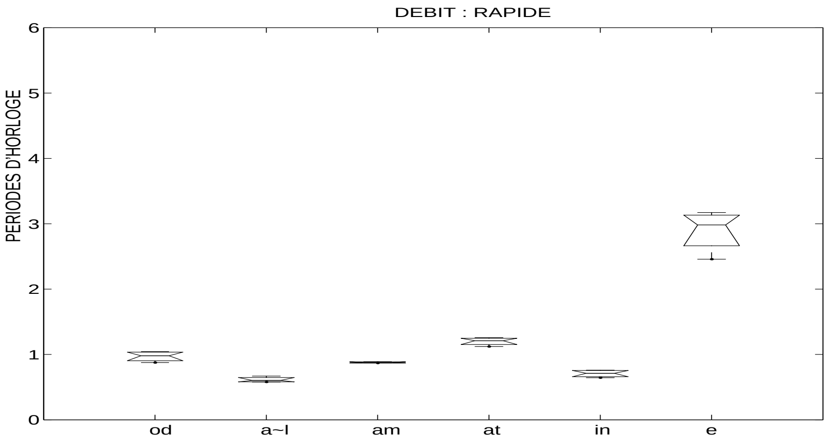

FIG. 5.6 - *Évolution du nombre de PH dans le corpus* PCenter*, débit* rapide*. Il s'agit du syntagme « Tôt dans la matinée», des phrases 1 à 5. Dans les 5 phrases de ce débit il y a un silence après « matinée».*

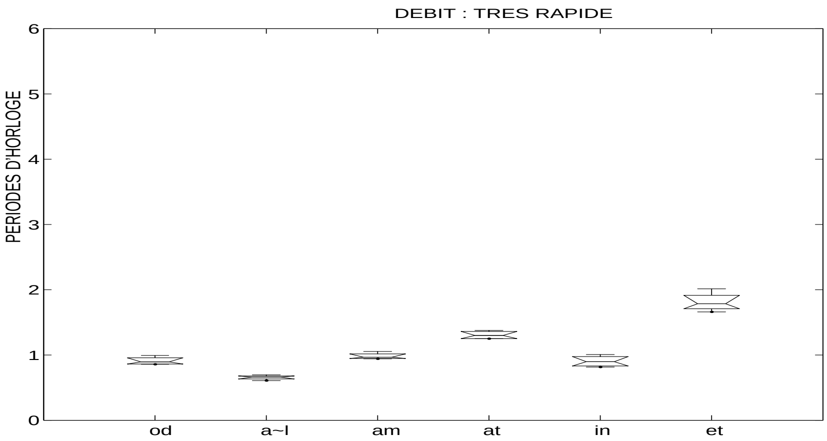

FIG. 5.7 - *Évolution du nombre de PH dans le corpus* PCenter*, débit* très rapide*. Il s'agit du syntagme « Tôt dans la matinée», des phrases 1 à 5. Il n'y a pas de silence après les GIPC.*

On constate, en analysant ces résultats, que la position accentuée possède une variance plus grande (ceci n'est pas dû seulement à la présence du silence, car il n'y en a pas un après « Tôt » dans les débits *lent*, *normal*, *rapide* et *très rapide*) que celle de la position inaccentuée. La variance du GIPC inaccentué est toujours réduite, dans tous les débits. Une variance réduite est révélatrice d'un contrôle plus précis de la part du système moteur, qui indiquerait le statut de la durée du GIPC inaccentué comme unité de référence du cycle rythmique. Cette hypothèse est renforcée par le fait qu'à la position « Tôt », il y a désaccentuation progressive par diminution du nombre de PH, lorsqu'on augmente la vitesse d'élocution (figures 5.3 à 5.7). De plus, au débit *très rapide* (figure 5.7), la variance à cette position est comparable à celles des positions inaccentuées. Le groupe accentuel semble se restructurer formant un groupe de 6 GIPC, culminant dans la réalisation de l'accent sur le /e/ de « matinée ».

Cet effet est visible même en réalisant l'analyse indépendamment de la position du syntagme dans la phrase, comme dans « les sommets du mont Karimo ». Les figures qui se suivent concernent la partie « les sommets du mont ». Regardez comment, peu à peu, l'accent sur /mɛ/ dans « sommet » semble s'effacer, avec la diminution du nombre de PH associé et la diminution de la variance (figures 5.8 à 5.12).

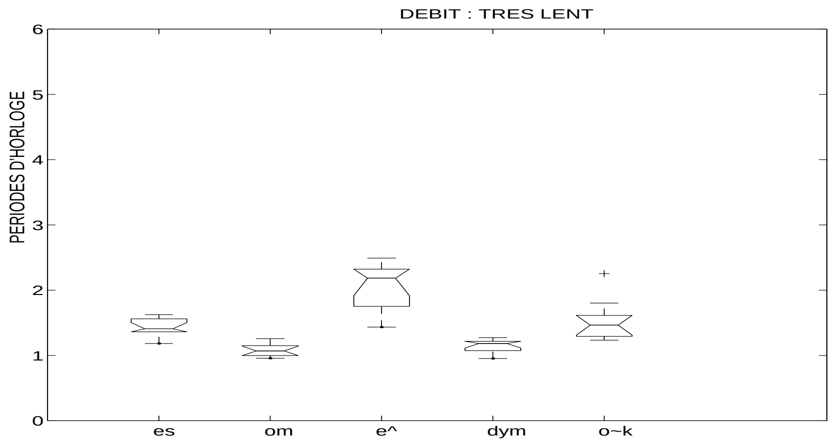

FIG. 5.8 - *Évolution du nombre de PH dans le corpus* PCenter*, débit* très lent. *Il s'agit du sous-groupe « les sommets du mont ».*

Cette approche de caractérisation de la durée par le nombre de PH nous a permis de faire une étude de la relation entre le nombre de GIPC d'un groupe accentuel et son nombre de PH. Cette analyse aurait permis de savoir si le phénomène pausal était associé à un nombre entier de PH où, en d'autres termes, si la pause, comprise comme *allongement final + silence éventuel*, (Duez 1987), était quantisée. Quoiqu'une certaine tendance vers la quantisation a pu être dégagée, comme d'ailleurs suggère le travail de

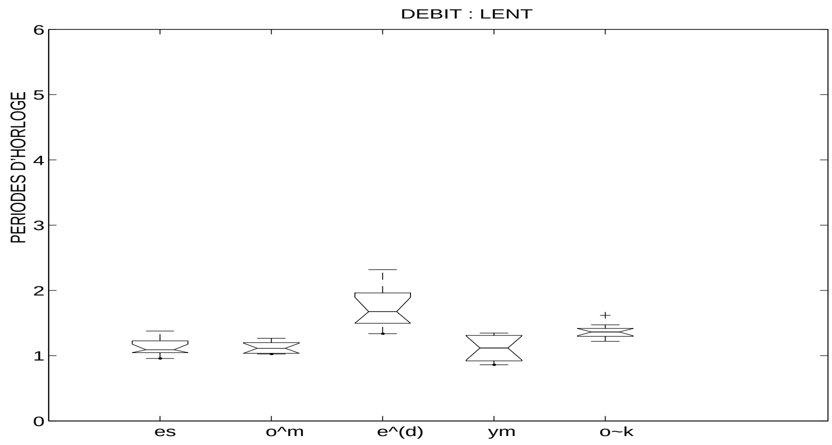

FIG. 5.9 - *Évolution du nombre de PH dans le corpus* PCenter*, débit* lent*. Il s'agit du sous-groupe « les sommets du mont» .*

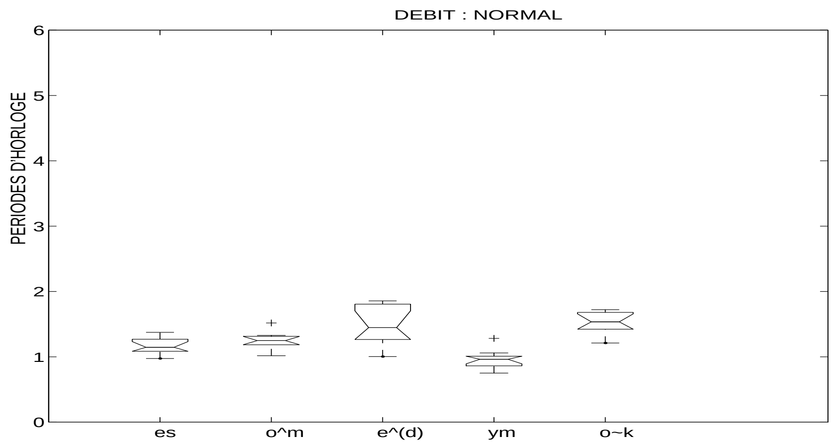

FIG. 5.10 - *Évolution du nombre de PH dans le corpus* PCenter*, débit* normal*. Il s'agit du sous-groupe « les sommets du mont» .*

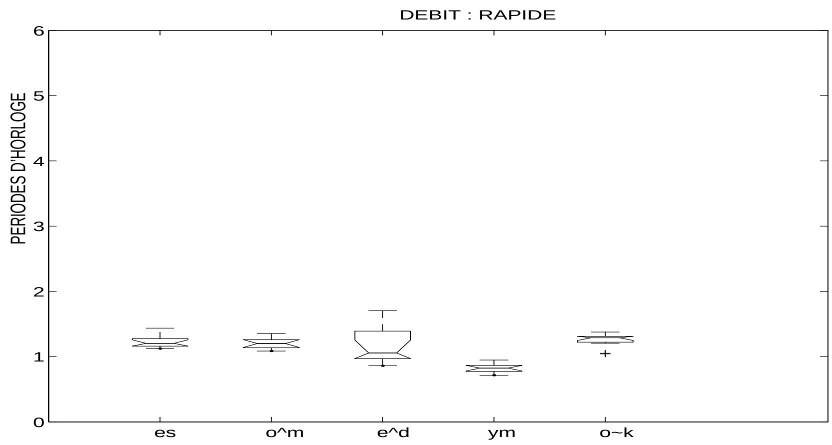

FIG. 5.11 - *Évolution du nombre de PH dans le corpus* PCenter*, débit* rapide*.*

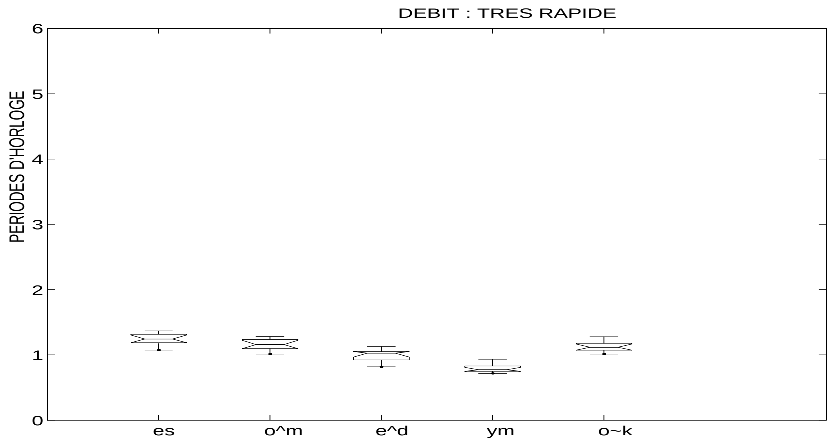

FIG. 5.12 - *Évolution du nombre de PH dans le corpus* PCenter*, débit* très rapide*. Il s'agit du sous-groupe « les sommets du mont». Il n'y a pas de silence après les GIPC.*

Fant et collègues (1991), aucune conclusion n'a pu être tirée. Ceci est discuté dans notre article « Characterisation of rhythmic patterns for text-to-speech synthesis» (Barbosa & Bailly 1994), reproduit ici en fin de volume.

La thèse, défendue par Duez (1987), que la pause fait partie du phénomène d'allongement final nous a inspiré la possibilité de générer le silence éventuel à partir du propre mécanisme d'allongement. Ceci a été possible grâce à la caractérisation des GIPC par les $z-scores$.

5.4 Émergence de la pause

5.4.1 Relation entre durées de la voyelle et du silence

Dans le but de savoir quelle durée doit être attribuée au silence dans le cadre du IIPC, nous avons étudié la relation entre le $z-score$ réel (z_r) de chaque voyelle et son $z-score$ virtuel (z_v), obtenu à partir de la durée de cette même voyelle additionée de la durée du silence qui éventuellement la suit. Dans ce corpus, puisque il s'agit de séquences CV, il y a toujours une voyelle avant un silence.

Pour des raisons de proximité des distributions des données à la fonction gaussienne, nous avons calculé le logarithme des valeurs de z_r et de z_v préalablement décalées, afin d'éviter le calcul du logarithme d'un numéro négatif. Le résultat se trouve dans la figure 5.13. Seulement les points correspondant à $z_r \neq z_v$ sont présentés.

Dans cette figure, la durée de la voyelle pour ce locuteur atteint des limites maxima d'allongement, au-delà desquelles le silence émerge par arrêt de la phonation.

La courbe de régression (avec corrélation de 73 %) est donnée par la formule :

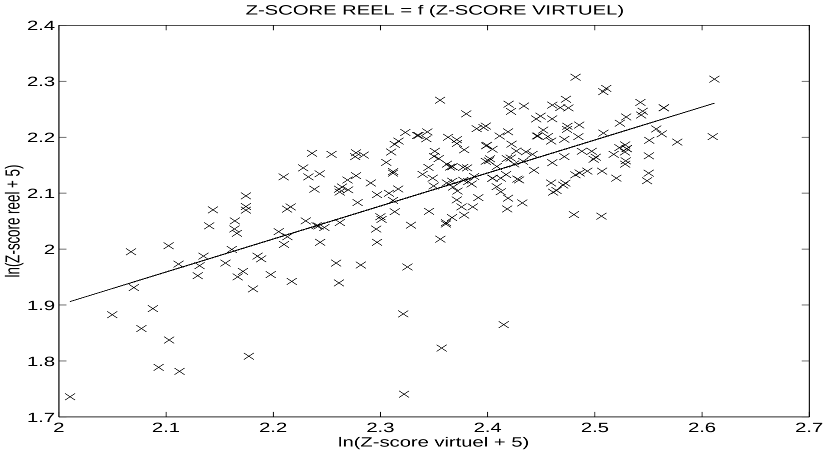

FIG. 5.13 - *Relation entre* $z-score$ *réel (voyelle) et* $z-score$ *virtuel (voyelle + silence) pour le corpus* PCenter, *tous débits confondus. La ligne de régression est donnée et son équation est :* $\ln(z_r + 5) = \ln(z_v + 5) \times 0,59 + 0,72$. *La ligne de régression rencontre la ligne* $z_r = z_v$ *à* $z_{critique} = 0,83$.

$$(z-score_{réel} + 5) = (z-score_{virtuel} + 5)^{0,59} \exp(0,72) \tag{5.1}$$

Cette ligne de régression coupe la ligne $z-score_{réel} = z-score_{virtuel}$, qui correspond à l'absence de silence après la voyelle, au point critique tel que $z-score_{critique} = 0,83$. En dessous de cette valeur il n'y pas de pause silencieuse après la voyelle.

Les points les plus éloignés de la ligne de régression correspondent surtout aux positions dans le débit *très rapide* où il y a un silence (phrases 12, 13, 14 et 15, après le mot « Karimo ») ou au débit *rapide*.

Ceci pourrait être une indication que le locuteur utilise, pour marquer la frontière prosodique, une stratégie différente entre les deux débits plus rapides et les trois autres. Une confirmation de cette hypothèse est donnée par la corrélation entre les durées intrinsèques de la voyelle et du silence qui la suit, pour chaque débit. Toutefois n'oublions pas qu'il s'agit seulement de quatre silences dans le débit *très rapide*. Le résultat se trouve dans le tableau ci-dessous.

Quoique les résultats des corrélations pour les débits *rapide* et *très rapide* ne soient pas très significatifs, on peut discerner deux tendances différentes entre ces débits et les trois autres. Ce résultat peut être mis en relation avec l'étude de Kohno (1992) sur deux sortes de mécanismes cognitifs pour le traitement des séquences sonores. En ayant mesuré la durée de ce qu'il appelle le *inter-beat interval* (IBI) — comparable à notre IIPC — il conclut que des traitements cognitifs globaux (holistiques) sont utilisés dans des conditions rythmiques rapides (IBI en dessous de 300 ms) et des traitements cognitifs

TAB. 5.5 - *Corrélation par méthode de Pearson entre les durées de la voyelle et du silence qui la suit, pour chaque débit du corpus* PCenter.

débit	corrélation	signifiance	degrés de liberté
très lent	0,22	$p < 0,03$	94
lent	0,31	$p < 0,01$	61
normal	0,27	$p < 0,2$	30
rapide	-0,34	$p < 0,1$	24
très rapide	0,01	$p < 1,0$	2

analytiques sont utilisés pour des durées au dessus de 400 ms, résultat souligné aussi par Dupoux (1989). Pour des tempos où l'IBI se trouve entre 300 et 400 ms, il existe une variation interindividuelle : chaque personne utiliserait soit un type de processus soit l'autre. Or, la plage de variation des deux débits plus rapides se trouve en dessous de 300 ms (cf. tableau 5.4, ce qui pourrait indiquer l'utilisation d'un processus holistique dans le traitement de la séquence sonore. La plage de variation du IIPC des autres trois débits se trouve proche de 300 ms jusqu'à plus de 400 ms, ce qui pourrait indiquer l'utilisation d'un processus analytique par le locuteur (vu le type de corrélation entre voyelle et silence ci-dessus).

L'étude de la relation entre $z-score$ de la voyelle et $z-score$ de la voyelle avec silence ajouté a permis l'intégration de ce mécanisme d'émergence de la pause dans le modèle de répartition.

5.4.2 Les Modèles de répartition simple et modifié

Le modèle de répartition, en sa première version, peut être présenté, d'après ce qui a été exposé dans le chapitre précédent par les pas suivants :

1. calcul du $z-score$ associé au GIPC (z_g) par la formule 4.1 :

$$Durée(GIPC) = \sum_{i=1}^{n} \exp(\mu_i + z_g \sigma_i) \tag{5.2}$$

   Le calcul est réalisé par récursion, à partir d'une valeur initiale affectée au $z-score$. Il est clair que pour ce calcul la durée du GIPC, le nombre et nature des phonèmes sont donnés au départ. Les indices i concernent les n phonèmes du GIPC et les valeurs (μ, σ) sont puisées dans le tableau 4.2, au chapitre précédent ;

2. calcul des durées phonémiques en imposant la valeur trouvée précédemment à la formule :

$$\exp(\mu_i + z_g \sigma_i) \tag{5.3}$$

Cette formule est appliquée dans le calcul de la durée de chacune des n réalisations du GIPC.

Cet algorithme sera désigné comme le **modèle de répartition simple**.

La répartition de la durée du GIPC entre ses éléments, avec éventuelle attribution d'une durée de silence, est basée sur l'analyse précédente (par régression linéaire) et la durée minimum de silence (cf. tableau 5.4), mesurée pour chaque débit. On suppose évidemment que l'on ne sait pas, en principe, si un silence fait partie ou non du GIPC en question. Dans notre expérience d'apprentissage automatique, le système qui génère les durées des GIPC n'a aucune information a priori sur la présence ou non de silence.

L'organigramme de la figure 5.14 illustre l'algorithme du modèle de répartition modifié pour intégrer l'émergence de la pause.

Les durées phonémiques et la pause silencieuse éventuelle correspondant à chaque GIPC d'une phrase seront calculées comme suit :

1. calcul du $z-score$ associé au GIPC (z_g) ;

2. si z_g est plus petit ou égal à $z_{critique}$, la procédure a fini : les durées phonémiques sont obtenues par la formule 5.3 du modèle de répartition simple ;

3. si z_g est plus est plus grand que $z_{critique}$, calcul du $z - score$ correspondant à la partie sonore (z_{sonore}) par la formule donnée par la ligne de régression ci-dessus : $z_{sonore} = (z_g + 5)^{0,59} \exp(0,72) - 5$;

4. calcul des durées phonémiques par la formule 5.3, avec z_{sonore} à la place de z_g ;

5. somme des durées phonémiques obtenues dans le pas précédent : la différence par rapport à la durée originale du GIPC est attribuée à la pause silencieuse ;

6. si la durée de la pause silencieuse est supérieure à la durée de silence minimum (tableau 5.4), la procédure est finie ;

7. si la durée de la pause silencieuse est inférieure ou égale à la durée de silence minimum, aucun silence ne sera attribué à ce GIPC et les durées phonémiques avec le $z-score$ z_g, par la formule 5.3.

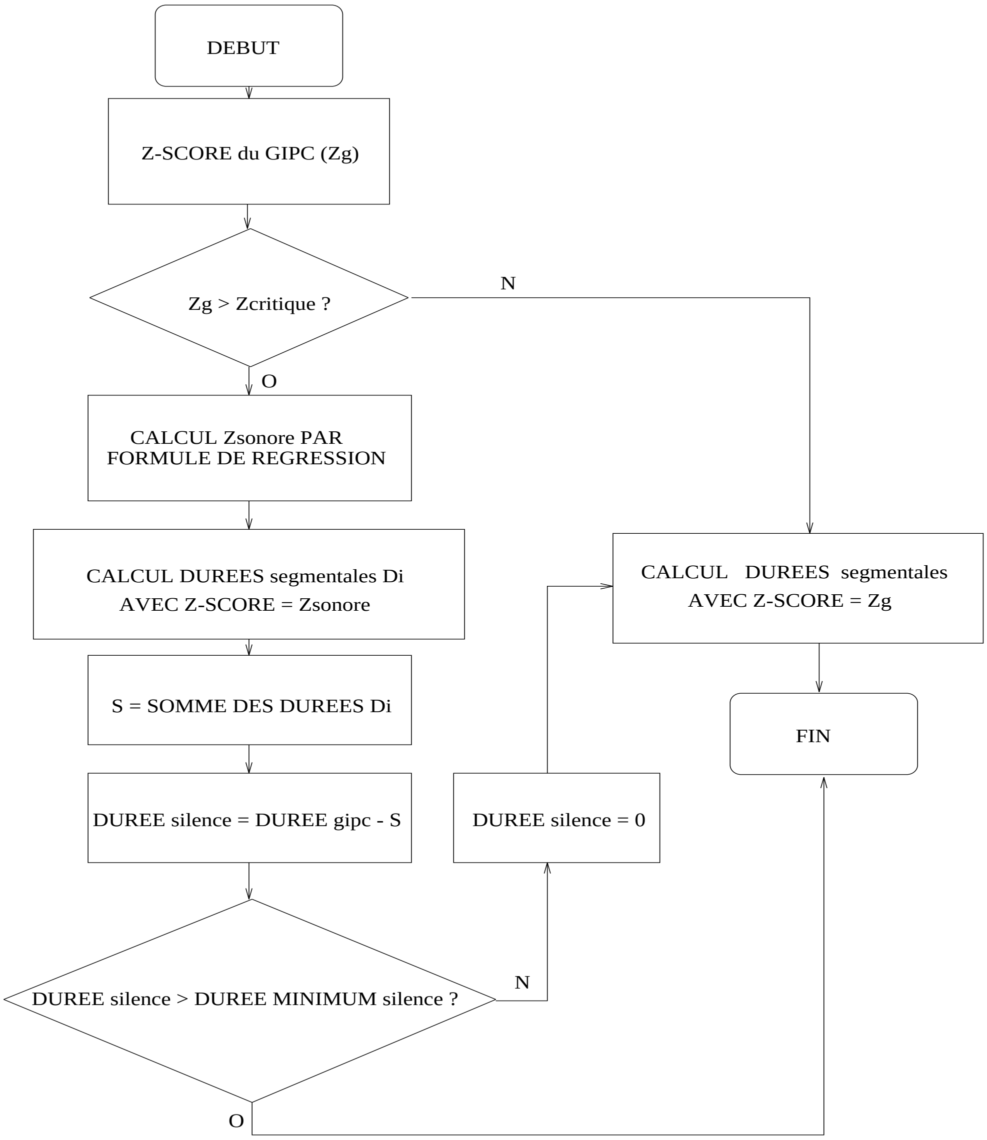

FIG. 5.14 - *Organigramme du modèle de répartition avec émergence de la pause.*

Cet algorithme sera désigné comme le **modèle de répartition modifié**.

Il convient maintenant de tester le modèle de répartition modifié en vue de la prédiction automatique.

5.4.3 Modèle de répartition : erreurs de prédiction

Au départ, la durée observée de chaque GIPC du corpus *PCenter* sera utilisée pour le calcul des durées phonémiques.

On analysera tout d'abord les erreurs de placement du silence par l'application du modèle de répartition modifié. Ces erreurs peuvent être de deux types : soit le modèle a prédit un silence et le GIPC observé n'en possède pas un, soit le modèle a omis un silence observé dans la phrase naturelle. À l'exception d'une erreur dans le débit *lent* du second type, tous les autres sont du premier type (surplacement de silences).

Les erreurs de placement sont reproduits dans le tableau 5.6.

TAB. 5.6 - *Nombre d'erreurs de placement du silence par le modèle de répartiton modifié. Le nombre de pauses silencieuses observées dans chaque débit du corpus* PCenter *est aussi donné. Il y a 285 GIPC par débit.*

débit	PH moyenne (ms)	nombre d'erreurs	nombre de silences observés
très lent	345	15	96
lent	317	13	64
normal	252	11	32
rapide	187	-	26
très rapide	150	1	4

Il est important de noter qu'aucune contrainte de localisation, comme une position accentuable, a été imposée pour le placement des silences : ils ont été placés par l'application du modèle de répartition modifié pour chaque GIPC.

La plupart des erreurs ci-dessus sont des erreurs de surplacement : le modèle a prédit un silence qui n'est pas observé dans la phrase naturelle. Cependant, ces silences ont été tous placés dans une position correspondante à une marque prosodique et de ce fait, accentuable. Noter que nous avons pris la décision de favoriser le placement de pause silencieuse au dessus d'une valeur minimale pour le silence qui pourra être affecté.

Les erreurs de prédiction pour les réalisations phonémiques et pour les silences sont données dans les tableaux 5.7 et 5.8.

TAB. 5.7 - *Erreurs de prédiction par débit et pour les réalisations phonémiques du modèle de répartition modifié. La moyenne (et l'écart-type) des erreurs a été calculée à partir de leurs valeurs absolues. La durée observée moyenne des réalisations par débit est aussi donnée ainsi que le rapport erreur moyenne/durée moyenne.*

	Partie sonore		
	moyenne erreur absolue : E (ms)	durée moyenne réalisation : D (ms)	erreur/durée E/D (%)
très lent	32 (29)	166	19
lent	31 (29)	157	20
normal	26 (22)	132	20
rapide	18 (19)	97	19
très rapide	13 (15)	78	16

TAB. 5.8 - *Erreurs de prédiction par débit et pour le silence du modèle de répartition modifié. La moyenne (et l'écart-type) des erreurs a été calculée à partir de leurs valeurs absolues. La durée moyenne des silences observés par débit est aussi donnée ainsi que le rapport erreur moyenne/durée moyenne.*

	Pauses silencieuses		
	moyenne erreur absolue : E (ms)	durée moyenne réalisation : D (ms)	erreur/durée E/D (%)
très lent	52 (51)	400	13
lent	81 (60)	242	34
normal	75 (60)	190	39
rapide	68 (47)	285	24
très rapide	106 (18)	106	100

On a voulu tester comment les unités syllabe, GIPC et VO (pour *vowel onset*, comprenant deux *onsets* vocaliques consécutifs et donc pas de prise en compte de la modification due à la pause) se comportent pour la prédiction de la durée. Les résultats se trouvent dans le tableau 5.9.

TAB. 5.9 - *Moyennes (et écarts-type) à partir des valeurs absolues des erreurs de prédiction pour trois classes au niveau phonémique (voyelles, consonnes et silences), dans le cadre de trois unités de programmation rythmique : la syllabe (SY), le groupe inter-*onset *vocalique (VO) et le GIPC (PC). Les valeurs de PH moyenne sont indiquées.*

débit	PH	*Voyelles*			*Consonnes*			*Silences*		
	(ms)	SY	VO	PC	SY	VO	PC	SY	VO	PC
tr. lent	345	34(32)	35(32)	34(29)	43(46)	42(43)	37(46)	67(58)	76(62)	55(53)
lent	317	38(40)	42(41)	36(32)	32(32)	34(31)	28(26)	73(52)	73(52)	73(56)
normal	252	26(26)	30(38)	29(23)	27(26)	28(26)	25(22)	54(37)	64(36)	56(35)
rapide	187	17(19)	18(19)	19(20)	21(25)	21(27)	19(21)	75(77)	68(99)	49(78)
tr. rapide	150	12(12)	11(13)	12(14)	13(13)	12(17)	11(12)	59(33)	93(28)	44(20)

Dans les débits *très lent*, *rapide* et *très rapide*, si l'on jette un regard sur la prédiction des silences, le GIPC fourni des erreurs absolues moyennes inférieures en comparaison avec celles obtenues avec les deux autres unités de programmation. Dans les deux autres débits, les valeurs pour la syllabe et le GIPC sont comparables. Quoique les prédictions pour les durées des consonnes et voyelles soient à peu près comparables en moyenne, les erreurs commises en positions accentuées, certes moins nombreuses, sont inférieures pour le GIPC, vu que la syllabe et le groupe VO comprennent des réalisations phonémiques avec un allongement très hétérogène (consonnes en position d'attaque pour la syllabe, consonnes après le silence pour le groupe VO) (Campbell 1992).

La conséquence perceptive des erreurs commises par l'algorithme de répartition sera présentée plus tard, sous la forme d'un test de perception, dans le cadre de notre modèle de génération de la durée.

Maintenant que l'on sait comment la durée du GIPC sera distribuée entre ses composants, il faut aborder le problème de la génération automatique de la durée du GIPC.

L'outil de travail pour la génération automatique est le réseau connexionniste, dont une introduction est donnée dans l'annexe E.

L'emploi des réseaux connexionnistes dans la génération des paramètres prosodiques est récent, mais on voit leur usage s'intensifier. La fréquence fondamentale est le paramètre le plus concerné dans les systèmes à apprentissage supervisé.

5.5 La Prédiction des paramètres prosodiques par réseau neuromimétique

Hormis le système de Todd, qui a proposé un réseau destiné à prédire – de façon intégrée – la durée et la hauteur de mélodies simples, les systèmes de Sagisaka (pour le japonais), Scordilis & Gowdy (pour l'anglais) et Traber (pour l'allemand), génèrent la courbe de fréquence fondamentale.

5.5.1 Le Système de Todd

À l'époque, le seul modèle connexionniste, à notre connaissance, qui ait proposé une génération simultanée, pour la durée et la fréquence fondamentale, a été réalisé par Todd (1989) pour des partitions de musique.

Dans le cadre d'un réseau séquentiel (figure 5.15), la mélodie est représentée par la séquence de valeurs fréquentielles discrètes. La durée est implicitement codée par la répétition des mêmes valeurs fréquentielles au cours de la séquence : la durée est ainsi représentée par un nombre entier d'une valeur temporelle minimum. Cette description est appropriée au caractère discret de la durée en musique. Pour la parole, une autre représentation doit être envisagée.

Les résultats ont été encourageants pour un nombre restreint de mélodies possédant des séquences temporelles simples. Pour être capable de généraliser pour des séquences plus longues, le réseau doit être capable de manipuler des rythmes et des séquences temporelles plus complexes : *ceci peut être amélioré par l'addition d'autres formes de contexte, telles qu'une horloge globale ou compteur additionnée aux entrées* (*ibid.*, p. 42).

5.5.2 Le Système de Sagisaka

Un réseau du type perceptron à trois couches (figure 5.16) est utilisé par cet auteur (Sagisaka 1990) pour la génération d'une forme générale de la fréquence fondamentale du japonais, dans le cadre de la synthèse de la parole. Cette forme générale est décrite par un contour ascendant suivi d'un contour descendant. Ce patron est associé à un syntagme mineur. Quand une suite de formes de ce type constitue un ensemble descendant, ce nouveau patron est associé à un syntagme majeur.

La forme générale de la courbe de fréquence fondamentale associée à un syntagme mineur est décrite à la sortie par trois points : positions initiale, maximale et finale. La description à cinq paramètres de l'entrée concerne cependant le syntagme majeur courant :

- type d'accent du syntagme mineur courant ;

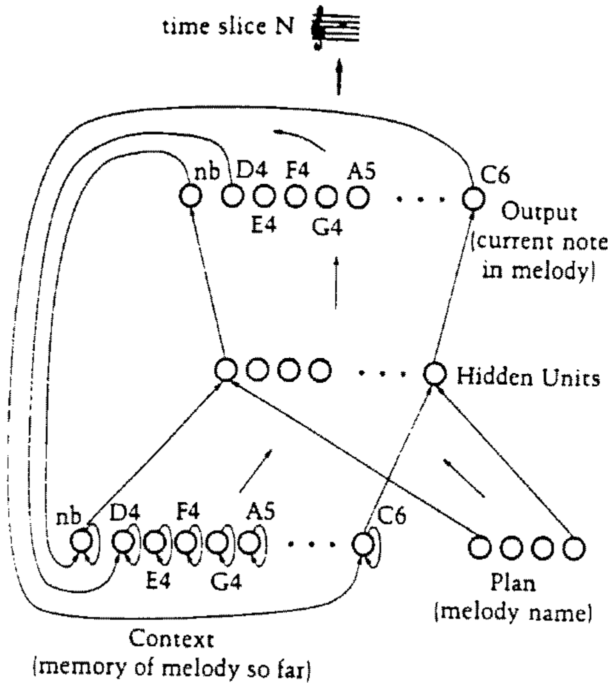

FIG. 5.15 - *Réseau neuromimétique utilisé par Todd (1989, p. 30) pour la génération simultanée de la durée et de la hauteur. Dans la figure, unités de contexte (mémoire de mélodie à long terme), unités de planification (nom de la mélodie), unités cachées et sortie (note actuelle dans la mélodie), unité de sortie avec note (n*e* *élément de taille* durée minimum*).*

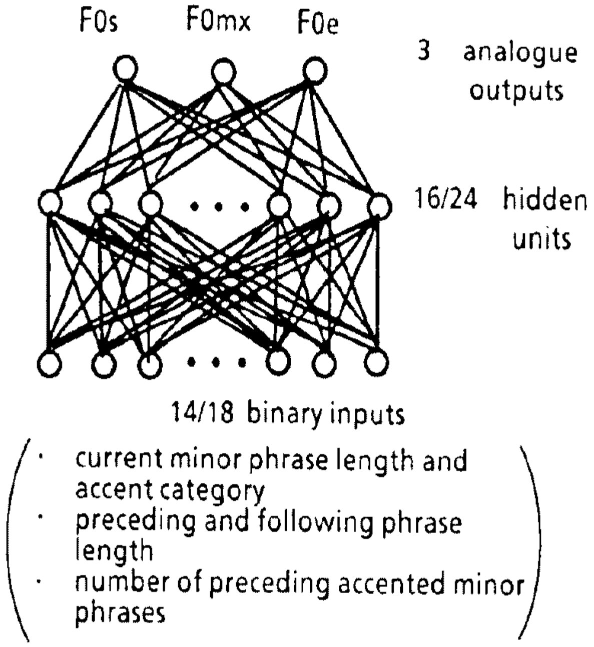

FIG. 5.16 - *Réseau neuromimétique utilisé par Sagisaka (1990, p. 326) pour la génération de la fréquence fondamentale. Voir texte principal pour description.*

- longueur du syntagme mineur courant ;
- nombre de syntagmes mineurs accentués précédant le syntagme majeur courant ;
- types d'accent à gauche et à droite de la position courante ;
- nombre de syntagmes précédents et suivants dans le syntagme majeur courant.

L'auteur dit que les erreurs de prédiction ont été comparables aux écarts-type des mêmes courbes de fréquence fondamentale naturelles avec les mêmes conditions d'entrée (des phrases différentes avec la même description en entrée, puisque la description de l'auteur ne fait pas intervenir le niveau phonémique). Les formes apprises sont cependant assez simples, constituées par une montée et une descente de la courbe de fréquence fondamentale.

5.5.3 Le Système de Scordilis & Gowdy

L'approche de Scordilis & Gowdy (1989) propose deux réseaux neuromimétiques. Le premier, le réseau macrophonémique, délivre la valeur de fréquence fondamentale associée à chaque phonème de la phrase et le deuxième, le réseau microphonémique, délivre la correction micromélodique appliquée à la première valeur, à chaque 10 ms au cours de la phrase. La langue étudiée est l'anglais.

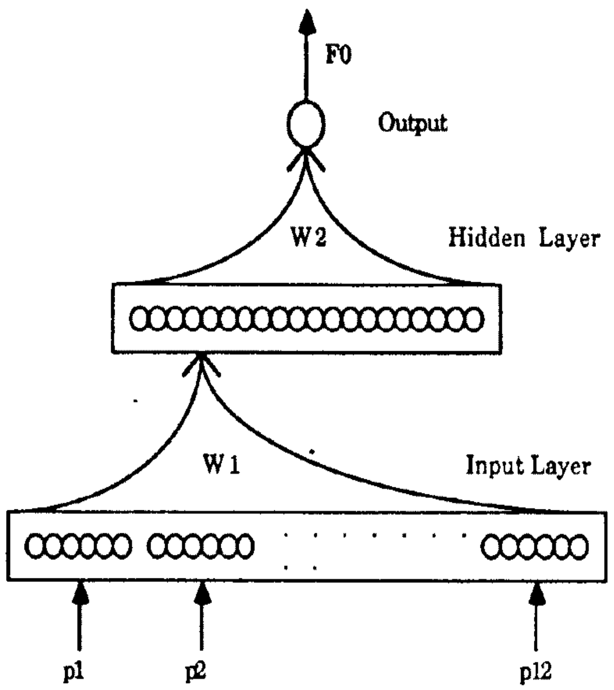

FIG. 5.17 - *Réseau macrophonémique utilisé par Scordilis & Gowdy (1989, p. 220) pour la génération de la fréquence fondamentale.*

Le réseau macrophonémique (figure 5.17) est un perceptron à une couche cachée. En entrée, il possède la description, selon les propriétés acoustiques, de 12 phonèmes avec une fonction d'activation linéaire appliquée à chaque neurone (72 neurones en entrée, six par phonème). La couche cachée possède un nombre variable de neurones et une fonction d'activation du type sigmoïde variant entre 0 et 1. Le neurone de sortie délivre la valeur de fréquence fondamentale associée au deuxième phonème de l'entrée (10 contextes gauches et un seul droit) par une fonction d'activation sigmoïde dont les limites correspondent aux valeurs minimum et maximum observées pour la fréquence fondamentale.

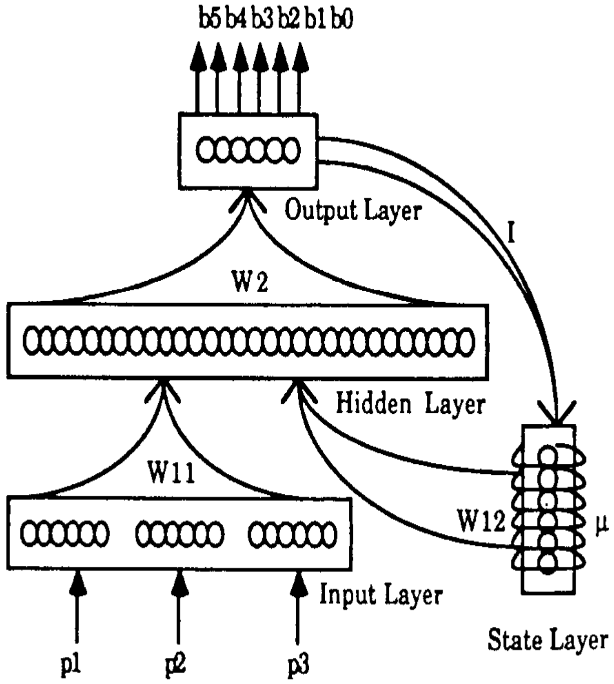

FIG. 5.18 - *Réseau microphonémique utilisé par Scordilis & Gowdy (1989, p. 220) pour la génération des variations locales de la fréquence fondamentale.*

Le réseau microphonémique (figure 5.18) est récurrent : les neurones de sortie sont connectés à une couche de neurones d'état, celle-ci connectée à la seule couche cachée. Chacun des trois phonèmes décrits en entrée est représenté par six neurones. La sortie possède six neurones (sous la forme de six bits) qui décrivent, dans le cadre d'un phonème, les variations de la fréquence fondamentale à partir du niveau moyen donné par le premier réseau.

Les erreurs de prédiction ont été de l'ordre de 4 Hz au maximum et les auteurs concluent que le degré de succès de ce type d'approche dépend de la taille et de la composition du corpus d'apprentissage. En ce qui concerne la durée ils écrivent : « La durée d'un phonème est en soi un problème intéressant mais elle n'est pas considérée dans ces expériences mais a été supposée connue. » (*ibid.*, p. 222).

On voit bien que le réseau macrophonémique a besoin d'intégrer une large information sur le passé (10 contextes à gauche) pour se passer d'une architecture récurrente.

5.5.4 Le Système de Traber

Cet auteur compare deux approches pour la génération des contours de fréquence fondamentale de la phrase allemande (Traber 1992). La première est une approche (que l'on appellera par « lexiques » : voir Aubergé 1991) qui emploie la concaténation de parties préalablement stockées du contour de fréquence fondamentale. La deuxième est l'approche par réseau neuromimétique.

Le réseau à quatre couches (figure 5.19) utilisé par l'auteur est récurrent (des boucles relient la deuxième couche cachée à l'entrée) et complètement connecté. La fonction d'activation associée à tous les neurones est la fonction sigmoïde.

À l'entrée, il y a une description phonologique de la phrase dont les éléments représentés appartiennent à deux catégories. D'abord ceux qui concernent la phrase comme un tout,

- quatre niveaux accentuels par rapport au syntagme ;
- frontières de syntagme et de mot ;
- marques indiquant type de syntagme (final ou non final) et position par rapport à l'accent de syntagme (avant ou après).

Ensuite, les éléments qui concernent la description phonologique de la syllabe dont le contour de fréquence fondamentale associé sera reproduit :

- voyelle longue ou courte ;
- voyelle haute ou basse, par rapport à la fréquence fondamentale intrinsèque ;
- voisement du contexte consonantique gauche et droit.

Ces deux codages sont associés à une syllabe par un système de fenêtrage qui parcours toute la phrase, syllabe par syllabe. La fenêtre correspondant au niveau phrastique a des informations qui s'étendent sur une région assez large (13 syllabes) sur le passé et l'avenir.

À la sortie, huit valeurs variant entre 0 et 1 représentent l'évolution fréquentielle de la syllabe correspondante. Les valeurs sont recodées entre 50 et 200 Hz, intervalle approprié au corpus d'apprentissage avec lequel le réseau a été alimenté.

L'auteur conclut que les deux méthodes de génération fournissent des contours de fréquence fondamentale de haute qualité à un système de synthèse de la parole à partir

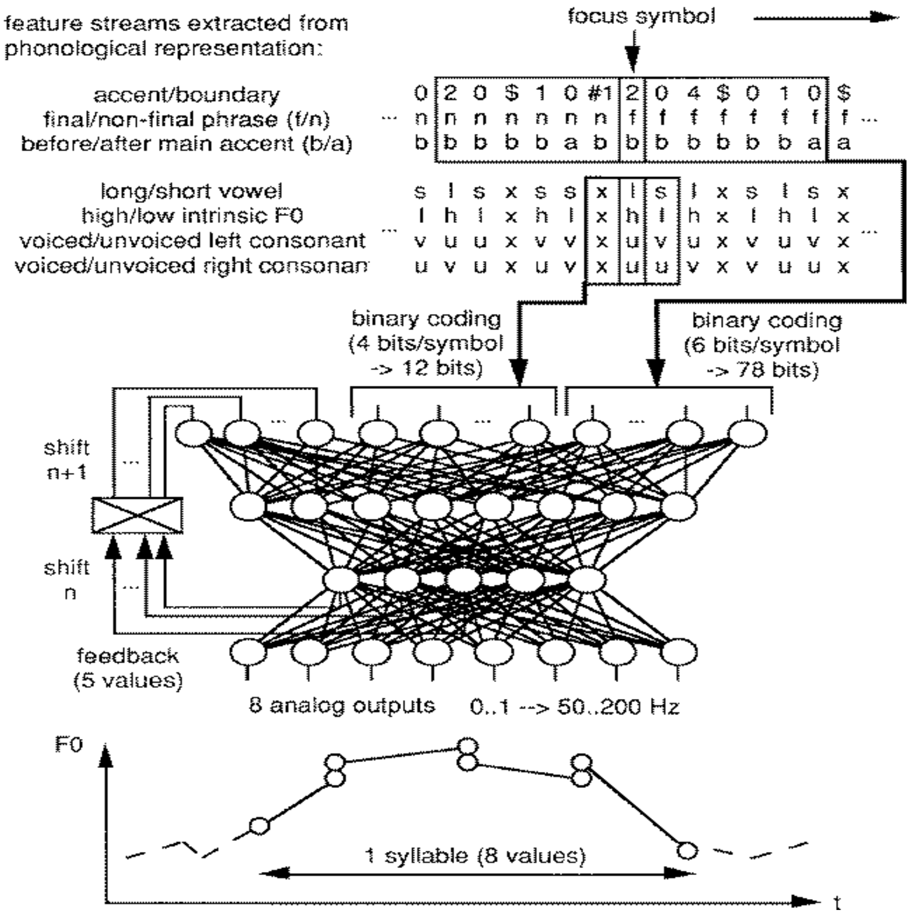

FIG. 5.19 - *Réseau neuromimétique utilisé par Traber (1992, p. 299) pour la génération de la courbe de fréquence fondamentale. L'entrée possède une fenêtre avec 6 symboles à gauche et 6 symboles à droite représentant des informations à trois valeurs au niveau phrastique (à long terme) et une fenêtre avec 3 symboles représentant quatre informations sur la nature de la voyelle, le contexte consonantique gauche et droit et la F0 intrinsèque. À la sortie, l'évolution de la fréquence fondamentale pour la syllabe courante.*

du texte. De plus, ces méthodes sont flexibles et facilement adaptables à d'autres besoins. L'approche réseau, cependant, a nécessité beaucoup moins d'effort humain pour une haute qualité de génération (cf. le travail prometteur de Morlec 1994, pour le français : la même procédure est utilisée et ses conclusions vont dans la même direction). La performance obtenue avec le réseau est une corroboration de la description phonologique adoptée, cependant, l'usage d'une fenêtre qui comprend un si large contexte n'est pas justifiable dans un réseau récurrent.

5.5.5 Discussion

Un perceptron semble être un bon choix pour l'apprentissage de classes assez simples, comme est le cas du système de Sagisaka, où des formes globales de la fréquence fondamentale sont apprises, sans faire appel à une description micromélodique de la phrase. Si l'on regarde le réseau macrophonémique de Scordilis & Gowdy, qui fourni des informations concernant la nature des phonèmes, on voit que la décision d'employer un perceptron multicouches entraîne une part importante d'information sur le passé : 10 phonèmes en contexte gauche.

Le choix d'un réseau récurrent nous paraît être le meilleur choix afin d'éviter un emploi excessif de la description du contexte gauche. Cette qualité des réseaux récurrents n'est pas suffisamment exploitée dans le réseaux de Traber. De plus, son système de fenêtrage peut être éliminé par le choix d'un réseau séquentiel (ce type de réseau tient compte d'une description au long de la phrase sous la forme de séquences. Chaque séquence peut être associée à une syllabe ou à un GIPC).

5.6 Génération automatique de la durée

L'approche connexionniste a été très peu utilisée pour la prédiction de la durée. Si l'on souhaite obtenir les paramètres prosodiques de façon intégrée il faut aussi penser à prédire la durée par un réseau connexionniste.

Dans cette section on présentera le modèle combiné pour la génération de la durée phonémique : apprentissage par réseau neuromimétique et modèle de répartition.

Notre modèle de génération de la durée s'est inspiré dans l'approche de Campbell (1992). Il fonctionne en deux étapes, comme est représenté schématiquement par la figure 5.20.

Dans la première étape, un réseau connexionniste délivre la durée du GIPC à partir de la description phonologique de la phrase et d'une mesure du débit. Puisque ce stade

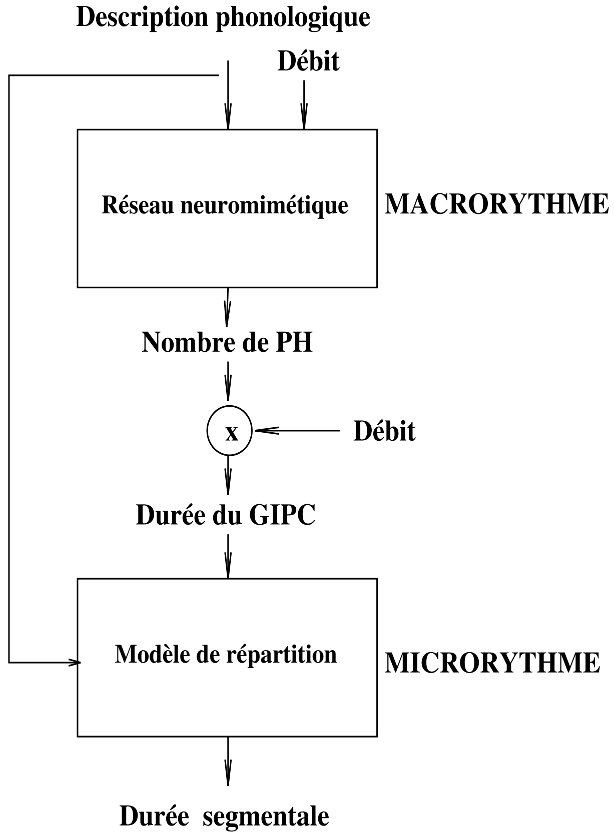

FIG. 5.20 - *Schéma du modèle de génération automatique de la durée phonémique.*

génère la durée du GIPC, l'unité de programmation rythmique, il est lié à la **composante macrorythmique**. Dans une deuxième étape, le modèle de répartition s'occupe de distribuer la durée du GIPC entre chacun des ses composants. Ces composants sont les réalisations phonémiques et les éventuels silences de l'énoncé. Puisque ce stade traite du niveau inférieur à celui de la structuration rythmique, il est lié à la **composante microrythmique**.

Le modèle de répartition modifié a été largement abordé dans ce travail. Il convient de décrire les choix concernant le réseau connexionniste.

5.6.1 Le Réseau connexionniste

À la différence du réseau adopté par Campbell, nous avons décidé d'intégrer deux acquis du paradigme connexionniste. La **récurrence** nous permet d'avoir une mémoire de ce qui vient de se passer, ce qui est intéressant si l'on met en relation avec le mécanisme d'accentuation progressif ou l'influence des phonèmes en contexte gauche sur la durée de la réalisation phonémique courante. Le besoin d'une représentation qui tenait compte du temps s'est fait nécessaire d'emblée. Ce temps est considéré de façon discrétisé tout au long de la phrase sous la forme séquentielle. De ce fait le réseau choisi est un **réseau**

récurrent et séquentiel. Ceci est nécessaire d'autant plus que l'on se propose à générer la structure rythmique dans le contexte de l'énoncé, ce qui n'est pas le cas de l'approche de Campbell, qui génère la syllabe de façon pratiquement isolée. Une autre différence par rapport à la méthode de Campbell est bien entendu la génération de la durée du GIPC à la place de celle de la syllabe. Il n'est plus nécessaire de revenir sur les raisons de ce choix.

L'architecture du réseau est représentée dans la figure 5.21.

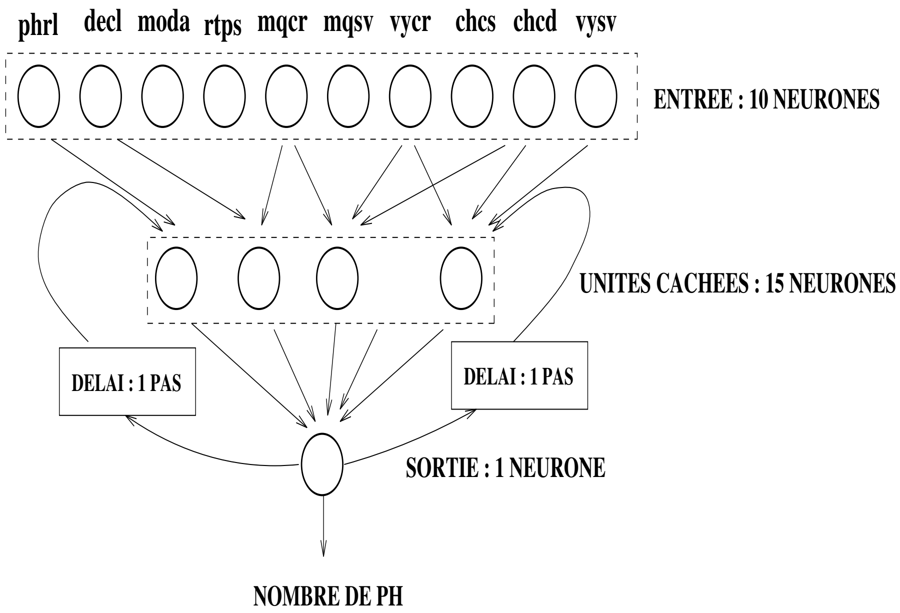

FIG. 5.21 - *Architecture du réseau connexionniste employée. Il s'agit d'un réseau complètement connecté. Toutes les connexions n'ont pas été représentées pour des raisons de clarté. La fonction d'activation appliquée dans les neurones de la couche cachée est la fonction sigmoïde variant entre -1 et 1. La valeur d'activation de la sortie est obtenue par la somme des valeurs d'activation de la couche antérieure.*

Le débit et la description phonologique sont décrits à l'entrée du réseau par des nombres réels variant continûment entre 0 et 1. Les dix neurones donnent une description de la phrase de façon séquentielle : une description est présentée au réseau à chaque pas de la séquence. Le temps est représenté par la séquence de GIPC de la phrase. Certains éléments de l'entrée changent à chaque pas et d'autres restent inchangés tout au long de la phrase. La description phonologique à l'entrée du système décrit, à chaque pas, le GIPC courant au niveau phonémique. Les neurones d'entrée sont :

phrl Il s'agit de la période d'horloge (voir plus haut) associée à la phrase et exprimée en secondes. Le maximum possible est une seconde. Elle est la même pour toute la séquence ;

decl Ligne de déclinaison. Sur la base d'un maximum de 100 syllabes, il s'agit d'une variable réelle qui commencent par le nombre de GIPC de la phrase divisé par le maximum au premier pas de la séquence et fini à 0 en fin de phrase (et fin de la séquence) ;

moda La modalité de la phrase. Elle est codée entre 0 et 1. Toutes les phrases du corpus d'apprentissage étaient assertives et ont été représentées avec cette entrée codée par 0. La valeur de 0,5 a été laissée pour la phrase impérative et celle de 1 pour la phrase interrogative ;

rtps Il s'agit d'une rampe discrète ou d'un compte à rebours évoluant entre deux marques prosodiques successives ou entre le début de la phrase et la première marque ou entre la dernière marque et la fin de la phrase. La valeur initiale dans chaque groupe intermarque est définie par le nombre de GIPC qui manquent jusqu'à la marque prosodique suivante plus proche divisé par le nombre de GIPC dans le groupe intermarque. La valeur finale, qui coïncide avec un GIPC dominé par une marque prosodique, est toujours nulle ;

mqcr Nature de la marque prosodique qui domine le GIPC courant : il s'agit de la marque qui le domine directement ou de la marque suivante la plus proche. Les marques prosodiques ont été codées entre 0 et 1 proportionnellement à leur hiérarchie déterminée par l'approche $z-score$ présentée plus haut ;

mqsv Nature de la marque prosodique qui suit la marque de l'item précédent ;

vycr Nature de la voyelle du GIPC courant. Toutes les voyelles du français ont été codées entre 0 et 1 selon un degré ascendant de durée intrinsèque : des plus fermées et intrinsèquement plus courtes jusqu'aux voyelles plus ouvertes et plus longues (les nasales) ;

chcs Charge consonantique. Codage entre 0 et 1 (pour un maximum de 8) du nombre de consonnes entre la voyelle courante et la voyelle suivante ;

cgcd Charge consonantique du coda correspondant à la syllabe courante. Le codage est aussi entre 0 et 1 sur la base d'un maximum de huit consonnes. Ce codage a été nécessaire puisque lors de l'émergence du silence il se met juste après le coda de la syllabe classique et jamais après l'éventuel attaque de la syllabe suivante. Cette information a été jugée nécessaire dans le cas d'un apprentissage concernant de phrases plus complexes que des séquences CV. Cette information est utilisée dans le modèle de répartition pour bien placer la pause silencieuse éventuelle sur la chaîne phrastique ;

vysv Nature de la voyelle qui débute le GIPC suivant. Le codage est le même que dans la voyelle courante (**vycr**).

On peut dire que les unités **phrl**, **decl** et **rtps** vont dans le sens d'intégrer au système une réalité biologique : l'hypothèse de l'existence d'une horloge interne, corroborée par la tendance à l'isochronisme dont la déviation produite par l'accent a une signification linguistique. En fait, l'intégration d'une horloge interne pour l'obtention de la durée dans le cadre d'un réseau connexionniste a déjà été suggérée par Todd (cf. section 5.5.1)

La seule couche cachée possède 15 neurones. Leur valeur d'activation est déterminée par une combinaison linéaire des valeurs des neurones à l'entrée et de la valeur d'activation décalée d'un pas du neurone de la sortie, suivi de l'application d'une fonction du type sigmoïde limitée entre -1 et 1.

La valeur d'activation du seul neurone de la sortie est déterminée par une combinaison linéaire des neurones de la couche cachée. La sortie que le réseau doit délivrer, et qu'il apprend par une règle de rétro-propagation adaptée aux réseaux séquentiels et récurrents (Rumelhart, Hinton & William 1986), est le nombre de périodes d'horloge, comme expliqué dans la section 5.3.

Ce réseau récurrent et séquentiel a été simulé dans le cadre du logiciel *xrbp* développé, à partir d'un programme d'un groupe de chercheurs en traitements parallèles et distribués (*PDP*, en anglais) (McClelland & Rumelhart 1988), par Laboissière (1992, chapitre 3) et utilisé largement dans le cadre du problème de l'inversion en parole (des formants jusqu'aux articulateurs). Le succès de ce logiciel pour la simulation des réseaux connexionnistes les plus divers est aussi confirmé par le travail de Lœvenbruck & Perrier (1993) où ils l'emploient pour la recherche d'invariants articulatoires dans un signal de grande variabilité (phenomènes de réduction vocalique). Il permet la simulation complète d'un réseau connexionniste arbitraire : définition de l'architecture, apprentissage par la règle de rétro-propagation, affichage et test.

Le corpus d'apprentissage utilisé est un sous-ensemble des phrases du corpus *PCenter*. Les résultats de la prédiction seront présentés dans la prochaine section.

5.6.2 La Prédiction automatique

Les phrases impaires du corpus *PCenter* (voir section A.2) dans les cinq débits ont constitué le **corpus d'apprentissage**, dans un total de 50 phrases. Les phrases paires (50 phrases) dans les cinq débits, le **corpus de test**. Rappelons que seules les phrases du corpus d'apprentissage sont présentées au réseau. Toutes les phrases ont été décrites selon le codage présenté dans la section précédente. L'apprentissage a été réalisé en présentant au réseau la paire entrée/sortie désirée. La sortie présentée est la séquence des durées du GIPC de chaque phrase, exprimée en termes de nombre de périodes d'horloge (*durée du GIPC/période d'horloge*).

Une fois appris (procédure fréquemment lourde en temps de machine : approximativement 200 000 pas, chaque 20 000 pas durant de 3 à 6 heures pour être complétés), le réseau est prêt à fonctionner : à chaque description donnée en entrée, il fourni une évolution de la durée au long de la phrase. Chaque durée de la séquence est exprimée en termes de PH.

À l'entrée de ce modèle, la période d'horloge (voir figure 5.20) est utilisée pour le calcul des durées de la séquence de GIPC de la phrase. Chaque durée est employée dans le calcul des durées phonémiques par le modèle de répartition modifié. Rappelons que le nombre et la nature des phonèmes sont nécessaires pour le calcul des $z-scores$ associés au GIPC et utilisés pour le calcul des durées phonémiques. Les phonèmes pris pour le calcul de la durée du premier GIPC de la phrase sont considérés être la voyelle et les consonnes éventuelles entre celle-ci et la voyelle suivante.

Dans ce stade on décide de la durée du silence éventuel dans le GIPC et on le place (si il existe) après la frontière syllabique (la syllabe concernée est celle contenant la voyelle du GIPC courant). Évidemment, s'il y avait des consonnes non occlusives sourdes dans le GIPC courant, elles doivent être considérées dans la distribution de la durée du GIPC suivant. Le processus continue jusqu'à la fin de la phrase.

Cette procédure a permis la prédiction des durées phonémiques à l'aide du réseau. Les résultats sont présentés dans le tableau 5.10.

Si l'on compare les deux corpora, celui d'apprentissage et celui de test, on voit qu'il y a une tendance à la généralisation : les écarts-type des réalisations phonémiques, notamment, sont comparables entre les deux corpora. Le placement inexact des silences s'accentue avec la durée de GIPC délivrée par le réseau (voir tableau 5.7). Ceci est dû à une position accentuelle mal apprise. Ce problème est illustré dans les figures 5.23 et 5.24 à la section suivante.

Si l'on compare le tableau ci-dessus avec le tableau des erreurs du seul modèle de répartition (tableau 5.7), on voit que ce dernier est la plus grande cause des erreurs du modèle de génération. Les erreurs introduites par le modèle de répartition sont dues à l'utilisation de l'hypothèse forte d'élasticité, à savoir l'allongement uniforme des éléments du GIPC.

Les erreurs introduites par le réseau peuvent théoriquement être baissées par la conti-

TAB. 5.10 - *Moyenne (et écart-type) en millisecondes à partir de la distribution des erreurs entre les durées observées et les durées prédites. Les durées présentées concernent les réalisations phonémiques et les silences. Les résultats concernent le corpus d'apprentissage et le corpus de test.*

	Apprentissage		*Test*	
débit	silences	réalisations phonémiques	silences	réalisations phonémiques
très lent	-41 (137)	-1 (49)	3 (207)	4 (56)
lent	70 (116)	-2 (40)	-82 (147)	6 (46)
normal	67 (122)	0 (37)	-105 (113)	5 (43)
rapide	-189 (84)	3 (30)	64 (144)	0 (28)
très rapide	-4 (170)	2 (25)	-49 (193)	-1 (23)

nuation du processus d'apprentissage (si l'erreur est minimum, le corpus d'apprentissage n'introduirait pas d'erreurs supplémentaires au modèle de génération) et d'un bon choix de phrases pour l'apprentissage (ceci pourrait améliorer la capacité du réseau à généraliser).

Quoique certaines erreurs trop grossières, introduites par le modèle de répartition, nous aient permis de corriger quelques erreurs dans la segmentation des phrases des trois corpora, la performance générale de la prédiction au niveau microrythmique reste une conséquence de l'hypothèse d'élasticité adoptée. Reste à savoir son incidence dans le processus de perception, ce qui sera analysé dans la section suivante. Un début de réponse a été donnée dans le test de perception précédemment présenté puisque les phrases utilisées dans ce test ont semblé avoir des durées assez naturelles.

5.7 La Pertinence perceptuelle du modèle de prédiction

Nous avons pensé à valider les durées prédites par notre modèle de génération en faisant un test de perception. Ce qui sera évalué principalement est la tendance du réseau à maintenir la structure rythmique de la phrase.

Comme dans le test de perception précédent, les stimuli ont été obtenus par une technique d'analyse-synthèse de haute qualité (Moulines 1992; Bailly, Barbe & Wang 1992) où seules les durées phonémiques ont été modifiées.

Deux versions de chaque phrase utilisée dans le test ont été comparées. La première, étiquetée version « modèle» , est obtenue en imposant les durées de chaque réalisation (silence ou son) calculées par notre modèle de génération combiné. La deuxième, étiquetée version « bruit », est obtenue en modifiant les durées phonémiques de la phrase naturelle d'une valeur aléatoire. Cette valeur aléatoire est issue d'une distribution gaussienne ayant les même moyenne et écart-type que la distribution, très proche de la distribution gaussienne (corrélation supérieur à 98 %), des erreurs de prédiction de notre modèle de génération (voir tableau 5.10). Une distribution gaussienne par débit et par catégorie de réalisation (une pour les pauses silencieuses et une autre pour les réalisations phonémiques) a été considérée pour l'obtention des valeurs aléatoires de la version « bruit».

Ce qui sera évalué est la tendance de la version « modèle» à maintenir les positions des *onsets* vocaliques de la phrase naturelle, avec une certaine marge d'erreur, contre les durées de la version « bruit», présentant la même marge d'erreur mais, non guidée par des contraintes du niveau macrorythmique.

Quinze sujets travaillant à l'époque à l'*ICP* ont participé à cette expérience de perception. Chaque session a duré entre 7 et 10 minutes. Dans un test du type ABBA les sujets ont été instruits à choisir la phrase qui leur semblait la plus naturelle[2]. Dix paires de phrases ont été utilisées dans le test, deux par débit. L'écoute pouvait être répétée une fois et une touche préalablement choisie pourrait être utilisée par l'auditeur pour indiquer que les deux phrases lui ont paru semblables. Les résultats sont présentés dans la figure 5.22.

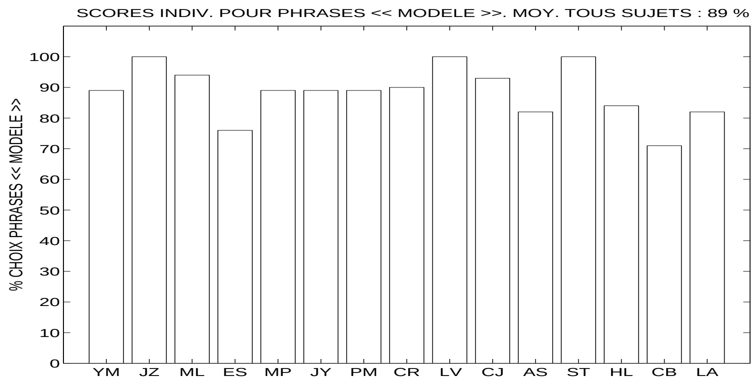

FIG. 5.22 - *Scores individuels pour le choix des phrases du type « modèle». Notez que les résultats sont très semblables entre les sujets.*

Tous les sujets confondus, le choix général pour la phrase du type « modèle» comme la plus naturelle, a été de 89 %. Dans 15 % des fois les sujets avaient un doute pour les

2. Cette question est relativement bien définie dans ce contexte car la qualité de l'analyse-synthèse la justifie. Voir pour cela commentaires du premier test, section 4.7.1, et la discussion dans cette section.

distinguer.

Quelques commentaires recueillis nous ont permis de savoir si ce qui était en train d'être évalué était bien des aspects liés à la durée ou aux conséquences de sa modification.

> Quand le rythme était continu ; quand la phrase « coulait », elle m'a semblé plus naturelle. *HL*
>
> Mon critère a été le rythme, cependant certains sons m'ont paru bizarres. *JY*
>
> Je les ai tranchées par rapport aux mots. S'ils étaient hachés au long de la phrase ou si le rythme n'était pas continu. *ML*
>
> Aucune des deux m'a paru naturelle, toutefois dans les 2/3 du temps j'ai tranché par rapport aux accents toniques en fin de mot. *CR*
>
> Des phrases avec changement de rythme en plein milieu de la phrase ne me semblent pas naturelles. *LV*

Les autres commentaires ont répété des aspects de l'interruption du rythme, présence gênante de « mots hachés » ou une « accentuation anormale ».

5.7.1 Discussion

De par les commentaires on s'aperçoit que certains aspects comme la présence de « sons bizarres » indiquerait, soit des problèmes du logiciel d'analyse-synthèse, soit des sons avec une durée anormale. Cependant le résultat général est très positif puisque les jugements ont, dans la plupart des cas, concerné des aspects globaux du rythme. Les résultats semblent montrer une préférence claire pour les phrases dont les durées sont données par notre modèle de génération.

Ceci confirme une préférence assez homogène (cf. scores intersujets dans la figure ci-dessus) pour un modèle qui maintienne la structuration rythmique représentée en termes de séquences de GIPC, un modèle dont le principal atout est la préservation de la composante macrorythmique de la phrase.

Si l'on regarde les phrases que les sujets ont eu le plus de doute pour prendre une décision ou qu'ils se sont décidés pour la phrase de la version « bruit » (figures 5.23 et 5.24), on se rend vite compte que la raison a pu bien être la surprédiction en position accentuée, ce qui entraîne des durées fort longues pouvant gêner l'auditeur.

Seule la continuation du processus d'apprentissage (étant donné que ces deux phrases font partie du corpus d'apprentissage) peut améliorer les prédictions pour ces deux phrases, si l'on garde la même architecture pour le réseau. Toutefois, on voit qu'il est en train d'apprendre car il a déjà appris que ces deux positions sont privilégiées.

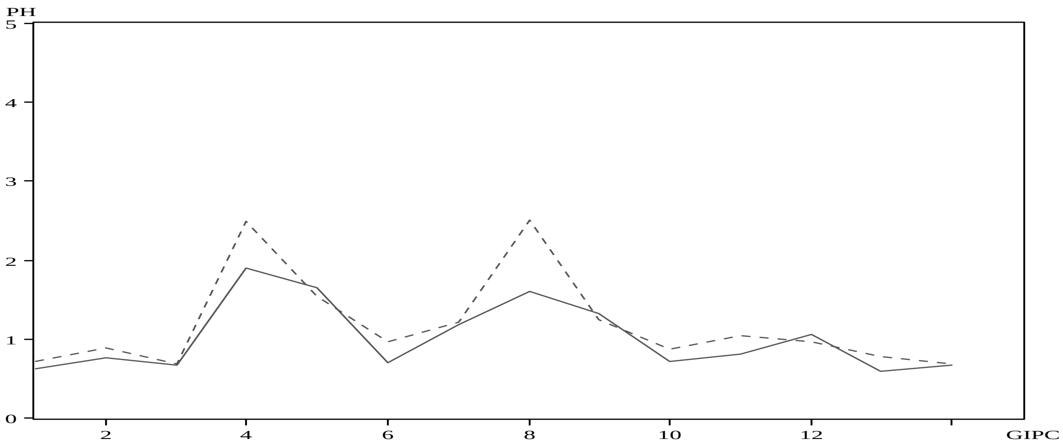

FIG. 5.23 - *Phrase 17 du corpus* PCenter *débit* normal. *Les lignes pleines représentent l'évolution de la durée, exprimées en PH, au long de la phrase naturelle (sortie désirée du réseau). Les lignes en traits discontinus représentent l'évolution de la durée prévue par le réseau pour cette phrase et exprimées en PH.*

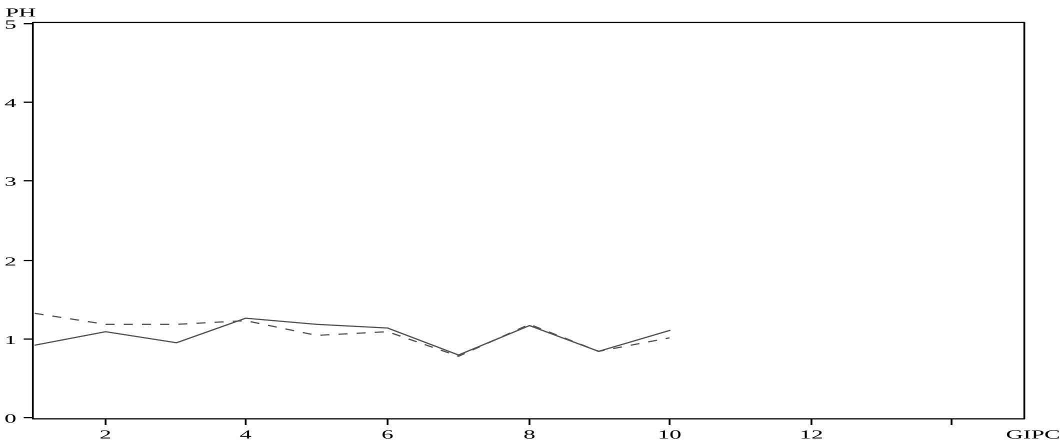

FIG. 5.24 - *Phrase 7 du corpus* PCenter *débit* très rapide . *Les lignes pleines représentent l'évolution de la durée, exprimées en PH, au long de la phrase naturelle (sortie désirée du réseau). Les lignes en traits discontinus représentent l'évolution de la durée prévue par le réseau pour cette phrase et exprimées en PH.*

5.8 Sommaire

Dans ce chapitre on a décrit la structuration rythmique du corpus *PCenter* à partir de l'approche *z–score*, ce qui a permis de confirmer une tendance à l'accentuation progressive dès le début du groupe accentuel.

Comme dans le chapitre précédent, les marques prosodiques présentent une hiérarchie qui peut les diviser en trois catégories au minimum. La force des marques corroborent les intuitions de la hiérarchie donnée par Bailly. Ceci a été possible grâce à la caractérisation des durées des GIPC par le $z-score$.

L'hypothèse de l'horloge interne a été exploitée et s'est montré fructueuse pour caractériser le débit de la phrase. Le GIPC inaccentué s'est montré être la bonne trace acoustique du repère représenté par l'horloge interne, en présentant une tendance à l'isochronisme. La réalité biologique de l'horloge interne corrobore l'existence d'une organisation macrorythmique de l'énoncé.

L'émergence de la pause dans un modèle de génération de la durée intégrant la part du macrorythme et celle du microrythme dans la détermination de la durée phonémique a pu être reproduite, grâce à l'approche $z-score$. Le silence éventuel au cours de l'énonciation est déterminé en fonction de l'hypothèse d'un silence minimum, attribué à une plage de variation de la période d'horloge, et des limites maximum du GIPC à s'allonger.

Un modèle de génération automatique est proposé par le biais d'un réseau connexionniste. Le réseau est capable de généralisation tout en préservant la composante macrorythmique des énoncés comme a été montré par un test de perception. Il ne suffit pas de fournir des durées avec un minimum d'erreur de prédiction, il est encore plus intéressant qu'elles soient structurées par des contraintes de niveau supérieur à celui du phonéme : la sensibilité perceptive des erreurs de prédiction n'est pas uniformement repartie sur la chaîne phrastique.

La performance générale du modèle de prédiction l'habilite à être intégré dans un système de synthèse à partir du texte pour la prédiction de la durée, tout en possibilitant la génération intégrée des autres paramètres prosodiques, dans le cadre de l'apprentissage automatique.

Conclusions et perspectives

Nous avons essayé de montrer au cours de notre travail la pertinence de la considération d'un événement psychoacoustique privilégié, le *Perceptual-center* (PC). Deux PC consécutifs définissent les frontières du groupe *inter-Perceptual-center* (GIPC), pris comme l'unité de programmation rythmique.

La durée du GIPC et la caractérisation de la composante microrythmique du locuteur par une paire de valeurs statistiques, a permis le calcul d'un indice d'allongement intrinsèque de ce groupe, le $z-score$. Ce même $z-score$ a permis de justifier le choix du GIPC comme unité rythmique de structuration des durées phonémiques : l'allongement de chaque réalisation est plus cohérent dans le cadre du GIPC que de celui de la syllabe. En fait, la rime et les consonnes de l'attaque se comportent différemment par rapport à l'allongement. Cet effet est augmenté en position accentuée, où la rime est beaucoup plus longue en durée que les consonnes de l'attaque.

Les valeurs de $z-score$ associées au GIPC ont permis de révéler la structuration macrorythmique de l'énoncé par l'identification des GIPC accentués et des frontières prosodiques associées à la durée.

Un test de perception nous a permis de valider la structure rythmique révélée par l'approche GIPC en montrant que le patron caractérisé par un ralentissement progressif de l'énonciation au long du groupe accentuel était nécessaire pour une perception naturelle de la phrase : le groupe accentuel en français n'est pas une suite de syllabes inaccentuées terminée par une syllabe accentuée, il est caractérisé par un véritable geste de ralentissement partant d'une durée de GIPC de référence très contrainte (cf. section 5.3) pour atteindre un niveau d'accentuation lié au degré de cohésion entre unités syntagmatiques (cf. section 5.1.3).

À partir des $z-scores$, on a pu calculer, au niveau microrythmique, la durée des réalisations phonémiques qui composent le GIPC.

Ces deux caractéristiques ont été intégrées dans un système de génération automatique de la durée complet : apprentissage automatique de la structure macrorythmique de la phrase et répartition de la durée du GIPC entre ses réalisations par un modèle de répartition.

En ce qui concerne l'apprentissage automatique, elle a permis de réaliser avec succès, comme démontre le test de perception que l'on a passé à quinze sujets, la génération de la durée au niveau macrorythmique tout en intégrant la notion d'horloge interne, ce qui est une corroboration de sa pertinence opérationnelle. On a montré en outre que cette horloge a une plausibilité biologique, en tant que point d'ancrage de la rythmicité.

Une mesure de l'horloge interne à partir de la durée du GIPC a possibilité la caractérisation du débit de chaque phrase de façon cohérente : le rôle du GIPC inaccentué comme repère d'isochronisme a pu être dégagé.

L'approche du $z-score$ associé au GIPC a possibilité la mise en œuvre d'un algorithme qui permet de générer, en plus des durées des sons, le silence éventuel au plan de la substance de l'expression. La pertinence de cette génération est évaluée de façon corollaire dans le test de perception que l'on a conçu pour tester l'efficacité du réseau à rendre compte de la structuration rythmique de la phrase. La pause n'est pas calculée phonologiquement et donnée a priori comme élément contextuel, comme on voit dans la plupart des modèles qui la considèrent, mais est obtenue par le processus même de génération des autres éléments de la deuxième articulation.

La prise en compte de la composante rythmique dans la génération a facilité l'obtention de la durée phonémique, sans le besoin de procédures lourdes comme l'usage de coefficients de modification de la durée selon le contexte et les catégories grammaticales des mots auxquels les phonèmes appartiennent.

Il est clair que les caractéristiques trouvées pour notre locuteur doivent être étendues à d'autres et à des textes plus variés, avec des phrases de diverses longueurs, dans le sens de vérifier son universalité, d'abord pour le français. Il nous semble que ce pas doit être effectué avant d'étendre une quelconque étude sur d'autres styles d'énonciation que la lecture.

Mais la suite naturelle de ce travail nous semble être l'intégration de ce modèle de prédiction de la durée avec la prédiction simultanée de la fréquence fondamentale, comme effectué dans l'approche de Todd (cf. section 5.5.1). Un travail de prédiction de la fréquence fondamentale qui promet un avenir intéressant a été déjà entamé par Morlec (1994).

Annexe A

Corpora de parole naturelle

A.1 Corpus *Bureau*

1 Rappelez Monsieur Dupont à son bureau jeudi.

2 Rappelez Monsieur Dupont à son bureau.

3 Rappelez Monsieur Dupont à quinze heures dix.

4 Rappelez Monsieur Dupont jeudi.

5 Rappelez Monsieur Dupont jeudi à quinze heures dix.

6 Rappelez Monsieur Dupont à quinze heures dix jeudi.

7 Rappelez Monsieur Dupont.

8 Rappelez Messieurs Dupont et Filippi à leur bureau jeudi à quinze heures dix.

9 Rappelez Messieurs Dupont et Filippi à leur bureau jeudi.

10 Rappelez Messieurs Dupont et Filippi à leur bureau.

11 Rappelez Messieurs Dupont et Filippi à quinze heures dix.

12 Rappelez Messieurs Dupont et Filippi jeudi.

13 Rappelez Messieurs Dupont et Filippi jeudi à quinze heures dix.

14 Rappelez Messieurs Dupont et Filippi à quinze heures dix jeudi.

15 Rappelez Messieurs Dupont et Filippi.

16 Rappelez à quinze heures dix.

17 Rappelez jeudi à quinze heures dix.

18 Rappelez à quinze heures dix jeudi.

19 Rappelez jeudi.

20 Rappelez à quinze heures dix Monsieur Dupont à son bureau.

21 Rappelez-le à quinze heures dix.

22 Rappelez-le à son bureau.

23 Veuillez rappeler Monsieur Dupont à son bureau.

24 Vous rappelez Monsieur Dupont à son bureau.

25 Vous rappelerez Monsieur Dupont à son bureau.

26 Vous devez rappeler Monsieur Dupont à son bureau.

27 Monsieur Dupont vous rappelera à quinze heures dix.

28 Monsieur Tapi, rappelez Monsieur Dupont à son bureau.

29 Monsieur Tapi, veuillez rappeler Monsieur Dupont à son bureau.

30 Monsieur Tapi, vous rappelez Monsieur Dupont à son bureau.

31 Monsieur Tapi, vous rappelerez Monsieur Dupont à son bureau.

32 Monsieur Tapi, vous devez rappeler Monsieur Dupont à son bureau.

33 Monsieur Tapi, Monsieur Dupont vous rappelera à quinze heures dix.

34 Messieurs Tapi et Filippi, rappelez Monsieur Dupont à son bureau.

35 Messieurs Tapi et Filippi, veuillez rappeler Monsieur Dupont à son bureau.

36 Messieurs Tapi et Filippi, vous rappelez Monsieur Dupont à son bureau.

37 Messieurs Tapi et Filippi, vous rappelerez Monsieur Dupont à son bureau.

38 Messieurs Tapi et Filippi, vous devez rappeler Monsieur Dupont à son bureau.

39 Messieurs Tapi et Filippi, Monsieur Dupont vous rappelera à quinze heures dix.

40 Rapidement, rappelez Monsieur Dupont à son bureau.

41 Rappelez rapidement Monsieur Dupont à son bureau.

42 Rappelez Monsieur Dupont rapidement à son bureau.

43 Rappelez Monsieur Dupont à son bureau rapidement.

44 Le plus rapidement possible, rappelez Monsieur Dupont à son bureau.

45 Rappelez le plus rapidement possible Monsieur Dupont à son bureau.

46 Rappelez Monsieur Dupont le plus rapidement possible à son bureau.

47 Rappelez Monsieur Dupont à son bureau le plus rapidement possible.

48 À quinze heures, rappelez Monsieur Dupont à son bureau.

49 Rappelez à quinze heures Monsieur Dupont à son bureau.

50 Rappelez Monsieur Dupont à quinze heures à son bureau.

51 Rappelez Monsieur Dupont à son bureau à quinze heures.

52 Après quinze heures quarante-cinq, appelez Monsieur Dupont à son bureau.

53 Rappelez après quinze heures quarante-cinq Monsieur Dupont à son bureau.

54 Rappelez Monsieur Dupont après quinze heures quarante-cinq à son bureau.

55 Rappelez Monsieur Dupont à son bureau après quinze heures quarante-cinq.

56 Quand c'est fini, rappelez Monsieur Dupont à son bureau.

57 Rappelez, quand c'est fini, Monsieur Dupont à son bureau.

58 Rappelez Monsieur Dupont quand c'est fini à son bureau.

59 Rappelez Monsieur Dupont à son bureau quand c'est fini.

60 Dès que vous aurez fini, rappelez Monsieur Dupont à son bureau.

61 Dès que vous aurez fini, rappelez Monsieur Dupont.

62 Rappelez dès que vous aurez fini, Monsieur Dupont à son bureau.

63 Rappelez Monsieur Dupont à son bureau dès que vous aurez fini.

64 Rappelez Monsieur Dupont dès que vous aurez fini.

65 Quand vous aurez fini votre réunion avec Monsieur Tapi, rappelez Monsieur Dupont à son bureau.

66 Rappelez Monsieur Dupont à son bureau quand vous aurez fini votre réunion avec Monsieur Tapi.

67 Rappelez, quand vous aurez fini votre réunion avec Monsieur Tapi, Monsieur Dupont à son bureau.

68 Rappelez Monsieur Dupont quand vous aurez fini votre réunion avec Monsieur Tapi.

69 Il est quinze heures quarante-cinq, rappelez Monsieur Dupont à son bureau.

70 Rappelez Monsieur Dupont à son bureau : il est quinze heures quarante-cinq.

71 À quinze heures quarante-cinq rappelez Monsieur Dupont à son bureau.

72 Rappelez à quinze heures quarante-cinq Monsieur Dupont à son bureau.

73 Rappelez Monsieur Dupont à quinze heures quarante-cinq à son bureau.

74 Rappelez Monsieur Dupont à son bureau à quinze heures quarante-cinq.

75 Rappelez et convoquez Monsieur Dupont à son bureau.

76 Vous voudrez bien rappeler Monsieur Dupont à son bureau.

77 Vous voudrez ne pas rappeler Monsieur Dupont à son bureau.

78 Vous voudrez bien ne pas rappeler Monsieur Dupont à son bureau.

79 Vous voudrez bien ne rappeler Monsieur Dupont qu'après quinze heures dix.

80 Contactez le président à son bureau à quinze heures quarante-cinq.

81 Contactez le président européen à son bureau à quinze heures quarante-cinq.

82 Contactez le nouveau président à son bureau à quinze heures quarante-cinq.

83 Contactez le nouveau président européen à son bureau à quinze heures quarante-cinq.

84 Contactez le Président-Directeur-Général à son bureau à quinze heures quarante-cinq.

85 Contactez le président de la société à son bureau à quinze heures quarante-cinq.

86 Contactez le président de la Société Générale à son bureau à quinze heures quarante-cinq.

87 Contactez le président de la société lyonnaise Dupéral à son bureau à quinze heures quarante-cinq.

88 Contactez le nouveau Président-Directeur-Général européen de la société lyonnaise Dupéral à son bureau.

A.2 Corpus *PCenter*

1 Tôt dans la matinée, Thibaud peint les sommets du Mont Karimo.

2 Tôt dans la matinée, tu peins les deux sommets du Mont Karimo.

3 Tôt dans la matinée, tes mains devinaient les sommets du Mont Karimo.

4 Tôt dans la matinée, tes mains visaient les sommets du Mont Karimo.

5 Tôt dans la matinée, ta main, posée sous le manteau, visait les sommets du Mont Karimo.

6 Thibaud peint les sommets du Mont Karimo.

7 Tu peins les deux sommets du Mont Karimo.

8 Tes mains devinaient les sommets du Mont Karimo.

9 Tes mains visaient les sommets du Mont Karimo.

10 Ta main, posée sous le manteau, visait les sommets du Mont Karimo.

11 Thibaud peint les sommets du Mont Karimo, tôt dans la matinée.

12 Tu peins les deux sommets du Mont Karimo, tôt dans la matinée.

13 Tes mains devinaient les sommets du Mont Karimo, tôt dans la matinée.

14 Tes mains visaient les sommets du Mont Karimo, tôt dans la matinée.

15 Ta main, posée sous le manteau, visait les sommets du Mont Karimo, tôt dans la matinée.

16 Tu m'avais dit : « Filé laissa sa maman pendant la matinée. »

17 Tu m'avais dit : « Filez les sous, vous pouvez baisser les mains. »

18 Tu m'avais dit : « Filé-Lessaint n'a pas changé. »

19 Tu m'avais dit : « Filé, Lessaint, Dupont n'ont pas changé. »

20 Tu m'avais dit : « Filé, laissez les tampons là-bas. »

Annexe B

Phrase 56 : test de perception

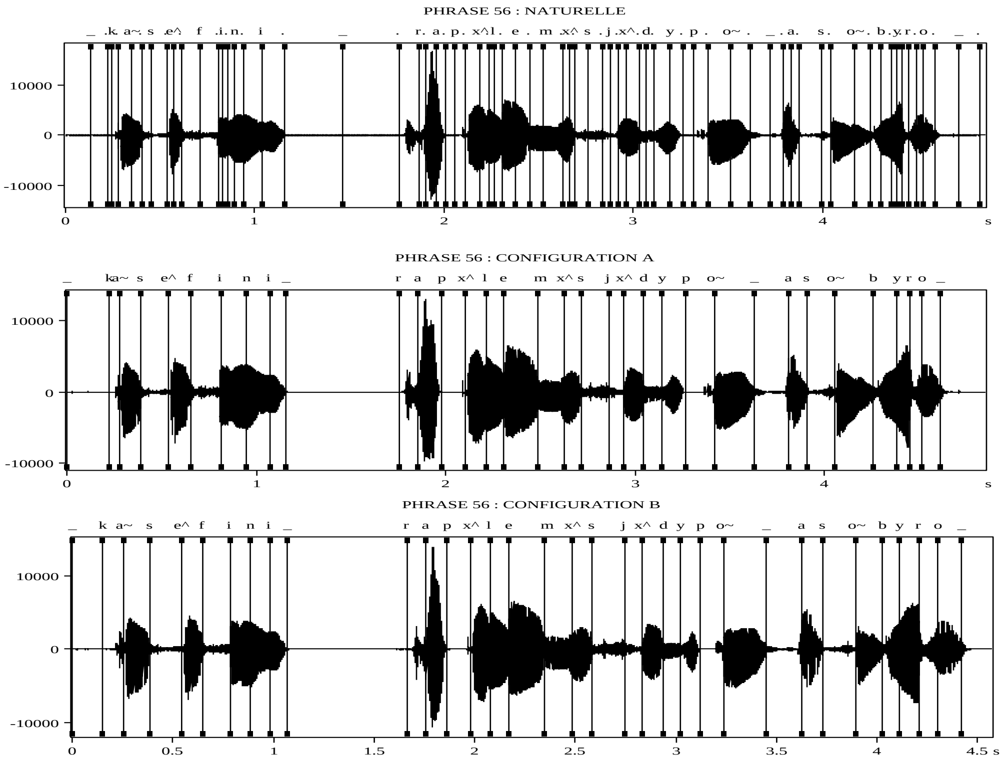

FIG. B.1 - *Exemple de phrase étiquetée : « Quand c'est fini, rappelez Monsieur Dupont à son bureau.» Le signal d'en-haut est la phrase naturelle avec étiquette phonémique au centre de la réalisation et frontières délimitées par un point. La phrase du milieu (configuration A du test de perception) est celle obtenue par analyse-synthèse avec les durées modifiées par le modèle de répartition. La phrase d'en-bas (configuration B du test de perception) est obtenue par analyse-synthèse en faisant $z - score = 0$ en position non accentuée (dans le syntagme final « à son bureau.» les $z-scores$ sont les mêmes que dans la phrase). Dans les deux dernières phrases il est indiqué par des étiquettes phonémiques le début de chaque réalisation.*

Annexe C

Le Modèle de marquage de Bailly

Le modèle de marquage est un modèle de transfert proposant une transformation au niveau de la forme, d'une structure du contenu vers une structure de l'expression.

Les modèles de transfert ainsi proposés exploitent normalement une relation non bijective entre un espace bidimensionnel, celui des arborescences ou structures de surface, qui sont une représentation de la structure syntactico-morphologique du texte, et un espace unidimensionnel, la projection de l'arbre sur l'axe temporel. Les relations entre les constituants au plan syntagmatique sont représentées de ce fait par un système de marques prosodico-syntaxiques, les marques relationnelles.

Bailly (1986) propose un marquage du texte issu de la projection sur l'axe syntagmatique d'un arbre syntaxique. Cet arbre est construit à partir des informations fournies par une grammaire de dépendance (Tesnière 1965; Hays 1964; Veillon 1970), exprimant les relations syntaxiques que les mots entretiennent entre eux. Dans une telle grammaire, ces relations énumèrent les classes lexicales dépendantes d'une autre qui peut les gouverner (cette dernière est appelée le gouverneur). Le sommet de cet arbre est en général le verbe principal de la phrase (cf. figure C.1).

Dans un tel système, chaque mot lexical non clitique (accentuable) reçoit une marque qui indique la relation entretenue par celui-ci et le non clitique subséquent. Le système de Bailly est basé sur deux approches théoriques. Tout d'abord il est une généralisation de la notion de frontière majeure et mineure, développée par Di Cristo (1981) : les marques de Bailly sont *fortes* ou *faibles* ; ensuite, le système est basé sur le fait signalé par Martin (1981) que les relations entretenues entre les mots sont quadrivalentes : indépendance (aucun lien de dépendance direct entre les deux mots), interdépendance (les deux mots sont dépendants du même gouverneur), dépendance gauche (le premier mot est dépendant du deuxième) et dépendance droite (le deuxième est dépendant du premier). Son système a évolué au cours de son expérience dans le domaine de la synthèse

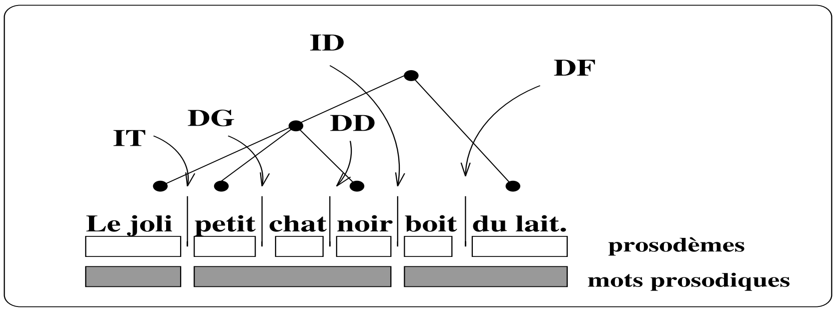

FIG. C.1 - *Exemple d'arbre de dépendance et de marquage prosodique. D'après Bailly (1989, p. 140).*

de la parole pour arriver à un ensemble de six marques relationnelles organisées en ordre décroissant de force :

ID indépendance.

Exemple : *Monsieur Tapie, **(ID)** rappelez Monsieur Dupont jeudi.*

GF dépendance gauche forte. Entre nom sujet et verbe subséquent.

Exemple : *Papa **(GF)** mange une pomme.*

IT interdépendance.

Exemple : *Pierre, **(IT)** Jean **(IT)** et Marie boivent du lait.*

DG dépendance gauche faible.

Exemple : *Le petit **(DG)** chat joue.*

DF dépendance droite forte. Entre le verbe et le complément subséquent.

Exemple : *Le chat noir boit **(DF)** du lait.*

DD dépendance droite faible.

Exemple : *Le chat **(DD)** noir boit du lait.*

L'auteur précise que ces marques sont révélatrices de relations locales entre les éléments de l'axe syntagmatique : la structure prosodique qui sera générée n'inclue pas la connaissance a priori de la structure complète de la phrase. Ceci conduit à des systèmes pour le marquage automatique moins coûteux et plus réalistes, au moins du point de vue du mécanisme d'énonciation des phrases lues puisque le locuteur procède par étapes en raisons des limitations de ses habitudes de lecture.

Annexe D

Méthodes et techniques de synthèse

> On s'est étonné que la machine parlante qui nous est venue, il y a quelques années d'Amérique (Barnum hatte die Maschine nach einer Amerikatournee 1875 nach Paris gebracht[1]), et qui a été exhibée au Grand-Hotêl fût d'une extrême complication, alors que le phonographe résolvait le problème d'une manière si simple : c'est que l'une de ces machines ne faisait que reproduire la parole, tandis que l'autre l'émettait, et l'inventeur de cette dernière machine avait dû, dans son mécanisme, mettre à contribution tous les organes, qui dans notre organisme, concourent à la production de la parole.
>
> *du* MONCEL *(1880)*

On voit dans ce passage où Köster (1973, p. 148) explique le fonctionnement de la machine de Joseph Faber (présentée pour la première fois en 1835 par ce professeur de mathématiques) que, très tôt, deux voies de recherche se sont constituées autour de la synthèse vocale. La première cherche à reproduire le mieux possible le signal de parole : nous l'appellerons le *faire-semblant*. La deuxième cherche à revenir aux causes de déclenchement du mécanisme de production phonatoire : nous l'appellerons le *faire-comme* (ICP 1994).

La première voie présente le danger de générer des monstres acoustiques, c'est-à-dire, des signaux acoustiques qui ne pourraient jamais être reproduits par un être humain. De nos jours on voit apparaître un intérêt grandissant pour une méthode-technique de synthèse dont la nomenclature ne s'est pas encore fixée : il s'agit de la synthèse par contrôle moteur, synthèse par systèmes dynamiques ou synthèse articulatoire (Coker 1968; Bailly, Laboissière & Schwartz 1991).

1. Barnum avait apporté la machine à Paris en 1875 après une tournée en Amérique

D.1 Méthodes de synthèse

Les méthodes de synthèse se divisent d'abord quant à la taille du vocabulaire employé. Les systèmes sont dits à **vocabulaire limité** ou à **vocabulaire illimité**. Dans le premier cas, le degré de manipulation effectué sur le signal de parole pour obtenir une bonne qualité d'écoute est assez réduit : l'unité utilisée généralement est le mot. Dès que de nouvelles phrases se font nécessaires, l'explosion du vocabulaire oblige l'adoption de méthodes dont l'unité manipulée est inférieure au mot.

Les méthodes de synthèse à vocabulaire illimité se divisent en **synthèse par règles**, **synthèse par concaténation d'unités stockées** et **synthèse par systèmes dynamiques**.

D.1.1 La Synthèse par règles

Son principe général est la modélisation des transitions entre les phonèmes sous la forme de règles. À partir d'une représentation formantique du signal de parole, ces règles décrivent, par exemple, l'évolution des formants entre deux *valeurs cibles* (Klatt 1987).

Dans un tel système on stocke très peu de données (les valeurs cibles) mais, le nombre de règles nécessaires à décrire les transitions formantiques peur devenir très grand et difficile à établir si l'on souhaite une reproduction fidèle de la parole humaine.

D.1.2 La Synthèse par unités stockées

Ce système propose la concaténation d'unités pré-stockées. Au lieu de modéliser la transition on la stocke dans des lexiques de signal. Après des tentatives infructueuses pour composer le signal de parole par concaténation d'unités de la taille du phonème (voir Aubergé 1991 pour un bref exposé de l'histoire de la synthèse), le **diphone** sera proposé comme unité minimum utilisée dans la concaténation :

> Un diphone est un élément sonore caractéristique de la transition entre deux phonèmes s'étendant de la partie stable d'un phonème à la partie stable du phonème suivant. *Emerard (1977)*

La tendance actuelle est d'augmenter de plus en plus la taille de l'unité à stocker, toujours dans le même soucis de préserver des transitions complexes, comme dans les *clusters* consonantiques ou dans le cas des semi-voyelles. Le terme de *polyson* a été proposé pour cette nouvelle unité (Barbe 1990; Bailly, Barbe & Wang 1992).

Dans un tel système, la mémoire exigée peut devenir grande, mais le nombre de règles de concaténation est réduit.

D.1.3 La Synthèse par systèmes dynamiques

Cette méthode se propose à modéliser l'organisation des gestes articulatoires ou des trajectoires acoustico-visuelles inspirées par le système biologique et les théories du contrôle moteur (Liberman & Mattingly 1985). Un autre aspect prometteur de cette méthode est l'intégration possible entre signal visuel et acoustique dans un espace articulatoire commun (Erber 1975; Summerfield 1987; Robert-Ribes, Escudier & Schwartz 1992). L'importance de la composante visuelle se fait sentir notamment dans les environnements bruités, où la récupération des cibles phonémiques est largement aidée par l'information fournie par les lèvres.

D.2 Techniques de synthèse

Pour rendre un signal acoustique audible il se fait nécessaire de piloter un synthétiseur de parole au moyen d'une technique appropriée. Nous avons organisé les techniques de synthèse dans un crescendo de complexité, allant des modèles fonctionnels aux modèles du conduit vocal.

D.2.1 La Technique PSOLA

La technique PSOLA (de l'anglais, *Pitch Synchronous Overlap and Add*), utilisée avec succès pour l'obtention de systèmes de haute qualité (Moulines 1992; Bailly, Barbe & Wang 1992), se propose à adapter le signal acoustique à différentes exigences prosodiques avec un minimum de distorsion spectrale. Le contrôle de la durée est obtenu par manipulation des signaux à court terme en augmentant ou en diminuant le nombre de périodes du signal. Par sa facilité d'utilisation, cette technique se prête bien aux techniques d'analyse-synthèse (modification des paramètres prosodiques du signal naturel tout en maintenant le niveau segmental d'origine).

D.2.2 La Technique LPC

La prédiction linéaire (*Linear Prediction Coding*) est une technique à l'origine utilisée pour réduire la quantité d'information en vu du codage et de la transmission à faible débit du signal de parole (Atal 1983).

Chaque échantillon du signal de parole est décrit comme une combinaison linéaire des échantillons précédents. Les coefficients de la combinaison sont obtenus par minimisation de l'erreur quadratique entre le signal original et le signal prédit sur une fenêtre appropriée.

Les coefficients sont réactualisés toutes les 5 à 20 ms. Un nombre de coefficients de 10 à 15 est nécessaire pour atteindre une bonne qualité de prédiction.

D.2.3 Le Synthétiseur à formants

On sait que l'information perceptive la plus importante dans un spectre de parole est donnée par les formants. La représentation spectrale est simplifiée par le codage des seules valeurs formantiques. Les formants sont représentées par leur fréquence centrale, leur largeur de bande et leur amplitude.

Une autre douzaine de paramètres permet la bonne caractérisation des zones non voisées et de l'effet du canal nasal.

La génération du signal de parole est effectuée par un modèle de où un générateur d'impulsions piloté par la fréquence fondamentale et un générateur de bruit constituent les signaux d'excitation.

Les deux dernières techniques présentent de défauts de qualité (signal bruyant), dus à une représentation trop simpliste du signal d'excitation (séparation complète entre une source *bruit* et une source voisée). D'autres techniques visant à réaliser le faire-comme discuté précédemment, ont connu le jour.

D.2.4 La Technique de simulation du conduit vocal

Cette technique cherche à reproduire le fonctionnement de l'appareil phonatoire. Les paramètres de commande du synthétiseur sont directement liés au mécanisme de production de la parole (positions des articulateurs buccaux, pression sub-glottique, tension des cordes vocales). Le signal de parole est obtenu à partir d'une simulation mathématique de l'écoulement de l'air dans un modèle du conduit vocal.

Toute la difficulté de cette technique se trouve justement dans l'étude du mécanisme de production (et perception, pour prendre en compte des paramètres pertinents) de la parole. Le besoin d'une recherche accrue dans ce domaine est primordial pour optimiser et améliorer la qualité finale du système.

L'avènement des systèmes dynamiques de synthèse rend difficile la séparation entre méthode et technique de synthèse. Le contrôle prosodique tend à être intrinsèque au système par la manipulation de nouveaux paramètres comme la durée de la commande motrice et la raideur de l'articulateur (Lœvenbruck & Perrier 1993).

Annexe E

Les Réseaux neuromimétiques

Quoique les idées fondamentales aient pu se développer dans les années 40 (McCulloch & Pitts 1943), l'étude des réseaux neuromimétiques (*neural nets*) n'a pris un nouvel essor qu'à partir de la deuxième moitié des années 80.

Un réseau neuromimétique, ou réseau de neurones formel, consiste physiquement dans un ensemble de processeurs élémentaires interconnectés, appelés nœuds ou neurones formels. Certains de ces neurones forment l'entrée du réseau et d'autres vont constituer sa sortie. La façon dont les neurones sont interconnectés définit l'architecture du réseau. Le point fort d'un réseau de neurones formel réside dans la connexion entre ses éléments, dont les noms de connexionisme, pour ce champ de recherche, et de réseau connexionniste, pour l'objet d'étude.

Représentons schématiquement un neurone par la figure E.1.

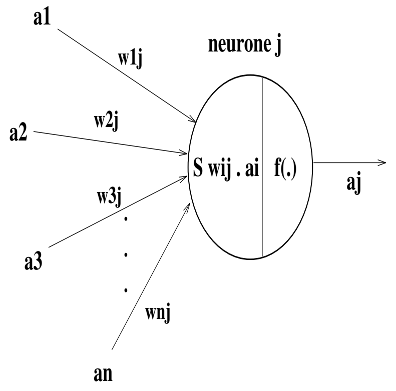

FIG. E.1 - *Le neurone formel et ses connexions dans un réseau schématique. Le terme $Sw_{ij}a_i$ représente la combinaison linéaire entre les poids w_{ij} et les valeurs d'activation a_i et le terme $f(.)$ représente la fonction d'activation. Le terme a_j est la valeur d'activation du neurone j.*

Cette unité reçoit en entrée un ensemble de connexions venant d'un sous-ensemble de neurones du réseau auxquels on associe des poids. Chaque poids représente l'influence du neurone afférent sur le neurone efférent. La valeur d'activation du neurone efférent est obtenue par un processus calculatoire qui peut être, dans la plupart des cas, une combinaison linéaire entre les valeurs d'activation des neurones afférents (pondérée par les poids des connexions) suivie de l'application d'une fonction différenciable limitée. Les valeurs d'activation des neurones de sortie sont les valeurs de sortie du réseau. Une fonction d'activation très répandue est la fonction sigmoïde (graphique dans la figure E.2), exprimée par :

$$f(x) = \frac{1}{1+\exp(-x)} \tag{E.1}$$

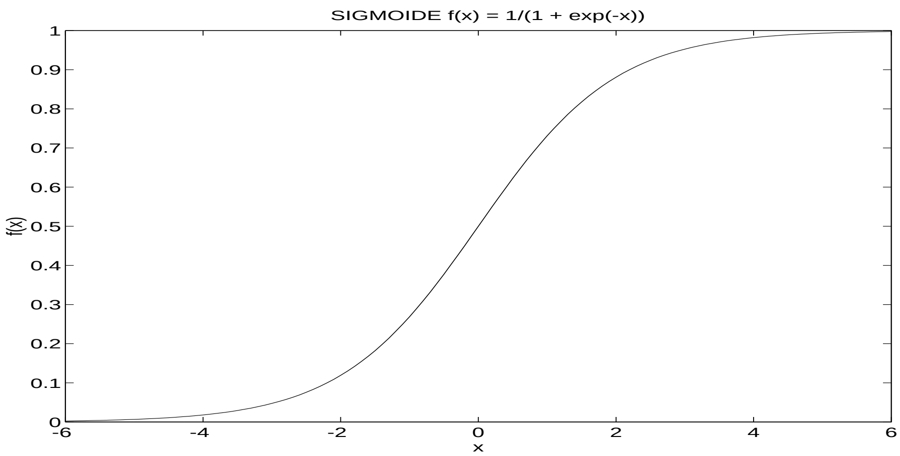

FIG. E.2 - *Fonction sigmoïde, utilisée largement comme fonction d'activation dans les réseaux neuromimétiques.*

La finalité d'un réseau neuromimétique est de simuler une fonction, ou un système physique, sans connaître sa description analytique, soit parce que le système ne possède pas une forme analytique, soit parce que les facteurs qui influent sur le résultat final sont multiples (ou mal connus) ou la façon par laquelle ils s'associent pour produire le phénomène observé n'est pas encore bien comprise.

Pour réaliser une telle fonction, un réseau à apprentissage supervisé (Lippmann 1987) doit apprendre à associer un espace d'entrée à l'espace de sortie au moyen de la présentation successive de paires d'entrée/sortie significatives de l'association en question. Ceci est réalisé par une règle d'apprentissage, qui réalise les modifications nécessaires dans les poids du réseau pour qu'il s'adapte au mieux aux paires d'exemples entrée/sortie présentés. L'apprentissage fini, le réseau est prêt à opérer. Plus les exemples sont représentatifs du phénomène sous-jacent, plus le réseau sera capable de généraliser : à une entrée qui n'a pas été considérée pendant la phase d'apprentissage, il fournira une sortie, souhaitée être une bonne approximation de la réponse du système simulé à la nouvelle entrée.

Si l'on caractérise un réseau par un triplet (architecture, caractéristiques du neurone, règle d'apprentissage), ces éléments doivent être déterminés pour chaque application de façon à maximiser la performance de la simulation. Pour répondre aux besoins des systèmes de reconnaissance de formes qui classent des centaines ou des milliers de patrons, il existe des algorithmes appelés à apprentissage incrémental, qui mettent en œuvre des architectures évolutives de réseau (Reilly, Cooper & Erlbaum 1982; Alpaydin 1990).

Le premier réseau connexionniste a été le **perceptron** (Rosenblatt 1959). Dans la figure E.3 on voit le schéma d'un perceptron multicouches. Chaque couche est représentée par un sous-ensemble de neurones formels. À l'entrée et à la sortie, les paires d'exemples sont fournies au réseau pendant l'apprentissage. Les couches intermédiaires (entre l'entrée et la sortie) sont appelées couches cachées. Leur valeur d'activation est dépendante de l'ensemble des valeurs des poids, de la manière avec laquelle les neurones sont connectés (architecture du réseau) et de la fonction d'activation employée. On dit que le réseau est complètement connecté si chaque neurone d'une couche est relié à tous les neurones de la couche précédente (en parcourant le réseau dans le sens entrée vers sortie).

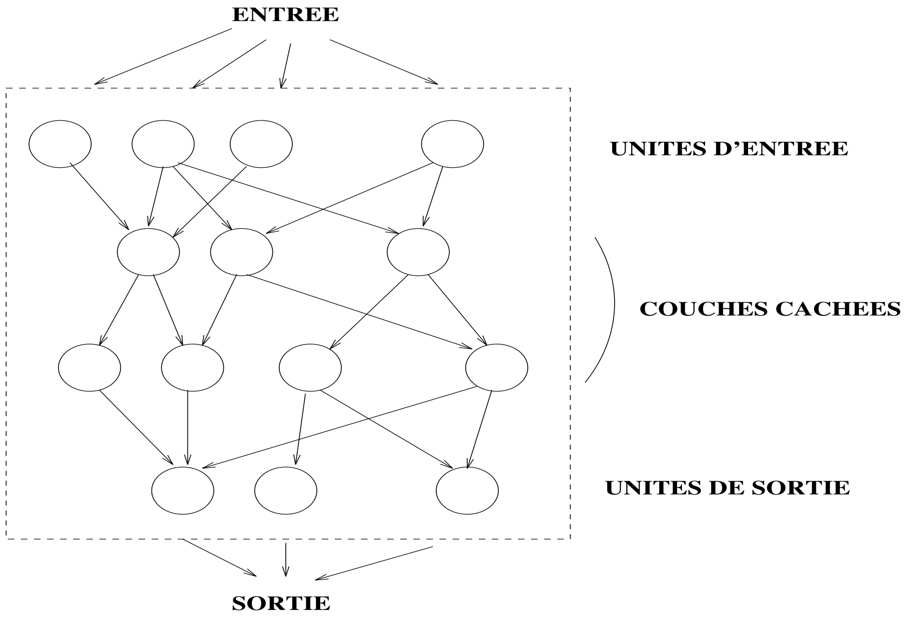

FIG. E.3 - *Un perceptron multicouches : schéma général. Quelques connexions n'ont pas été établies pour des raisons de clarté.*

Des études mathématiques axées sur les régions de décision délimitées par les perceptrons (pour les problèmes de reconnaissance des formes), ont montré qu'un perceptron avec une couche cachée est capable de délimiter des régions convexes et qu'un perceptron avec deux couches cachées est capable d'établir des régions de décision de complexité arbitraire, dépendant uniquement du nombre de neurones utilisé (Lippmann 1987). C'est pourquoi il est inutile d'utiliser un nombre de couches cachées qui soit supérieur à deux. Cependant, deux couches cachées peuvent entraîner une augmentation considérable du nombre de connexions entre les neurones, ce qui rend l'apprentissage d'un tel système parfois fort lent. Un réseau à une couche cachée est capable de bien généraliser pour la plupart des cas, comme on verra plus loin dans les travaux de Todd (1989), Sagisaka (1990) et Scordilis & Gowdy (1989).

Une règle d'apprentissage très employée, la règle de rétro-propagation d'erreur (*back-propagation rule*) (Rumelhart, Hinton & William 1986) compare la sortie observée avec celle obtenue par l'application du processus calculatoire propre à l'architecture et aux fonctions définies dans le réseau. Cette règle calcule une fonction d'erreur quadratique en sortie, dont le gradient est utilisé pour réaliser la modification de l'ensemble des poids du réseau. Ceci constitue le premier pas d'apprentissage. La règle se poursuit jusqu'à ce qu'une valeur minimum (préétablie) pour l'erreur quadratique globale soit atteinte ou qu'un nombre maximum de pas soit exécuté.

À partir du perceptron, on peut réaliser des modifications sur la façon de connecter les neurones, de façon à obtenir quelques réseaux couramment utilisés dans la littérature. Si l'on rajoute des connexions non seulement dans le sens entrée vers sortie mais aussi de la sortie vers l'entrée, on obtient un **réseau récurrent**. Il est fréquent d'avoir des connexions qui partent des neurones de la sortie vers un sous-ensemble de neurones appelés unités d'état (*state units*) qui activent une ou plusieurs couches cachées. Ces unités d'état bouclent sur elles-mêmes fonctionnant ainsi comme une espèce de mémoire du système.

L'intérêt pour ces types de réseaux remonte aux travaux de McCullouch & Pitts (1943) et Kleene (1956). L'une des raisons qui justifient d'un tel intérêt se trouve dans le fait que ces types de réseau sont capables de simuler des systèmes dynamiques assez complexes (Parlos, Chong & Atiya 1994).

En essayant d'élargir le champ d'application du paradigme de l'apprentissage supervisé aux problèmes de systèmes à excès de degrés de liberté, Jordan (1986) a développé un programme de recherche dont l'une des pierres de touches est la proposition d'un **réseau séquentiel**, pour permettre l'apprentissage et le stockage de séquences. Ce que l'on fait apprendre au réseau est une séquence de valeurs pour la sortie, qui fonctionnent ainsi comme un échantillonnage temporel de la sortie d'un système dynamique.

Références bibliographiques

D. Abercrombie (1967). *Elements of general phonetics.* Edinburgh University Press.

C. Abry, D. Autesserre, C. Barrera, C. Benoît, L.-J. Boë, J. Caelen, G. Caelen-Haumont, M. Rossi, R. Sock & N. Vigouroux (1985). Un modèle de congruence relationnel pour la synthèse de la prosodie du français. In *14es Journées d'Étude du Groupe Communication Parlée*, Paris, pp. 156–163. G.A.L.F.

G. D. Allen (1970). Temporal structure in speech production. *J. Acoust. Soc. Am. 47*, 58(A).

G. D. Allen (1972). The location of rhythmic stress beats in English I & II. *Language and Speech 15*, 72–100,179–195.

G. D. Allen (1973). Segmental timing control in speech production. *Journal of Phonetics 1*, 219–237.

G. D. Allen (1975). Speech rhythm: its relation to performance universals and articulatory timing. *Journal of Phonetics 3*, 75–86.

G. D. Allen (1981). Suprasegmental constraints on segmental representation: Research involving speech production. In J. L. T. Myers & J. Anderson (Eds.), *The cognitive representation of speech*, pp. 151–159. New York: North-Holland.

E. Alpaydin (1990). *Neural models of incremental supervised and unsupervised learning.* Ph. D. thesis, École Polytechnique Fédérale de Lausanne.

B. S. Atal (1983). Efficient coding of LPC parameters by temporal decomposition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, Boston, USA, pp. 81–84.

V. Aubergé (1991). La synthèse de la parole : « des règles aux lexiques ». Thèse de troisième cycle, I.N.P.G./Université Pierre Mendès France.

G. Bailly (1986). Un modèle de congruence relationnel pour la synthèse de la prosodie du français. In *15es Journées d'Étude sur la Parole*, Aix-en-Provence, France, pp. 75–78. Organisée par le GALF.

G. Bailly (1989). Integration of rhythmic and syntactic constraints in a model of generation of French prosody. *Speech Communication 8*, 137–146.

G. Bailly, T. Barbe & H. Wang (1992). Automatic labelling of large prosodic databases: tools, methodology and links with a text-to-speech system. In G. Bailly & C. Benoît (Eds.), *Talking Machines: Theories, Models and Designs*, pp. 323–333. Elsevier B.V.

G. Bailly, R. Laboissière & J.-L. Schwartz (1991). Formant trajectories as audible gestures: an alternative for speech synthesis. *Journal of Phonetics 19*(1), 9–23.

T. Barbe (1990). Méthodologie et outils pour la mise en œuvre automatique d'une synthèse de parole de haute qualité. Thèse de troisième cycle, I.N.P.G.

P. Barbosa & G. Bailly (1994). Characterisation of rhythmic patterns for text-to-speech synthesis. *Speech Communication 15 (1-2)*, 127–137.

K. Bartkova & C. Sorin (1987). A model of segmental duration for speech synthesis in French. *Speech Communication 6*, 245–260.

F. Beaugendre & A. Lacheret-Dujour (1993). Automatic generation of French intonation based on a perceptual study and morpho-syntactic information. In *Proceedings of European Conference on Speech Communication and Technology*, Volume 2, Berlin, pp. 1219–1222.

F. Bell-Berti, C. E. Gelfer, M. Boyle & C. Chevrie-Muller (1990). Neurological factors in speech timing. *J. Acoust. Soc. Am. 88, Suppl. 1*, S128. 120^{th} Meeting: Acoustical Society of America.

C. Benoît (1990). An intelligibility test using semantically unpredictable sentences: Towards the quantification of linguistic complexity. *Speech Communication 9*, 293–304.

E. Benveniste (1951). La notion de « rythme » dans son expression linguistique. *J. Psychol. Norm. Path. 44*, 401–411.

H. Bergson (1968). *Durée et simultanéité*. Paris: Presses Universitaires de France.

V. Berthier, C. Abry & T. Lalouache (1991). Coordination du geste et de la parole dans la production d'un instrument traditionnel. In *Proceedings of the XII^{th} International Conference on Phonetic Sciences*, Volume 4, Aix-en-Provence, France, pp. 34–37. Université de Provence.

F. Bimbot (1988). Synthèse de la parole : des segments aux règles, avec utilisation de la décomposition temporelle. Thèse de troisième cycle, ENST.

D. Bolinger (1989). *Intonation and its uses.* London: Edward Arnold.

M.-C. Botte, G. Canévet, L. Demany & C. Sorin (1989). *Psychoacoustique et perception auditive.* Paris: Les Éditions INSERM.

L. Breiman, J. H. Friedman, R. A. Olshen & C. J. Stone (1984). *Classification and Regression Trees.* Monterey, CA, EEUU: Wadsworth & Brooks.

C. P. Browman & L. M. Goldstein (1985). Dynamic modeling of phonetic structure. In V. A. Fromkin (Ed.), *Phonetic Linguistics: Essays in honor of Peter Ladefoged*, pp. 35–53. Academic Press, Inc.

G. Caelen-Haumont (1978). Strcutures prosodiques de la phrase énociative simple et étendue. Thèse de troisième cycle, Toulouse.

Calliope (1989). *La parole et son traitement automatique.* Paris: Masson.

W. N. Campbell (1992). *Multi-level Timing in Speech.* Ph. D. thesis, University of Sussex.

W. N. Campbell (1993). Automatic detection of prosodic boundaries in speech. *Speech Communication 13*, 343–354.

R. Carlson, B. Granström & D. H. Klatt (1979). Some notes on the perception of temporal patterns in speech. In *9^{th} International Conference of Phonetic Sciences*, Copenhague, Danemark.

L. A. Chistovich (1980). Auditory processing of speech. *Language and Speech 23*(1), 67–73.

L. A. Chistovich & E. A. Ogorodnikova (1982). Temporal processing of spectral data in vowel perception. *Speech Communication 1*, 45–54.

A. Classe (1939). *The Rhythm of English Prose.* Oxford: Blackwell.

A. Cohen, R. Collier & T. Hart (1982). Declination: Construct or intrinsic feature of speeck pitch? *Phonetica 39*, 254–273.

C. H. Coker (1968). Speech synthesis with a parametric articulatory model. In J. L. Flanagan & L. R. Rabiner (Eds.), *Speech Symposium, reprinted in Speech Synthesis*, pp. 135–139. Dowden, Hutchinson and Ross, Stroudsburg, PA.

R. Collier (1992). A comment on the prediction of prosody. In G. Bailly & C. Benoît (Eds.), *Talking Machines: Theories, Models and Designs*, pp. 205–208. Elsevier B.V.

A. M. Cooper, D. H. Whalen & C. Fowler (1988). The syllable's rhyme affects its P-center as a unit. *Journal of Phonetics 16*, 231–241.

L. A. Cooper (1982). Strategies for visual comparison and representation: individual differences. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence*. Hillsdale, N.J.: Lawrence Erlbaum.

W. Cooper & J. Sorensen (1981). *Fundamental Frequency in Sentence Production*. Heidelberg: Springer.

W. A. Cooper, J. M. Paccia & S. G. Lapointe (1978). Hierarchical coding in speech timing. *Cognitive Psychology 10*, 154–177.

O. Creutzfeldt, F.-G. Hellweg & C. Schreiner (1980). Thalamo-cortical transformation of responses to complex auditory stimuli. *Exp. Brain Res. 39*, 87–104.

D. Crystal (1969). *Prosodic systems and intonation in English*. Cambridge, UK: Cambridge University Press.

D. Crystal (Ed.) (1985). *A dictionnary of Linguistics and Phonetics*. Basil Blacwell in association with André Deutsch.

T. H. Crystal & A. S. House (1988). Segmental durations in connected speech signals. *J. Acoust. Soc. Am. 83*, 1553–1573.

T. H. Crystal & A. S. House (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *J. Acoust. Soc. Am. 88*(1), 101–112.

R. M. Dauer (1983). Stress-timing and syllable-timing re-analyzed. *Journal of Phonetics 11*, 51–62.

K. de Jong, M. E. Beckman & J. Edwards (1993). The interplay between prosodic structure and coarticulation. *Language and Speech 36*(2-3), 197–212.

M. de Staël (1932). *De l'Allemagne*. Éditions Garnier. II, IX.

P. Delattre (1966). A comparison of syllable length conditioning among languages. *IRAL 4*, 183–198.

F. Dell (1984). L'accentuation dans les phrases en français. In F. Dell, D. Hirst & J. Vergnaud (Eds.), *Forme sonore du langage*, pp. 65–122. Hermann.

J. Derrida (1968). Sémiologie et grammatologie. In *Information sur les sciences sociales*.

A. Di Cristo (1981). Aspects phonétiques et phonologiques des éléments prosodiques. *Modèles Linguistiques 3*, 5–24.

A. Di Cristo (1985). *De la microprosodie à l'intonosyntaxe*. Ph. D. thesis, Universitè de Provence.

G. Dogil (1984). Nonlinear phonology in the area of speech and grammar pathology. In H. van der Hulst & N. Smith (Eds.), *Advances in nonlinear phonology*, pp. 359–376. Foris, Dordrecht.

G. Dogil (1987). Prototypical speech events and speech perception. In *Proceedings of the XIth International Congress of Phonetic Sciences*, Volume 3, Tallin, Estonia, pp. 360–365.

G. Dogil & G. Braun (1988). *The PIVOT model of speech parsing.* Wien: Verlag.

T. du Moncel (1880). *Le téléphone, le microphone et le phonographe, Chapitre 3.* Paris.

J. Dubois, M. Giacomo, L. Guespin, C. Marcellesi, J.-B. Marcellesi & J.-P. Mével (1994). *Dictionnaire de linguistique et des sciences du langage.* Paris: Larousse.

O. Ducrot & T. Todorov (1972). *Dictionnaire encyclopédique des sciences du langage.* Paris: Éditions du Seuil.

D. Duez (1987). *Contribution à l'étude de la structuration temporelle de la parole en français.* Ph. D. thesis, Université de Provence.

E. Dupoux (1989). Identification des mots parlés - détection de phonèmes et unité prélexicale. Thèse de Doctorat, EHESS.

M. Duras (1993). *Écrire.* Paris: Éditions Gallimard. p. 65.

J. Edwards & M. E. Beckman (1988). Articulatory timing and the prosodic interpretation of syllable duration. *Phonetica 45*, 156–174.

J. Edwards, M. E. Beckman & J. Fletcher (1991). The articulatory kinematics of final lenghthening. *Journal of the Acoustical Society of America 89*(1), 369–382.

P. A. Eling, J. C. Marschall & G. P. van Galen (1980). Perceptual centres for Dutch digits. *Acta Psychologica 46*, 95–102.

F. Emerard (1977). *Synthèse par diphones et traitement de la prosodie.* Ph. D. thesis, Université de Grenoble III.

N. P. Erber (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders 40*, 481–482.

G. Fant (1973). *Speech Sounds and Features.* Cambridge, USA: M. I. T.

G. Fant (1991). Units of temporal organization. stress groups versus syllables and words. *Proceedings of the XIIth International Conference on Phonetic Sciences 1*, 247–250.

G. Fant & A. Kruckenberg (1989). Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR 2*, 1–80.

G. Fant, A. Kruckenberg & L. Nord (1991). Durational correlates of stress in Swedish, French and English. *Journal of Phonetics 19*, 351–361.

Fletcher & Munson (1933). Loudness, its definition, measurement, and calculation. *J. Acoust. Soc. Am. 5*, 1982.

C. A. Fowler (1979). "Perceptual centres" in speech production and perception. *Perception and Psychophysics 25*, 375–388.

C. A. Fowler, D. H. Whalen & A. M. Cooper (1988). Perceived timing is produced timing: a reply to Howell. *Perception and Psychophysics 43*, 94–98.

P. Fraisse (1974). *La psychologie du rythme*. Paris: Presses Universitaires de France.

D. B. Fry (1955). Duration and intensity as physical correlates of linguistic stress. *J. Acoust. Soc. Am. 27*, 765–768.

H. Fujisaki & H. Kawai (1988). Realization of linguistic information in the voice fundamental frequency contour of the spoken japanese. pp. 663–666. IEEE International Conference on Acoustics, Speech and Signal Processing.

H. Fujisaki & H. Sudo (1972). A generative model for the prosody of connected speech in japanese. Newton, USA, pp. 140–143. IEEE International Conference on Acoustics, Speech and Signal Processing.

E. Gårding (1983). A generative model of intonation. In A. Cutler & D. R. Ladd (Eds.), *Prosody: Models and Measurements*. New York: Springer-Verlag.

P. Garde (1968). *L'Accent*. Paris: Presses Universitaires de France.

W. R. Gardner (1974). *The processing of information and structure*. Potomac, Md: Erlbaum.

J. P. Gee & F. Grosjean (1981). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology 15*, 411–458.

K. Gibran (1990). *Le Prophète*. Paris: Éditions Albin Michel. Traduction de Marc de Smedt, p. 19.

G. M. Glasgow (1952). A semantic index of vocal pitch. *Speech Monographs 19*, 64–68.

F. Goldman-Eisler (1968). *Psycholinguistics: Experiments in Spontaneous Speech*. London: Academic Press.

M. Grammont (1965). *Petit traité de versification française*. Paris: Armand Collin.

F. Grosjean & J.-Y. Dommergues (1983). Les structures de performance en psycholinguistique. *L'Année Psychologique 83*, 513–536.

F. Grosjean & L. G. H. Lane (1979). The pattern of silence: performance structures in sentence production. *Cognitive Psychology 11*, 58–81.

P. Guillaume (1979). *La Psychologie de la forme.* Paris: Flammarion.

D. G. Hays (1964). Dependency theory: a formalism and some observations. *Language 46*, 511–525.

D. Hirst (1993). Peak, boundary and cohesion characteristics of prosodic grouping. In D. House & P. Touati (Eds.), *Proc. ESCA Workshop on Prosody*, Lund, Sweden, 27-29 September, pp. 32–37.

D. J. Hirst (1977). *Intonative features.* The Hague: Mouton.

L. T. Hjelmslev (1968). *Structure fondamentale du langage.* Paris: Éditions de Minuit. Traduction de l'œuvre en danois, Prolégomènes à une théorie du langage, Copenhague.

C. F. Hockett (1942). A system of descriptive phonology. *Language 18*, 3–21.

A. House (1961). On vowel duration in English. *J. Acoust. Soc. Am. 33*, 1174–1178.

P. Howell (1988). Prediction of P-center location from the distribution of energy in the amplitude envelope: I & II. *Perception and Psychophysics 43*, 90–93,99.

P. Howell & S. Scott (1992, June). Alternative theoretical perspectives for explaining the basis of perceptually isochronous rhythms. In C. Auxiette, C. Drake & C. Gérard (Eds.), *Fourth Rhythm Workshop: Rhythm Perception and Production*, pp. 141–149. Bourges, France: Ville de Bourges.

ICP (1994). Équipe synthèse : du faire-semblant au faire-comme. *Rapport d'Activité de l'Institut de la Communication Parlée*, 49–50.

M. Jordan (1986). Serial order: a parallel distributed processing approach. Technical Report ICS 8604, Institute for Cognitive Science, La Jolla: University of California.

D. Katz (1950). *Gestalt Psychology. Its nature and Signifiance.* New York: Ronald Press.

H. Kellenberger (1932). *The Influence of Accentuation on French Word Order.* Princeton.

E. Keller (1994). The serial application of high-level constraints on timing variables. Speech Maps Research Project 6975, Université de Lausanne, Lausanne, Suisse.

E. Keller, B. Zellner, S. Werner & N. Blanchoud (1993). The prediction of prosodic timing: rules for final syllable lengthening in French. In D. House & P. Touati

(Eds.), *Proc. ESCA Workshop on Prosody*, Lund, Sweden, 27-29 September, pp. 212–215.

D. H. Klatt (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics 3*, 129–140.

D. H. Klatt (1976). Linguistic uses of segmental duration in English : acoustic and perception evidence. *J. Acoust. Soc. Am. 59*, 1208–1221.

D. H. Klatt (1979). Synthesis by rule of segmental durations in English sentences. In B. Linblom & S. Ohman (Eds.), *Frontiers of Speech Communications Research*, pp. 287–299. New York: Academic Press.

D. H. Klatt (1987). Review of text-to-speech conversion for English. *J. Acoust. Soc. Am. 82*, 737–793.

D. H. Klatt & W. E. Cooper (1975). Perception of segment duration in sentence contexts. In A. Cohen & S. G. Nooteboom (Eds.), *Structure and Process in Speech Perception*, pp. 69–89. New York: Springer.

S. C. Kleene (1956). Representation of events in nerve nets and finite automata. In C. E. Shannon & J. McCarthy (Eds.), *Automata Studies*, pp. 3–42. Princeton, NJ: Princeton University Press.

K. J. Kohler (1986). Invariability and variability in speech timing: from utterance to segment in German. In J. Perkell & D. H. Klatt (Eds.), *Invariance and Variability in Speech Processes*, pp. 268–298. Erlbaum Hillsdale.

K. J. Kohler (1987). The linguistic functions of F0 peaks. In *Proceedings of the XIth International Conference on Phonetic Sciences*, Volume 3, Tallinn, Estonia, pp. 149–152.

K. J. Kohler (1991). Prosody in speech synthesis: The interplay between basic research and TTS application. *Journal of Phonetics 19*(1), 121–138.

M. Kohno (1992). Two mechanisms of processing sound sequences. In Y. Tohkura, E. Vatikiotis-Bateson & Y. Sagisaka (Eds.), *Speech, perception, production and linguistic structure*, pp. 287–293. IOS Press.

J.-P. Köster (1973). *Historische Entwicklung von Syntheseapparaten.* Hamburg: Helmut Buske Verlag Hamburg.

V. A. Kozhevnikov & L. A. Chistovich (1965). Speech articulation and perception. In *Joint Publications Research Service*, pp. 543.

R. Laboissière (1992). Préliminaires pour une robotique de la communication parlée : inversion et contrôle d'un modèle articulatoire du conduit vocal. Thèse de troisième cycle, I.N.P.G.

D. R. Ladd (1983). Phonological features of intonational peaks. *Language 59*, 721–759.

D. R. Ladd (1988). Declination "reset" and the hierarchical organization of utterances. *Journal of the Acoustical Society of America 84*(2), 530–544.

D. R. Ladd & N. Campbell (1991, August). Theories of prosodic structure: evidence from syllable duration. In *Proceedings of the XIIth International Conference on Phonetic Sciences*, Aix-en-Provence, France, pp. 290–293.

G. D. Lame (1990). P centers and temporal distortion. *J. Acoust. Soc. Am. 88, Suppl. 1*, S128. 120th Meeting: Acoustical Society of America.

W. Lawrence (1974). The phoneme, the syllable, and the parameter track. In *Proc. Speech Communication Seminar*, Stockholm.

W. A. Lea (1974). Prosodic aids to speech recognition: IV. A general strategy for prosodically-guided speech understanding. Univac Report PX10791, Sperry Univac, DSD, St. Paul, Minnesota, USA.

I. Lehiste (1970). *Suprasegmentals*. Cambridge, Massachussets: MIT Press.

I. Lehiste (1977). Isochrony reconsidered. *Journal of Phonetics 5*, 253–263.

I. Lehiste (1978). Temporal organization and prosody. perceptual aspects. In *Joint Meeting of A.S.A. and A.S.J.*, Volume 1, Honolulu, pp. 1–17.

I. Lehiste, J. P. Olive & L. A. Streeter (1976). The role of duration in disambiguating syntactically ambiguous sentences. *J. Acoust. Soc. Am. 60*, 1199–1202.

H. C. Leiner, A. L. Leiner & R. S. Dow (1991). The human cerebro-cerebellar system: its computing, cognitive, and language skills. *Behavioural Brain Research 44*, 113–128.

H. C. Leung & V. W. Zue (1984). A procedure for automatic alignment of phonetic transcriptions with continuous speech. In *Proceedings of the IEEE ICASSP*, Volume 1, San Diego, pp. 2.7.1–2.7.4.

A. M. Liberman & I. G. Mattingly (1985). The motor theory of speech perception revisited. *Cognition 21*, 1–36.

A. M. Liberman & M. Studdert-Kennedy (1978). Phonetic perception. In R. Held, H. W. Leibowitz & H. L. Teuber (Eds.), *Handbook of sensory physiology*, Volume VIII: Perception, pp. 431–461. New York: Springer-Verlag.

M. Liberman & J. Pierrehumbert (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff & R. Oerhle (Eds.), *Language Sound Structure*, pp. 157–233. Cambridge, MA, U.S.A.: MIT Press.

M. Liberman & A. Prince (1977). On stress and linguistic rhythm. *Linguistic Inquiry 8*, 249–336.

B. Lindblom (1986). Phonetic universals in vowel systems. *Experimental Phonology*, 13–44.

B. Lindblom (1990). Explaining phonetic variation: A sketch of the H & H theory. In H. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling*, pp. 403–440. Dordrecht: Kluwer.

R. Lippmann (1987). An introduction to computing with neural nets. *IEEE on Acoustics, Speech and Signal Processing Magazine*, 4–22.

R. R. Llinás (1989). The role of the intrinsic electrophysiological properties of central neurones in oscillation and resonance. In A. Goldbeter (Ed.), *Cell to Cell Signalling: From Experiments to Theoretical Models*, pp. 3–16. Academic Press.

R. Llorca (1984). Éléments d'analyse du rythme de la parole en français. Thèse de troisième cycle, Université de Franche-Comté, Besançon, France.

J. Locke (1983). *Phonological acquisition and change*. New York: Academic Press.

H. Lœvenbruck & P. Perrier (1993). Vocalic reduction: prediction of acoustic and articulatory variabilities with invariant motor commands. In *Proceedings of the 3rd European Conference on Speech Communication and Technology*, pp. 83–85. Berlin, Allemagne.

I. Maddieson (1983). *Patterns of sounds*. Cambridge, UK: Cambridge University Press.

R. C. Major (1981). Stress-timing in Brazilian Portuguese. *Journal of Phonetics 9*, 343–351.

S. M. Marcus (1975). Perceptual centres. Unpublished fellowship dissertation, King's College, Cambridge, UK.

S. M. Marcus (1976). Perceptual centres. Unpublished PhD thesis, Cambridge University.

P. Martin (1973). Les problèmes de l'intonation : recherches et application. *Langue Française 19*, 4–32.

P. Martin (1981). L'intonation est-elle une structure congruente à la syntaxe ? In *L'Intonation : de l'acoustique à la sémantique*, pp. 234–271. Paris: Klincksieck.

P. Martin (1982). Phonetic realisations of prosodic contours in French. *Speech Communication 1*(3-4), 283–294.

A. Martinet (1991). *Éléments de linguistique générale* (3e ed.). Paris: Armand Collin.

J. Mates (1992, June). Timing and corrective mechanisms in sinchronization of motor acts with a sequence of stimuli. In C. Auxiette, C. Drake & C. Gérard (Eds.), *Fourth Rhythm Workshop: Rhythm Perception and Production*, pp. 43–48. Bourges, France: Ville de Bourges.

J. L. McClelland & D. E. Rumelhart (1988). *Explorations in Parallel Distributed Processing: a handbook of models, programs and exercises.* The MIT Press.

W. S. McCulloch & W. Pitts (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics 5*, 115–133.

C. Meireles (1982). *Viagem/Vaga Música.* Rio de Janeiro: Editora Nova Fronteira. Poème Herança, Coleção Poiesis.

P. Mertens (1993). Intonational grouping, boundaries, and syntactic structure in French. *Working Papers of Lund University 41*, 156–159.

P. Monnin & F. Grosjean (1993). Les structures de performance en français : caractérisation et prédiction. *L'Année Psychologique 93*, 9–30.

Y. Morlec (1994). Génération automatique de la prosodie du français. Rapport de DEA, I.N.P., Grenoble.

A. C. Morris (1992). Analyse informationnelle du traitement de la parole dans le système auditif et le noyau cochléaire - application à la reconnaissance des occlusives voisées du français. Thèse de troisième cycle, I.N.P.G., France.

J. Morton, S. Marcus & C. Frankish (1976). Perceptual centers (p-centers). *Psychological Revue 83*(5), 405–408.

E. Moulines (1992). Synthesis models: a discussion. In G. Bailly & C. Benoît (Eds.), *Talking Machines: Theories, Models, and Designs*, pp. 7–12. Elsevier Science Publishers B.V.

W.-A. Munson (1947). The growth of auditory sensation. *J. Acoust. Soc. Am. 19*, 584–591.

Y. Nishinuma & S. Santi (1992). Comment percevons-nous la dernière syllabe d'une phrase ? Relations entre intensité et durée dans la discrimination de segments vocaliques. *Travaux de l'Institut de Phonétique d'Aix 14*, 149–162.

S. G. Nooteboom (1991, August). Some observations on the temporal organisation and rhythm of speech. In *Proceedings of the XIIth International Conference on Phonetic Sciences*, Volume 1, Aix-en-Provence, France, pp. 228–237.

S. G. Nooteboom, J. P. L. Brokx & J. J. de Rooi (1976). Contributions of prosody to speech perception. In *Studies in Language Perception*. Proceedings of the Symposium on Language Perception, International Congress of Psychology, Paris.

L. Nord, A. Kruckenberg & G. Fant (1990). Some timing studies of prose, poetry and music. *Speech Communication 9*, 477–483.

J. D. O'Connor (1965). The perception of time intervals. Progress Report 2, Phonetics Laboratory, University College, London, UK.

J. D. O'Connor & G. F. Arnold (1973). *Intonation of Colloquial English*. London: Longman. 2^{e} ed.

J. J. Ohala (1970). Aspects of the control and production of speech. Workin Papers en Phonetics 15, UCLA Phonetics Laboratory, Los Angeles.

J. Olive & L. Nakatani (1974). Rule-synthesis of speech by word concatenation: A first step. *J. Acoust. Soc. Am. 55*, 660–666.

D. O'Shaughnessy (1981). A study of French vowel and consonant durations. *Journal of Phonetics 9*, 385–406.

D. O'Shaughnessy (1984). A multispeaker analysis of durations in read French paragraphs. *J. Acoust. Soc. Am. 76*(6), 1664–1672.

A. G. Parlos, K. T. Chong & A. F. Atiya (1994). Application of the recurrent multilayer perceptron in modeling complex process dynamics. *IEEE Transactions on Neural Networks 5*(2), 255–266.

V. Pasdeloup (1987). Analyse acoustique de la structuration rythmique du français oral. In *16^{e} Journée d'Étude sur la Parole*, Hammamet, pp. 259–262.

V. Pasdeloup (1990). Modèle de règles rythmiques du français appliqué à la synthèse de la parole. Thèse de troisième cycle, Université de Provence, Aix-en-Provence.

V. Pasdeloup (1992). Durée inter-syllabique dans le groupe accentuel en français. In U. L. de Bruxelles (Ed.), *19es Journées d'Études sur la Parole*, pp. 531–536.

J. Pierrehumbert (1980). *The phonetics and phonology of English intonation*. Ph. D. thesis, MIT.

J. Pierrehumbert (1981). Synthetizing intonation. *J. Acoust. Soc. Am. 70*(4), 985–995.

K. L. Pike (1945). *The intonation of American English.* Ann Arbor: University of Michigan Press.

B. Pompino-Marschall (1989). On the psychoacoustic nature of the p-center phenomenon. *Journal of Phonetics 17*, 175–192.

B. Pompino-Marschall (1991). The syllable as a prosodic unit and the so-called p-centre effect. Forschungsberichte 29, Institut für Phonetik und Sprachliche Kommunikation der Universität München, Munich, Allemagne.

M. Proust (1954). *Du côté de chez Swann.* Paris: Éditions Gallimard. Collection Folio, p. 123.

K. A. Rapp (1971). A study of syllable timing. Quaterly Progress and Status Report 1/1971, Speech Transmission Laboratory, Stockolm, Sweden.

D. L. Reilly, L. N. Cooper & C. Erlbaum (1982). A neural model for category learning. *Biological Cybernetics 45*, 35–41.

M. Rifaterre (1979). *La Production du texte.* Paris: Éditions du Seuil.

M. D. Riley (1992). Tree-based modelling of segmental durations. In G. Bailly & C. Benoît (Eds.), *Talking Machines: Theories, Models, and Designs*, pp. 265–273. Elsevier Science Publishers B.V.

P. Robert (1993). *Le Nouveau Petit Robert.* Paris: Dictionnaires Le Robert.

J. Robert-Ribes, P. Escudier & J.-L. Schwartz (1992). Modèles d'intégration audition-vision dans la perception des voyelles : une étude neuromimétique. In *Actes du Cinquième Colloque de l'ARC*, Nancy, France, pp. 85–100.

E. Rosch (1973). Natural categories. *Cognitive Psychology 4*, 328–350.

R. Rosenblatt (1959). *Principles of neurodynamics.* New York: Spartan Books.

M. Rossi, A. Di Cristo, D. Hirst, P. Martin & Y. Nishinuma (1981). *L'intonation : de l'acoustique à la sémantique.* Paris: Klincksieck.

D. E. Rumelhart, G. E. Hinton & R. J. William (1986). Learning representation by back-propagating errors. *Nature 323*, 533–536.

Y. Sagisaka (1990). On the prediction of global fo shapes for Japanese text-to-speech. *IEEE International Conference on Acoustics, Speech and Signal Processing 1*, 325–328.

H. Schütte (1978). Ein Funktionsschema für die Wahrnehmung eines gleichmäßigen Rhythmus in Schallimpulsfolgen. *Biological Cybernetics 29*, 49–55.

M. Scordilis & J. Gowdy (1989). Neural network based generation of fundamental frequency contours. *IEEE International Conference on Acoustics, Speech and Signal Processing 1*, 219–222.

S. K. Scott (1993). *Perceptual Centres in Speech - An Acoustic Analysis.* Ph. D. thesis, University College London.

E. Selkirk (1980). The role of prosodic categories in English word stress. *Linguistic Inquiry 11*, 563–605.

A. Semjen, H.-H. Schulze & D. Vorberg (1992, June). Temporal control in the coordination between repetitive tapping and periodic external stimuli. In C. Auxiette, C. Drake & C. Gérard (Eds.), *Fourth Rhythm Workshop: Rhythm Perception and Production*, pp. 73–78. Bourges, France: Ville de Bourges.

Y. Shen & G. G. Peterson (1962). Isochronism in English. *Studies in Linguistics, Occasional Papers 9*, 1–36.

R. N. Shepard (1984). Ecological constraints on internal representation: Resonant kinematics of perceiving, imaging, thimking, and dreaming. *Psychological Review 91*, 417–447.

P. Smolensky (1992). IA connexionniste, IA symbolique et cerveau. In D. Andler (Ed.), *Introduction aux sciences cognitives*, Folio-Essais, Chapter 2, pp. 77–106. Gallimard. Collection Folio-Essais.

K. N. Stevens (1991). The contribution of speech synthesis to phonetics: Dennis Klatt's legacy. In *Proceedings of the XIIth International Confrence on Phonetic Sciences*, Volume 1, pp. 28–37.

K. N. Stevens & S. E. Blumstein (1978). Invariant cues for place of articulation in stop consonants. *J. Acoust. Soc. Am. 64*(5), 1358–1368.

Q. Summerfield (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: the psychology of lipreading*, pp. 3–51. Lawrence Erlbaum.

L. Tesnière (1965). *Éléments de syntaxe structurale.* Paris: Klincksieck.

G. Thinès (1990). Gestaltisme. In *Encyclopædia Universalis*, pp. 434–439.

P. Todd (1989). A connectionist approach to algorithmic composition. *Music Journal 13*(4), 27–43.

P. Touati (1987). *Structures prosodiques du suédois et du français : profils temporels et configurations tonales.* Lund University Press.

C. Traber (1992). Fo generation with a database of natural fo patterns and with a neural network. In G. Bailly & C. Benoît (Eds.), *Talking Machines: Theories, Models, and Designs*, pp. 287–304. Elsevier Science Publishers B.V.

B. Tuller & C. A. Fowler (1981). The contribution of amplitude to the perception of isochrony. Status Report on Speech Research SR-65, Haskins Laboratories.

M. T. Turvey, R. C. Schmidt & L. Rosenblum (1990). Clock and motor components in absolute coordination of rhythmic movements. Status Report on Speech Research SR-101/102, Haskins Laboratories.

J. Vaissière (1971). Contribution à la synthèse par règles du français. Thèse de troisième cycle, Grenoble III.

J. Vaissière (1980). La structuration acoustique de la phrase française. In *Annali della Scuola Normale Superiore di Pisa X*, Volume 2, pp. 529–560.

J. Vaissière (1983). Language-independent prosodic features. In A. Cutler & D. R. Ladd (Eds.), *Prosody: Models and Measurements*, pp. 53–66. New York: Springer-Verlag.

N. Vallée (1994). Systèmes vocaliques : de la typologie aux prédictions. Thèse de troisième cycle, Grenoble III.

J. P. H. van Santen (1992). Deriving text-to-speech durations from natural speech. In G. Bailly & C. Benoît (Eds.), *Talking Machines: Theories, Models, and Designs*, pp. 275–285. Elsevier Science Publishers B.V.

J. P. H. van Santen (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language 8*, 95–128.

G. Veillon (1970). Modèles et algorithmes pour la traduction automatique. Thèse d'état, Grenoble I.

N. Vigouroux & J. Caelen (1985). Segmentations en vue de l'organisation d'une base de données acoustiques et phonétiques. In *14e Journée d'Étude sur la Parole*, Paris, pp. 152–155.

P. Viviani & N. Stucchi (1992). Biological movements look uniform: evidence of motor-perceptual interaction. *Journal of Experimental Psychology: Human Perception and Performance 18*(3), 603–623.

B. J. Wenk & F. Wioland (1982). Is French really syllable-timed? *Journal of Phonetics 10*, 193–216.

D. H. Whalen (1990). Coarticulation is largely planned. *Journal of Phonetics 18*(1), 3–35.

C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf & P. J. Price (1992). Segmental durations in the vicinity of prosodic boundaries. *J. Acoust. Soc. Am. 91*(3), 1707–1717.

N. Willems, R. Collier & J. T. Hart (1988). A synthesis scheme for British English intonation. *J. Acoust. Soc. Am. 84*(4), 1250–1261.

I. H. Witten (1977). A flexible scheme for assigning timing and pitch to synthetic speech. *Language and Speech 20*, 240–260.

H. Woodrow (1951). Time perception. In S. Stevens (Ed.), *Handbook of Experimental psychology*, pp. 1224–1236. New York:Wiley.

D. W. Zaidel (1984). Les fonctions de l'hémisphère droit. *La Recherche 15*(153), 332–340.

Table des matières

Liste des tableaux

Table des figures