# Pistes pour le contrôle d'un robot parlant capable de réduction vocalique

Hélène Loevenbruck

Thèse présentée par

**Hélène LŒVENBRUCK**

pour obtenir le titre de

**DOCTEUR**

**de**

**l'INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE**

(Arrêté ministériel du 30 mars 1992)

**(Spécialité : Sciences cognitives)**

# Pistes pour le contrôle d'un robot parlant capable de réduction vocalique

Soutenue le 28 mai 1996, devant la Commission d'Examen

**JURY :**

| | |
|---|---|
| S. Gentil | Président |
| D. J. Ostry | Rapporteur |
| R. Chatila | Rapporteur |
| E. Vatikiotis-Bateson | Examinateur |
| C. Abry | Examinateur |
| P. Perrier | Examinateur, directeur de thèse |

U.R.A. - C.N.R.S. n°368
icp
INSTITUT DE LA COMMUNICATION PARLEE

# Remerciements

*J'ai appris il y a peu que je suis née scientifiquement à Marseille au début des années 1990 de deux pères extraordinaires que j'honore et remercie ici très sincèrement, Pascal Perrier, mon directeur de thèse et Christian Abry, directeur de l'équipe articulatoire au sein de laquelle j'ai fait mes premiers pas en parole.*

*Merci à Pascal Perrier qui m'offrit un jour ce défi de travailler sur les cibles et qui a su être toujours vigilant à ce que je continue de le relever. Je le remercie pour son enthousiasme, sa générosité, sa disponibilité, son énergie et la passion avec laquelle il vit son métier d'enseignant-chercheur. Je suis fière et heureuse d'être, avec Christophe et Yohan mes frères de thèse —que je remercie au passage pour leur bon cœur—, membre de sa petite famille scientifique. Je remercie Pascal, aux belles moustaches brassenssiennes, d'avoir su trouver chaque fois les mots justes et les conseils stimulants, de m'avoir incitée à faire les rencontres ou les voyages utiles au bon moment, d'avoir su me redonner confiance, grâce à son recul et sa volonté, lorsque mes doutes prenaient le dessus.*

*Merci aux deux rapporteurs Raja Chatila et David J. Ostry, d'avoir accepté de se pencher sur ce travail. À Raja Chatila pour s'être investi dans un domaine qui ne lui était pas familier et à David Ostry pour le temps qu'il a consacré à ce mémoire nonobstant la langue dans laquelle il était écrit. Merci aussi à David de m'avoir accueillie deux mois au sein de son laboratoire à Montréal où j'ai découvert qu'un bon capteur de mouvement peut s'imposer devant un bon microphone. Je remercie David pour toutes les questions qu'il sait poser et qui ont fait certainement progresser ce travail.*

*Je suis très reconnaissante envers Sylviane Gentil, présidente du jury, pour toute la minutie avec laquelle elle a lu cette thèse. Elle a su apporter un éclairage d'automaticienne tout-à-fait intéressant aux chapitres concernant l'optimisation. Je la remercie pour son ouverture, sa gentillesse, et sa disponibilité mais aussi pour sa rigueur scientifique et pour les discussions fructueuses que nous avons pu avoir avant et après la soutenance de thèse.*

*Je remercie Eric Vatikiotis-Bateson qui a accepté d'être membre du jury, malgré la fâcheuse conséquence d'être astreint à lire 250 pages de français. Il a su dégager clairement les points forts et les points faibles de cette thèse et c'est grâce à son recul que j'oriente maintenant avec confiance mes travaux vers une voie que je n'avais pas envisagée alors.*

*Je remercie infiniment Christian Abry d'avoir accepté de lire ce mémoire, mais surtout d'avoir su, en tant que directeur de l'équipe articulatoire, toujours me guider et m'éviter les pièges encourus par les ingénieurs hâtifs. Je remercie Christian de ne pas garder pour lui son savoir immense et d'opérer sans cesse une maïeutique géniale, par voie verbale, électronique, ou poétique.*

*Je tiens à remercier les deux directeurs de l'Institut de la Communication Parlée, le premier Jean-Marc Dolmazon pour m'y avoir accueillie et le second, Pierre Escudier, pour m'avoir encouragée à y rester grâce à son enthousiasme communicatif, ses passions et son ouverture d'esprit remarquable.*

*Je remercie également Pierre Escudier de m'avoir acceptée au sein du passionnant DEA de Sciences Cognitives de Grenoble et au sein de l'orchestre universitaire dont il est le président efficace, énergique et sympathique. Il est la preuve joyeuse que musique et science sont bien sœurs.*

*Je remercie Jean-Luc Schwartz pour sa diction parfaite qui a grandement facilité l'optimisation des données. Mais je le remercie surtout pour ses conseils avisés, sa vision scientifique à long terme, sa compréhension en profondeur des phénomènes scientifiques mais aussi humains...*

*Je tiens à remercier mes trois "grandes sœurs" qui par leur douceur ont su me rendre la vie douillette à l'ICP. Je remercie Joëlle Miguet pour sa finesse, sa gentillesse et le temps qu'elle a su me consacrer dans les moments difficiles. Merci à Francisca Bustarret pour sa joie de vivre et ses sourires qui font chaud au cœur et à Monique Revil pour ses petites attentions délicates et discrètes et ses boutures qui sont maintenant devenues de belles plantes à la maison.*

*Je remercie Nadine Bioud qui sait dompter la mêlée de photocopieuse, agrapheuse et relieur récalcitrants et sait transformer un essai de document en un mémoire de thèse digne de ce nom.*

*Je remercie Nino Medves pour son efficacité et sa disponibilité qui m'ont permis de rester sereine lorsque des difficultés techniques apparaissaient.*

*Je remercie aussi Rafael Laboissière qui eut la patience de m'expliquer les secrets et les ficelles du fameux logiciel rbp et Gérard Bailly qui sait encore l'exploiter. Merci à Christophe Vescovi pour son aide et ses conseils lors de l'utilisation du logiciel SIMOND.*

*Je tiens à remercier tous les chercheurs, thésards et stagiaires de l'ICP, "permanents" et "provisoires", pour leur précieux conseils et toutes leurs idiosyncrasies qui font qu'il est doux de travailler à l'ICP.*

*Merci à Thierry Guiard-Marigny, Maryline Letranchant, Denis Coté et David Ostry d'avoir participé à l'expérience avec l'Optotrak et accepté de parler la bouche pleine d'un appareillage ridicule.*

*Je remercie les stagiaires de l'ICP et les amis qui ont bien voulu perdre une heure de leur temps précieux pour subir les tests perceptifs. Merci à Hélène Fulchiron, Fabienne Libaud, Mariette Bessac, Philippe Pichon, Stéphane Fehrenbach, Pascale Giraudet, Cédric Chosson, Cédric Vieau, Jean-Marc Boedt, David Bouvier, Eddy Zongo. Et merci à Christophe Savariaux et Emmanuel Tessier pour la mise en place matérielle et logicielle de l'expérience.*

*Je remercie Kcurb et Kiki, mes parents chéris, qui les premiers m'ont fait connaître Daniel Jones et André Martinet et m'ont enseigné les nuances de l'accent d'insistance et l'accent lexical en anglais. Je les remercie pour tout ce qu'ils m'ont donné et insufflé. Merci d'avoir su communiquer assez d'énergie à la flèche pour qu'elle puisse espérer atteindre sa cible...*

*Je remercie Pistoulette l'artiste et Ptit-Riri-kia-grandi, pour leur affection et pour leurs rappels tendres et malicieux qu'être l'aînée ne dispense pas de croire en ses rêves et ne permet pas de donner des leçons ineptes.*

*Et je dédie les quelques pages qui suivent à mon Philou, tendre humaniste éclairé, sans qui ce pavé girait incomplet au fond d'un ordinateur obsolète. Merci d'avoir su croire contre vents et marées en l'aboutissement de ce travail, d'avoir su t'effacer quand il le fallait et être présent quand j'en avais tant besoin, balayer mes doutes stériles et écouter mes angoisses. Merci pour les beaux dessins et tableaux qui ornent ce texte aride. Merci d'avoir participé à mes joies et mes effervescences. Merci d'avoir vécu cette thèse avec moi.*

# TABLE DES MATIÈRES

## LISTE DES ILLUSTRATIONS

# INTRODUCTION

"Peut-être est-ce parce qu'il ne savait pas la musique qu'il avait pu éprouver une impression aussi confuse, une de ces impressions qui sont peut-être pourtant les seules purement musicales, inétendues, entièrement originales, irréductibles à tout autre ordre d'impressions. Une impression de ce genre pendant un instant est pour ainsi dire *sine materia*. Sans doute les notes que nous entendons alors, tendent déjà, selon leur hauteur, et leur quantité, à couvrir devant nos yeux des surfaces de dimensions variées, à tracer des arabesques, à nous donner des sensations de largeur, de ténuité, de stabilité, de caprice. Mais les notes sont évanouies avant que ces sensations soient assez formées en nous pour ne pas être submergées par celles qu'éveillent déjà les notes suivantes ou même simultanées."

Marcel Proust. *Du Côté de chez Swann.*

## La parole et ses partitions

À partir de la même suite de notes (fréquences) musicales, il est possible de produire une quantité infinie de pièces de musique différentes, en jouant sur le tempo (*adagio, allegro, andante,* ...), le rythme (croches, noires, triolets, ...), les accords ou les notes simples, les nuances (*piano, fortissimo, descrescendo*), les attaques (*sforzando, smorzando*). La musique est variabilité, mais on sait la reproduire et le chef d'orchestre connaît les paramètres qu'il faut ajuster pour produire l'effet souhaité.

La parole, compétence cognitive des plus remarquables, mais aussi des plus partagées, est elle aussi variabilité. Il est possible de parler rapidement/lentement, fort/doucement, de mettre plus ou moins d'emphase, d'accentuation. Sur cette plasticité de la parole, les mots de B. Lindblom [1988] valent sûrement mieux que les nôtres :

"*Everyday experience indicates that speaking is a highly flexible process. We are capable of varying our style of speech from fast to slow, soft to loud, casual to clear, intimate to public. We speak in different ways when talking to foreigners, babies, computers and hard of hearing persons. And we change our pronunciations as a function of the social rules that govern speaker-listener interactions.*"

Comme Proust qui s'interroge sur la matérialité de la musique, on peut s'interroger sur la matérialité de la parole. Est-il possible de relier les caprices, les nuances du discours à quelques paramètres contrôlables? Peut-on récupérer, à partir d'un signal de parole hautement variable, des régularités comportementales? Peut-on expliquer comment s'ajuste la communication parlée, à l'aide d'un petit nombre de paramètres clefs? Cette question de la récupération des commandes —générées par le système nerveux central— qui caractérisent le message phonémique et règlent la prosodie, s'inscrit naturellement dans le projet européen ESPRIT "Speech MAPS" (*Mapping of Action and Perception in Speech*) auquel nous avons participé, au sein de l'Institut de la Communication Parlée (ICP), ces trois dernières années et dont un des objectifs essentiels était l'étude des stratégies de contrôle d'un robot parlant simulant de façon anthropomorphique le processus allant de l'encodage moteur à la génération du signal acoustique de parole. Notre recherche a consisté, par le recours à un modèle dynamique simplifié des articulateurs de la parole, en l'inférence de stratégies motrices aptes à décrire les signaux acoustiques et articulatoires fort variés, mesurés dans certaines tâches vocaliques. Nous espérons, par les travaux que nous présentons ici, contribuer, tant soit peu, à la compréhension des relations entre les niveaux centraux (moteurs) et périphériques (articulatoires et acoustiques) dans la production de la parole et tester quantitativement certaines théories perceptives sur la récupération d'éléments invariants dans la dynamique des mouvements articulatoires.

Dans le premier chapitre, nous présentons le débat sur l'invariance et la variabilité en parole et montrons comment, de l'invariance poursuivie par les premiers phonéticiens, on peut passer à une variabilité assumée. Nous introduisons alors notre démarche, consistant en l'évaluation d'un cadre d'hypothèses par une modélisation quantitative. Le deuxième chapitre esquisse le cadre de notre travail. Nous y discutons de l'existence des cibles en parole, traitons des principaux modèles de contrôle de la parole à ce jour et proposons un cadre général pour le contrôle de la production de la parole, exploitant nos hypothèses sur la nature des cibles. Le robot parlant est contrôlable, il reste à le rendre capable de communication parlée. Le troisième chapitre est alors consacré à une expérience de synthèse adaptative au cours de laquelle le robot parlant se montre capable de flexibilité. C'est au quatrième chapitre que nous mettons en œuvre la récupération des cibles invariantes et des paramètres variables, dans trois conditions d'élocution différentes et à partir de deux modèles dynamiques de la langue et la mandibule. Des tests perceptifs sont menés, à la suite d'une expérience de récupération à l'aide d'un premier modèle simple, pour s'assurer de la pertinence du petit nombre de paramètres responsables de variabilité prosodique. Une deuxième expérience de récupération de cibles est élaborée, à partir d'un modèle biomécanique plus sophistiqué et plus réaliste. Nous concluons sur la validité de nos hypothèses initiales et donnons quelques pistes pour aller plus loin encore...

*"Là ci darem la mano"*
et *andiam allegretto!*

# CHAPITRE I

# Invariance *vs* Variabilité

*"Ellowen Deeowen"*

Salman Rushdie

# 1.1 De l'invariance poursuivie à la variabilité assumée

> HIGGINS *[with the roar of a wounded lion]* Stop. Listen to this, Pickering. This is what we pay for as elementary education. This unfortunate animal has been locked up for nine years in school at our expense to teach her to speak and read the language of Shakespear and Milton. And the result is Ahyee, Bə-yee, Cə-yee, Də-yee. *[To Eliza]* Say A, B, C, D.
> LIZA *[almost in tears]* But I'm saying it. Ahyee, Bə-yee, Cə-yee -
> HIGGINS. Stop. Say a cup of tea.
> LIZA. A cappətə-ee.
> HIGGINS. Put you tongue forward until it squeezes against the top of your lower teeth. Now say cup.
> LIZA. C-c-c - I cant. C-Cup.
> PICKERING. Good. Splendid, Miss Doolittle.
> HIGGINS. By Jupiter, she's done it at the first shot. Pickering: we shall make a duchess of her.
>
> Bernard Shaw. *Pygmalion.*

Les phonéticiens furent d'abord des taxinomistes qui tentaient de déterminer les traits caractéristiques pertinents des sons, aux niveaux acoustique ou articulatoire. Selon Monin [1991], les voyelles ont été, de l'antiquité jusqu'au milieu du XVIIIe siècle, simplement classées dans l'ordre arbitraire et linéaire de l'alphabet : a, e, i, o, u. Quelques tentatives de classement selon la durée des voyelles ou le degré d'aperture apparaissent au XVIe siècle mais restent limitées à la description de certaines voyelles. En 1653, le mathématicien J.Wallis propose un schéma de classement des voyelles selon le lieu d'articulation et le degré d'ouverture qui s'applique aux voyelles en général, et non pas seulement à celles de l'anglais. Les observations articulatoires et acoustiques se font plus précises au cours du XVIIIe siècle, et en 1781, Hellwag, dans son mémoire de médecine intitulé *De Formatione Loquelæ* (traduit en français dans le *Bulletin de la Communication Parlée* [1991]), propose la première représentation triangulaire des voyelles, établie, selon l'auteur, à partir de critères physiologiques, articulatoires et auditifs :

*"La première des voyelles, la base des autres ou, si elles sont disposées en échelle, le centre, c'est le* a. *À partir du* a *monte une double échelle qui se termine, sur les derniers degrés, par* i *et* u. *Entre les degrés extrêmes et leurs homologues inférieurs existent des degrés intermédiaires. La relation qui existe entre ces degrés et ces points intermédiaires à la base peut être représentée par le schéma symétrique suivant :*

u ü i

o ö e

a° ä

a

*La voyelle* o *se situe, au milieu, entre le* u *et le* a°, a° *entre le* o *et le* a; *de la même façon,* e *se trouve entre le* i *et le* ä, ä *entre le* e *et le* a. *Par* ü, *on passe de* u *à* i, *par* ö *de* o *à* e. *On peut représenter le point par lequel on passe de* a° *à* ä. *Entre ces degrés que j'ai décrits, on peut, à l'infini, en interpoler d'autres que des nations différentes par leurs langues et par les variétés des langues produisent en permanence quand elles parlent. Toutes les voyelles et toutes les diphtongues que la langue humaine a produites ne peuvent-elles pas être déterminées quasi mathématiquement en fonction de ces degrés?".*

Selon Boë & Perrier [1988] et Boë, Gabioud, Perrier, Schwartz & Vallée [1995], la structure triangulaire originale de Hellwag marque un tournant dans l'histoire des représentations vocaliques. La classification des voyelles en termes de positions horizontale et verticale de la langue est, d'après Catford [1981], officialisée à la suite de travaux de A. Melville Bell [1867], publiés dans son traité intitulé "*Visible Speech: the Science of Universal Alphabetics*". Cette classification, reprise par Henry Sweet [1877], et systématisée par Daniel Jones [1922, 1940] dans son schéma des voyelles cardinales, est à l'origine de la plupart des descriptions vocaliques actuelles.

Henry Sweet [1899] (le pygmalion de la pièce de Bernard Shaw) définit en ces termes les méthodes d'étude en phonétique :

*"The first business of phonetics is to describe the actions of the organs of speech by which sounds are produced, as when we describe the relative positions of tongue and palate by which (s) is produced. This is the* organic *side of phonetics. The* acoustic *investigation of speech-sounds, on the other hand, describes and classifies them according to their likeness to the ear, and explains how the acoustic effect of each sound is the necessary result of its organic formation, as when we call (s) a hiss-sound or sibilant, and explain why it has a higher pitch —a shriller hiss—than the allied hiss-consonant (ʃ) in* she."

L'étude des voyelles reposait alors sur une investigation des sensations proprioceptives fournissant des informations sur la position globale de la langue. Selon ces analyses, les voyelles formaient un diagramme de forme triangulaire dans un repère dont le premier axe représentait le degré d'ouverture de la mâchoire et le deuxième, le degré d'antériorité du point le plus élevé de la langue. Avec le développement de méthodes d'analyse acoustique fiables (spectrographes acoustiques), les phonéticiens ont eu le moyen d'établir des mesures du timbre des voyelles. La description physiologique approximative était secondée par une mesure objective des fréquences de résonance vocaliques (formants). Représentées selon les valeurs de leurs deux premiers formants, les voyelles formaient alors un triangle similaire au triangle articulatoire. Un premier parallèle se dessinait entre classifications articulatoire et acoustique (Joos, [1948], Delattre [1948]). L'étude des transitions formantiques a permis d'autre part une classification des consonnes selon leur

lieu d'articulation. Ces différentes démarches procèdent finalement d'une même intention : trouver les corrélats physiques de la commande linguistique, qu'ils s'agissent d'indices acoustiques ou articulatoires, qui permettent de classer et d'identifier les sons d'une langue. Ceci suppose en fait l'existence d'une certaine invariance physique, que l'on peut définir à la manière de Cooper, Liberman, Borst [1951] :

*"By examining numerous spectrograms of the same sounds, spoken by many persons and in a variety of contexts, an investigator can arrive at a description of the acoustic features common to all of the samples, and in this way make progress toward defining the so-called invariants of speech, that it, the essential information-bearing sound elements on which the listener's identifications critically depend."*.

Cependant, cette idée d'invariance est confrontée à l'observation d'une grande variabilité du substrat physique de la parole :

*"In the elementary case of a word containing a consonant-vowel-consonant phoneme structure, a speaker's pronunciation of the vowel within the word will be influenced by his particular dialectal background; and his pronunciation of the vowel may differ both in phonetic quality and in measurable characteristics from that produced in the word by speakers with other backgrounds. [...] Variations are observed when a given individual makes repeated utterances of the same phoneme. A very significant property of these variations is that they are not random in a statistical sense, but show trends and sudden breaks or shifts in level, and other types of nonrandom fluctuations."* Peterson & Barney [1952].

Si les sons de la parole arborent tant de variabilité inter- et intra- locuteur, comment la communication d'un message linguistique est-elle possible? Quels sont les éléments qui masquent les traits invariants annoncés par les premiers phonéticiens? Ces invariants existent-ils? Ce problème de la variabilité qui dissimulerait l'invariance est encore et toujours d'actualité et sa définition actuelle est donnée en ces termes, par Lindblom, Perkell et Klatt, dans la préface des actes du *Symposium on Invariance and Variability of Speech Processes* [Perkell & Klatt, 1986] :

*"Thus, a major theme motivating applied as well as fundamental speech research efforts continues to be the quest for an integrated theory which accounts for all aspects of the speech code. Presumably such a theory would incorporate definitions of fundamentals units of speech communication and their relationship to signal characteristics and observable behavior, along with a comprehensive and successful accounting of* invariance and variability of speech processes. *What is the nature of invariance? What are the sources and function of speech variability? [...] Variability in the acoustic manifestations of a given utterance is substantial and arises from many sources. These include: (a) recording conditions [...], (b) within speaker variability (breathy/ creaky voice quality, changes in*

*voice fundamental frequency, speaking-rate related undershoot in articulatory targets, slight statistical variation in articulation that may lead to significant acoustic changes, nasality propagation into non-nasals sounds) and (c) cross-speaker-variability (differences in vocal tract anatomy, dialect, detailed articulatory habits). [...] On the other hand, the underlying nature and information-transmitting function of speech communication argue compellingly for some kind of invariance."*

Les points de vue sur l'invariance et la variabilité sont fort nombreux et variés. Certains suggèrent que l'invariance peut-être trouvée dans le signal acoustique lui-même (Stevens & Blumstein [1978]), d'autres prétendent qu'elle se cache dans les trajectoires articulatoires (Fujimura [1986]), d'autres encore estiment que c'est au niveau des "gestes" qu'il faut chercher (Liberman & Mattingly [1985], ou Fowler [1986]). À l'opposé, d'autres rejettent l'idée d'invariance physique et suggèrent que c'est dans le contrôle d'une variabilité qu'il faut rechercher les corrélats de la tâche linguistique ("adaptive variability", Lindblom [1990]). Les positions sur l'invariance et la variabilité peuvent aussi être distinguées suivant qu'elles s'inscrivent dans une étude de la production ou de la perception de la parole (Lindblom [1995]). Nous présentons succinctement quelques éléments de ce débat en tentant de les rattacher à trois courants principaux : invariance acoustique, invariance articulatoire et invariance abandonnée.

### 1.1.1 Invariance acoustique

**La théorie de l'Invariance Acoustique**

Les premières recherches systématiques des propriétés acoustiques qui permettraient d'identifier le lieu d'articulation des consonnes occlusives ont été soldées par des échecs. Il s'avérait en effet que les invariants acoustiques ne tenaient pas lorsque l'environnement vocalique était modifié. Pour un même lieu d'articulation consonantique, Cooper, Delattre, Liberman, Borst & Gerstman [1952] remarquent que l'identification du spectre d'explosion varie d'un environnement vocalique à un autre :

*"most of the subjects heard a rising second formant transition as b and [...] falling transitions might be heard either as g or d, depending on the vowel."*

De même, les points de départ des transitions des deux premiers formants (*locus*) varient en fonction de la voyelle et, de plus, la direction de la transition du deuxième formant dépend du contexte vocalique (Delattre, Liberman & Cooper [1955]). Stevens & Blumstein [1978] estiment cependant que l'invariant peut se trouver au niveau acoustique et qu'il faut rechercher des propriétés acoustiques plus globales ("intégrées") des lieux d'articulation. À l'aide d'expériences d'identification perceptive du lieu d'articulation dans des stimuli CV synthétiques (où C est une consonne occlusive, /b/, /d/ ou /g/, et V une

voyelle quelconque), ces auteurs montrent qu'il est possible d'identifier des indices invariants au sein même du signal acoustique. Ils suggèrent que l'identification du lieu d'articulation par le système auditif s'appuie sur la forme *globale* du spectre à la détente de la consonne, incluant le spectre d'explosion *et* les valeurs formantiques du début du voisement. Les labiales (/b/) présentent un spectre *diffus* (au sens de Jakobson, Fant et Halle [1963], *i.e.* étalement de l'énergie sur une large bande de fréquences) et *descendant* (les amplitudes des formants diminuent avec leur ordre), les alvéolaires (/d/) un spectre *diffus* et *montant* et les vélaires (/g/) un spectre *compact* (concentration de l'énergie dans une zone étroite du spectre) avec un pic proéminent dans les fréquences moyennes de 1 à 3 kHz.

Blumstein [1986] rappelle clairement les deux principes de la **théorie de l'Invariance Acoustique** présentée par Stevens & Blumstein [1978, 1981] (et Blumstein & Stevens [1979, 1980, 1981]) pour les occlusives et les nasales, selon laquelle l'invariance dans le signal acoustique émerge par une intégration de diverses propriétés acoustiques :

*"[The theory of acoustic invariance] has been guided by two claims. The first is that there is acoustic invariance in the speech signal corresponding to the phonetic features of natural language. That it, it is hypothesized that the speech signal is highly structured in that it contains invariant acoustic patterns for phonetic features, and these patterns remain invariant across speakers, phonetic contexts, and languages. The second claim is that the perceptual system is sensitive to these invariant properties. That is, it is hypothesized that the perceptual system can use these invariant patterns to provide the phonetic framework for natural language, and to process the sounds of speech in ongoing perception."*

Blumstein [1989] précise d'autre part que la théorie de l'Invariance Acoustique trouve un élément corroborant dans la Théorie Quantique de la parole présentée par Stevens [1972, 1989]. Selon cette théorie, il existe, dans les espaces acoustique ou articulatoire, des zones de stabilité et des zones de variation, en fonction desquelles la production de la parole se structurerait :

*"The multidimensional space that depicts acoustic-articulatory or auditory-acoustic relations, rather than showing continuous and monotonic variations, exhibits quantal attributes characterized by rapid changes in state over some regions and less abrupt variations or greater stability over other regions. [...] during the time the articulatory structures are close to the target state specified by a particular feature, some change in this configuration or state can occur without a significant modification in the relevant attribute of the sound pattern."*

Stevens suggère que la forme quantique des relations entre paramètres auditifs et acoustiques ou acoustiques et articulatoires permet de relier paramètres acoustiques, auditifs et articulatoires aux traits distinctifs phonologiques classiques (Jakobson *et al.* [1963]).

**Invariance acoustique ou auditive ?**

L'idée qui semble se dégager de ces travaux sur l'invariance acoustique est que l'auditeur reconnaît dans le signal acoustique des formes invariantes. Mais est-ce le signal acoustique lui-même qui est invariant, ou bien l'invariance est-elle plutôt auditive?

La remarque de Lindblom [1988] à ce sujet est assez instructive :

*"Incidentally, let us note that, if it exists, acoustic invariance is a rather strange notion since talkers can only monitor it through their senses and listeners can only access it through their hearing system. Why should sensory feedback and auditory transduction be assumed to impose negligible transformation of the acoustic signal? Is it the case that what people really mean when they talk about acoustic invariance is in fact 'auditory' invariance?"*

Selon Lindblom [1988], l'effet Traunmüller [1985] —l'augmentation groupée du fondamental (F0) et du premier formant (F1) annihile l'effet acoustique de l'augmentation de F1 seul et permet de conserver la qualité de la voyelle initiale— et les résultats de Schulman [publiés en 1989] —les voyelles prononcées à très haute voix, et donc présentant une augmentation de F0, voient leur premier formant augmenter— sont la marque d'une *constance* comportementale visant à préserver un invariant à un niveau de représentation *auditif.*

L'hypothèse selon laquelle l'invariance est auditive est reprise par Kluender, Diehl et Wright [1988]. Ces auteurs défendent l'hypothèse que l'augmentation de la durée des voyelles précédant des consonnes voisées (par rapport à la durée des voyelles précédant des consonnes non-voisées) est *intentionnelle* et qu'elle vise à *renforcer* l'indice acoustique de durée de fermeture qui distingue les consonnes voisées (/b/, durée de fermeture courte) des non-voisées (/p/, durée de fermeture plus longue) : une voyelle plus longue fait paraître l'intervalle de fermeture qui suit plus court et donc appartenant à une consonne plus voisée. Diehl & Walsh [1989] étendent cette hypothèse à la distinction entre occlusives (/b/) et spirantes bilabiales (*glides* /w/) dans des stimuli CV naturels et synthétiques. La durée de la voyelle qui suit la consonne est un indice sur la nature de la consonne : occlusive si durée longue, bilabiale si durée courte. Diehl & Walsh interprètent ce résultat à l'aide d'un principe général de *contraste durationnel.* Le fait que la durée soit un indice distinctif aussi bien pour les stimuli naturels que synthétiques est, selon ces auteurs, une preuve que le principe de *contraste durationnel* est un mécanisme de l'audition en général, c'est-à-dire qu'il n'est pas spécifique à la parole et qu'il est donc plus acceptable que les principes de normalisation du débit de parole invoqués par Miller & Liberman [1979] pour expliquer le même type de résultat. Diehl & Kluender [1989] proposent le terme de "*auditory enhancement*" pour qualifier ce processus de renforcement de la perception des indices acoustiques par l'ajustement de la durée. Les expériences menées par Kluender [1991] sur

des cailles japonaises, entraînées à répondre différemment selon le caractère voisé ou non-voisé de l'occlusive présentée, indiquent que ces oiseaux reproduisent les mêmes différences de catégorisation, selon la valeur de départ du premier formant, que les êtres humains. Selon Kluender, ces similitudes de comportement sont une justification du caractère universel de certains aspects de la perception de la parole et laissent penser que l'invariant auditif n'impliquerait aucun traitement cognitif :

*"[...] human speech perception may rely on quite general mechanisms of audition and categorization."*

Les travaux de Assman, Nearey et Andruski sur la perception des voyelles (Assmann, Nearey & Hogan [1982], Nearey et Assmann [1986], Nearey [1989], Andruski et Nearey [1992]) sont aussi fondés sur l'idée que l'invariance est auditive. L'hypothèse de ces auteurs est que les cibles des voyelles, en termes de formants acoustiques, et les changements spectraux inhérents à chaque voyelle permettent l'identification des voyelles en contexte /bVb/. Nous reviendrons plus en détail, au paragraphe 2.1, sur les implications de ces études dans le débat sur l'existence des cibles.

### 1.1.2 Invariance articulatoire

En alternative à la proposition d'invariance acoustique, les tenants de l'investigation articulatoire proposent que le signal acoustique ne soit que le vecteur de l'invariant situé au niveau articulatoire. Cette conception motrice de la parole est exprimée par Stetson [1928] :

*"Speech is rather a set of movements made audible than a set of sounds produced by movements."*

**La Théorie Motrice de Liberman & Mattingly**

Liberman et ses collègues des laboratoires *Haskins* (Liberman, Cooper, Shankweiler & Studdert-Kennedy [1967]) contestent clairement l'hypothèse selon laquelle l'invariant doit se lire dans le signal acoustique lui-même :

*"[...] there is typically a lack of correspondence between acoustic cues and perceived phoneme, and [...] it appears that perception mirrors articulation more closely than sound."*

Selon Liberman *et al.*, l'invariant doit être recherché au niveau des "commandes motrices" :

*"The most that we can expect is that some subset of [the features that comprise a phoneme in a particular context], and so of the neural signals to the muscles [...], will be invariant with the phoneme; there will be then for each subphonemic feature characteristic neuromotor 'markers', implicating only one or a few component parts of the system,*

*perhaps only the contraction of a single specific muscle. These characteristic components of the total neural activity we will refer to as 'motor commands'."*

Dans une révision de cette première mouture de la *Théorie Motrice*, Liberman & Mattingly [1985] précisent leurs arguments contre les théories de l'invariance auditive. Ils recensent quelques faits incompatibles avec de telles théories :

1. Les indices acoustiques font montre de *variabilité contextuelle.*

2. La *multiplicité* et la *variété* des indices acoustiques d'une même catégorie compliquent considérablement la *description* acoustique.

3. La coarticulation fait qu'il y a *chevauchement* d'informations dans le signal acoustique, alors que perceptivement ces informations sont bien *séparées.* Il n'existe pas de correspondance directe entre les segmentations phonétique et acoustique.

4. Le *contexte*, quoique variable, *contribue* à la perception des catégories phonétiques.

Pour Liberman et Mattingly ces faits sont la preuve qu'il est vain de rechercher l'invariant dans le signal acoustique et ils proposent que les véritables objets de la perception soient les **gestes phonétiques** (articulatoires) planifiés par le locuteur :

*"[...] the objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands that call for movements of the articulators through certain linguistically significant configurations. These gestural commands are the physical reality underlying the traditional phonetic notions —for example, "tongue backing", "lip rounding", and "jaw raising"—that provide the basis for phonetic categories. They are the elementary events of speech production and perception."*

Et ils donnent plus loin une définition précise du geste phonétique :

*"[...] A phonetic gesture [...] is a class of movements by one or more articulators that results in a particular, linguistically significant deformation, over time, of the vocal-tract configuration."*

Ainsi selon eux, l'invariance se trouve précisément dans le geste phonétique. C'est la première conjecture de la Théorie Motrice. La Théorie Motrice permet alors de rendre compte des faits énumérés ci-dessus, incompatibles avec les théories auditives :

1. La *variabilité contextuelle* s'explique par des chevauchements de geste.

2. La *description* est simplifiée : un même geste pour de *multiples* indices acoustiques.

3. Le *chevauchement* des informations acoustiques est une conséquence de la *réalisation* physique de gestes phonétiques *discrets* mettant en jeu des articulateurs communs et qui coarticulent.

4. Le contexte est source d'information sur le processus d'articulation et par là même sur les gestes phonétiques impliqués.

La seconde conjecture de la Théorie Motrice est un corollaire de la première et établit que la production et la perception de la parole sont intimement liées :

*"[...] for language, perception and production are only different sides of the same coin. [...] On the one side of the module, the motor gestures are not the means to sounds designed to be congenial to the ear; rather, they are, in themselves, the essential phonetic units. On the other side, the sounds are not the true objects of perception, made available for linguistic purposes in some common auditory register; rather they only supply the information for immediate perception of the gestures."*

Contrairement à leur description *behavoriste* de 1967, selon laquelle le lien perception-production était construit par l'apprentissage, Liberman & Mattingly affirment ici que ce lien est inné et qu'il existe un "module" spécialisé, le module phonétique, correspondant à la perception et la production des gestes phonétiques. Ils s'inspirent en cela des conceptions de Fodor [1983], qui propose l'existence de modules neuronaux responsables de tâches spécifiques et fournissant aux processus cognitifs centraux les représentations des objets ou des événements d'une classe (d'un domaine) donnée. Selon eux, le module phonétique très spécialisé, qui contrôle un processus perceptif *non cognitif,* est peu soumis aux interventions des processus centraux et est, de ce fait, beaucoup plus rapide dans ses computations que des processus centraux moins spécialisés.

**La Perception Directe**

Le fait que Liberman et Mattingly supposent que la perception ne nécessite pas d'association arbitraire entre le signal acoustique et les catégories phonétiques ni de progression également arbitraire entre une étape auditive et la désignation phonétique suprême, laisse croire qu'ils se rapprochent d'une conception *directe* de la perception. Ils citent même Studdert-Kennedy [1976] :

*"[...] the phonetic category 'names itself'".*

Cependant, ces auteurs se distinguent explicitement de la théorie de la *Perception Directe* de Gibson [1966]. Selon eux, dans la théorie de Gibson, les événements perçus sont les mouvements effectifs des articulateurs, alors que dans la Théorie Motrice, ce sont les gestes, tels que le locuteur les a planifiés. D'autre part, contrairement à Gibson, ils ne pensent pas que les gestes soient lisibles *directement* dans le signal acoustique, la multiplicité des configurations du conduit vocal associées au même signal acoustique prohibant cette lecture directe. L'inférence des gestes se fait par le biais du module phonétique :

*"[..] the processes of speech perception are, like other linguistic processes, inherently computational and quite indirect. If perception seems nonetheless immediate, it is not because the process is in fact straightforward, but because the module is so well-adapted to its complex-task."*

L'idée de la *Perception Directe* est reprise par Carol Fowler [1986], des mêmes laboratoires *Haskins*, dans sa présentation d'une approche de la perception fondée sur les *événements*. S'appuyant sur un paradigme de la perception en général, elle établit un parallèle entre perception visuelle et auditive qui permet de mieux accepter que l'objet de la perception soit articulatoire et non acoustique. En effet, si pour la vision, l'*événement* distal est la scène visuelle, pour la parole, c'est le conduit vocal articulant. En vision, le *médium* de l'information est le flux lumineux, en parole, c'est le signal acoustique. Ayant acquis la *structure* de l'événement à percevoir, le médium transmet une information sur les propriétés de celui-ci en stimulant les organes sensoriels de l'auditeur ou de la personne qui perçoit. Le médium permet donc la perception *directe* des événements distaux. *L'invariance* est donc à rechercher dans les gestes coordonnés (cf. Fowler [1980]) qui donnent forme au conduit vocal.

L'approche de Fowler présente de nombreuses similitudes avec celle de Liberman et Mattingly. Toutefois, Fowler rappelle que dans le cadre de la Perception Directe, le signal acoustique est totalement transparent aux objets à percevoir, les gestes articulatoires, alors que pour Liberman & Mattingly, les objets de la perception sont les structures de contrôle, récupérées par le biais du module phonétique, et non pas directement les gestes des articulateurs.

L'idée que le signal acoustique fournit directement l'information phonétique, sans traitement perceptif préalable, pourrait susciter un rapprochement avec l'hypothèse de l'Invariance Acoustique. Cependant Fowler [1986] se distingue clairement des tenants de cette théorie. Elle réfute leur hypothèse de départ selon laquelle les segments phonétiques seraient des lots de traits distinctifs (comme diffus + descendant, cf. plus haut) essentiellement statiques. Pour elle les segments phonétiques correspondent à des gestes (dynamiques) articulatoires coordonnés. De plus, alors que, selon Stevens & Blumstein, l'auditeur est censé porter son attention sur les parties les moins coarticulées du signal, Fowler estime que les conséquences acoustiques de la coarticulation sont des sources d'information capitales.

**Le modèle des *Icebergs* de Fujimura : le bien-fondé réaffirmé de l'invariance articulatoire**

Les théories de l'invariance articulatoire proposées par les laboratoires *Haskins*, sont corroborées par l'observation d'invariants du mouvement. Ayant remarqué que les mouvements des articulateurs cruciaux pour la consonne produite (comme la lèvre supérieure pour les labiales) présentent des allures constantes lorsque les aspects temporels sont modifiés par des modulations prosodiques, Fujimura [1987] propose le terme d'*icebergs* pour représenter ces formes articulatoires constantes, solides qui semblent flotter

librement dans le temps par rapport aux évolutions des autres articulateurs. Une définition précise en est donnée dans Fujimura [1986] :

*"The iceberg pattern, a hypothetical elementary gesture, is a relatively invariant articulatory movement pattern (for a given speaker, let us say), into or out of consonantal occlusion, out of or into a specific (perhaps only stressed) syllable nucleus."*

Fujimura [1991] insiste sur l'intérêt que présente la Théorie Motrice pour les chercheurs en parole et, par une analyse rigoureuse de l'organisation temporelle des séquences articulatoires, veut précisément évaluer la justesse de cette théorie. En effet selon lui, si elle se révélait exacte, la Théorie Motrice serait particulièrement puissante pour comprendre la structure intrinsèque de la parole, puisque la représentation motrice discrète serait reliée de façon approximativement linéaire à la représentation phonologique (contrairement à la relation non-linéaire et complexe entre représentation acoustique et phonologique, ne serait-ce qu'en raison de la coarticulation). Alors, par l'analyse des composantes de la représentation motrice, donc des comportements articulatoires, on obtiendrait des informations sur la nature du message linguistique :

*"In my interpretation, the motor theory is an attempt to represent the auditory or some cognitive perceptual patterns of speech in terms of the units in the motor-level representation of the message. The claim is in essence, if I am correct, that the information at this level is representable in such a way that a phonological representation has an approximately linear mapping into this motor-level representation. If this claim can be maintained, it seems at least to me that our study of speech organization can be reduced to components of tractable forms."*

La Théorie Motrice, présentée par Liberman et ses collègues et étayée par Fujimura, a également donné lieu à la Phonologie Articulatoire, proposée par Browman & Goldstein, qui postule que la structure phonologique réside dans l'organisation de gestes articulatoires.

### 1.1.3 Oublier l'Invariance?

> "Nous lisons de deux manières : le mot nouveau ou inconnu est épelé lettre après lettre; mais le mot usuel et familier s'embrasse d'un seul coup d'œil, indépendamment des lettres qui le composent; l'image de ce mot acquiert pour nous une valeur idéographique." *F. de Saussure [1916].*

Hockett dans une conférence récente sur "la vie et la mort du phonème" (Université Stendhal, 1994) transpose ce constat de Saussure sur la lecture, à l'écoute. Selon lui cette conception est *gestaltiste* : nous n'écoutons pas les détails, mais la forme globale, et pourtant nous parvenons à comprendre le message. Ainsi, selon Hockett, la variabilité, à laquelle l'auditeur s'attend, ne fait que préciser ou entamer une forme globale à percevoir. L'invariance physique n'a pas lieu, mais l'auditeur perçoit des formes, que l'auditeur façonne selon les besoins de la communication.

Dans cette idée que la fonction du langage est communicative, Martinet [1970] insiste sur le but premier que nous poursuivons lorsque nous parlons : nous faire comprendre. L'invariant n'est pas physique, il est dans la communication :

*"[...] la fonction fondamentale du langage humain est de permettre à chaque homme de communiquer à ses semblables son expérience personnelle. Par "expérience", il faut entendre tout ce que l'homme ressent ou perçoit, que le stimulus soit interne ou externe, que cette "expérience" prenne la forme d'une certitude, d'un doute, d'un désir ou d'un besoin. La communication à autrui pourra prendre la forme d'une affirmation, d'une question, d'une demande ou d'un ordre, sans cesser d'être communication. Que le langage serve de support à la pensée, nul n'en doute, que nous l'utilisions souvent, moins pour communiquer, que pour nous débonder, comme le faisait le barbier du roi Midas, la chose est claire et les bavards sont là pour nous le rappeler. Mais quel que soit l'emploi que nous fassions du langage, qu'il nous serve à ordonner ou à clarifier notre pensée, ou que nous parlions pour nous exprimer au sens propre du terme, nous nous comporterons toujours comme s'il fallait nous faire comprendre d'autrui."*

Si le but essentiel de la parole est donc bien de faire *percevoir* un message, il existe une exigence qui lui est opposée, celle de minimiser l'effort du *locuteur*. C'est donc plus d'une simple interaction entre le locuteur et l'auditeur qu'il s'agit, c'est d'une *négociation*, d'une tractation.

C'est ce constat qui est à la base des travaux récents de Lindblom sur la variabilité. Lindblom [1988] estime que les phénomènes de *réduction contrôlée* —l'objectif acoustique est plus ou moins atteint pour une même contrainte de diminution de la durée car il semble que le locuteur soit capable de sur- ou sous-articuler— ou les phénomènes de *compensation* du "bite-block" (mâchoire maintenue à une position fixe par une cale)

démontrent les capacités *réorganisationnelles* du système de production de la parole dans le but, explicité par les premiers linguistes, de "se faire comprendre d'autrui" :

*"speech production is a highly versatile process and sometimes appears strongly listener-oriented."*

Les faits que l'espace vocalique des voyelles rétrécisse en parole spontanée (*casual speech* ou *hypospeech*) et au contraire s'étende, pour une diction claire (*clear speech* ou *hyperspeech*), que les contrastes de durée entre voyelles (pour l'anglais par exemple) s'estompent en parole spontanée ou que les contrastes sur le temps de démarrage du voisement entre consonnes voisées et non voisées soient plus prononcés en diction claire, sont des preuves que la variabilité phonétique est de nature *systématique* et se définit selon un axe hyper/hypo.

Le signal de parole ne contient pas toute l'information, il interagit avec des connaissances mémorisées chez l'auditeur et qui sont présupposées de la part du locuteur. Lindblom [1990] explique ainsi comment la parole reste intelligible :

*"Physically ambiguous information is disambiguated and incomplete stimulus information is restored. It appears as if the signal-complementarity processes modulate the input and shape the percept in a most tangible way."*

Il indique en outre que ces processus complémentaires au signal ne sont pas incompatibles avec les théories de Perception Directe (Gibson, Fowler) ainsi que le suggère Shepard [1984] avec l'idée de *résonance*. Le signal de parole "résonne avec" des connaissances internes. On retrouve là les hypothèses des psychologues de la *Gestalt,* évoquées aussi, nous l'avons dit plus haut, par Hockett : l'auditeur reconnaît des formes, qui apparaissent après un traitement cognitif du signal de parole.

Lindblom [1988] poursuit cette idée que l'invariance n'est pas physique et en réponse aux trois questions :

*"Is phonetic invariance articulatory, acoustic or auditory?"*, il propose :

*"[...] we have asked our three questions the wrong way. Invariance cannot be seen as a phonetic problem. It is not a signal analysis problem at all. For the invariance of linguistic categories is ultimately to be defined only at the level of listener comprehension."*

Toute théorie digne d'expliquer l'invariance linguistique doit être capable de rendre compte de ces deux aspects fondamentaux et contradictoires de la communication :

*"[...] speech production is shaped primarily by two forces: plasticity and economy. Plasticity is evident when listener-oriented control is called for. Economy is manifest in reductions and other talker-oriented simplifications."*

Le locuteur adapte son énonciation aux conditions ambiantes :

*"Intra-speaker phonetic variation is genuine and arises as a consequence of the speaker's adaptation to his judgment of the need of the situation."*

Il joue de la variabilité pour optimiser sa production ou pour communiquer une information supralinguistique (niveau du discours, emphase, émotion, etc.). En conclusion, Lindblom plaide pour un abandon de l'invariance phonétique et pour une concentration sur la variabilité adaptative, systématique et contrôlée (nous reviendrons en détail sur cette notion d'adaptation au chapitre III). La variabilité est *assumée*, en ce sens qu'elle sert la négociation locuteur-auditeur et qu'elle est source d'information sur les aspects supra-linguistiques du message.

Si la théorie de la variabilité adaptative de Lindblom est séduisante par bien des aspects, elle n'évoque toutefois aucune hypothèse sur la représentation de la tâche du locuteur dans son espace proximal. Elle se limite ainsi à la perception sans éclairer la production.

## 1.2 Notre démarche : l'évaluation d'un cadre d'hypothèses par une modélisation quantitative

L'exposé du débat sur l'invariance et la variabilité en parole ne permet pas de trancher définitivement sur la question. Il n'existe en effet aucun élément décisif en faveur de l'une ou l'autre des théories.

L'invariance acoustique est certes difficilement défendable, mais l'idée d'une invariance auditive, ne faisant pas intervenir la cognition, est étayée par des tests perceptifs chez les animaux (Kluender [1991], Kuhl & Padden [1983]). Cependant, en termes de contrôle moteur, la caractérisation auditive pose le problème de l'encodage de l'invariant dans l'espace proximal du locuteur. En effet, le temps de réponse de l'information auditive afférente semble trop long pour qu'un contrôle en ligne de l'accomplissement de son action soit envisageable pour le locuteur.

Les théories sur l'invariance motrice ou gestuelle sont séduisantes. Elles offrent, en effet, une transposition immédiate de l'invariant dans l'espace de contrôle moteur du locuteur ; elles permettent de comprendre la variabilité acoustique (Browman & Goldstein [1990], Saltzman & Munhall [1989], cf. paragraphe 2.2.1) ; elles proposent un cadre pour expliquer comment le locuteur classifie et identifie les régularités dans le substrat physique variable (Liberman & Mattingly [1985], Fowler [1980]). Toutefois, elles sont peu convaincantes pour expliquer certains résultats en parole perturbée (Savariaux, Perrier & Orliaguet [1995]) ou pathologique (Morrish [1988]) qui plaident pour une représentation au moins partiellement auditive de la tâche motrice du locuteur.

La théorie de la Variabilité Adaptative est puissante car elle déplace le centre du problème, de la recherche d'une invariance, à l'exploitation optimale de la variabilité ; elle est attrayante car elle permet de comprendre comment la négociation entre exigences

motrices et exigences perceptives peut s'opérer. Mais c'est une théorie résolument perceptive qui ne propose rien de quantitatif sur la façon dont le locuteur transpose dans son propre espace de contrôle cette gestion de la variabilité.

Notre démarche consiste à formuler un certain nombre d'hypothèses relatives à ce problème de l'invariance phonologique *vs* la variabilité physique et à bâtir un modèle quantitatif permettant de les évaluer. Nous justifierons nos choix en détail au chapitre II, mais donnons dès maintenant les axes principaux de notre cadre d'hypothèses :

1. Les mouvements de la parole ne sont pas contrôlés pas à pas mais sont produits vers des cibles.

2. Le phonème est l'unité de base de la commande linguistique.

3. À chaque phonème est associé un objectif perceptif, qui possède une projection bien identifiée dans l'espace moteur du locuteur.

4. Cette projection correspond à un ensemble de configurations articulatoires qui exploite les possibilités de compensation de l'appareil de production de la parole.

5. Parmi toutes les configurations possibles, le locuteur en choisit une seule, spécifique à chaque phonème dans un contexte phonétique donné ; cette configuration peut cependant être modifiée si les conditions de production de la parole empêchent sa réalisation (fumeur de pipe, cale ou *bite-block*, tube labial).

6. Dans un contexte donné, la variabilité est due à des variations des paramètres dynamiques du mouvement et non pas des configurations cibles.

Pour tester ces hypothèses nous proposons de bâtir un modèle de production de la parole allant du contrôle moteur au signal acoustique et exploitant des stratégies de production reliées aux commandes linguistiques : un *robot parlant*. Nous proposons d'inclure des modèles développés et testés à l'Institut de la Communication Parlée. La transformation des positions des divers articulateurs impliqués en une coupe sagittale du conduit vocal est ainsi effectuée par le modèle articulatoire de Maeda ; le passage de cette description à deux dimensions à un "tuyau vocal" (depuis la glotte jusqu'aux lèvres) à trois dimensions est pris en charge par le modèle géométrique de Perrier, Boë & Sock [1992] ; enfin le signal acoustique, émanant de ce tuyau vocal, est obtenu grâce au modèle de Badin & Fant [1984].

Notre tâche est donc la modélisation fonctionnelle de la dynamique des articulateurs en jeu et la spécification des variables de contrôle moteur. Nous avons ainsi à rendre compte du rôle de contrôle du système nerveux central aussi bien que de la dynamique autonome de l'appareil vocal. Il s'agit de tenter de relier, d'une part les intentions du locuteur aux paramètres spatio-temporels du contrôle central et, d'autre part l'aisance et la fluidité avec lesquelles l'articulation se produit à la dynamique naturelle du système musculo-squelettique. L'utilisation de paramètres de contrôle idoines doit nous permettre

de rendre compte des phénomènes de variabilité adaptative et le modèle de la dynamique articulatoire doit incarner les propriétés cinématiques et dynamiques des articulateurs. Il nous faut par conséquent proposer des commandes motrices qui permettent d'ajuster la communication parlée, et tenter de les relier avec les commandes linguistiques (phonémiques et prosodiques).

Dans cette perspective, l'hypothèse du Point d'Équilibre, présentée par Feldman [1966] pour le contrôle des mouvements des membres, est particulièrement attirante. En effet, cette hypothèse permet de différencier l'objectif à atteindre par les articulateurs —le Point d'Équilibre, que nous relierons au phonème— et la manière d'atteindre cet objectif —que nous associerons à l'ajustement prosodique—. Nous développerons ces points en détail au chapitre II.

# CHAPITRE II

# Quel cadre pour le contrôle moteur?

A noir, E blanc, I rouge, U vert, O bleu: voyelles,
Je dirai quelque jour vos naissances latentes :
A, noir corset velu des mouches éclatantes
Qui bombinent autour des puanteurs cruelles,

Golfes d'ombre; E, candeurs des vapeurs et des tentes,
Lance des glaciers fiers, rois blancs, frissons d'ombelles ;
I, pourpres, sang craché, rire des lèvres belles
Dans la colère ou les ivresses pénitentes ;

U, cycles, vibrements divins des mers virides,
Paix des pâtis semés d'animaux, paix des rides
Que l'alchimie imprime aux grands fronts studieux ;

O, suprême Clairon plein des strideurs étranges,
Silences traversés des Mondes et des Anges :
— O l'Oméga, rayon violet de Ses Yeux !

Arthur Rimbaud

## 2.1 De l'existence de cibles vocaliques

La notion de cible émane de cette volonté originelle de transposer les représentations phonologiques idéales dans les espaces physiques qui constituent le canal de transmission du message linguistique. Les cibles sont les représentants spatio-temporels physiques des phonèmes abstraits. L'appareil de production de la parole viserait des configurations (articulatoires, formantiques ou autres) canoniques, ou plus simplement des cibles. L'auditeur pourrait, selon certains auteurs, récupérer l'objectif visé et non pas simplement la configuration effectivement atteinte. La nature exacte des cibles diffère selon les auteurs (articulatoire, acoustique, auditive, etc.), nous aborderons cette question au paragraphe 2.1.1. Leur contenu informatif est également discuté dans la littérature.

Lindblom [1963] donne une définition explicite du concept de "cible", en termes acoustiques et articulatoires :

*"[...] the term target can be given an explicit definition in terms of the asymptotic values of the first-formant frequencies of a given vowel. A target was found to be independent of consonantal context and duration and can thus be looked upon as an invariant attribute of the vowel. [...] Since the speed of articulatory movements is [...] limited [owing to intrinsic physiological constraints], the extent to which articulators reach their target positions depends on the relative timing of the excitation signals. As a vowel becomes shorter, there is less and less time for the articulators to complete their "on-" and "off-glide" movements within the CVC syllable. [...] the speech organ fail, as a result of the physiological limitations, to reach the positions that they assume when the vowel is pronounced under ideal steady-state conditions. In the acoustic domain, this is paralleled by undershoot in the formant frequencies relative to the bull's-eye formant pattern."*

Öhman [1967] s'appuie sur la notion de cible articulatoire pour élaborer un modèle numérique de la coarticulation, selon lequel la forme globale du conduit vocal s'obtient à partir d'une combinaison linéaire de formes cibles et de fonctions coarticulatoires. Les résultats du modèle reproduisent convenablement les données cinéradiographiques sur la forme du conduit vocal.

Selon MacNeilage [1970], l'hypothèse que la parole est spécification de cibles en termes de configuration du conduit vocal est souvent reprise dans la littérature (Stevens & House [1963], Halle & Stevens [1964], Ladefoged [1967]). Cependant, le terme de "cible" évoque simplement la tendance qu'ont les articulateurs à approcher une position déterminée pour un phonème donné, dans diverses conditions. Aucune hypothèse sur la réalité neurologique d'un mécanisme de contrôle de la parole par cible n'est encore envisagée à l'époque. MacNeilage propose alors un modèle de l'*Ordonnancement Sériel*

(ou plus simplement "séquencement"), s'appuyant explicitement sur la notion de cible. Son hypothèse est ainsi formulée :

*"[...] there are a number of lines of evidence which suggest that speech production is controlled, in part, by the specification of targets in an internalized space coordinate system."*

Cette notion de représentation spatiale interne est issue d'hypothèses sur le comportement visuo-moteur, selon lesquelles le pointage de cible, sans aide visuelle, nécessite la capacité de se représenter les cibles en termes de coordonnées spatiales internes. Selon le modèle proposé par MacNeilage, l'information phonologique, sur la séquence de parole à produire, est fournie au système de coordonnées spatiales qui traduit cette information en une série de cibles spatiales à atteindre. Une série de requêtes est alors adressée au mécanisme de contrôle du système moteur qui génère des patrons de commandes motrices, appliqués aux muscles, permettant aux articulateurs d'atteindre les cibles spécifiées, dans l'ordre requis. MacNeilage discute de la nature des cibles, qui sont généralement supposées être liées aux positions effectivement atteintes par les articulateurs, dans des conditions de production idéales.

MacNeilage revient sur ces notions en 1980 et rectifie ses premières hypothèses sur la nature spatiale des cibles. Le fait que plusieurs configurations articulatoires puissent correspondre au même segment phonétique est contradictoire avec la notion de cible spatiale. D'autre part, il lui semble que tous les objectifs articulatoires ne sont pas spécifiés en termes de *positions*, certaines diphtongues étant plutôt définies par une *vitesse* spécifique du mouvement articulatoire. Il en vient alors à la notion de cible *auditive* :

*"Auditory properties of speech production are obviously primary in the sense that the auditory information provided by our language community is by far the main source of goals for our acquisition of speech production."*

Sur le plan de la production des voyelles, ce concept de cible a permis l'élaboration de nombreux modèles intéressants. Les Points Attracteurs de la Dynamique de Tâche des laboratoires *Haskins*, les Points de Passage (*Via Points*) des laboratoires *ATR*, sont quelques-uns des succès de l'approche par cible ; nous les présenterons en détail au paragraphe 2.2.

L'idée de cibles paraît donc séduisante et peut expliquer les comportements observés chez des locuteurs. Cependant se pose la question de la relation entre ces éventuelles cibles et la commande phonologique. Stevens & House [1963], puis Mac Neilage [1970] ou Houde [1968], ont en effet relevé la grande variabilité des caractéristiques physiques (acoustiques ou articulatoires) mesurées pour un même phonème dans des conditions de production variables. La parole étant produite pour être perçue, il nous semble qu'une bonne démarche pour étudier l'existence de cibles en parole, soit d'examiner leur pertinence perceptive. En effet, si les cibles existent et si elles constituent une représentation

remarquable de la relation entre la chaîne phonologique et les signaux physiques, cette représentation doit être partagée par le locuteur et l'auditeur. Dans le cas contraire, à quoi bon servirait, pour le locuteur, de s'évertuer à produire une information qui ne serait pas pertinente pour l'auditeur? Cette démarche, exploitant l'interaction perception-action, est classique dans la littérature.

Le débat s'est engagé dans les années 1970, à la suite d'une étude de Strange, Verbrugge, Shankweiler & Edman [1976] indiquant que les voyelles en contexte sont mieux identifiées que les voyelles isolées. La notion de cible a été alors fortement ébranlée. De nombreux travaux sur l'identification des voyelles ont suivi, apportant de l'eau aux moulins des deux camps.

## 2.1.1 Récupération perceptive des cibles

Pour rendre compte des variabilités inter- et intra-locuteur, les tenants de l'hypothèse de l'identification par cible ont recours à la notion de transformation qui projette les sons dans un espace interprétable par l'auditeur.

### 2.1.1.1 Le cas mono-locuteur

Lindblom [1963] propose un premier modèle décrivant la variabilité intra-locuteur. Nous reviendrons en détail sur ce modèle et sur ceux qu'il a engendrés, dans le chapitre sur la réduction vocalique (3.2), aussi nous bornons-nous ici à présenter succinctement l'hypothèse de Lindblom [1963] : les articulateurs atteignent plus ou moins leurs positions cibles canoniques, selon le temps qui leur est imparti. Les fréquences formantiques correspondantes suivent le même principe.

En complément à ce premier modèle de production décrivant le phénomène d'*undershoot* des cibles (ratage des cibles par défaut), Lindblom & Studdert-Kennedy [1967] proposent un modèle perceptif selon lequel la cible spectrale prévue pour la voyelle peut être récupérée par extrapolation, en tenant compte de la durée. Le modèle d'*undershoot* en production a désormais son pendant en perception : le modèle d'*overshoot* (récupération par excès, extrapolation) perceptif.

Kuwabara [1985] présente une méthode de normalisation des effets de la coarticulation s'inspirant de l'approche de Lindblom & Studdert-Kennedy [1967]. Partant de résultats perceptifs qui indiquent que les voyelles sont mieux identifiées lorsqu'elles sont présentées au sein du contexte dans lequel elles ont été prononcées que lorsque l'on tronque le contexte, Kuwabara propose une procédure de modification des caractéristiques acoustiques des voyelles. Les deux premiers formants F1 et F2 sont modifiés en appliquant

une formule de convolution tenant compte des informations spectrales passées et futures. L'ambiguïté sur les voyelles est alors réduite : dans le plan des deux premiers formants transformés, les groupes de voyelle sont bien mieux séparés que dans le plan traditionnel F1/F2.

Une extension de la procédure de normalisation de Kuwabara est proposée par Akagi [1993] qui présente un modèle des effets de la coarticulation sur la qualité des voyelles. Le modèle est capable de prédire des positions cibles spectrales, correspondant aux positions observées pour les voyelles isolées, à partir de trajectoires spectrales réduites, en utilisant des interactions entre paires de pics spectraux.

Dans la lignée du premier modèle de Lindblom [1963], qui suggère que des formules de correction peuvent permettre de compenser les effets de contexte et de récupérer la qualité phonétique de la voyelle, Broad & Clermont [1987] élaborent différents modèles mathématiques de contours formantiques rendant compte des effets systématiques du contexte consonantique sur la voyelle centrale dans des séquences CVC.

Nearey [1989] développe une expérience de perception s'inspirant directement du paradigme d'*overshoot* perceptif de Lindblom & Studdert-Kennedy [1967]. Des stimuli synthétiques, correspondant à des séquences du type /bVb/, /dVd/ ou /V/, sont présentés aux auditeurs, de façon séparée (un seul type de séquence) ou mixte (les trois types de séquences sont présentés aléatoirement). Pour la partie vocalique des stimuli, le formant F2 varie continûment de 900 à 1800 Hz, de sorte que la voyelle centrale parcourt trois catégories phonétiques (/ɒ/, /ʌ/, /ɛ/). Les *loci* initiaux des consonnes sont caractéristiques de /b/ et /d/, F2 valant 2000Hz pour /d/ et 700Hz pour /b/. Pour la partie transitoire une interpolation polynomiale d'ordre 6 est utilisée entre *loci* consonantiques et cibles vocaliques. Les résultats perceptifs indiquent que les frontières d'identification des voyelles présentées en contexte sont décalées, par rapport aux frontières des voyelles présentées en isolation, dans la direction prévue par le modèle d'*overshoot*, *i.e.* dans la direction du *locus* consonantique (frontières plus basses en fréquence pour /bVb/ et plus élevées pour /dVd/). Cependant les effets d'*overshoot* observés sont plus faibles que ceux reportés par Lindblom & Studdert-Kennedy [1967] et l'auteur conclut que la compensation n'est peut-être que partielle :

*"We must bear in mind the possibility of* partial perceptual compensation *combined with* partial loss of contrast *in studying this problem."*

L'idée sous-jacente est que le système perceptif n'est pas contraint à retrouver absolument une cible pour assurer la compréhension, car il existe dans le message linguistique des éléments complémentaires à l'information phonémique. Ainsi une partie du contraste phonétique peut être endommagée sans mettre en péril la communication.

Beautemps [1993] propose de rechercher, à partir des trajectoires formantiques, des paramètres permettant l'identification des voyelles [a] et [ɛ] en contexte [iVi], dans diverses conditions de débit et d'accentuation. Il étudie trois modèles d'analyse des trajectoires formantiques dans le but de récupérer la cible vocalique perdue. Le premier exploite des paramètres globaux calculés sur la trajectoire complète (vitesse de la transition, corrélée avec la valeur formantique maximale, et durée relative du plateau de la voyelle inconnue) et permet efficacement de discriminer les données, *i.e.* de maintenir l'opposition entre [a] et [ɛ]. Le deuxième s'appuie sur des paramètres locaux nécessitant la connaissance d'un nombre restreint (environ 4) de points consécutifs sur la trajectoire (plutôt centrés après le pic de vitesse). Le troisième implique une identification de systèmes dynamiques et met en œuvre une véritable *inversion* perceptive (cf. paragraphe 3.4 pour plus de détails sur l'inversion). Selon Beautemps, il est ainsi possible d'extraire du signal acoustique des paramètres discriminants pour l'identification de voyelles.

Dans le même ordre d'idée Pitermann [1996], tente d'extraire une cible formantique invariante à partir de trajectoires formantiques du type de celles étudiées par Beautemps, dans un grand nombre de conditions de débit et d'accentuation. Son approche, fondée sur l'inversion de modèles dynamiques, ne permet de réduire qu'une faible partie de la variabilité des valeurs quasi stationnaires des formants. Ses résultats semblent par conséquent peu compatibles avec l'hypothèse de cibles formantiques invariantes quels que soient le débit et l'accentuation. Des *régions* formantiques cibles seraient plus adaptées.

### 2.1.1.2 Le cas multi-locuteur

Selon Nearey [1989], parmi les différentes sources de variabilité spectrale, pour une voyelle donnée, la nature du locuteur (masculin *vs* féminin, enfant *vs* adulte) est la plus importante. Il convient donc de considérer maintenant quelques méthodes de réduction ou d'interprétation de la variabilité inter-locuteur.

Nearey [1989] présente une revue des théories sur la normalisation inter-locuteur et distingue, selon la terminologie de Ainsworth [1975], les théories *intrinsèques* et *extrinsèques*. Les théories extrêmes de normalisation intrinsèque suggèrent que lorsqu'une représentation paramétrique adéquate des propriétés spectrales de la voyelle est employée, utilisant certaines transformations sur le fondamental et les formants, il est alors possible de distinguer les voyelles entre elles (Miller [1953], Peterson [1961], Miller [1984,1989], Syrdal [1984]). Parallèlement, les théories extrêmes de normalisation extrinsèque supposent que l'auditeur rétablit un cadre de classification des voyelles à partir d'informations disséminées sur l'ensemble des voyelles d'un locuteur (Joos [1948], Ladefoged & Broadbent [1957], Gerstman [1968], Nordström & Lindblom [1975], Nearey [1978]). Il

**Théories extrinsèques**

Joos [1948] postule que la qualité phonétique d'une voyelle dépend de la relation entre les valeurs formantiques de cette voyelle et celles des autres voyelles prononcées par le même locuteur. L'auditeur ne porte pas son attention uniquement sur les valeurs des fréquences formantiques d'une voyelle donnée mais sur les relations entre ces fréquences et la gamme de fréquence qui semble être caractéristique du locuteur.

Ladefoged & Broadbent [1957] confirment l'hypothèse de Joos par une expérience au cours de laquelle des auditeurs, à qui l'on présente un mot à identifier, sont nettement influencés par la gamme de fréquences formantiques de la phrase qui précède ce mot. Selon ces auteurs, il semble donc que :

*"[...] the linguistic information conveyed by a given vowel is largely dependent on the relations between the frequencies of its formants and the frequencies of the formants of other vowels occurring in the same auditory context."*

Une évaluation quantitative de diverses procédures de normalisation extrinsèque (Gerstman [1968], Harshman [1970], Lobanov [1971], Nearey [1977]) est proposée par Disner [1980]. Elle évalue, pour chaque procédure, le degré de suppression de la variance des données, obtenues pour six langues germaniques. Elle montre cependant que les procédures les plus efficaces (celles de Nearey et de Lobanov) dans la suppression de la variance introduisent des artefacts et altèrent la qualité relative des voyelles d'une langue à l'autre : les voyelles normalisées d'une langue donnée ne sont pas comparables avec celles d'une autre langue.

Nearey [1989] remarque que les théories extrinsèques posent un problème de référence : si chaque voyelle est définie par rapport à toutes les autres, comment peut-on espérer l'identifier? Toutefois il admet que ce problème est en partie résolu si l'on considère que certaines voyelles d'amorçage, ou certaines informations acoustiques, permettent de calibrer un système vocalique (voir par exemple Perrier, Apostol & Payan [1995], pour une normalisation exploitant des critères articulatoires, extraits des voyelles extrêmes [i], [a], [y] et [u]). Mais il rappelle en outre que Assmann, Nearey & Hogan [1982] obtiennent des taux d'identification élevés pour des voyelles prononcées par différents locuteurs et présentées aux auditeurs dans un ordre aléatoire, sans phrase de calibration. Leurs résultats perceptifs sur des voyelles tronquées (pour lesquelles les informations fournies par la durée et les transitions formantiques sont supprimées) indiquent que le recouvrement spectral entre les différentes catégories de voyelle n'est pas si important que le suggère par exemple Joos [1948]. Les théories de pure normalisation extrinsèque ne permettent pas d'expliquer ces résultats.

**Théories mixtes**

Selon Nearey [1989], l'abondance des questions irrésolues dans la littérature sur les deux types de normalisation (intrinsèque et extrinsèque) indique que des positions intermédiaires doivent être recherchées.

Un compromis sur la question de la normalisation est proposé par Ryalls & Liebermann [1982]. Ils montrent que l'identification de voyelles synthétiques est influencée par des modifications du fondamental. De meilleures identifications sont obtenues pour des fréquences plus basses du fondamental. Les auteurs interprètent ces résultats en termes de spectres de la fonction de transfert du conduit vocal : un spectre plus dense (correspondant à un F0 plus bas) facilite l'extraction des formants. Ils indiquent en outre que le fondamental joue un rôle secondaire dans la normalisation. Selon ces auteurs, la normalisation s'appuie essentiellement sur les valeurs formantiques intrinsèques. Ils notent toutefois, que certaines voyelles (/i/, /ɑ/ et /u/) sont généralement mieux identifiées que d'autres et servent probablement de références acoustiques extrinsèques.

Le point de vue de Holmes [1986] est également assez nuancé. Il présente des expériences de synthèse, mettant en œuvre des décalages formantiques systématiques, et conclut de ses résultats perceptifs que l'hypothèse d'un simple décalage formantique est insuffisante mais qu'une procédure de normalisation complexe, et probablement adaptée à chaque voyelle, permettrait de récupérer la qualité phonétique des voyelles isolées.

S'inspirant des travaux de Ainsworth [1975], Nearey [1989] propose une expérience permettant d'évaluer le poids des différents facteurs intrinsèques et extrinsèques dans la perception des voyelles. Les facteurs intrinsèques manipulés sont le fondamental et les formants F3 et F4. Le facteur extrinsèque consiste en un ensemble de valeurs des deux premiers formants (F1 et F2) pour un groupe de voyelles d'amorçage. Les trois différents facteurs peuvent prendre deux types de valeurs synthétiques, celles d'un adulte masculin moyen ou celles d'un enfant, ce qui donne huit conditions expérimentales possibles. Les résultats statistiques sur l'identification des voyelles indiquent que les facteurs extrinsèques sont dominants, mais qu'il est impossible de négliger les facteurs intrinsèques. En effet, le fondamental F0 a un effet considérable sur F1 et moindre sur F2. Les formants F3 et F4 n'affectent que très légèrement F2. Nearey en conclut que les deux types d'information ont leur rôle à jouer dans la perception vocalique :

*"We clearly need models that are sensitive to both intrinsic and extrinsic effects of speaker variation."*.

#### 2.1.1.3 En résumé

Les idiosyncrasies des locuteurs et la coarticulation posent des questions troublantes sur la nature et l'existence des cibles. Comment prétendre que la production de voyelle est mouvement vers des cibles, si ce qui semble être une cible varie considérablement avec le locuteur, le contexte phonétique ou la condition prosodique?

Une première hypothèse consiste à rappeler que la cible n'est qu'un objectif visé et qu'elle est rarement atteinte en condition normale de parole. Les variations observées ne sont donc que les conséquences de mouvements, plus ou moins achevés, vers des cibles. Si la cible n'est pas atteinte, il faut expliquer comment l'information linguistique est malgré tout récupérée. Différents modèles complémentaires de production et de perception sont envisagés dans la littérature. Dans la partie 2.2, nous en présentons deux, parmi les plus importants, mettant en œuvre diverses hypothèses sur la nature des cibles et la distribution de l'information. Il apparaît ainsi que la notion de cible abstraite est tout-à-fait compatible avec les effets concrets de variabilité.

Toutefois, il est des théories qui contestent clairement ce concept de cible et selon lesquelles l'information utile à l'auditeur, et véhiculée par le locuteur, est purement dynamique. Considérons donc maintenant ce point de vue divergent et les questions qu'il suscite, avant d'exposer notre hypothèse.

### 2.1.2 Pertinence perceptive des transitions

#### 2.1.2.1 De la supériorité des voyelles en contexte sur les voyelles isolées

**L'expérience de Strange *et al.* [1976]**

La notion de cible est gravement remise en question par les travaux des chercheurs de l'Université du Minnesota (Strange, Verbrugge, Shankweiler & Edman [1976]). Partant des expériences de Fairbranks & Grubb [1961], qui fournissent des pourcentages d'erreur relativement élevés dans des tâches d'identification de voyelles isolées, et de l'étude comparative de Fujimura & Ochiai [1963], qui indique que les voyelles privées de leur contexte sont moins intelligibles que les voyelles en contexte, Strange *et al.* s'interrogent sur le bien-fondé des descriptions acoustiques classiques en termes de cibles spectrales :

*"Could it be that the acoustic complexities introduced by syllabic structure better serve the requirements of the perceptual apparatus than do quasi-steady-state formants? If so, then it is surely inappropriate to characterize the cues for vowel identity in terms of static points in a space defined by the first two-formants."*

Pour répondre à cette question Strange *et al.* proposent deux expériences d'identification de voyelle mettant en jeu des voyelles prononcées en isolation et des voyelles en contexte consonantique, dans des conditions mono- ou multi-locuteur. Dans la première expérience, le contexte est fixé (/pVp/) et l'identité de la consonne est connue par les auditeurs. Les erreurs d'identification sont trois fois supérieures pour les voyelles prononcées en isolation que pour les voyelles en contexte. Les voyelles sont mieux identifiées lorsqu'elles ont été produites par un seul locuteur que lorsque les productions des 15 locuteurs sont mélangées aléatoirement. Cependant, l'augmentation du pourcentage d'erreur dû à l'absence de contexte est du même ordre en conditions mono- et multi-locuteur. Il n'est donc pas possible de conclure que le contexte consonantique ne sert qu'à fournir des indices de normalisation inter-locuteur. Les erreurs d'identification de voyelles isolées ont, selon Strange *et al.*, une explication plus fondamentale : les transitions, vers ou depuis les consonnes, fournissent une information décisive sur la voyelle :

*"[...] the experiment provides no evidence that coarticulated consonants facilitate identification by enabling the listener to recalibrate for each new talker. Coarticulated consonants are integral to the specification of vowels whether a talker is familiar or not."*

Au cours d'une deuxième expérience, la nature de la consonne varie aléatoirement, mais les scores d'identification restent du même ordre. Il est donc impossible d'arguer que la connaissance de l'identité de la consonne est le facteur qui améliore l'identification des voyelles en contexte, plutôt que la transition formantique. La conclusion de Strange *et al.* ébranle fortement l'hypothèse du caractère essentiel de la cible :

*"[...] cues that are ordinarily regarded as consonantal contribute regularly to the perception of the vowel. [...] no single, temporal cross section of a syllable conveys as much vowel information to a perceiver as is given in the dynamic contour of the formants.".*

Un certain nombre d'études complètent ces premiers travaux et vont dans le sens de l'hypothèse que l'information qui permet d'identifier les voyelles est dynamique (Shankweiler, Verbrugge & Studdert-Kennedy [1978], Verbrugge & Isenberg [1978], Gottfried & Strange [1978]).

**La théorie de l'*Overshoot* Perceptif**

Des études menées dans d'autres laboratoires viennent confirmer la primauté des aspects dynamiques en perception de voyelles. On note que la théorie de l'*overshoot* perceptif de Lindblom & Studdert-Kennedy [1967], présentée au 2.1.1.1, et les techniques qui s'en sont inspirées (Kuwabara [1985]), peuvent s'inscrire dans le cadre de la perception dynamique, puisque c'est la dynamique de la transition qui permet de récupérer l'information sur la voyelle. On peut classer dans la même veine, les travaux de Huang [1992], Di Benedetto [1989a, 1989b], Akagi [1990, 1993]. Ainsi, les travaux de Strange ne

semblent pas une réelle remise en cause de la théorie des cibles, puisqu'ils peuvent être expliqués dans ce cadre, à la fois par la notion de perception systémique (normalisation extrinsèque) et par celle d'*overshoot* perceptif.

Ce n'est pas le point de vue de Van Son [1993] qui remet en question la théorie de l'*Overshoot* perceptif. Il se fonde pour cela sur l'expérience de perception, décrite par Pols & Van Son [1993]. Reprenant le paradigme de Lindblom & Studdert-Kennedy [1967], Pols & Van Son utilisent des stimuli synthétiques de voyelles isolées et en contexte, avec des formants F1 et F2 stylisés, *i.e.* paraboliques ou à niveau constant, les sommets des paraboles correspondant aux niveaux constants. Pour une voyelle donnée, les identifications diffèrent selon que le stimulus est parabolique ou à niveau constant. Ainsi par exemple, si les auditeurs perçoivent /i/ lorsque F1 est constant, ils perçoivent /y/ lorsque F1 est parabolique, c'est-à-dire qu'ils perçoivent une voyelle de F1 *inférieur* (selon l'échelle des voyelles qu'ils proposent, le long de l'axe F1) lorsque le stimulus est parabolique. Le même schéma est observé pour le formant F2. La courbure de la trajectoire formantique induirait donc un *undershoot* perceptif plutôt qu'un *overshoot* [1]. Les auteurs concluent qu'aucun phénomène de compensation de l'*undershoot* en production (par un *overshoot* perceptif) ne se manifeste pour les stimuli à formants paraboliques, que la voyelle soit isolée ou en contexte consonantique. Il semble donc que la courbure des trajectoires aggrave les effets de la coarticulation, plutôt qu'elle ne les compense.

**Des limites expérimentales**

Cependant Van Son lui-même [1993] pondère ses conclusions en reconnaissant qu'il existe probablement un mécanisme permettant de compenser l'*undershoot* de cible, en production, dû à la coarticulation. Ce mécanisme, qu'il nomme *coarticulation inverse*, ne fonctionne pas uniquement sur l'allure spectro-temporelle de la voyelle, mais sur l'ensemble de la syllabe, voire au-delà. Il en conclut que la principale raison, pour laquelle Pols & Van Son [1993] ne parviennent pas à répliquer le phénomène d'*overshoot* perceptif, est que les stimuli synthétiques utilisés ne contiennent pas suffisamment d'information sur le contexte consonantique.

Par ailleurs Macchi [1980] répond précisément à l'article de Strange *et al.* [1976] par une expérience similaire d'identification de voyelles isolées et en contexte consonantique. Les conditions expérimentales sont améliorées par rapport à la première expérience de

[1] Cependant nous remarquons que dans l'expérience de Pols & Van Son [1993], un seul formant suit une courbe parabolique, alors que tous les autres sont fixés. Dans l'expérience originale de Lindblom & Studdert-Kennedy [1967], les trois premiers formants suivent tous une trajectoire parabolique. Il se pourrait donc aussi que les auditeurs soient troublés par ces stimuli très peu naturels et perdent leurs repères formantiques habituels. D'ailleurs, Pols & Van Son indiquent que dans le cas où les formants F1 et F2 varient tous les deux, le phénomène d'*undershoot* perceptif est moins régulier pour F1 et absent pour F2...

Strange *et al.* Les tests d'écoute sont de meilleure qualité audio, les locuteurs et les auditeurs sont de même dialecte régional, les artefacts dus à des problèmes d'orthographe sont supprimés par la mise en œuvre d'une tâche consistant à faire rimer la voyelle entendue avec un mot simple, cette tâche étant utilisée dans les deux tests (voyelles isolées et en contexte), contrairement à la première étude utilisant des tâches différentes dans les deux cas. Les pourcentages d'erreur d'identification sont alors globalement beaucoup plus faibles que dans l'expérience de Strange *et al.* (de l'ordre de 2%, au lieu de 31%, pour les *voyelles isolées* prononcées par un même locuteur et 8%, au lieu de 43%, pour les *voyelles isolées* produites par différents locuteurs). D'autre part, il n'est plus possible de distinguer l'identification des voyelles *isolées* de celles des voyelles en *contexte*, alors que l'étude de Strange *et al.* [1976] donnait des pourcentages d'erreurs très différents (dans le cas mono-locuteur, 10% pour les voyelles en contexte *vs* 31% en isolé, et dans le cas multi-locuteur, 17% en contexte *vs* 43% en isolé). L'auteur conclut que :

*"[...] consonantal coarticulation is not a necessary condition for accurate identification of naturally produced vowels."*

Dans une étude sur l'identification des voyelles isolées, évoquée au 2.1.1.2, Assmann, Nearey & Hogan [1982] confirment que les voyelles peuvent être remarquablement bien identifiées, même en l'absence de contexte, et que des difficultés orthographiques dans le choix des réponses à cocher sont une source d'augmentation des erreurs des auditeurs. En outre, les voyelles tronquées, pour lesquelles les informations sur la durée et les transitions formantiques sont supprimées, sont bien identifiées, ce qui suggère que les chevauchements spectraux des différentes catégories de voyelles ne sont pas si larges. Cependant les auteurs notent que les voyelles entières sont mieux identifiées que les voyelles tronquées, et soulignent que les informations dynamiques réduisent le chevauchement spectral des voyelles.

Dans le but d'évaluer dans quelle mesure les voyelles peuvent être considérées comme dynamiques, Harrington & Cassidy [1994] mettent en œuvre des classificateurs automatiques, à base de réseaux de neurones artificiels, pour analyser les informations utiles à la classification de voyelles en contexte /CVd/. Selon eux, si les indices sur l'identité de la voyelle sont dynamiques, alors des classifications à partir de différentes tranches de signal, prises dans les transitions consonantiques et au milieu du noyau vocalique, devraient donner de meilleurs résultats que des classifications utilisant une seule tranche au milieu de la voyelle. Au contraire, dans une théorie à cibles, les voyelles monophtongues devraient être suffisamment déterminées par l'information au milieu de la voyelle mais les diphtongues, supposées comporter deux cibles, devraient bénéficier de l'apport de tranches de signal supplémentaires. Les résultats des classifications automatiques indiquent que les monophtongues de l'anglais australien, sont identifiées avec la même précision (de l'ordre

de 90%), que l'information consiste en une ou plusieurs tranches de signal. Les diphtongues par contre sont mieux identifiées lorsque des tranches supplémentaires sont utilisées. D'autre part, à l'aide d'un classificateur utilisant un réseau de neurones récurrent et, par là même, sensible à l'ordre temporel des tranches spectrales, Harrington & Cassidy suggèrent que le séquencement temporel des trois tranches n'apporte aucune information supplémentaire sur l'identité des voyelles.

L'ensemble de ces résultats semblent donc conforter les théories à cibles au détriment des théories à transitions.

### 2.1.2.2 Le paradigme des stimuli à centres silencieux

En réponse aux critiques sur les premiers tests d'identification sur voyelles isolées et en contexte, les chercheurs de l'Université du Minnesota introduisent un nouveau paradigme, dit des "centres silencieux", qui semble cette fois sonner le glas du concept de cibles (Strange, Jenkins & Edman [1978], Strange, Jenkins & Johnson [1983]). Strange *et al.* [1983] notent que les expériences menées précédemment dans leur laboratoire ont le défaut de ne pas séparer les effets dus à la production, des effets dus à la perception, puisque l'on compare les scores d'identification de voyelles qui sont issues de productions différentes. Les auteurs présentent donc une nouvelle expérience d'identification de voyelles dans laquelle des syllabes CVC, dont on a supprimé ou modifié certains paramètres spectraux et temporels, sont présentées aux auditeurs. Les stimuli varient dans le type d'information qu'ils véhiculent. Les stimuli "à centres silencieux", où le noyau vocalique est remplacé par du silence de durée égale à celle de la voyelle, contiennent l'information dynamique spectrale de l'ensemble du geste vocalique ainsi que l'information temporelle sur la durée de la séquence. Les stimuli "à transitions tronquées", où le noyau vocalique est conservé intégralement mais où les transitions sont supprimées, contiennent l'information sur la cible vocalique et sur la durée de la voyelle. Les stimuli "à transitions tronquées et noyau raccourci", où le noyau vocalique n'est que partiellement conservé, contiennent l'information sur la cible et une information temporelle dégradée, et les stimuli "à centres silencieux raccourcis" ou "allongés", pour lesquels la durée du silence est inférieure ou supérieure à celle de la voyelle, contiennent l'information dynamique mais pas d'information sur la cible ni sur la durée intrinsèque de la voyelle. Enfin les stimuli "à partie initiale", composés uniquement de la transitions CV, et "à partie finale" (transition VC) ne contiennent que l'information dynamique sur le début ou la fin du geste. Les scores d'identification, en condition mono-locuteur, indiquent que les voyelles dans des stimuli à centres silencieux sont identifiées avec une bonne précision (6% d'erreur), ce qui confirme l'hypothèse que l'information dynamique suffit pour identifier des voyelles

coarticulées. Les stimuli à partie initiale ou finale donnent les scores les plus faibles (plus de 45% d'erreur), ce qui semble indiquer que l'information dynamique est distribuée sur l'ensemble de la voyelle. Les stimuli à centres silencieux raccourcis ou allongés sont identifiés avec un précision de l'ordre de celle des stimuli à centres silencieux intacts. Il semble donc que le spectre acoustique au début et à la fin de la syllabe CVC fournit une information considérable qui est bien plus qu'une simple information de durée. D'autre part, les stimuli à transitions tronquées sont bien identifiés (8% d'erreur) mais les scores s'affaiblissent pour les stimuli à transitions tronquées et noyau raccourci (21%). Les mouvements formantiques au sein du noyau vocalique sont donc probablement une source d'information dynamique non négligeable sur l'identité vocalique. Les auteurs concluent de ces résultats que l'information dynamique véhiculée par les parties transitoires et l'information temporelle sur la durée intrinsèque de la voyelle, contribuent ensemble à l'identification des voyelles en contexte CVC. Par contre, les cibles spectrales statiques semblent fournir une information limitée. Les résultats des tests perceptifs en condition multi-locuteur sont similaires, les pourcentages d'erreurs étant globalement plus élevés qu'en condition mono-locuteur. Les auteurs formulent la conclusion suivante :

*"[...] vowels, as gestures, are differentiated by their timing with respect to adjacent segments and syllables, as well as by the positioning of the tongue during the relatively sustained vocalic portion of the syllable. The perceiver must identify the intended vowels on the basis of information in the acoustic pattern about the timing of the gesture as well as the vocal tract state attained. [...] perceivers can utilize [...] abstract acoustic parameters in identifying vowels even when static vowel targets are completely missing from the signal."*

Cependant, si ces résultats confirment l'importance des aspects dynamiques dans la spécification des voyelles, ils ne contredisent pas la notion de cibles. En effet Verbrugge et Rakerd [1986] concèdent que les scores d'identification élevés obtenus par Strange *et al.* [1983] pour les stimuli à centres silencieux peuvent être expliqués de deux façons. On peut en effet concevoir que les auditeurs utilisent l'information dynamique contenue dans ces stimuli pour extrapoler les trajectoires formantiques et récupérer la cible excisée. Ou bien, on peut considérer à l'opposé que les voyelles sont des événements articulatoires, des gestes, et que l'information véhiculée par les régions dynamiques est complémentaire et radicalement différente de l'information de cible. Pour tester ces deux hypothèses divergentes, Verbrugge & Rakerd proposent un test d'identification sur des stimuli à centres silencieux hybrides, où les portions initiale et finale correspondent à des syllabes produites respectivement par un locuteur et une locutrice. Selon les auteurs, dans un cadre théorique d'extrapolation de cible, les stimuli hybrides devraient perturber les auditeurs, puisque, les locuteurs masculin et féminin ayant des conduits vocaux de taille et de forme différentes, les portions syllabiques tendent vers des cibles distinctes. Au contraire, dans un

cadre de perception d'événement, les stimuli hybrides devraient être acceptables perceptivement, puisque des locuteurs parlant un même dialecte sont censés produire des voyelles avec les mêmes schémas articulatoires et acoustiques. Or les résultats expérimentaux indiquent que les stimuli hybrides sont identifiés avec un score équivalent à celui des stimuli à centres silencieux originaux. Par conséquent, les auteurs affirment que :

*"[...] the vowel information in dynamic regions of a syllable is largely invariant across talkers. It is highly unlikely that this dynamic information subserves the perceptual extraction of any sort of acoustic target, since targets are highly variant across talkers. It is much more likely that the information is indicative of a characteristic articulatory style that is common to productions of the same vowel by talkers of the same dialect."*

Dans une seconde expérience, les auteurs analysent les jugements des auditeurs sur le nombre de locuteurs qu'ils croient entendre lors de l'écoute des stimuli hybrides et non-hybrides. Les stimuli hybrides sont perçus comme ayant été produits par un seul locuteur, dans 75% des cas (et 82% pour les stimuli à centre silencieux non-hybrides). Les sujets jugent en majorité que l'intonation varie pour les stimuli hybrides, et qu'elle est constante pour les stimuli à centre silencieux non-hybrides. Il semble donc que les auditeurs intègrent perceptivement les deux portions syllabiques et les affectent à une source commune. La conclusion des auteurs est :

*"[...] a dialect's vowels can be characterized by higher-order variables (patterns of articulatory and spectral* change*) that are independent of a specific talkers's vocal tract dimensions."*

### 2.1.2.3 La Spécification Dynamique

Strange [1989] résume de façon élégante les hypothèses de l'approche dite "*Spécification Dynamique*" :

*"[...] vowels are conceived of as characteristic gestures having intrinsic timing parameters (Fowler, 1980). These dynamic articulatory events give rise to an acoustic pattern in which the changing spectrotemporal configuration provides sufficient information for the unambiguous identification of the intended vowels".*

Elle présente une nouvelle série d'expériences d'identification mettant en œuvre le paradigme des centres silencieux qui analysent les contributions relatives de trois sources d'information acoustique : (1) l'information sur la cible présente dans le noyau vocalique de la syllabe, (2) l'information spectrale dynamique véhiculée par les transitions consonantiques initiale et finale, (3) l'information temporelle sur la durée intrinsèque de la voyelle. Les résultats indiquent que les auditeurs utilisent ces trois types d'information pour identifier les voyelles en contexte CVC dans des phrases porteuses. De plus, ils montrent

que la contribution des paramètres dynamiques est primordiale, alors que celle du noyau vocalique est insuffisante.

Strange discute en outre de la nature exacte de l'information dynamique qui permet de spécifier les voyelles en contexte. Elle envisage deux hypothèses.

La première est inspirée par Nearey & Assmann [1986] et expliquerait les bons scores d'identification obtenus pour les stimuli à centres silencieux et à transitions tronquées, en postulant que ces stimuli contiennent l'information sur les "*Changements Spectraux Inhérents à la Voyelle*" (VISC, pour la version anglaise). Nearey & Assmann démontrent en effet que les voyelles produites en isolation sont, elles aussi, bien identifiées dans les stimuli à centres silencieux, alors qu'aucune information de coarticulation n'est évidemment présente. Selon Strange, il existerait donc des indices sur l'identité de la voyelle, présents dans les parties initiale et finale de la voyelle, donc aussi dans les stimuli à centres silencieux. Ces indices seraient de deux types : ils fourniraient de l'information sur la cible, vers laquelle tendent les transitions CV et de laquelle partent les transitions VC, ainsi que de l'information sur les VISC, reflétée, par exemple, par les relations entre la fin de la transition CV et le début de la transition VC. Mais Strange conclut que des expériences supplémentaires doivent être menées pour vérifier la justesse de cette première explication. Des expériences sur stimuli hybrides (cf. plus loin) sembleraient en effet la contredire.

La seconde hypothèse est suggérée par les analyses acoustiques menées par Lehiste & Peterson [1961] sur des voyelles en contexte CVC. Selon ces auteurs, des différences systématiques sont observées dans la structure temporelle des trajectoires formantiques des voyelles tendues et relâchées. Les syllabes contenant des voyelles tendues présentent des transitions CV et VC relativement courtes et des noyaux vocaliques relativement longs, par rapport à la durée totale de la syllabe. Les trajectoires formantiques correspondantes sont donc symétriques, de part et d'autre du noyau vocalique. Par contre, les syllabes contenant des voyelles relâchées présentent des transitions CV légèrement plus longues, des transitions VC nettement plus longues et des noyaux vocaliques relativement courts : les trajectoires formantiques sont asymétriques. De telles différences de structure ont été mises en œuvre par Huang [1985] et Di Benedetto [1989a, 1989b] dans des stimuli synthétiques et ont donné des résultats perceptifs concluants. Strange suggère que les différences de forme des trajectoires formantiques, liées aux caractéristiques temporelles des gestes d'ouverture et de fermeture dans les syllabes coarticulées, fournissent probablement une information primordiale pour l'identification de la voyelle et permettent d'expliquer les bons scores obtenus pour les stimuli à centres silencieux. Elle ajoute enfin que les expériences sur stimuli hybrides, proposées par Verbrugge & Rakerd [1986] et répliquées par Jenkins & Strange [1987], cautionnent cette seconde hypothèse. En effet, dans les stimuli hybrides, l'information véhiculée par les VISC est fortement dégradée, à cause des différences entre

les locuteurs, alors que la structure de la trajectoire temporelle est, elle, peu altérée. Ce serait donc cette information qui serait primordiale pour la spécification des voyelles.

### 2.1.3 Et si la cible était spatio-temporelle?

Andruski & Nearey [1992] discutent l'hypothèse de la Spécification Dynamique selon laquelle les parties transitoires des syllabes CVC contiennent des indices dynamiques sur l'identité de la voyelle qui ne sont pas disponibles dans les voyelles isolées et qui sont parfois supérieurs aux indices propres à la voyelle. Deux séries d'expériences perceptives sont menées pour évaluer le type d'informations utilisées : informations inhérentes à la voyelle ou informations coarticulatoires.

Dans la première expérience, des stimuli hybrides à centres silencieux fabriqués à partir de syllabes /bVb/ ou de voyelles isolées, sont présentés aux auditeurs. Les scores d'identification étant de même ordre, que le stimulus original soit une syllabe /bVb/ ou une voyelle isolée, les auteurs suggèrent que les auditeurs utilisent le même type d'information et la même méthode pour identifier les deux types de stimuli. Cette expérience réfute en outre l'hypothèse de Verbrugge & Rakerd [1986], selon laquelle, nous l'avons dit plus haut, des stimuli hybrides sont inacceptables dans une théorie fondée sur l'extraction de cibles. Les stimuli hybrides ne troublent pas les locuteurs, puisqu'ils contiennent de l'information inhérente à la voyelle (l'information qu'ils transportent est la même en présence ou non de contexte consonantique).

La deuxième expérience met en œuvre des stimuli synthétiques simplifiés, fabriqués à partir des stimuli naturels à centres silencieux. Quatre points sont mesurés sur les trajectoires formantiques des stimuli naturels : le démarrage de la première consonne, la fin de la transition CV (début du noyau vocalique), le début de la transition VC et la fin de la dernière consonne. Des interpolations linéaires entre les quatre points, respectant les durées observées pour le stimuli naturel, forment les trajectoires formantiques d'un premier type de stimuli. D'autres stimuli ne contiennent que les parties centrales des premiers stimuli synthétiques, les parties transitoires étant supprimées. Selon l'hypothèse de la Spécification Dynamique, la structure dynamique complexe des stimuli à centres silencieux véhicule une information complémentaire à l'information inhérente à la voyelle. Si tel était le cas, les simplifications radicales effectuées dans les stimuli synthétiques devraient induire des changements notables dans les scores d'identification. Or les résultats perceptifs indiquent que les stimuli synthétiques fournissent une information phonétiquement équivalente à celle des stimuli /bVb/ naturels à centres silencieux.

Les résultats des expériences de Andruski & Nearey suggèrent que les auditeurs utilisent des indices similaires pour identifier les voyelles isolées et les voyelles en contexte

/bVb/, lorsque les stimuli sont à centres silencieux. L'information perceptive fournie par les marges des voyelles en contexte CVC n'est donc pas nécessairement de nature coarticulatoire. Les marges des voyelles semblent plutôt conserver l'information inhérente à la voyelle. Les auteurs déduisent de leurs travaux que :

*"coarticulatory cues appear to play at best a minor role in the perception of vowels in /bVb/ context, while vowel-inherent factors dominate listener's perception."*

Cette hypothèse est testée à nouveau par Nearey [1995] sur une expérience d'identification de voyelles avant, à base de stimuli synthétiques, utilisant des trajectoires F1 et F2 linéaires formées sur les parties clefs des voyelles. Les résultats indiquent que les VISC sont bien une source d'information capitale pour l'identification de voyelles.

L'étude de Bailey, Bevan & Burr [1995] éclaire d'un point de vue différent les résultats de Nearey. Bailey *et al.* partent en effet d'une analogie avec la perception de scènes visuelles : le mouvement permet de discriminer figure et arrière-plan. Ils montrent que la modulation des fréquences formantiques de voyelles synthétiques permet de mieux identifier les voyelles en condition bruitée, ce qui pourrait être une explication pour l'importance informative des transitions.

L'analyse de Hillenbrand, Getty, Clark & Wheeler [1995] confirme le rôle des VISC dans l'identification de voyelles isolées. Hillenbrand *et al.* étendent les travaux classiques de Peterson & Barney [1952] et étudient les voyelles en contexte /hVd/ produites par divers locuteurs masculins, féminins, adultes et enfants. Ils montrent que si les voyelles se chevauchent dans un plan classique F1/F2, il est toutefois possible d'obtenir de bonnes ségrégations en incluant des informations de durée et de changements spectraux. Ils estiment que leurs résultats sont cohérents avec les hypothèses de la littérature sur les voyelles de l'anglais américain :

*"the vowels of American English [may be] more appropriately viewed not as points in phonetic space but rather as trajectories through phonetic space."*

Les travaux de Lindblom, Brownlee, Davis & Moon [1992] sur les "transformations de la parole" soulignent aussi la complémentarité des informations sur la cible et sur la dynamique inhérente à la voyelle (ou la diphtongue). Du point de vue du locuteur, une analyse acoustique de diphtongues issues d'un corpus de parole naturelle (conversation informelle) indique que les trajectoires formantiques sont prévisibles et que leurs variations, leurs transformations par rapport à une forme canonique unique, sont systématiques. Elles peuvent être reconstruites assez précisément à partir de la durée de la diphtongue et des informations formantiques à ses frontières :

*"The observed variants can be seen as continuous transforms derivable from a single canonical form and from contextual information in the signal and as occuring on a single continuum of reduction."*

Du point de vue de l'auditeur, les transformations mettent en jeu les styles de parole. Les travaux sur différents styles (citation et parole très claire) menés par Moon [1991] sont insérés dans l'analyse de Lindblom *et al.* Différentes formes, en termes de trajectoires formantiques, apparaissent suivant le style de parole et la durée. Quel que soit le niveau de bruit ajouté au signal, les auditeurs reconnaissent toujours mieux les formes qui ont été prononcées plus clairement. De plus, une formule mathématique permet de décrire systématiquement les trajectoires formantiques mesurées. Nous reviendrons sur ces travaux au paragraphe 3.2.4, dans notre étude sur la réduction vocalique.

### 2.1.4 Notre proposition : la production de voyelles esquisse des mouvements vers des cibles

En l'absence de résultats décisifs pour ou contre l'existence de cibles, nous proposons de prendre parti pour l'hypothèse la mieux adaptée à un cadre simple et efficace du contrôle de la production. La notion de cible est séduisante pour la simplicité de sa relation avec la phonologie. D'autre part, il existe une théorie du contrôle moteur, certes débattue mais qui a remporté de beaux succès, qui suggère que les mouvements appris soient accomplis de cible en cible : la théorie du Point d'Équilibre, sur laquelle nous reviendrons en détail au paragraphe 2.3.1. C'est pourquoi nous proposons de retenir la notion de cible, qui ne serait pas toujours atteinte, mais vers laquelle le mouvement se dirigerait, avec suffisamment d'indices pour que l'auditeur puisse la détecter dans la dynamique du mouvement.

La production de voyelles consisterait donc à viser des cibles. Nous suggérons que le locuteur vise des cibles, images physiques (non nécessairement uniques) des unités phonologiques, et que l'auditeur sait repérer des indices dans la dynamique du signal physique (acoustique, visuel, moteur, etc.), lui permettant de deviner les cibles visées par le locuteur. Mais un phénomène supplémentaire est en jeu, il faut coarticuler et non pas épeler. Il faut donc savoir passer harmonieusement et rapidement d'une cible à une autre. Et puisque l'esquisse du mouvement, avant sa réalisation totale, donne des informations capitales sur les intentions de celui qui l'a initié, on peut imaginer que la production de voyelles *esquisse des mouvements vers des cibles successives*. Le schéma suivant (2.1) donne une image de ce que peut être l'esquisse de cibles dans la production de trois phonèmes successifs. Les trois phonèmes-cibles sont représentés par les trois plateaux en trait tireté. Le mouvement effectif, en trait plein, s'engage vers les différentes cibles sans jamais les atteindre. L'allure globale du mouvement est donc déterminée par les cibles successives et la spécification du mouvement effectif est complétée par des paramètres dynamiques.

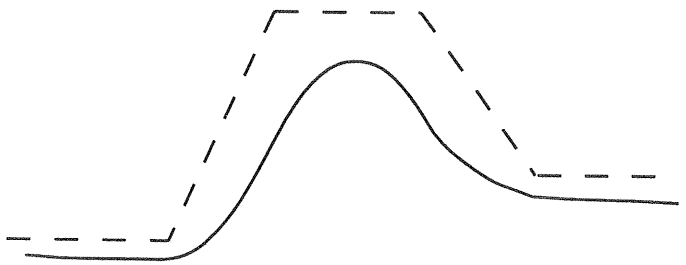


*Figure 2.1.* L'esquisse de cibles dans la production de trois phonèmes successifs. Le trait tireté correspond aux trois phonèmes-cibles, le mouvement effectif est représenté en trait continu.

Nous proposons d'étudier dans la suite de ce chapitre, comment une telle hypothèse s'inscrit dans notre cadre général pour le contrôle des mouvements de la parole.

## 2.2 Revue critique d'approches du contrôle de la production de la parole

Nous proposons un cadre général pour le contrôle du robot parlant qui s'appuie sur la notion de cible, en la reliant à l'hypothèse du Point d'Équilibre de Feldman (cf. 2.3.1.2), et espérons donner ainsi une vue parcimonieuse mais précise d'une variabilité articulatoire et acoustique assumée. Nous décrivons le passage des commandes motrices au signal acoustique par deux modèles. Le modèle dynamique (cf. 3.6.2) fournit une trajectoire articulatoire à partir de commandes motrices et met en œuvre l'hypothèse du Point d'Équilibre. Le modèle cinématique permet le passage des trajectoires articulatoires au signal acoustique. Il est décrit au paragraphe 3.5.1.
Avant de donner un canevas pour le contrôle du robot parlant, qui permette de mettre en œuvre le cadre d'hypothèses que nous avons présenté au chapitre I, présentons une revue des principaux modèles de contrôle de la parole, impliquant la notion de cible.

### 2.2.1 L'approche *Task Dynamics* des laboratoires *Haskins*

**Le modèle dit *Task Dynamics***

L'approche *Task Dynamics* est proposée, au sein des laboratoires *Haskins,* en alternative aux approches linguistiques traditionnelles qui placent les mesures statiques des configurations articulatoires ou des paramètres acoustiques au premier rang de la description phonétique, considérant comme secondaire l'étude de l'évolution du mouvement d'une configuration à l'autre. Les unités de base de l'approche *Task Dynamics* sont dynamiques, ce sont les *gestes* articulatoires.

Les gestes correspondent à la notion de *structure coordinative* proposée par le physiologiste russe Bernstein [1967][2] pour résoudre le problème de l'excès de degrés de liberté à contrôler par le système nerveux central. Les *structures coordinatives* gèrent la coordination des différents degrés de liberté et agissent afin qu'ils se comportent fonctionnellement comme une seule unité (cf. aussi R.A. Schmidt [1982]). Ce principe de *synergie,* défini par Bernstein, s'applique naturellement au contrôle moteur en parole. Les mouvements de la parole semblent en effet être *organisés* selon des principes de plus haut niveau qui mettent en jeu des interactions systématiques entre les articulateurs [Gracco, 1994]. Dans le modèle de la Dynamique de Tâche (*Task Dynamics*), développé par Saltzman et ses collègues (Saltzman & Kelso [1983], Saltzman [1986], Kelso, Saltzman & Tuller, [1986]), les gestes sont définis de façon abstraite en termes de *tâches* de parole (*speech tasks*) associant, *via* les structures coordinatives, les articulateurs par groupes fonctionnels. La *tâche* peut ainsi correspondre à l'établissement d'une constriction (ou pincement) en un endroit particulier du conduit vocal (par exemple au niveau des lèvres pour la production de labiales, ou bien entre le dos de la langue et le palais pour certaines voyelles) et avec un degré bien précis.

Les *intentions gestuelles* sont encodées sous forme d'attracteur du second ordre dans l'espace des variables dites "du conduit vocal" (*tract variables*). Le système dynamique représentant chaque *variable du conduit* correspond à un *point attracteur* de cette variable (à comparer avec la dynamique d'un pendule amorti ou d'un système masse-ressort amorti, dont les mouvements s'atténuent progressivement pour atteindre un *point* d'équilibre stable). Le mouvement d'une *variable du conduit* est ainsi modélisé par une équation linéaire du second ordre avec amortissement :

$$m\ddot{x} + b\dot{x} + k(x - x_0) = 0$$

où $m$ est la masse (fixée à 1), $b$ est l'amortissement, $k$ est la raideur, $x_0$ est la position de repos de la *variable du conduit* considérée et $x$, $\dot{x}$ et $\ddot{x}$ sont ses position, vitesse et accélération. Par conséquent, le contrôle du mouvement d'une *variable du conduit* se fait en spécifiant les valeurs des *paramètres* de l'équation dynamique du second ordre régissant ce mouvement : la position de repos, la raideur et l'amortissement. Les valeurs de ces paramètres peuvent être estimées empiriquement ou bien par une méthode d'optimisation. La position de repos est associée à la notion de cible, elle détermine la position vers laquelle le système tend. Sa valeur peut être fournie empiriquement par des descriptions articulatoires classiques (radiographies du conduit vocal). La raideur et l'amortissement déterminent la durée du mouvement vers la position de repos (la pseudo-pulsation dépendant de ces deux paramètres). Leur rapport caractérise aussi le régime du système qui

[2]Il semble que le terme ait été employé pour la première fois par Easton [1972] mais la notion sous-jacente apparaît chez Bernstein dès 1940 (traduction de [1967]).

peut être *sous-amorti* (ou pseudo-périodique), *critique*, ou *sur-amorti* (ou apériodique). Si l'on considère que le facteur d'amortissement est faible, alors il est possible de déterminer empiriquement la raideur comme le carré du rapport entre la vitesse maximale et l'amplitude du mouvement. La signification physique de ces divers paramètres est discutée plus loin (cf. paragraphe intitulé "Les gestes et les variations prosodiques"). Le temps d'établissement d'un geste ainsi que sa durée sont déterminés par les paramètres dynamiques présentés ci-dessus. Ces paramètres étant intrinsèques aux points attracteurs des *variables du conduit*, le temps d'établissement et la durée sont définis de façon *implicite* dans la dynamique de l'attracteur, dans l'espace des variables de contrôle.

Saltzman et Munhall ont proposé en 1989 l'existence de deux niveaux de coordination distincts mais agissant ensemble. La coordination temporelle des gestes entre eux est déterminée par le niveau dit *inter-gestuel* et la coopération entre les articulateurs est définie au niveau *inter-articulateur*.

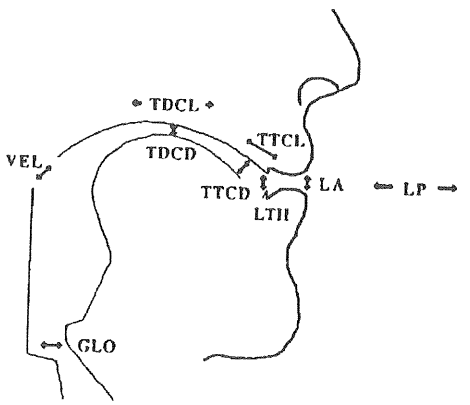
Commençons par décrire le **niveau inter-articulateur**.

Les *tâches* de constriction sont décrites à ce niveau par des équations dynamiques du second ordre qui caractérisent les variations temporelles des *variables du conduit*. Dans le modèle de la Dynamique de Tâche, ce sont les mouvements de ces *variables du conduit* qui sont spécifiés et non pas ceux des articulateurs individuels. À l'origine de la modélisation (Saltzman & Munhall [1989]), les intentions gestuelles, exprimées en termes de *variables du conduit*, étaient indépendantes du contexte. Récemment Saltzman (communication personnelle) a proposé que ces intentions soient déterminées en fonction du contexte, et a pour cela élaboré un système de "planification" fondé sur le modèle de Jordan [1986]. Les *variables du conduit* évoluent dans le temps vers ces positions intentionnelles et leurs trajectoires effectives dépendent d'une part des caractéristiques du système du second ordre définissant les attracteurs et d'autre part de la superposition des différents gestes qui simultanément peuvent influencer la valeur de ces variables (cf. plus loin le niveau inter-gestuel).

En parallèle, il existe des "variables articulatoires", coordonnées des *performances gestuelles*, dans l'espace des mouvements des articulateurs du conduit vocal. Ces *variables articulatoires*, qui définissent les mouvements des articulateurs du conduit vocal, sont également utilisées par un modèle du conduit vocal qui génère le signal acoustique correspondant [Rubin, Baer, Mermelstein, 1981].

Chaque type de constriction est en général associé, dans le plan sagittal, à deux *variables du conduit* : une pour la *position* de la constriction le long de l'axe longitudinal du conduit vocal et l'autre pour le *degré* de constriction, mesuré perpendiculairement à l'axe longitudinal. De même, chaque *variable du conduit* est associée à un sous-ensemble donné de *variables articulatoires*. Dans le cas de la production d'une labiale, par exemple,

les *variables du conduit* concernées sont la protrusion des lèvres (LP) et l'ouverture aux lèvres (LA). La *variable du conduit* LP est associée à la *variable articulatoire* LH (mouvements horizontaux des lèvres) et la *variable du conduit* LA correspond à trois *variables articulatoires* : JA (angle mandibulaire), ULV (mouvements verticaux de la lèvre supérieure) et LLV (idem, lèvre inférieure). La figure 2.2 montre la correspondance entre les *variables du conduit* vocal et leurs composantes (les *variables articulatoires*).




MODEL ARTICULATOR VARIABLES

| TRACT VARIABLES | LH ($\Phi_1$) | JA ($\Phi_2$) | ULV ($\Phi_3$) | LLV ($\Phi_4$) | TBR ($\Phi_5$) | TBA ($\Phi_6$) | TTR ($\Phi_7$) | TTA ($\Phi_8$) | V ($\Phi_9$) | G ($\Phi_{10}$) |
|---|---|---|---|---|---|---|---|---|---|---|
| LP ($Z_1$) | ● | | | | | | | | | |
| LA ($Z_2$) | | ● | ● | ● | | | | | | |
| TDCL ($Z_3$) | | ● | | | ● | ● | | | | |
| TDCD ($Z_4$) | | ● | | | ● | ● | | | | |
| LTH ($Z_5$) | | ● | | | | | | | | |
| TTCL ($Z_6$) | | ● | | | ● | ● | ● | ● | | |
| TTCD ($Z_7$) | | ● | | | ● | ● | ● | ● | | |
| VEL ($Z_8$) | | | | | | | | | ● | |
| GLO ($Z_9$) | | | | | | | | | | ● |

*Figure 2.2. Variables du conduit* et *variables articulatoires* du modèle *Task Dynamics.* La figure de gauche représente une coupe sagittale du conduit vocal où les degrés de liberté des *variables du conduit* sont indiqués par des flèches. Le tableau de droite décrit la correspondance entre *variables du conduit* (en ligne) et *variables articulatoires* (en colonne). Signification des abréviations pour les *variables du conduit* : LP/LA : protrusion/ouverture des lèvres ; TD/TT : dos/apex de la langue ; CL/CD : position/degré de la constriction ; LTH : position de l'incisive inférieure ; VEL : voile du palais ; GLO : glotte. Pour les *variables articulatoires* : LH : position horizontale des lèvres ; JA : angle de la mandibule ; ULV/LLV : position verticale de la lèvre supérieure/inférieure ; TBR/TBA : position radiale/angulaire du corps de la langue ; TTR/TTA : position radiale/angulaire de l'apex ; V : voile; G : glotte. D'après Saltzman & Munhall [1989].

Le passage des *variables du conduit* aux *variables articulatoires* met en œuvre une transformation cinématique. Les équations du mouvement en termes de *variables du conduit* ont ainsi leurs transformées dans l'espace des *variables articulatoires.* Illustrons, avec un exemple tiré de Kelso *et al.* [1986], comment le niveau inter-articulateur permet de gérer la coordination entre les articulateurs pour une tâche donnée. À partir de valeurs déterminées des paramètres dynamiques pour les *variables du conduit* LA et LP (ouverture et protrusion des lèvres), et à partir de positions et vitesses initiales données de la mâchoire et des lèvres, les équations du mouvement transformées pour les *variables articulatoires* génèrent des mouvements articulatoires coordonnés qui permettent de réaliser une tâche donnée (par exemple fermeture bilabiale) spécifiée au niveau des *variables du conduit.* Ainsi, on peut passer d'une configuration initiale où les lèvres sont ouvertes et peu

protruses (notée *a* sur la figure 2.3) à une configuration où les lèvres sont fermées et relativement protruses (*b*). Si la position de la mandibule est "gelée" sur place pendant le geste de fermeture labiale, la compensation de la perturbation est immédiate sur les lèvres supérieure et inférieure (*i.e.* il n'est pas nécessaire de reparamétrer le système pour compenser) et la fermeture bilabiale est bien réalisée (position notée *c* sur la figure). Le modèle de la Dynamique de Tâche permet donc de simuler des effets de compensation observés empiriquement, sans modification des paramètres dynamiques liés aux variables de contrôle du conduit.

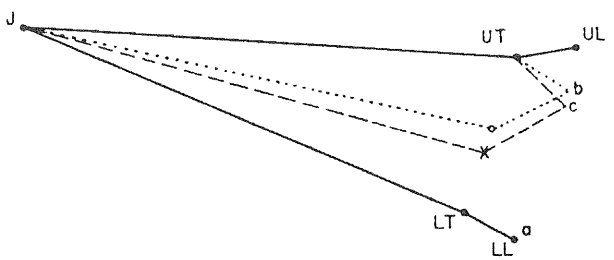



*Figure 2.3.* Configurations articulatoires simulées par le modèle *Task Dynamics* pour une tâche de fermeture bilabiale. a : configuration initiale (trait plein), b : configuration finale en condition non-perturbée (trait pointillé), c : configuration finale en condition perturbée (mandibule bloquée, trait tireté). D'après Kelso *et al.* [1986].

Considérons maintenant le **niveau inter-gestuel** et la notion de **constellation gestuelle.**

La coordination des différents gestes entre eux (par exemple entre les gestes de constrictions des lèvres et du dos de la langue pour une séquence V-b-V) est décrite au niveau inter-gestuel. Les gestes, ayant une durée propre et intrinsèque, sont en effet susceptibles de se chevaucher, à la fois temporellement (gestes actifs simultanément) et spatialement (gestes utilisant les mêmes *variables du conduit*). Les gestes impliqués dans une énonciation donnée, sont coordonnés pour former une organisation plus large qui constitue, selon Browman & Goldstein, la structure phonologique de l'énonciation considérée. Il ne s'agit pas bien entendu de définir une organisation pour chaque énonciation possible. Il existe des principes généraux qui définissent comment les gestes sont organisés ou synchronisés. Browman & Goldstein ont proposé le terme de *constellation* pour exprimer ces coordinations inter-gestuelles. Il faut définir d'une part comment les gestes sont synchronisés entre eux (c'est-à-dire l'instant auquel ils démarrent les uns par rapport aux autres) et d'autre part quels gestes font partie de la même constellation. Les principes de coordination, définis par Browman & Goldstein [1987] sont les suivants :

1. Tout geste vocalique est synchronisé par rapport au premier geste consonantique de la séquence consonantique associée (*i.e.* formant une syllabe avec la voyelle).

2. Le premier geste consonantique d'une séquence de consonnes est synchronisé par rapport au début du geste vocalique, si la séquence associée à la voyelle est au début de la syllabe (et par rapport à la fin du geste vocalique si la séquence est finale).

3. Chaque début de geste consonantique d'un groupe de consonnes est synchronisé par rapport à la fin du geste consonantique précédent.

4. Le premier geste consonantique d'une séquence de consonnes intervenant entre deux gestes vocaliques est associé (et donc synchronisé par rapport) aux deux gestes vocaliques.

L'intervalle temporel pendant lequel un geste influence le mouvement des articulateurs est défini, au niveau inter-gestuel, en fonction d'un troisième type de variables, les "variables d'activation", qui peuvent s'interpréter comme la force avec laquelle le geste considéré façonne les mouvements du conduit vocal. À chaque geste particulier correspond une *variable d'activation* propre. La figure 2.4 explicite les relations entre les trois systèmes de variable pour des constrictions bilabiale et vélaire.

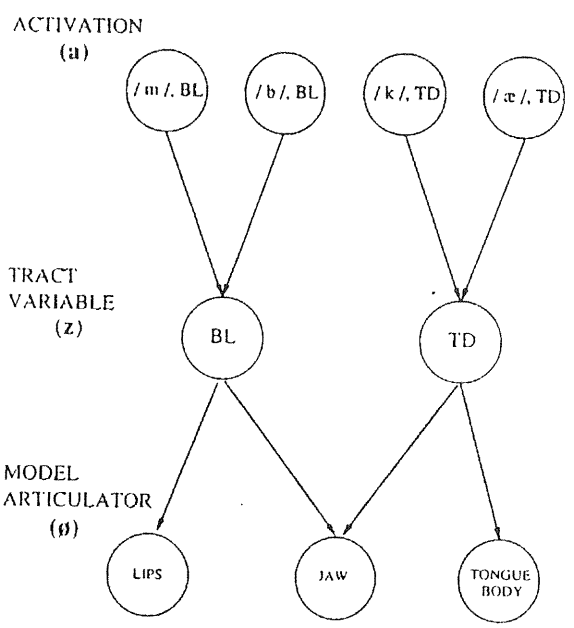



*Figure 2.4.* Relations entre les trois systèmes de coordonnées du modèle *Task Dynamics* pour des constrictions bilabiales (BL) et vélaires (TD, dos de la langue). D'après Saltzman & Munhall [1989].

L'orchestration de l'activité des gestes est exprimée à l'aide d'une partition gestuelle (*gestural score*) qui représente les domaines d'activation de chaque geste au cours du temps et dont les différentes portées correspondent aux différentes *variables du conduit*. La

superposition des différents gestes se lit ainsi aisément. Au sein de chaque domaine, les valeurs des paramètres dynamiques sont fixées et définissent ainsi le geste concerné. La figure 2.5 donne un exemple de partition gestuelle pour la séquence /pʌb/, en anglais. Les domaines d'activation des gestes sont représentés par des rectangles. Les trajectoires des *variables du conduit* obtenues avec le modèle dynamique du second ordre sont figurées dans chaque portée par le trait fin continu. Remarquons que lorsqu'aucun geste n'est actif pour un articulateur donné, celui-ci se déplace en direction de sa position neutre (celle occupée pour la production de la voyelle neutre *schwa*).

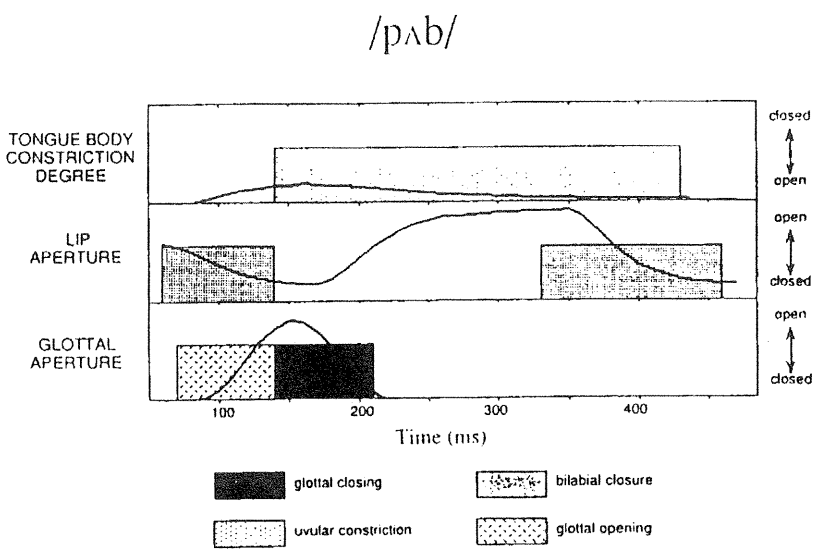



*Figure 2.5.* Partition gestuelle pour la séquence /pʌb/ selon le modèle *Task Dynamics*. La hauteur des rectangles, qui correspondent aux intervalles d'activation des gestes, vaut 0 (pas d'activation) ou 1 (activation complète). Les trajectoires du second ordre de chaque variable du conduit impliquée sont représentées par les traits continus. D'après Saltzman & Munhall [1989].

### La Phonologie Articulatoire

La *Phonologie Articulatoire* a été proposée par Browman & Goldstein en 1984 (Browman & Goldstein [1985], [1986]). Elle exploite le modèle de la Dynamique de Tâche pour établir le lien entre niveaux linguistique et articulatoire. Elle est reliée à la Théorie Motrice de Liberman & Mattingly, développée dans les mêmes laboratoires, et selon laquelle, nous l'avons vu précédemment, l'analyse des gestes articulatoires doit renseigner sur la nature profonde du message linguistique. Browman & Goldstein expliquent ainsi l'importance, pour la phonologie, de l'étude du mouvement des articulateurs :

*"[..] setting out to characterize articulator movement directly leads not to noise but to organized spatiotemporal structures that can be used as the basis for phonological*

*generalizations as well as accurate physical description. In our view, then, a phonetic representation is a characterization of how a physical system (e.g., a vocal tract) changes over time.".* Browman & Goldstein [1985].

Le terme de *Phonologie Articulatoire* vient de cette affirmation que la structure phonologique réside précisément dans l'organisation des actions articulatoires, motrices mises en jeu par la parole :

*"[...] much is missed when the line between phonological patterning and physical processes is drawn too firmly. The strong form of our view proposes that phonological structure resides in the organization of the physical actions involved in speaking. Thus, we call the approach we have been pursuing an 'articulatory phonology'.".* Browman & Goldstein [1990].

Selon eux, les gestes, qui correspondent à des événements discrets se déroulant lors du processus de production de parole, sont la passerelle entre unités articulatoires et unités phonologiques primitives :

*"Articulatory phonology attempts to describe lexical units in terms of these events and their interrelations, which means that gestures are basic units of contrast among lexical items as well as units of articulatory action."* Browman & Goldstein [1992].

Le cadre général, proposé par les laboratoires *Haskins,* est résumé par le schéma 2.6.

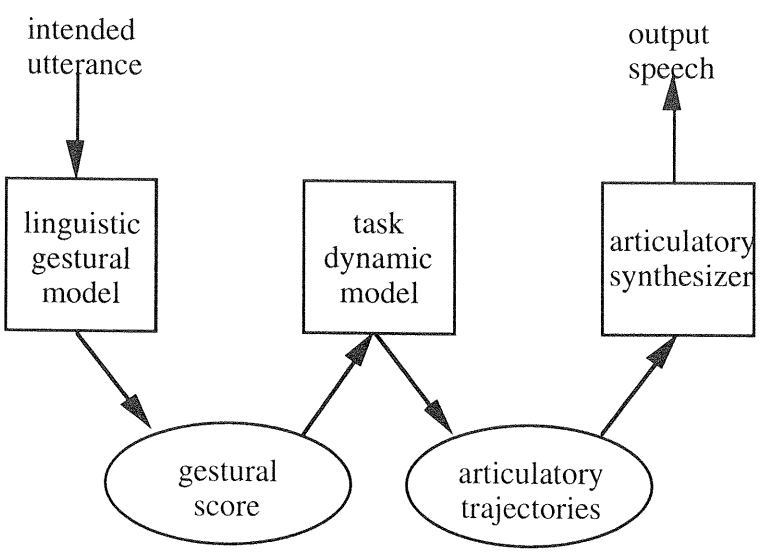



*Figure 2.6.* Le modèle gestuel des laboratoires *Haskins*. D'après Browman & Goldstein [1992].

En résumé, la séquence linguistique à produire est transformée en partition gestuelle par le modèle gestuel linguistique qui utilise les principes de phase, pour coordonner les gestes les uns par rapport aux autres. Cette partition gestuelle est fournie en entrée au modèle de la Dynamique de Tâche qui en déduit les trajectoires des *variables du conduit*

vocal puis des *variables articulatoires*. Le synthétiseur articulatoire, développé par Rubin *et al.*, génère alors, à partir de ces trajectoires articulatoires, le signal acoustique de parole.

**Les gestes et les variations prosodiques**

Saltzman & Munhall [1989] donnent quelques éléments sur la manière de faire varier le débit d'élocution au sein du modèle gestuel. Selon eux, les données articulatoires de Hardcastle [1985] ou Stetson [1951] indiquent que le chevauchement temporel entre gestes associés à des segments adjacents s'accroît avec l'augmentation du débit. Les gestes peuvent ainsi se chevaucher complètement lorsque le débit est très rapide. Ces observations peuvent s'interpréter dans le cadre théorique des laboratoires *Haskins* par des variations dans les relations de phase entre les intervalles d'activation des gestes. Browman & Goldstein [1990] ont ainsi simulé un phénomène de réduction syllabique, au cours duquel le mot anglais /beret/ se dégrade en /bray/, simplement en augmentant le chevauchement des deux gestes impliqués (fermeture bilabiale et geste apical). Il existerait donc des paramètres de contrôle au niveau **inter-gestuel** permettant d'ajuster ces relations de phase.

D'autre part au niveau **inter-articulateur** diverses études ont été menées aux laboratoires *Haskins* pour rendre compte des effets d'accentuation et de débit sur la durée et l'amplitude des gestes. Nous proposons de présenter ces travaux en détail, au risque regrettable d'appesantir l'exposé, afin de mieux différencier notre démarche de celle qu'entreprennent les chercheurs des laboratoires *Haskins*.

L'étude des activités électromyographiques effectuée par Tuller, Harris & Kelso [1982] indique que le débit de parole et l'accentuation ont des effets différenciés sur l'activité musculaire :

*"With a shift from destressed to stressed syllable production, all muscles directly associated with vowel production [...] increased in both duration and peak amplitude of activity."*

*"The effects of increases in rate of speech were less consistent across muscles and speakers. As speaking rate increased, muscle activity changed in one of three ways: (1) decrease in duration with no change in peak amplitude, (2) decrease in duration with an increase in peak amplitude, and (3) no change in duration but an increase in peak amplitude. In only one case [...] did muscle activity decrease in both duration and peak amplitude."*

Une analyse quantitative des variations des mouvements articulatoires en fonction du débit et de l'accentuation a été proposée par Kelso, Bateson, Saltzman & Kay [1985]. À partir de mesures des mouvements de la lèvre inférieure enregistrés pour des séquences de parole répétitives (successions de /ba/ ou /ma/) ces auteurs analysent les *variations cinématiques* en fonction du geste (ouverture/fermeture), de l'accentuation ou du débit. Ils

montrent ainsi que, de façon générale, les gestes accentués présentent des amplitudes et des durées plus importantes que les gestes non-accentués. De même, les gestes produits à débit rapide sont accomplis avec des amplitudes et des durées plus faibles que les gestes produits à débit lent. L'étude de la vitesse maximale (*Vm*) en fonction des divers paramètres montre que les gestes accentués sont produits avec des *Vm* plus élevées que les gestes non accentués. Par contre, l'effet du débit de parole sur *Vm* n'est pas régulier suivant les sujets observés. La pente de la relation entre *Vm* et l'amplitude (*Amp*) est sensiblement proportionnelle à la racine carrée de la raideur du système du second ordre représentant les mouvements, lorsque le système est faiblement amorti. Cette pente est en général plus importante pour les gestes non-accentués que pour les gestes accentués. Kelso *et al.* en concluent que les gestes accentués sont accomplis avec des raideurs plus *faibles* que les gestes non-accentués. En outre, dans une petite majorité des cas, la pente est légèrement plus élevée pour les débits plus rapides. Enfin, ces auteurs remarquent que la pente varie lorsque l'amplitude varie alors que l'accentuation est la même. Pour rendre compte de ces changements de raideur, lorsque l'amplitude varie, quelle que soit la catégorie d'accentuation, Kelso *et al.* proposent deux possibilités. La première est de considérer que des raideurs différentes sont choisies, par exemple, pour des gestes non-accentués, à amplitudes faibles, et des gestes accentués à amplitudes plus élevées. La deuxième possibilité est que le système du second ordre soit non linéaire. Pour éclairer leur démonstration, ils considèrent le cas d'une raideur globale qui induirait une force de rappel *linéaire* et une force de rappel *cubique*, dont la somme décroît de façon non linéaire avec la distance par rapport à la position d'équilibre ($F_r = -kx + ex^3$). Ainsi des mouvements à amplitude plus faible correspondent naturellement à des raideurs moyennes (moyenne de la dérivée de la fonction *Vm/Amp*) plus élevées sur l'ensemble du mouvement. Pour évaluer cette dernière hypothèse, ils observent la pente de la fonction accélération/amplitude. En effet, alors que les paramètres de commande ne changent pas, la raideur *linéaire k* d'un système non-amorti peut être estimée à partir de la pente de cette fonction, au point médian du mouvement. L'analyse de cette nouvelle pente montre que la raideur *linéaire* est plus importante pour les accentuations faibles que pour les accentuations fortes et, mais dans une moindre mesure et de façon moins générale, pour les débits rapides que pour les débits lents.

En résumé, dans le cadre de la modélisation *Task Dynamics*, Kelso *et al.* proposent donc de rendre compte de la variabilité de l'amplitude gestuelle d'un même geste CVC, en jouant sur la raideur, sans modifier la position de l'attracteur, dans l'espace des variables de contrôle du conduit vocal. Ainsi le mouvement d'ouverture/fermeture qu'ils étudient est représenté par une oscillation autour d'une position d'équilibre attractrice, qui reste constante pour tout le mouvement CVC. L'accentuation est associée à une diminution de la

raideur et l'accélération du débit, pour une même condition d'accentuation, correspond à une augmentation de la raideur.

Bateson et Kelso [1993] élargissent l'expérience de Kelso *et al.* à d'autres langues en comparant les effets de l'accentuation et du débit en anglais, français et japonais. Leurs résultats montrent que le système dynamique du second ordre, où la raideur et la position d'équilibre varient, permet de rendre compte des mouvements observés pour la lèvre inférieure. Ils observent, en français et en anglais, que les gestes accentués ont des amplitudes, des durées et des vitesses maximales plus élevées que les gestes non-accentués. La pente de la relation entre vitesse maximale et amplitude du mouvement est, ici encore, inversement proportionnelle à la durée moyenne du mouvement. Elle est plus élevée pour les gestes non-accentués de faible amplitude que pour les gestes accentués et plus larges. Par ailleurs, et contrairement aux premières conclusions de Kelso *et al.*, l'accentuation est fortement corrélée avec l'amplitude du mouvement et dépendrait donc du deuxième paramètre du système dynamique : la position d'équilibre. En anglais, l'accentuation se reflète à la fois au niveau temporel (les gestes accentués sont plus lents) et au niveau spatial (les gestes accentués vont plus loin). Les auteurs en déduisent :

*"[...] English stress distinctions may be correlated with both stiffness and equilibrium position."*

En français, il semble que l'amplitude et la durée soient moins liées, la durée étant moins altérée pour les mouvements non-accentués que l'amplitude (surtout pour les gestes de fermeture). La conclusion de Bateson & Kelso est alors :

*"[...] for French, equilibrium position may be the only parameter that consistently varies with stress. It is possible, then, that equilibrium position is the primary underlying parameter governing stress distinction in both languages."*

Remarquons ici que l'accent étudié par ces auteurs pour le corpus français n'est pas un accent d'emphase, mais un accent naturel (et faible) sur les fins de mots. Ceci expliquerait le faible effet sur la durée en condition non-accentuée.

Pour Bateson & Kelso, les faits conjoints que l'amplitude et la durée covarient et que l'amplitude et la pente de la courbe *Vm/Amp* varient de façon inverse, peuvent s'interpréter par l'hypothèse que les paramètres de raideur et de position d'équilibre ne sont pas indépendants, mais interviennent de façon coordonnée. Ils proposent alors, puisque le système ne peut être du type non-amorti (pour lequel la durée du mouvement est indépendante de l'amplitude), d'utiliser un système du second ordre linéaire mais amorti. L'amortissement permet en effet de réduire la vitesse maximale pour une amplitude donnée et d'augmenter la durée. Si l'amortissement augmente lorsque l'amplitude du mouvement augmente, il est alors possible de rendre compte des covariations observées.

Ainsi, pour Bateson & Kelso, la cible vocalique est liée à la position de repos d'un système linéaire du second ordre peu amorti et la trajectoire du mouvement de la consonne à la voyelle correspond à la première arche de la réponse indicielle du système. Dans ces conditions, le contrôle du temps est "intrinsèque" et s'effectue en manipulant la raideur du système, une durée plus longue étant associée à une raideur plus faible.

**Notre position**

Ces auteurs montrent qu'à partir d'un petit nombre de paramètres, il est possible de rendre compte du comportement spatio-temporel du système articulatoire lorsque le débit et l'accentuation sont modifiés. Ils proposent aussi que les contrôles du débit et de l'accentuation soient liés à des canaux différents. Cependant, dans quelle mesure le choix de la manière dont doivent varier la raideur, l'amortissement et la position d'équilibre est-il dépendant du modèle choisi? Quels sont les arguments déterminants pour, à l'intérieur du cadre modélisateur proposé, justifier le choix de tel ou tel modèle? Peut-on associer spécifiquement paramètres du modèle et paramètres prosodiques? D'autre part, il nous paraît gênant de considérer que les consonnes ne correspondent qu'à des oscillations autour de la position d'équilibre des voyelles. Cette hypothèse est d'ailleurs incohérente avec le modèle de Saltzman & Munhall [1989], pour lequel les consonnes sont bien associées à des cibles dans l'espace des tâches.

Notre point de vue est qu'il faut tenter de relier plus précisément les effets dus au débit et à l'accentuation, à un ou plusieurs paramètres du système dynamique, dont les rôles s'inscriraient de façon cohérente dans le cadre d'un schéma général de contrôle de la production. Il nous paraît également nécessaire de clarifier le contrôle de l'évolution des paramètres au cours du temps.

Une autre approche pour le contrôle moteur en parole est celle qu'ont proposée les laboratoires *ATR* au Japon. Présentons maintenant cette approche, orientée neurosciences et robotique.

## 2.2.2 L'approche *Via Points* des laboratoires *ATR*

Les travaux de Bateson sur le contrôle de la parole par Points de Passage (*Via Points*) émanent d'une approche particulièrement intéressante, s'inspirant à la fois des conceptions des laboratoires *Haskins* sur la Dynamique de Tâche et de l'approche neuronale et robotique des laboratoires *ATR*. Le modèle permettant de passer des commandes motrices aux trajectoires des articulateurs de la parole a été développé aux laboratoires *ATR* et était initialement appliqué aux mouvements du bras. Nous proposons d'introduire d'abord les concepts liés à ce modèle original puis de revenir en détail sur l'application des Points de Passage à la parole.

**Le modèle de Kawato *et al.***

Kawato, Furukawa & Suzuki [1987] examinent le contrôle et l'apprentissage des mouvements volontaires sous l'angle des neurosciences et de la robotique. Le modèle computationnel (au sens de Marr [1982], cf. 3.4), qu'ils proposent pour le contrôle des mouvements volontaires, distingue trois niveaux. Le premier niveau consiste en la **détermination de la trajectoire** à suivre, parmi toutes celles possibles, pour atteindre l'endroit désiré (le *but* du mouvement). Cette trajectoire est établie en visant l'endroit à atteindre, dont les coordonnées spatiales sont fournies par le système visuel; elle est donc exprimée dans l'espace des tâches. La détermination de la trajectoire est par conséquent fondée sur des critères purement cinématiques. Le deuxième niveau effectue un changement de repère pour passer des coordonnées dans l'espace des tâches (espace distal) aux coordonnées dans l'espace proximal (angles des articulations ou longueurs des muscles). C'est la phase de **transformation des coordonnées** ou **inversion cinématique**. Au dernier niveau, les commandes motrices (couple, par exemple), qui permettent de coordonner l'activité d'un certain nombre de muscles afin de réaliser la trajectoire désirée, sont générées. C'est la phase de **génération des commandes motrices,** qui correspond à une **inversion dynamique**. À chacun de ces niveaux apparaissent des problèmes *mal-posés* (cf. 3.4) : plusieurs trajectoires sont possibles dans l'espace des tâches pour aller d'un point à un autre ; il existe un excès de degrés de liberté du bras pour une configuration spatiale donnée ; plusieurs tensions musculaires correspondent à une même amplitude angulaire.

Uno, Kawato & Suzuki [1987] se sont penchés sur la phase de détermination de trajectoire. La même année, Kawato *et al.* s'appuient sur des notions physiologiques pour proposer un réseau de neurones hiérarchique expliquant la phase de génération des commandes motrices. Ils proposent qu'un **modèle interne de la dynamique** du système musculo-squelettique soit peu à peu appris en associant les commandes motrices aux mouvements qui en découlent. Une fois que le modèle interne de la dynamique est

construit, les commandes motrices sont mises à jour par correction de l'erreur prédite par ce modèle, donc à travers une boucle de rétroaction moins longue que la boucle de rétroaction sensorielle physiologique. D'autre part, un second modèle interne est acquis durant l'apprentissage. Ce modèle interne dit **de la dynamique inverse** du système musculo-squelettique est construit à partir des associations entre trajectoires désirées et commandes motrices. Il permet de raccourcir le calcul complexe des commandes motrices. Ces deux modèles internes sont représentés par un réseau de neurones distribué-parallèle. Une fois acquis, ils permettent de généraliser le contrôle à des mouvements différents de ceux qui figurent dans le corpus d'apprentissage.

En 1990, une vue d'ensemble des trois phases est présentée par Kawato, Maeda, Uno & Suzuki. Le modèle computationnel, qui permet le passage direct du but du mouvement dans l'espace distal aux commandes motrices, est utilisé pour les mouvements acquis. Pour les mouvements plus difficiles ou moins bien maîtrisés, les trois phases présentées ci-dessus sont successivement mises en jeu. Kawato *et al.* [1990] proposent un modèle de réseau de neurones pour le modèle computationnel direct des mouvements du bras.

Pour résoudre les problèmes *mal-posés* des trois niveaux décrits ci-dessus, les auteurs suggèrent, dans la lignée de Flash & Hogan [1985], d'ajouter un critère de minimisation du changement du couple. Le modèle direct de génération de trajectoires fournit les commandes motrices, en termes de couples, qui minimisent la fonction de coût relative au changement de couple, parmi celles qui satisfont la dynamique (relation entre position, vitesse et accélération, dans l'espace proximal), la cinématique (relation entre coordonnées proximales et distales) et les conditions du mouvement (points initiaux et finaux, *points de passage* (*via points*) et obstacles à éviter).

Durant une phase d'apprentissage, les **modèles directs cinématique** (passage des coordonnées proximales aux coordonnées distales) et **dynamique** (passage des couples aux positions dans l'espace proximal) sont acquis à partir de données sur un modèle de bras de robot à l'aide d'une technique de rétropropagation du gradient de l'erreur entre trajectoires désirée et obtenue. Les deux modèles sont inclus dans un seul réseau à structure répétitive en cascade, permettant de donner une représentation spatiale du temps.

Durant la phase de génération de trajectoires (*pattern generating*), l'erreur à rétropropager correspond cette fois aux conditions du mouvement (points finaux, *points de passage*, etc.) et ce sont les couples qui sont inférés. Les changements d'état du réseau visent à minimiser l'erreur sur les conditions du mouvement, tout en réduisant les changements de couple. Le réseau tend alors vers un état stable, qui est un état à énergie minimum. On obtient alors les **modèles inverses cinématique et dynamique**.

Le modèle computationnel général de Kawato *et al.* [1990] pour le contrôle des mouvements volontaires est mis en œuvre spécifiquement dans le domaine de la parole par

un certain nombre de chercheurs des laboratoires *ATR*. Considérons donc maintenant cette approche particulière.

**Le contrôle de la parole par points de *passage* (*Via Points*)**

Bateson, Hirayama & Kawato [1991] utilisent le réseau de neurones en cascade de Kawato *et al.* [1990], pour acquérir le modèle direct dynamique permettant de passer des commandes motrices aux trajectoires articulatoires observées en parole. Ce modèle dynamique s'insère dans un **modèle global de production de la parole**, dont les entrées seraient les chaînes phonémiques. Les auteurs proposent que chaque phonème soit associé à une cible ou un point de *passage* (*via point*), spécifié dans l'espace des tâches (au sens des laboratoires *Haskins*, cf. plus haut). Le caractère lisse et régulier (fluide) des mouvements serait issu d'une part d'un contrôle central actif et d'autre part de l'action passive de la biomécanique. Les auteurs envisagent de rattacher la fluidité à une origine unique, correspondant à des contraintes d'élocution, comme le débit et le style de parole. Notons que les points de *passage* et les contraintes de lissage sont fonctionnellement opposés puisque le lissage tend à gommer les aspérités du mouvement reliant les points de *passage*.

Bateson *et al.* [1991] font un premier pas vers l'établissement de ce modèle global de production de la parole en mettant en place le modèle dynamique direct. Celui-ci est acquis à l'aide d'un apprentissage associant les positions articulatoires, vitesses et EMG aux changements de position (vitesse) et de vitesse (accélération), ces données ayant été mesurées expérimentalement pour des séquences de parole itératives. Il est ensuite inclus dans un réseau de neurones à structure en série (cascade), afin de vérifier sa capacité à produire des trajectoires articulatoires continues. Lorsque les activités EMG (commandes motrices) sont fournies, le réseau est capable de générer des trajectoires articulatoires proches de celles mesurées. Les auteurs disposent donc d'un modèle dynamique direct représentatif de la dynamique fonctionnelle d'un certain nombre d'articulateurs de la parole et fournissant des informations sur les propriétés viscoélastiques du système (pulsation propre et amortissement).

L'année suivante, Hirayama, Bateson, Kawato et Honda [1992] élargissent ce premier modèle à des séquences de parole plus complexes et comprenant divers styles de parole, ainsi qu'à un nombre plus élevé de muscles et de dimensions articulatoires étudiés. D'autre part, le **modèle dynamique direct** des relations entre commandes motrices et trajectoires articulatoires est complété en aval par un modèle de transformation des trajectoires articulatoires en produit acoustique (**modèle acoustique direct**). Il est en outre complété en amont par un réseau capable de générer des commandes motrices à partir d'informations discrètes reliées à la commande linguistique, comme les cibles articulatoires associées à chaque phonème dans l'espace des tâches (au sens des *Haskins*) et les paramètres de style et

de débit de parole. Des contraintes de lissage (du type minimisation des changements de commandes motrices) sont utilisées pour résoudre le problème de l'excès de commandes motrices associées à un type d'information linguistique. Les points de *passage* et les contraintes de lissage interagissent de façon à ce que les trajectoires correspondant à un débit de parole rapide soient plus fluides et aient plus tendance à rater les cibles définies par les points de *passage* que les trajectoires produites à débit lent. Le réseau global utilisé par Hirayama *et al.* [1992] et Hirayama, Bateson, Kawato et Jordan [1992] pour générer des commandes motrices puis des trajectoires articulatoires à partir du débit de parole et des cibles articulatoires (points de *passage*) est celui de Kawato *et al.* [1990] présenté ci-dessus. Des séquences de parole itératives (succession de /ba/) sont produites.

En 1993, Bateson, Hirayama, Wada et Kawato [1993] s'intéressent à la modélisation de séquences de parole plus naturelles. Les séquences de parole naturelles impliquent l'activité de beaucoup plus d'articulateurs que les séquences itératives étudiées précédemment (mettant principalement en jeu les lèvres et la mâchoire). Il devient alors nécessaire de séparer le modèle dynamique direct en trois modèles : un modèle pour la mâchoire seule, un pour le complexe mâchoire+lèvre inférieure et enfin un pour le complexe mâchoire+langue.

D'autre part, pour la parole itérative, il suffisait de deux cibles (une pour chaque phonème de /ba/) s'alternant régulièrement dans le temps. Les seuls articulateurs concernés étant la mâchoire et les lèvres, les cibles de passage pouvaient être spécifiées indifféremment dans l'espace des articulateurs (positions des lèvres et de la mâchoire) ou dans l'espace des tâche (ouverture aux lèvres). De plus, les différences de débit pouvaient s'expliquer par des valeurs différentes de la pulsation propre (de l'oscillateur harmonique décrivant le système dynamique) et des contraintes de lissage. Bateson *et al.* [1993] en sont donc venus à développer un système d'estimation automatique des points de passage pour des configurations articulatoires complexes. En partant des trajectoires articulatoires réelles, les points de passage sont calculés de façon systématique, en utilisant un critère de minimisation du *jerk* et sont approximativement en nombre égal à celui des phonèmes. Bateson, Tiede, Wada, Gracco et Kawato [1994] améliorent la technique d'estimation de points de passage en optimisant la fenêtre d'analyse, le critère d'erreur (qui détermine le nombre de points de passage et l'ajustement de la trajectoire articulatoire calculée à celle donnée) et la distribution des poids relatifs des articulateurs, introduisant ainsi une influence des facteurs prosodiques dans la détermination des points de passage.

**Notre position**

L'idée, avancée par Bateson et ses collègues, selon laquelle la parole serait une tâche optimisée vers des cibles, est séduisante et nous y adhérons. Cependant, au delà de ce principe, la question est posée de la réalité d'un système de contrôle de la parole mettant en jeu un modèle dynamique inverse. La parole met en effet en jeu des mouvements qui durent parfois moins de 50ms. Comment peut-on envisager que la procédure d'inversion dynamique puisse être opérationnelle dans un tel laps de temps ? Par ailleurs, nous regrettons que l'estimation des points de passage mettent finalement au même niveau les facteurs de type phonémique et ceux de type prosodique. Une telle approche n'éclaircit pas la relation entre invariance phonémique et variabilité articulatori-acoustique. Dans le modèle que nous présentons ci-après, nous proposons un cadre pour contourner ce type de problème.

## 2.3 Proposition d'un schéma général pour le contrôle de la production de la parole

Nous présentons dans cette partie le schéma général de contrôle qui nous permettra d'évaluer nos hypothèses sur la production des voyelles. La notion de cible, qui est au cœur de ce schéma, est mise en œuvre par l'application d'une théorie majeure en contrôle moteur, l'hypothèse du Point d'Équilibre de Feldman.

### 2.3.1 Les variables de contrôle moteur

L'hypothèse du Point d'Équilibre, qui est indéniablement une théorie majeure dans le domaine du contrôle moteur (cf. par exemple le numéro spécial de la revue *Behavioral and Brain Sciences* de décembre 1995, qui lui est consacré), a été d'abord proposée pour les mouvements du bras à une articulation (Feldman [1966], [1986]). Elle s'applique désormais à de nombreux domaines : mouvements pluri-articulaires du bras (Feldman, Adamovitch, Ostry et Flanagan [1990], Flanagan, Ostry et Feldman [1993]), mouvements oculaires (Feldman [1981]), mouvements de la mandibule en mastication et en parole (Flanagan, Ostry, Feldman [1990], Laboissière, Ostry et Feldman [1996]). Elle s'appuie sur l'idée fondamentale que les mouvements des membres résultent de modifications des paramètres du contrôle central, qui déplacent le point d'équilibre du système moteur. Mais rappelons d'abord quelques principes élémentaires du contrôle des articulateurs de la parole (Perrier & Ostry [1994], Perrier, Ostry & Laboissière [1996]).

### 2.3.1.1 Le contrôle de l'activité musculaire

Le mouvement des membres résulte de l'activation de motoneurones stimulés par des commandes centrales (information *efférente*). Le maintien de la position d'un membre en fonction d'une consigne centrale et l'entretien du tonus musculaire sont assurés par l'arc réflexe monosynaptique spinal (la boucle gamma). Cette boucle met en œuvre deux sortes de motoneurones dont les noyaux appartiennent à la substance grise de la moelle épinière (corne antérieure) : les motoneurones alpha et gamma (MNα et MNγ).

Les MNα sont responsables de la contraction de fibres extrafusales de larges diamètres qui constituent la partie essentielle d'un muscle, tandis que les MNγ innervent les fuseaux musculaires, petites structures fusiformes situées parallèlement aux fibres extrafusales. Ces fibres fusiformes, de faible diamètre, n'ont pas d'effet direct sur la contraction des muscles, mais, sensibles à l'étirement au niveau des mécanorécepteurs annulo-spiralés, elles peuvent, à travers les fibres neurosensitives Ia et II, directement connectées aux mécanorécepteurs, délivrer une information *afférente* à la partie postérieure de la moelle, qui sera transmise aux MNα. Un étirement de la région centrale des fibres fusiformes (région équatoriale), dû soit à une activation centrale des MNγ, soit à un étirement des longues fibres musculaires, contribue à l'activation des MNα et ainsi indirectement à la contraction musculaire. L'activation des MNα, suivie de contraction musculaire, est par conséquent due à la fois à une commande centrale efférente et à un *feedback* afférent. La figure 2.7 représente cette boucle gamma.

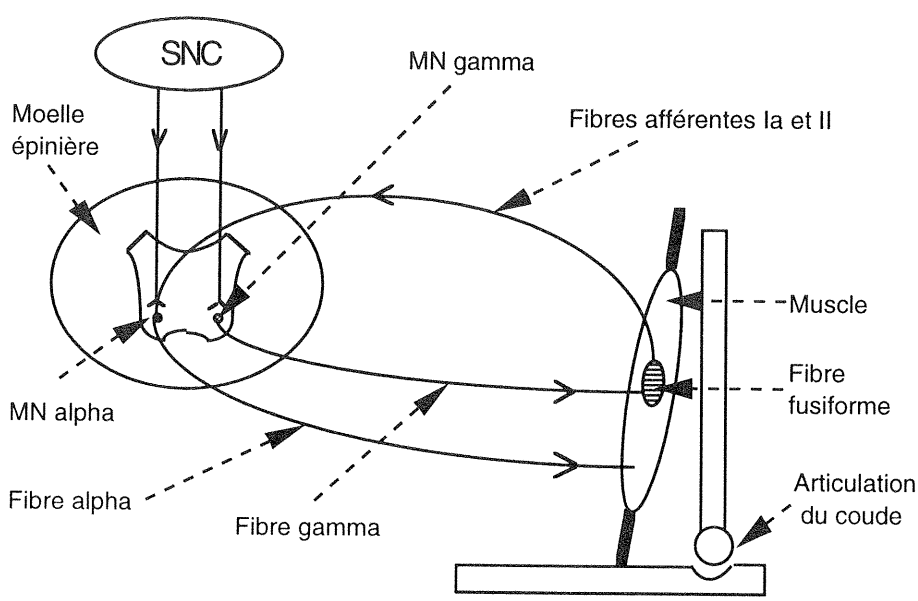



*Figure 2.7.* Représentation schématique du contrôle de la contraction musculaire *via* la boucle gamma. D'après Perrier & Ostry [1994].

Lors d'un mouvement volontaire, l'activation alpha est donc accompagnée d'un renforcement de l'activité gamma ce qui, en augmentant le tonus, prépare le mouvement et permet des mouvements harmonieux. Pour une consigne centrale donnée et constante (du type contraction du muscle fléchisseur du bras pour porter un livre), le rôle de la boucle gamma dans le maintien de la posture, suite à une perturbation (ajout soudain d'un deuxième livre), est le suivant : l'information d'étirement est transmise le long des fibres Ia, elle entraîne ainsi l'activation des MNα et génère finalement une nouvelle contraction du muscle. L'étirement est ainsi contrebalancé (le bras soutient les deux livres). Le niveau de contraction musculaire n'est donc pas directement contrôlé par le système nerveux central, il est la conséquence d'une consigne centrale et d'une rétroaction sensitive, fonction de la longueur du muscle. Remarquons déjà que l'activation musculaire ne peut pas être considérée comme le principal paramètre contrôlé pour le mouvement des articulateurs (cf. la question de Bizzi, Hogan, Mussa-Ivaldi, Gistzter [1992] et cf. plus loin 2.3.1.3).

### 2.3.1.2 L'hypothèse du Point d'Équilibre

L'hypothèse de Feldman [1966] est que les mouvements résultent de déplacements du point d'équilibre du système musculaire. Asatryan et Feldman [1965] avaient remarqué que, pour une consigne centrale déterminée, le comportement des muscles peut être décrit par une relation invariante et unique entre la force et la longueur musculaire (ou bien le couple et l'angle), qu'ils nommèrent la "caractéristique invariante" (CI, cf. figure 2.8).

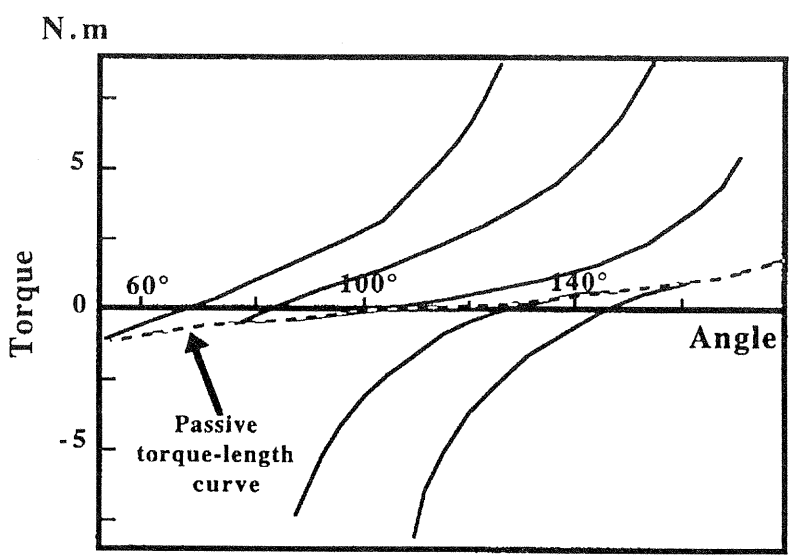



*Figure 2.8.* Caractéristiques invariantes de l'avant-bras. Les différentes courbes correspondent à différentes positions initiales du membre et différentes commandes centrales. D'après Feldman [1986].

Pour une charge externe donnée (poids d'un livre...) et pour une consigne centrale donnée (activation centrale des MN$\alpha$ et MN$\gamma$), il existe une longueur *unique* du muscle (déterminée par la CI) pour laquelle la force musculaire générée peut compenser la charge externe. Une autre position d'équilibre pourra être atteinte, avec la même charge externe, si le comportement du muscle suit une autre CI. Le système Nerveux Central (SNC) peut donc spécifier la position d'équilibre désirée en sélectionnant une courbe CI appropriée. De plus, comme on peut le remarquer sur la figure 2.8, il existe une longueur spécifique du muscle, $\lambda$, en deçà de laquelle la force musculaire générée par le SNC est nulle. Le comportement du muscle est alors réduit à celui d'un corps élastique passif. Ce paramètre $\lambda$ peut donc être considéré comme le *seuil de recrutement musculaire* et le choix de ce seuil détermine complètement la CI. Feldman a donc proposé l'hypothèse suivante :

*Le Système Nerveux Central détermine la position d'équilibre atteinte par le membre pour une charge donnée, en spécifiant le seuil de recrutement $\lambda$, via l'activation centrale des motoneurones $\alpha$ et $\gamma$.*

La forme exponentielle des CI reflètent un principe de taille (*size principle*) : lorsque la différence entre la longueur réelle du muscle et le seuil $\lambda$ augmente progressivement, des unités motrices de taille de plus en plus importante sont recrutées, ce qui augmente la force en conséquence. La figure 2.9 clarifie le mécanisme fondamental du contrôle de la posture et du mouvement, dans le cas très simple d'un membre à un degré de liberté et un seul muscle, subissant une force gravitationnelle externe déterminée (Flanagan, Feldman et Ostry [1992]). Le panneau (A) représente quatre postures du membre (a, b, c, d), le panneau B fournit les niveaux de dépolarisation des motoneurones pour ces quatre cas (a, b, c, d) et les panneaux C et D montrent les relations correspondantes entre force et longueur.

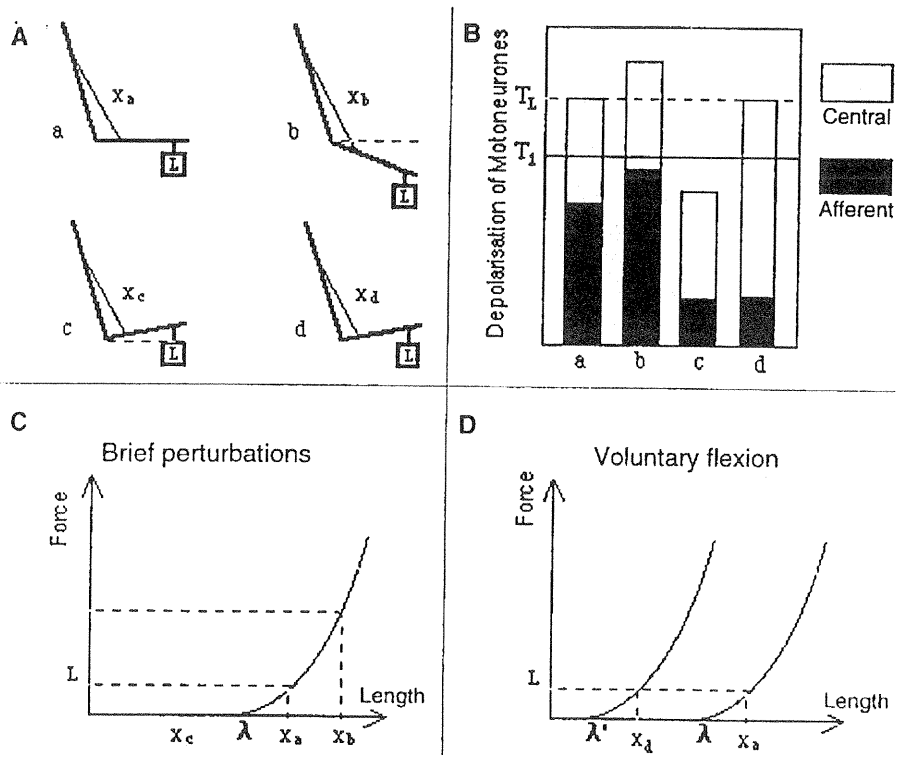



*Figure 2.9.* Contrôle de la posture et du mouvement dans le cas simplifié d'un membre à un seul degré de liberté et un seul muscle. Pour les postures a et d, le membre est à l'équilibre. Les postures c et d représentent des perturbations en extension et en flexion, respectivement. Après une perturbation, l'activation afférente et la force musculaire augmentent en b et diminuent en c. Une flexion volontaire du membre (de a à d) correspond à une augmentation de l'activation centrale et à un déplacement du seuil de recrutement (de λ à λ'). Voir texte pour plus de détail. D'après Flanagan *et al.* [1992].

**Maintien de la posture (déplacement le long d'une CI)**

Au point noté (a), une combinaison d'informations afférentes et efférentes produit l'activation nécessaire pour la génération d'une force $L$ qui permet de compenser une charge externe, à la longueur musculaire $x_a$. Si, suite à une perturbation, le membre change de posture pour se retrouver en (b), alors que l'activation centrale reste la même que pour (a), le muscle est étiré et a pour longueur $x_b$. L'activation afférente (qui dépend de la longueur musculaire et de l'activation centrale des MNγ) augmente et la force générée dans le muscle devient supérieure à la force nécessaire pour compenser la charge ($L$). Ceci entraîne donc un mouvement dans la direction de la posture (a). Dans la posture (c), la longueur du muscle $x_c$ est inférieure au seuil, mais on conserve la même activation centrale qu'en (a). L'activation afférente est diminuée et la force générée ne compense plus la force due à la charge. La charge entraîne donc le membre vers la position d'équilibre du (a).

**Mouvement volontaire (changement de CI)**

Pour déplacer volontairement le membre de la posture (a) à la posture (d), la contribution centrale augmente et l'on passe de la CI correspondant au seuil $\lambda$ à la CI de seuil $\lambda'$. Le seuil de recrutement étant diminué, la différence entre la longueur réelle des muscles et la longueur de seuil augmente, ce qui a pour effet de recruter plus de motoneurones et de diminuer la longueur du muscle (contraction). Le membre se trouve dans une nouvelle posture, pour laquelle la longueur du muscle est $x_d$ et la force ($F_d = L$) compense toujours bien la charge. En conclusion, le contrôle du mouvement est initié par un changement de seuil $\lambda$ et dépend à la fois de l'activation efférente et afférente.

La figure 2.10 montre comment s'organise le contrôle du mouvement pour un système à deux muscles antagonistes (Perrier, Ostry & Laboissière [1996]), où les notions de force et de longueur sont remplacées par des notions de couple et d'angle.

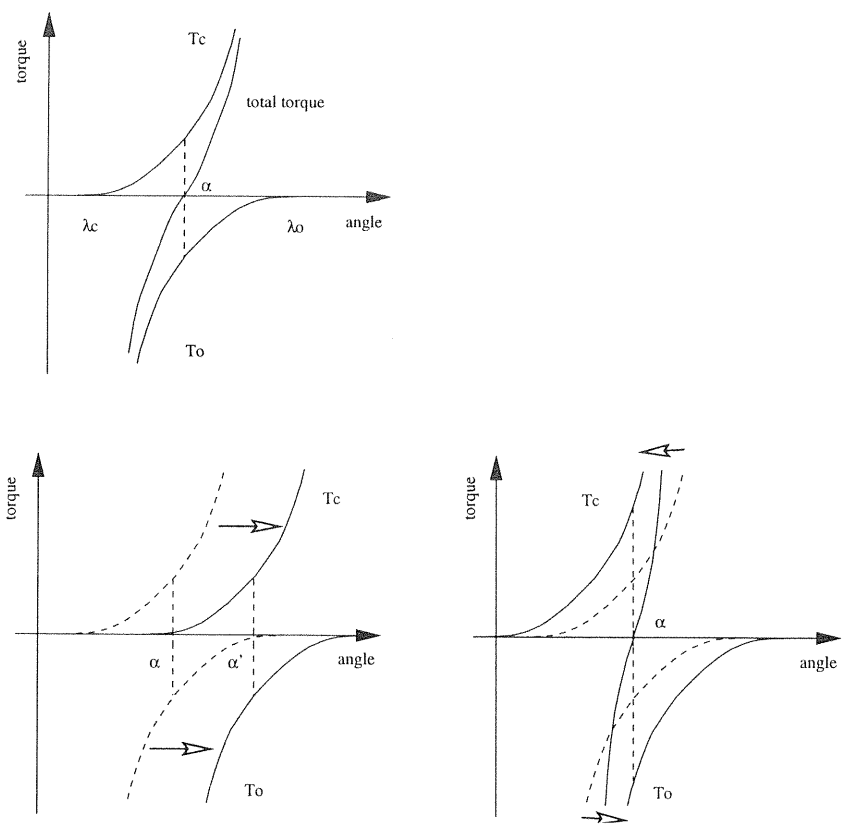



*Figure 2.10.* Contrôle du mouvement pour un système à deux muscles antagonistes. D'après Perrier, Ostry & Laboissière [1996].

Dans le premier panneau, le muscle fléchisseur (ou qui ferme, en anglais *Closer*) produit un couple $T_c$, et le muscle extenseur (ou qui ouvre, en anglais *Opener*) un couple $T_o$, dans la direction opposée. La sélection des seuils $\lambda$ pour ces deux muscles, permet de spécifier un *angle d'équilibre* $\alpha$ (pour lequel le couple total, sans force externe, est nul).

Ces seuils permettent aussi de contrôler la *cocontraction musculaire*, représentée par la pente de la courbe de couple total, correspondant à la somme des deux couples $T_c$ et $T_o$. Si les seuils $\lambda$ des deux muscles sont déplacés dans la même direction (cf. panneau inférieur gauche de la figure 2.10), l'angle d'équilibre passe de $\alpha$ à $\alpha'$ et la cocontraction est inchangée. De façon complémentaire, si les seuils sont déplacés d'une même quantité dans des directions opposées (cf. panneau inférieur droit), la cocontraction augmente, alors que l'angle d'équilibre n'est pas affecté.

**En résumé, les combinaisons de seuils $\lambda$ permettent de contrôler à la fois l'angle d'équilibre et la cocontraction musculaire.**

### 2.3.1.3 Pour ou Contre l'hypothèse du Point d'Équilibre (PE)?

L'hypothèse du Point d'Équilibre de Feldman, dite aussi "modèle λ", ne fait pas l'unanimité parmi la communauté des chercheurs en contrôle moteur (Stein [1982], Schmidt [1988] et le numéro spécial de la revue *Behavioral and Brain Sciences,* Feldman & Levin [1995]).

Tout d'abord, il existe une hypothèse concurrente, utilisant aussi la notion de Point d'Équilibre : celle de Bizzi *et al.* [1992], dite "modèle α". Ces auteurs minimisent l'importance de l'information afférente transmise par les fibres Ia et Ib (cf. 2.3.1.1) et supposent que l'action du SNC est de contrôler la raideur (ou cocontraction) des muscles *via* essentiellement l'activation des motoneurones alpha (d'où le nom du modèle). Dans le modèle α, l'activation afférente *via* les fibres Ia et Ib étant négligée, les courbes des forces en fonction de la longueur n'ont pas la forme exponentielle du modèle λ, mais sont rectilignes. Enfin le modèle α n'explique pas comment une même activité EMG peut être observée pour deux positions d'équilibre différentes. Si le modèle α est plus simple, il nous semble toutefois que l'hypothèse de Feldman, qui respecte mieux la physiologie, est plus puissante.

D'autre part, certains chercheurs (Atkeson et Hollerbach [1985], Hollerbach & Flash [1982]), reconnaissent que les modèles de Point d'Équilibre sont les meilleurs pour rendre compte de la façon dont une articulation atteint une position *finale*, mais mettent toutefois en cause la capacité de ce type de modèle à expliquer *la phase de démarrage* des mouvements, particulièrement pour les mouvements rapides. D'autres (Schmidt, Sherwood, Zelaznik et Leikind [1985]) estiment que la coordination intermusculaire intervenant dans les systèmes à plusieurs articulations n'est pas expliquée de façon satisfaisante par les modèles à Point d'Équilibre.

Cependant, Flanagan, Ostry et Feldman [1990] ont appliqué le modèle λ au contrôle des mouvements à plusieurs articulations. Ils montrent ainsi qu'en manipulant la direction

et la vitesse de déplacement des points d'équilibre, il est possible de rendre compte des trajectoires cinématiques et électromyographiques observées expérimentalement. En outre, Flanagan, Feldman et Ostry [1992], se sont penchés avec succès sur le contrôle des mouvements rapides de pointage de cible. Ces résultats, ainsi que des travaux sur le contrôle d'un articulateur important en parole, la mâchoire (Laboissière, Ostry & Feldman [1996], Perrier, Ostry et Laboissière [1996]), nous invitent fortement à adopter l'hypothèse PE.

Le concept de Point d'Équilibre est particulièrement intéressant dans notre étude, puisqu'il permet de concrétiser la notion de cible idéale (position d'équilibre) vers laquelle les articulateurs tendraient pour un phonème donné. Les intentions, les *gestes* planifiés, peuvent ainsi s'exprimer en termes de trajectoires virtuelles d'un point d'équilibre à un autre. La sélection d'un niveau de cocontraction ainsi que l'agencement temporel des transitions de point d'équilibre à point d'équilibre permettent de produire une multitude de trajectoires virtuelles susceptibles de rendre compte de la variabilité observée en parole.

### 2.3.2 Notre schéma général de contrôle

Esquissons à présent un schéma pour le contrôle du robot parlant. Les perturbations (tube labial, cale maintenant la mâchoire à une position fixe, etc.), dont l'étude sort du cadre de notre thèse, ne sont délibérément pas prises en compte dans ce schéma. Les différentes phases qui interviennent sont les suivantes :

**I.** Spécification de la tâche.

**II.** Transposition de la tâche dans un espace proximal du locuteur.

**III.** Planification. Cette phase met en jeu des représentations internes des relations entre positions articulatoires et effets perceptifs ; elle prend en compte autant que possible l'état du système périphérique, autant que les délais de *feedback* le permettent ; elle minimise un critère dépendant du locuteur, tant que cela est possible, *i.e.* n'affecte pas la réussite de la tâche perceptive. Remarquons que c'est principalement à ce niveau que les perturbations externes sont prises en compte.

**IV.** Génération des commandes motrices. S'appuyant sur la théorie PE, cette phase prend en compte les commandes prosodiques et met en jeu un modèle inverse des relations entre positions spatiales d'équilibre et commandes motrices posturales (certaines perturbations du système sont aussi intégrées à ce niveau).

**V.** Exécution de la tâche. C'est la phase de génération du mouvement et du son par envoi des commandes motrices aux muscles de l'appareil phonatoire et du conduit vocal.

Le *feedback* auditif est utilisé pour vérifier *a posteriori* la réussite de la tâche et réactualiser en permanence les différentes représentations internes, directes ou inverses, mises en jeu dans la production de la parole.

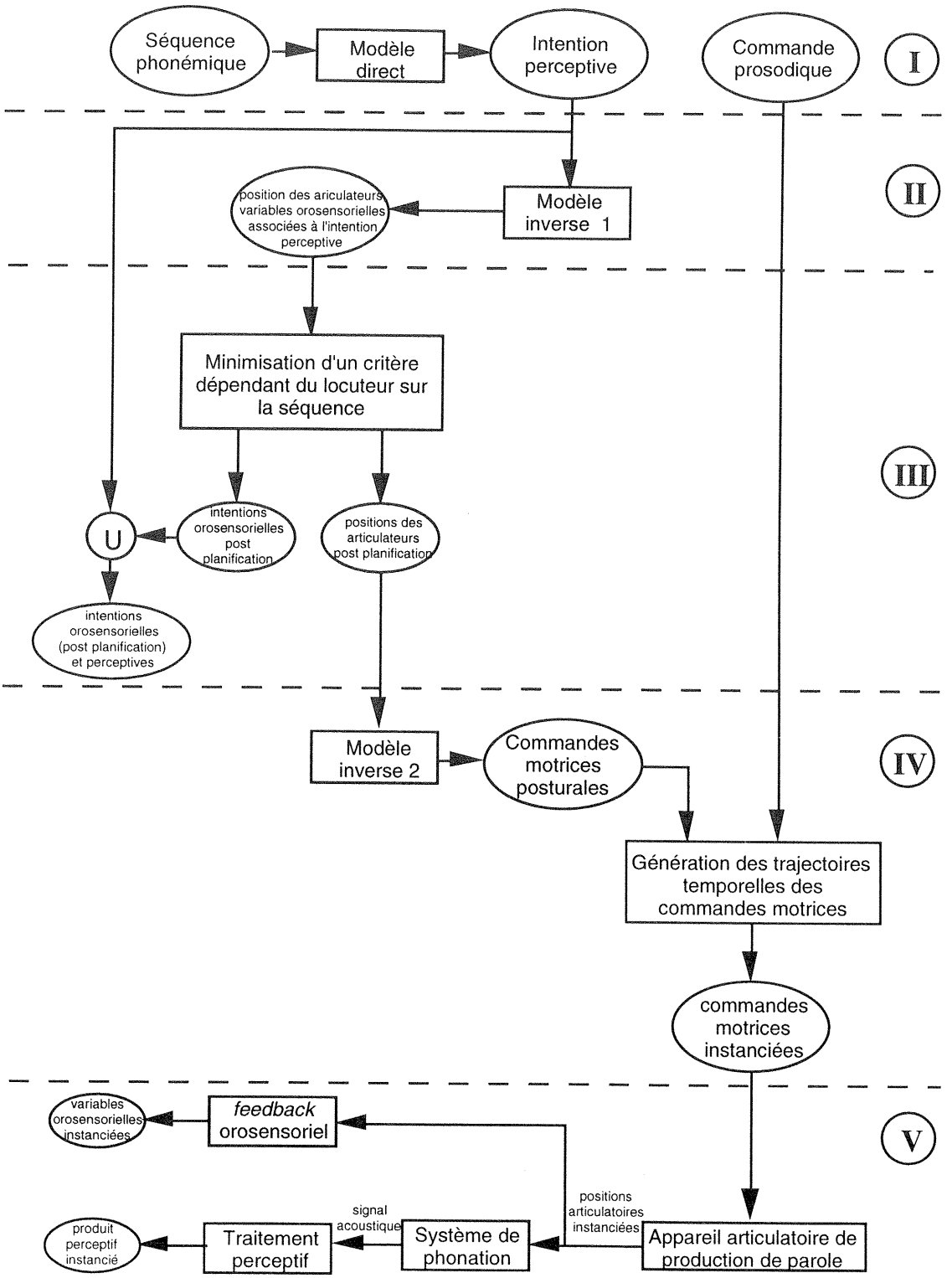



*Figure 2.11.* Schéma général de contrôle de la production de la parole.
I. Spécification de la tâche.
II. Transposition de la tâche dans un espace proximal.
III. Planification.
IV. Génération des commandes motrices.
V. Exécution de la tâche.

## 2.4 Bilan

Récapitulons à présent sur le schéma général de contrôle que nous avons proposé, en précisant comment nous nous situons par rapport aux modèles cités au paragraphe 2.2.

### 2.4.1 Notre position par rapport au modèle des *Haskins*

Notre approche présente un certain nombre de similitudes avec le modèle développé par les chercheurs des laboratoires *Haskins*. Toutefois, il existe des divergences importantes que nous allons présenter maintenant.

Les articulateurs que nous considérons sont ceux d'un modèle articulatoire du conduit vocal qui permet de relier un certain nombre de paramètres articulatoires à des configurations formantiques (cf. 3.5.1). Ils correspondent approximativement aux *variables articulatoires* du modèle *Task Dynamics* des laboratoires *Haskins*. Nous proposons de représenter chacun de ces articulateurs physiques par un modèle du second ordre. En effet, s'il est reconnu qu'un modèle dynamique du second ordre rend assez bien compte de la cinématique des mouvements observés pour le bras ou la mâchoire (ce point est discuté plus en détail au 3.6), nous estimons qu'il est plus naturel d'appliquer ce modèle directement à un articulateur physique plutôt qu'à un geste abstrait. Les notions de force et de masse mises en jeu par un tel modèle nous semblent mal adaptées à des *variables du conduit* abstraites (cf. à ce propos le modèle de Kröger [1993] mettant en œuvre la notion de geste de Browman et Goldstein mais appliquant le modèle dynamique directement aux articulateurs, sans passer par les "tâches" de Saltzman et Kelso). Pour nous, l'abstraction est au niveau des commandes de ce modèle. D'autre part, le modèle du second ordre choisi diffère quelque peu de celui des laboratoires *Haskins* puisque nous considérons ici un modèle *distribué* qui symbolise les effets des groupes de muscles agonistes et antagonistes (cf. 3.6.2).

Nous supposons que la trajectoire temporelle de la position d'équilibre est définie par des paramètres spécifiant la durée des paliers d'équilibre et des transitions d'une position d'équilibre à une autre. Nous proposons de déterminer si ces durées présentent des variations systématiques lors de modulations prosodiques et s'il est possible de les associer à des commandes motrices pour le contrôle d'un articulateur. Nous n'excluons donc pas *a priori*, contrairement aux chercheurs des laboratoires *Haskins*, la possibilité que le temps soit une variable contrôlée. Nous reviendrons en détail sur nos positions sur ce problème du contrôle éventuel du *timing* au paragraphe 4.3.3.

### 2.4.2 Notre approche comparée à l'approche computationnelle de Kawato *et al.*

Contrairement à Kawato, Bateson et collègues, nous n'adoptons pas la notion de modèle dynamique inverse dans notre modèle de contrôle de la parole. Les commandes motrices sont obtenues directement à partir des intentions posturales, grâce à la correspondance qu'offre l'Hypothèse du Point d'Équilibre, pour un champ de forces externes donné, entre positions dans l'espace et commandes motrices. Les transitions entre positions d'équilibre sont spécifiées directement sous la forme d'une trajectoire dite "virtuelle" (Hogan, [1984]). Ce type de contrôle, fondé sur l'Hypothèse du Point d'Équilibre et préconisé par de nombreux chercheurs (Flash & Hogan [1985], Flash [1987], Flanagan, Ostry & Feldman [1993]), permet de déduire automatiquement le couple (ou la force musculaire) de la raideur musculaire et de la différence entre trajectoires virtuelle et réelle. Les simulations de Flanagan, Ostry & Feldman [1993] sur un modèle de bras à deux articulations permettent de reproduire, à l'aide de trajectoires virtuelles simples, les trajectoires de bras complexes, observées au cours d'expériences de pointage de cibles, fixées ou déplacées.

Katayama et Kawato [1993] discutent cette proposition de trajectoires virtuelles. Ils montrent que les trajectoires linéaires simples, observées dans l'espace des tâches pour les mouvements de point à point rapides ou à faible raideur, ne peuvent être expliquées que par des trajectoires virtuelles très complexes. Ils estiment que, pour ce type de mouvement, le contrôle par trajectoire virtuelle n'est pas justifié puisqu'il requiert la planification de trajectoires très complexes. La récupération du modèle inverse dynamique leur semble alors mieux adapté.

Cependant l'estimation de trajectoires virtuelles dépend très fortement des caractéristiques du modèle dynamique utilisé, et en particulier des modèles musculaires (explicitement différents pour les modèles de Flanagan *et al.* et Katayama et Kawato). En la matière, la démonstration de Kawato et de ses collègues n'est donc pas décisive. À l'inverse, l'idée de la mise en œuvre d'un modèle inverse dynamique nous paraît, nous l'avons dit, peu défendable pour la production de la parole, compte tenu des durées mises en jeu dans cette tâche motrice (à comparer avec des tâches de pointage qui durent,couramment, quelques centaines de millisecondes). Nous préférons donc l'hypothèse d'un contrôle par trajectoire virtuelle simple reliant entre elles différentes positions d'équilibre successives, puisqu'elle permet d'envisager de représenter directement, au niveau des variables de contrôle moteur, les commandes phonémiques et prosodiques.

Le système dynamique choisi, qui produit des trajectoires articulatoires à partir de commandes motrices, est un système du second ordre dont on sait qu'il se rapproche d'un

système à minimisation de *jerk* (dérivée de l'accélération; cf. Nelson [1983] et paragraphe 3.6.2). On retrouve là l'idée de minimisation d'un critère d'effort présente dans le modèle global de Kawato *et al.* Cependant ce critère n'est pas global sur l'ensemble de la trajectoire, mais s'applique seulement d'une cible à la suivante. De plus, il concerne les articulateurs pris individuellement.

Venons-en maintenant à l'apprentissage des deux modèles directs cinématique et dynamique. Le premier, qui régit le passage de l'articulatoire à l'acoustique, est appris à partir d'un dictionnaire de correspondances par une technique de rétropropagation de l'erreur entre trajectoires désirée et obtenue, fort similaire aux techniques d'apprentissage proposées par Kawato *et al.* pour l'acquisition des modèles cinématique et dynamique. Le second est tout simplement imposé comme étant du second ordre. Il ne résulte pas d'un apprentissage, mais plutôt de l'observation que les mouvements des articulateurs de la parole présentent des caractéristiques cinématiques du second ordre (cf. 3.6.2).

Enfin il convient de préciser nos choix en ce qui concerne l'implémentation des deux modèles. Nous avons conservé, pour le modèle cinématique, l'implémentation connexionniste mise en œuvre pour l'apprentissage. En effet ce type d'organisation permet de simuler, par des organisations inter-cellulaires, les interactions éventuelles entre articulateurs dans la production d'un signal acoustique. Ainsi si le modèle articulatori-acoustique ne fournit pas d'hypothèse sur les coordinations inter-articulateurs, celles-ci peuvent être éventuellement définies ("à la main") par la structure du réseau (cf. 3.5.3).

Pour le modèle dynamique, le choix connexionniste ne s'imposait pas, les relations entre commandes motrices et trajectoires articulatoires se déduisant aisément de la résolution d'une équation différentielle du second ordre à second membre non constant, par une méthode classique (Runge-Kutta, cf. 3.6.3). Précisons ici que notre but est avant tout de rendre compte de l'aspect fonctionnel du contrôle afin de préciser l'origine de certaines modulations prosodiques. Nous cherchons à déterminer les rôles précis d'un certain nombre de paramètres du contrôle moteur en parole, mais ne cherchons pas à les affecter à tel ou tel élément neuronal réel, nos paramètres n'étant probablement que des symboles. L'approche de Kawato *et al.* [1987] est intéressante et élégante puisqu'elle identifie clairement les zones corticales dévolues à certaines tâches. Mais il nous semble inutile voire dangereux de prolonger systématiquement la métaphore neuronale lorsqu'elle ne fait que compliquer la représentation.

### 2.4.3 Les Points de Passage *vs* les Points d'Équilibre

La figure suivante donne le schéma de modèle de production de la parole proposé par Bateson *et al.* [1993].

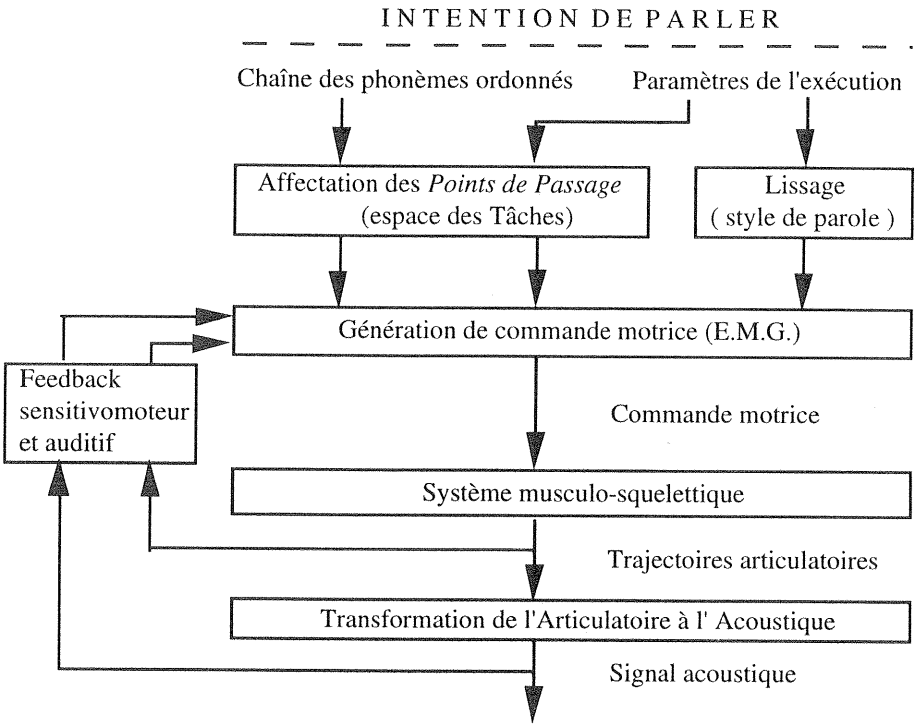



*Figure 2.12.* Le schéma de production de la parole proposé par Bateson *et al.* [1993].

On l'aura compris, notre modèle se différencie de celui-ci, d'abord parce que la phase de génération des commandes motrices n'implique pas la mise en jeu d'une inversion dynamique ; elle n'est pas séparée de la phase d'affectation des points de passage et de spécification des contraintes de lissage. En effet, comme nous l'avons expliqué plus haut, nous avons choisi le contrôle par trajectoire virtuelle, qui évite le calcul du modèle inverse générant les commandes motrices et permet une représentation de l'unité de commande linguistique dans l'espace des commandes motrices et non pas seulement dans l'espace de réalisation physique de la parole. D'autre part, nous nous plaçons dans le cadre théorique de Feldman [1986] (cf. paragraphe 2.3) dans lequel le système nerveux central spécifie des points d'équilibre et non des activités EMG. Les commandes motrices sont donc directement les points d'équilibre (images, pour des forces externes données, des points de passage) et les paramètres prosodiques de raideur et de *timing* (équivalents des contraintes de lissage). Notons que les points d'équilibre ne dépendent pas, dans notre cadre d'hypothèses, des conditions prosodiques : ils sont constants pour une séquence phonémique donnée. La figure suivante reprend le canevas de Bateson *et al.* pour notre schéma du contrôle.

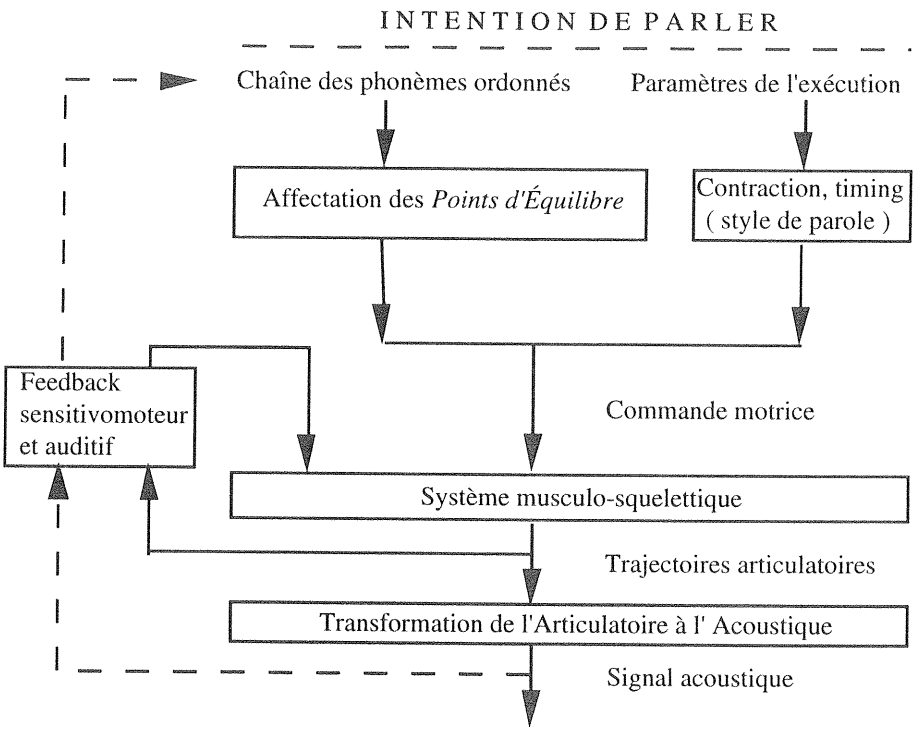



*Figure 2.13.* Notre schéma de production, sur le canevas de Bateson *et al.* [1993].

La rétroaction sensori-motrice est implicitement prise en compte dans le modèle dynamique du second ordre où la force à générer pour produire du mouvement dépend de la différence entre la position actuelle de l'articulateur et la position d'équilibre. Il nous semble d'autre part que la flèche de rétroaction auditive doit se prolonger jusqu'au niveau de l'intention, afin de permettre une réorganisation articulatoire complète, comme dans le cas de perturbations par tube labial ou cale (*bite-block*). Enfin, nous séparons de façon plus radicale que Bateson *et al.* les commandes prosodiques des commandes phonémiques (on remarquera que la flèche allant des paramètres de l'exécution vers les points de passage n'existe plus dans notre schéma). Le nombre de points d'équilibre affectés à un phonème (un, en fait) ne varie pas selon les conditions prosodiques. C'est la durée des plateaux (qui peut tendre vers zéro), la durée des transitions entre plateaux et le niveau de cocontraction qui permettent de moduler la trajectoire articulatoire en fonction des exigences prosodiques. Nous reviendrons en détail au chapitre 4 sur la signification précise de ces divers paramètres prosodiques.

Nous proposons, dans une première phase (chapitre III), de simuler le spécimen classique de variabilités acoustique et articulatoire qu'est la réduction vocalique, en manipulant les paramètres d'intention prosodique du schéma de contrôle. Nous tenterons dans une deuxième phase (chapitre IV) de rattacher précisément les différents paramètres prosodiques de notre schéma au débit et à l'accentuation. Ainsi, à l'aide de ce cadre

général, nous espérons d'une part, donner une vision économique du contrôle en parole, en désignant un petit nombre de paramètres pertinents, et d'autre part, en capturant certaines régularités, proposer une unification des aspects phonétiques —et versatiles— et des aspects phonologiques —et constants.

# CHAPITRE III

# Synthèse Adaptative

Ce Chapitre fait l'objet d'un article dans *Journal of Phonetics* (Perrier, Lœvenbruck & Payan [1996]).

# 3.1 Introduction

La redondance qui caractérise les organismes vivants permet une certaine flexibilité de leurs comportements. Schmidt [1988] indique les conséquences de cette redondance pour le contrôle moteur :

*"An important concept in biology and evolution is that organisms are structured with a great deal of redundancy, or duplication, so that various parts of the central nervous system can be destroyed with little or no loss in behavioral capabilities. Redundancy appears to have application to the motor behavior as well, as there seems to be a number of ways that the system can perform a certain action, with performance being only slightly impaired when the primary system is fatigued or damaged."*

C'est la redondance, ou plasticité, qui permet à l'être humain d'*adapter* ses comportements à son environnement social et biologique. Pour le sociologue E. Morin [1984], la société humaine est auto-organisatrice en ce sens qu'elle essaie, en évoluant, d'échapper aux (ou d'assumer les) perturbations de l'environnement naturelles (sécheresse, famine, épidémie, etc.) ou non naturelles (guerres, conflits), aux dégradation des artefacts (maisons, outils, machines) et aux conflits entre individus, groupes ou classes, tout en maintenant une organisation invariante (lois, structures).

Le terme "adaptatif" a été introduit en biologie pour dénoter la plasticité comportementale montrée par un organisme dans sa lutte pour survivre dans un environnement nouveau ou changeant (Sommerhof [1950]). Cette notion est reprise, dans son sens biologique, en robotique et en automatique, pour des contrôleurs présentant des capacités similaires à modifier leurs stratégies comportementales face à des changements imprévisibles du système contrôlé ou de ses entrées (Gaines [1969]).

Du point de vue physiologique, de nombreuses activités du corps humain fonctionnent de manière adaptative. Les modulations de l'activité cardiaque sont un exemple parlant (cf. par exemple Houdas [1990]). La fièvre, les émotions, l'exercice physique, le stress, le tabac, font augmenter la fréquence cardiaque. Le cœur *adapte* ses activités en fonction des besoins de l'organisme. Il existe en effet dans le quatrième ventricule, situé dans le bulbe rachidien, des zones nerveuses centrales *régulant* l'activité cardiaque. Le centre *cardio-modérateur,* situé dans la région du plancher, est responsable de la diminution de la fréquence cardiaque et de la pression artérielle. Les zones latérales et médullaires ont elles des effets sur le diamètre des vaisseaux sanguins et permettent ainsi d'augmenter le débit cardiaque, ce sont les zones cardio-vasculaires *excitatrices* du cœur et des vaisseaux. Les zones nerveuse centrales agissent par l'intermédiaire de nerfs moteurs, parasympathiques et orthosympathiques. Le rythme cardiaque résulte donc d'un *équilibre*

entre l'automatisme du myocarde, l'action modératrice du système parasympathique et celle accélératrice du système orthosympathique.

L'*adaptation* (du latin *ad* (à) *aptus* (apte), participe passé de *apere* (lier, attacher)) signifie, au sens biologique, physiologique, psychologique ou sociologique, la modification d'un organisme vivant le rendant apte à son milieu, à sa situation. En ce sens, tout comme le comportement social ou cardiaque le comportement "parlé" est adaptatif, c'est l'hypothèse de Lindblom [1988], que nous avons présentée au chapitre I. Cette adaptativité est possible car, comme la plupart des phénomènes biologiques, le langage est redondant (cf. Hockett [1965] et Lindblom *et al.* [1992]). C'est ce qui le protège des dégradations éventuelles du signal :

*"[...] language is redundant at all levels of structure. It codes information in multiple and overspecified ways, thereby acquiring strong protection against signal degradation and providing a mechanism for dealing with partial and reduced signal information."* (Lindblom *et al.* [1992]).

L'idée que la parole est adaptative est déjà présente chez MacNeilage [1970] qui parle de *variabilité contrôlée* :

*"[...] the essence of the speech production process is not an inefficient reponse to invariant central signals, but an elegantly controlled variability of response to the demand for a relatively constant end."*

Lindblom *et al.* [1992] précisent ce point de vue sur les adaptations en-ligne que connaît la parole :

*"[...] the transforms that speech signals undergo may reflect the speaker's attempt to adapt his articulatory/acoustic output to various on-line functional social and communicative demands. Varying on a moment-to-moment basis, those demands are monitored by the speaker. They tune phonetic performance producing signals varying on a continuum poor-to-rich in physical clues about the linguistic structure of the utterance."*

Dans sa théorie H&H (pour *Hyperspeech & Hypospeech*), Lindblom [1990] désigne les deux exigences fondamentales et contradictoires sous-jacentes à ce comportement adaptatif : l'exigence de minimisation du coût articulatoire du point de vue du locuteur et l'exigence d'un minimum de contraste perceptif du point de vue de l'auditeur. L'adaptation résulte donc d'une interaction, ou même négociation, locuteur/auditeur. La parole varie selon un continuum H&H :

*"In the ideal case, the speaker [...] dynamically tunes the production of its elements to the short-term demands for either output-oriented control (hyperspeech) or system-oriented control (hypospeech). What he/she needs to control is -not that linguistic units are actualised in terms of physical invariants (higher-order or whatever) - but that their signal*

*attributes possess sufficient contrast, that is discriminative power that is sufficient for lexical access."*

Selon Lindblom, lorsque les contraintes de réception sont dominantes, l'articulation du locuteur s'efforce d'être précise et soignée, les hyperformes apparaissent, tandis que lorsque ce sont les contraintes de production qui prédominent, le locuteur cherche à minimiser son effort et les hypoformes sont observées. Lindblom donne une classification des facteurs responsables de la variation intra-locuteur dans un schéma que nous reprenons dans la figure 3.1.

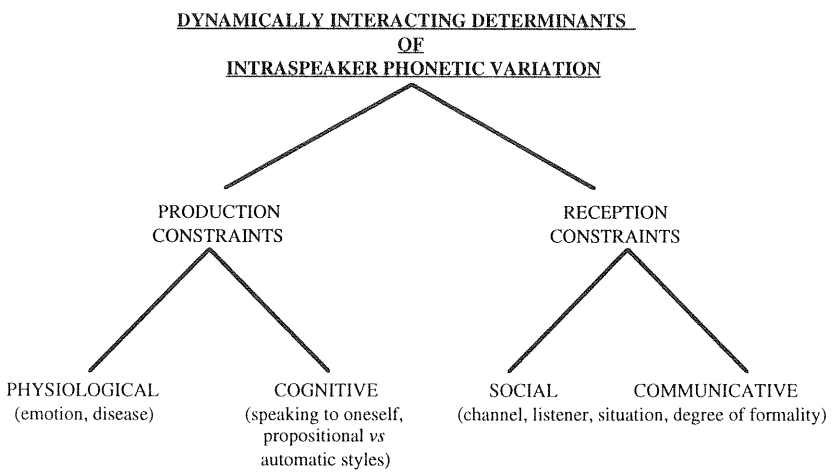



*Figure 3.1.* Classification des facteurs de la variation intra-locuteur. D'après Lindblom [1990].

Notre robot parlant, s'il veut être un tant soit peu compétitif face à l'être humain, doit faire montre d'une telle *adaptivité* ou *flexibilité*. Un bel exemple de flexibilité intra-locuteur est le phénomène classique de *réduction vocalique*, au cours duquel la réalisation formantique des voyelles est dégradée. Dans ce chapitre, nous présentons un ensemble de simulations au cours desquelles le robot adapte son énonciation aux contraintes de débit ou de précision, en ajustant certains paramètres ou commandes centrales. Ainsi à partir d'une même séquence de parole, le robot génère des produits acoustiques variés, qualifiables de hypo- à hyper-articulés, ou de réduits à contrastés. À la lumière de cette expérience de *synthèse adaptative* —où la variabilité est simulée par un ajustement de commandes prosodiques, sans modification des commandes phonémiques— nous cherchons à différencier l'*objectif* à atteindre de la *manière* d'atteindre cet objectif.

La notion de réduction vocalique est précisée au paragraphe 3.2 et les paramètres qui la modulent sont décrits.

Nous présentons ensuite au paragraphe 3.3 le corpus acoustique qui met en œuvre un phénomène manifeste de réduction vocalique.

La méthode d'inversion en chaîne du signal acoustique pour obtenir les commandes centrales est décrite au 3.4. Les deux étapes qui la composent font l'objet des paragraphes 3.5 et 3.6. La première inversion, qui vise à récupérer la trajectoire articulatoire à partir du signal acoustique, utilise un modèle articulatoire du conduit vocal décrivant les relations entre les espaces articulatoire et acoustique en parole. La seconde étape, qui permet d'inférer les commandes motrices à partir de la trajectoire articulatoire, s'appuie sur un modèle fonctionnel simple des articulateurs de la parole, qui code les objectifs articulatoires en termes de points d'équilibre et paramétrise le mouvement, en direction de ces objectifs, à l'aide de commandes temporelles et dynamiques.

Une synthèse des formants acoustiques à partir des commandes centrales obtenues par l'inversion en chaîne est mise en œuvre au paragraphe 3.7 pour vérifier la pertinence de ces commandes.

Enfin l'expérience de synthèse adaptative est présentée au paragraphe 3.8. On montre alors qu'en altérant les commandes motrices prosodiques obtenues par inversion chaînée, le robot parlant est capable de générer de la variabilité acoustique. Il est ainsi possible de simuler des effets de réduction vocalique comparables à ceux que l'on observe dans le corpus enregistré.

## 3.2 La réduction vocalique

### 3.2.1 Intérêt et définition de la réduction vocalique

La réduction vocalique est un cas exemplaire de variabilité intra-locuteur. À ce titre, elle a été étudiée par de nombreux chercheurs en parole. Les formes affaiblies des phonèmes lorsqu'ils apparaissent dans la parole continue et non-accentuée ont intéressé les phonéticiens et le phénomène de réduction vocalique a d'abord été examiné sous l'angle de l'*accentuation*.

**Le triangle vocalique**

La réflexion linguistique et phonétique sur la structure phonémique des langues européennes a conduit à diverses descriptions et classifications des voyelles à partir de voyelles de références choisies arbitrairement (Gleason [1955]). Parmi toutes ces classifications, celle de D. Jones [1940], qui s'appuie sur un ensemble de huit voyelles de référence, les *voyelles cardinales,* est couramment utilisée, car elle présente l'avantage de relier les propriétés articulatoires et acoustiques. Dans ce système, les critères utilisés sont, du point de vue articulatoire, la hauteur de la langue, la position (avant ou arrière) du point le plus élevé de la langue et l'arrondissement des lèvres, et du point de vue acoustique, les

valeurs des deux premiers formants F1 et F2. L'augmentation de F1 correspond à une ouverture de la bouche et un abaissement de la langue, celle de F2 à une avancée de la langue. Les huit voyelles cardinales, représentées dans le plan F1/F2 (ou hauteur/avancée) forment un triangle dont les extrémités correspondent aux voyelles /i,a,u/ et dont le centre correspond à la voyelle neutre *schwa*, cf. figure 3.2. Sur cette même figure, les autres voyelles du français ont été représentées, elles sont incluses dans le triangle (ou pentagone) cardinal (Delattre [1948]).

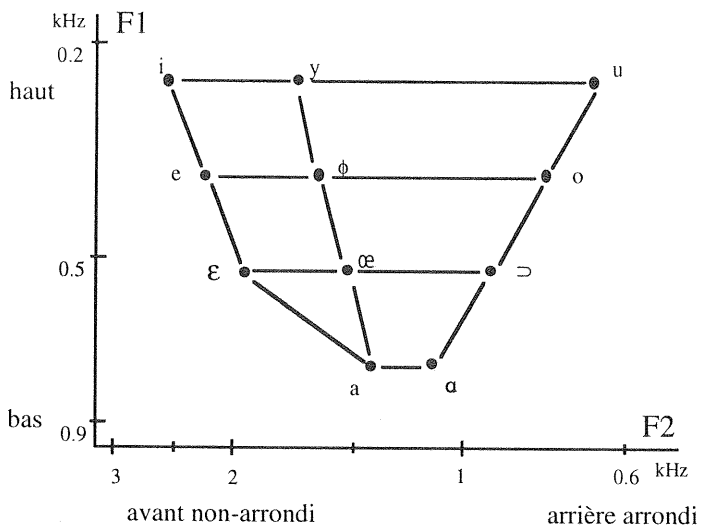



*Figure 3.2.* Triangle vocalique des voyelles du français, d'après Delattre [1948].

**La réduction vocalique**

Cette classification des voyelles ne fait que donner des points de repère. L'analyse des aspects dynamiques de la production des voyelles, *i.e.* l'analyse des voyelles prononcées au sein d'une conversation en débit normal et non pas isolées et spécialement articulées (dans le cadre d'un laboratoire), a révélé la grande variabilité à la fois des valeurs des formants et des formes du conduit vocal (Tiffany [1959], Shearme & Holmes [1962], Delattre [1969], Stålhammar, Karlsson & Fant [1973], Koopmans Van Beinum, [1980], etc.). Deux phénomènes sont découverts : la *centralisation* (ou *neutralisation)* et la *coarticulation* ou (*assimilation contextuelle*). D'une part, les voyelles, en position syllabique non-accentuée, ont tendance à occuper une position plus centrale dans le plan vocalique F1/F2, elles se déplacent en direction de la position neutre du *schwa*. On parle alors de *centralisation*. D'autre part, les contrastes intervocaliques (*i.e.* distances dans le plan F1/F2) qui peuvent être observés sur les voyelles isolées peuvent varier fortement lorsque ces voyelles sont prononcées à débit normal et en parole continue. Il s'agit là de *coarticulation*.

### 3.2.2 Un premier modèle de réduction vocalique : Lindblom [1963]

À l'époque où Lindblom propose son modèle, s'il est reconnu qu'accentuation et durée de la voyelle sont intimement liées, peu d'études existent sur les variations acoustiques dues au changement de débit de parole. Les recherches se concentrent plutôt sur les effets de la réduction d'accentuation. Lindblom [1963] est le premier à examiner de façon systématique la dynamique de la production des voyelles dans diverses conditions d'accentuation et de *débit*. Il observe ainsi les variations de fréquences formantiques lorsque le débit de parole augmente ou que le degré d'accentuation diminue, sur la voyelle /u/ par exemple, dans trois contextes consonantiques différents ([dud], [bub], [gug]). Ses résultats montrent que ces deux modifications prosodiques induisent un déplacement de la configuration formantique depuis le coin supérieur droit du triangle (position standard du /u/, cf. figure 3.2) vers le centre : il est en présence de réduction vocalique, essentiellement caractérisée par une augmentation de F1 et de F2.

La suggestion de Lindblom repose sur l'hypothèse que la production de la parole vise à atteindre des cibles successives correspondant à des phonèmes successifs. Dans cette perspective, la réduction vocalique correspondrait à un *undershoot* (ratage) de la cible vocalique planifiée. Selon lui, la réduction vocalique dépend de trois facteurs :

- le contexte consonantique adjacent,
- la cible vocalique planifiée,
- la durée de la voyelle.

Il formule cette idée de façon compacte en décrivant la relation entre $F2_0$ (valeur du formant F2 au point central de la réalisation temporelle de la voyelle), la durée, le contexte consonantique et la nature de la voyelle :

$$F2_0 = k(F2_i - F2_t).e^{-aT} + F2_t$$

où - $F2_t$ : est la cible de F2 à atteindre pour la voyelle considérée,

- k et a : dépendent du contexte consonantique (mais pas de la voyelle),

- $F2_i$ : est la valeur de F2 au début du segment vocalique; elle dépend de la voyelle et de la consonne précédente,

- T : est la durée de la voyelle.

Une relation similaire est donnée pour F1.

Selon ce modèle, pour un contexte donné, ce serait une réduction de la durée de la voyelle (qu'elle soit due à une augmentation du débit ou à une réduction de l'accentuation) qui empêcherait le système articulatoire d'exécuter le geste complet requis : les trajectoires articulatoires et formantiques n'atteignent pas leurs cibles (*undershoot*).

Citons Lindblom lui-même :

*"[...] it is immaterial whether a given length of the vowel is produced chiefly by the tempo or the degree of stress. Duration seems to be the main determinant of the reduction."*

### 3.2.3 Révisions du premier modèle

De nombreux modèles de réduction vocalique ont été proposés depuis et ont ouvert le débat sur la nature de la *cause primitive* de ce phénomène.

D'abord, comme le fait remarquer Nord [1986], Lindblom lui-même [1968] introduit dans son modèle un facteur de "force" en plus du facteur de durée. Il se rapproche là du point de vue de Delattre [1969, cité par Nord [1986]] sur la question :

*"stress and tempo are the primary determinants of vowel reduction."* et *"duration a product of stress and tempo and therefore a second determinant of vowel reduction."*

Harris [1975], à partir d'une étude sur les mécanismes de modification de la durée, et Nord[1975] dans une première analyse de la réduction vocalique en suédois, suggèrent tous deux que l'accentuation, et non la durée, serait le premier déterminant dans la façon d'atteindre les cibles acoustiques.

Kuehn & Moll [1976] mesurent les effets du débit de parole sur la vitesse et l'amplitude de mouvements articulatoires (apex et dos de la langue, lèvre inférieure), recueillis par cinéradiographie, et montrent que le degré de réduction peut varier suivant les locuteurs :

*"With an increase in speaking rate, some speakers did not change velocity or actually reduced velocity and thus decreased displacement, while others increased velocity which resulted in less "undershoot" of the articulatory position achieved for the phone at a slower rate."*

Dans une étude de syllabes CVC et CVCVC prononcées dans différentes conditions de débit et d'accentuation, Gay [1978] remet clairement en cause l'idée originale de Lindblom [1963] et montre que les cibles acoustiques des voyelles (les valeurs formantiques au point central) ne varient pas nécessairement en fonction du débit de parole. Toutefois, si pour les voyelles accentuées, seule la durée est affectée par l'augmentation du débit, il n'en est pas de même pour les voyelles non-accentuées dont l'amplitude globale, la fréquence fondamentale et la couleur (F1/F2) sont à la fois réduites. Gay est donc conduit à une conclusion *opposée* à celle de Lindblom [1963] :

*"Because the tendency for formant frequencies to be reduced toward the neutral vowel schwa occurs only for an unstressed vowel, even if it is of the same duration as its stressed counterpart, the present data suggest that the degree of reduction is linked to stress, regardless of the relative or absolute duration of the segment."*

Nord [1986] précise les conséquences acoustiques de la réduction vocalique et distingue deux types de phénomène : centralisation (neutralisation) et coarticulation (assimilation). On peut en effet observer soit une "neutralisation" des propriétés articulatoires de la voyelle, induisant une "centralisation" dans le triangle acoustique (vers la position neutre du *schwa*), soit une plus grande influence des phonèmes adjacents, *i.e.* une moins grande résistance à "l'assimilation" par le contexte, qui se traduit par des effets de "coarticulation" marqués. Il conclut de son analyse sur des mots de deux syllabes, qu'en suédois, la centralisation est observée plutôt en position finale et la coarticulation plutôt en position inter-consonantique. Il note par ailleurs que les voyelles non-accentuées présentent des phénomènes de réduction vocalique, quelle que soit leur durée :

*"[...] irrespective of their duration, unstressed vowels coarticulate strongly with context: in non-final syllable position with surrounding phonemes and in final syllable position with a neutral position corresponding to a centralized schwa vowel."*

Engstrand [1988] confirme cette hypothèse à partir d'une étude de l'activité articulatoire observée pour des modifications de débit et d'accentuation. Ses données cinéradiographiques et acoustiques sur des séquences VCV (où les voyelles sont choisies parmi /i/, /a/ et /u/) prononcées dans diverses conditions de débit et d'accentuation, indiquent que c'est l'accentuation, et non le débit de parole, qui joue sur les caractéristiques spectrales et articulatoires des voyelles étudiées :

*"The spectral characteristics of the tense vowels /i a u/ were significantly influenced by stress but not by speaking rate [...] stressed vowels displayed narrower tongue constrictions of the oral tract than did unstressed vowels at both speaking rates studied".*

Ces travaux indiquent clairement que le premier modèle de Lindblom qui, dans sa version extrême, considère la réduction comme dépendant primitivement de la durée, ne permet pas de décrire de façon satisfaisante les données empiriques.

### 3.2.4 La synthèse de Lindblom *et al.* [1992]

En 1992, Lindblom, Brownlee, Davis & Moon ont révisé le modèle original de réduction en y introduisant le *style* de parole (le discours clair ou informel, les formes isolées, etc.) :

*"[...] vowel reduction seems to be more than simply "durationally induced contextual assimilation"."*

Moon [1991] a en effet examiné les effets de réduction liés à la durée pour des mots d'une, de deux ou trois syllabes, dont la première est toujours accentuée, et pour deux styles de parole : confortable et hyper-articulé. Pour rendre compte des effets spectraux,

Moon réutilise la formule mathématique originale de Lindblom [1963] (cf. 3.2.2) en ajoutant une nuance : les valeurs de k, a et $F2_t$ dépendent cette fois du style.

Lindblom *et al.* en viennent ainsi à reformuler la première hypothèse de Lindblom :

*"Reduction processes can be seen as contextual assimilations durationally induced, but, within certain limits, speakers appear capable of controlling the precise degree of reduction."*

### 3.2.5 Un point de vue différent

Les propositions précédentes sont remises en question par Van Bergem [1993] et Pols & Van Son [1993] qui, comme nous l'avons indiqué au chapitre II, réfutent l'hypothèse d'une production de la parole orientée vers des *cibles*. Ils suggèrent que la production de la parole consiste à générer des caractéristiques pertinentes dans les parties *dynamiques* des trajectoires formantiques (plutôt que dans les parties stables). Selon eux, la durée de la voyelle ne doit pas être considérée comme la cause primitive de la réduction vocalique, mais comme une des conséquences du contrôle des transitions, différent selon le style de parole :

*"The stressed vowel tokens were generally longer and less reduced [...] than the unstressed ones [...]. However vowel duration alone was not enough to explain those differences. It is probably the other way round: stress, context and speaking style result in certain formant and duration changes, and are for the greater part actively controlled by the speaker."* (Pols & Van Son [1993]).

Dans ce débat, notre but est de proposer un modèle quantitatif d'une production des voyelles s'appuyant sur des cibles invariantes et permettant de générer de la variabilité à partir de ces cibles, en contrôlant la durée, le contexte et le style de parole. Pour cela, un corpus de réduction vocalique a été enregistré, mettant en œuvre diverses conditions de débit et d'accentuation.

## 3.3 Corpus

Notre corpus consiste en la séquence [iai] dans la phrase porteuse "il **y a** immédiatement éternué" enregistrée pour un locuteur Français masculin (JLS) (cf. pour plus de détail sur la mise en place du corpus, Schwartz, Beautemps, Arrouas & Escudier [1992] et Beautemps [1993]). Trois conditions d'énonciation sont étudiées permettant de faire varier le débit de parole et l'accentuation. Ce corpus a été choisi parce qu'il met en œuvre un geste articulatoire allant de la voyelle antérieure non-arrondie haute qu'est le [i], à la voyelle moins postérieure non-arrondie basse qu'est le [a]. La direction du geste articulatoire est donc de l'extrémité gauche vers le bas du triangle vocalique.

La voyelle [a] existe-t-elle? Existe-t-il une forme absolue vers laquelle tendent les voyelles [a] de tous les contextes phonémiques? Daniloff & Hammarberg [1973] répondent par l'affirmative à cette question en supposant l'existence de *formes canoniques*, invariantes, idéales et non-coarticulées.

"*[The best approximation of an ideal canonical form occurs] when a segment is produced in isolation in a sustained manner, or when the sound is produced in a context assumed to be minimally coarticulatory.*"

Fowler [1980] réfute complètement cette hypothèse. Pour elle, les formes canoniques incluent une dimension temporelle et toute production tient compte des productions antérieures et postérieures "*[...] phonological segments are considered essentially or canonically four dimensional.*"

Nous adhérons au point de vue de Fowler et considérons avec elle que les cibles doivent tenir compte du contexte. C'est pourquoi il est important de préciser que deux types de variabilité peuvent être évoqués à propos de notre corpus :

- la variabilité due au *contexte phonémique* : la production de la voyelle [a] tient compte de l'environnement symétrique des voyelles [i]. Par la présence de ces [i], la configuration articulatoire est différente par exemple de celle qui viserait à produire la voyelle [a] entourée de voyelles [u] (cf. les documents cinéradiographiques sur les voyelles du français de l'Institut de Phonétique de Strasbourg : Bothorel, Simon, Wioland & Zerling [1986]).

- la variabilité *prosodique* : la vitesse d'élocution et l'accentuation influencent chacune à leur façon les configurations articulatoires effectives.

C'est à cette dernière variabilité que nous nous intéressons dans le corpus présenté ici.

**Consignes données au locuteur**

L'accent que nous considérons ici est un accent d'*emphase*. Il est demandé au locuteur d'insister particulièrement sur l'auxiliaire "a". Dans la première condition (*lente accentuée*), on a demandé au locuteur de parler lentement et d'accentuer la voyelle [a] ; dans la deuxième condition (*lente non-accentuée*), l'instruction est de parler lentement sans accentuer particulièrement le [a] ; enfin dans la dernière condition (*rapide accentuée*), l'instruction est de parler rapidement en accentuant la voyelle [a].

Pour la vitesse d'élocution, la consigne est de parler posément dans les cas lents et le plus vite possible (tout en restant compréhensible) dans le cas rapide. On évite ainsi d'avoir un débit trop lent pour les premiers cas et un débit à peine accéléré pour le cas rapide, ce vers quoi les locuteurs tendent naturellement. Les séquences [iai] des deux premières conditions durent respectivement 440 ms et 400 ms pour les cas accentué et non accentué.

Ce qui donne un débit de 6 à 7 voyelles par seconde. La séquence du cas rapide accentué dure 280 ms soit un débit de 10 voyelles par seconde.

**Acquisition des données**

Le signal acoustique a été enregistré en chambre sourde et numérisé à la fréquence d'échantillonnage de 10 kHz. Les formants sont obtenus par analyse LPC à 16 coefficients (Beautemps [1993]). L'extraction de la séquence [iai] se fait à partir des formants, mais on vérifie que le résultat est correct du point de vue acoustique. Les absences de détection de formants, dues à un niveau d'énergie trop faible, sont corrigées par interpolation linéaire. Un premier essai d'inversion à partir de formants non lissés ayant donné des trajectoires articulatoires très bruitées, un lissage avec une fonction *spline* est effectué sur les quatre premiers formants.

La figure 3.3 donne les sonagrammes obtenus dans les trois conditions d'élocution enregistrées, dans l'ordre : lente accentuée, lente non-accentuée, et rapide accentuée. Les formants corrigés et lissés sont représentés par les croix blanches.

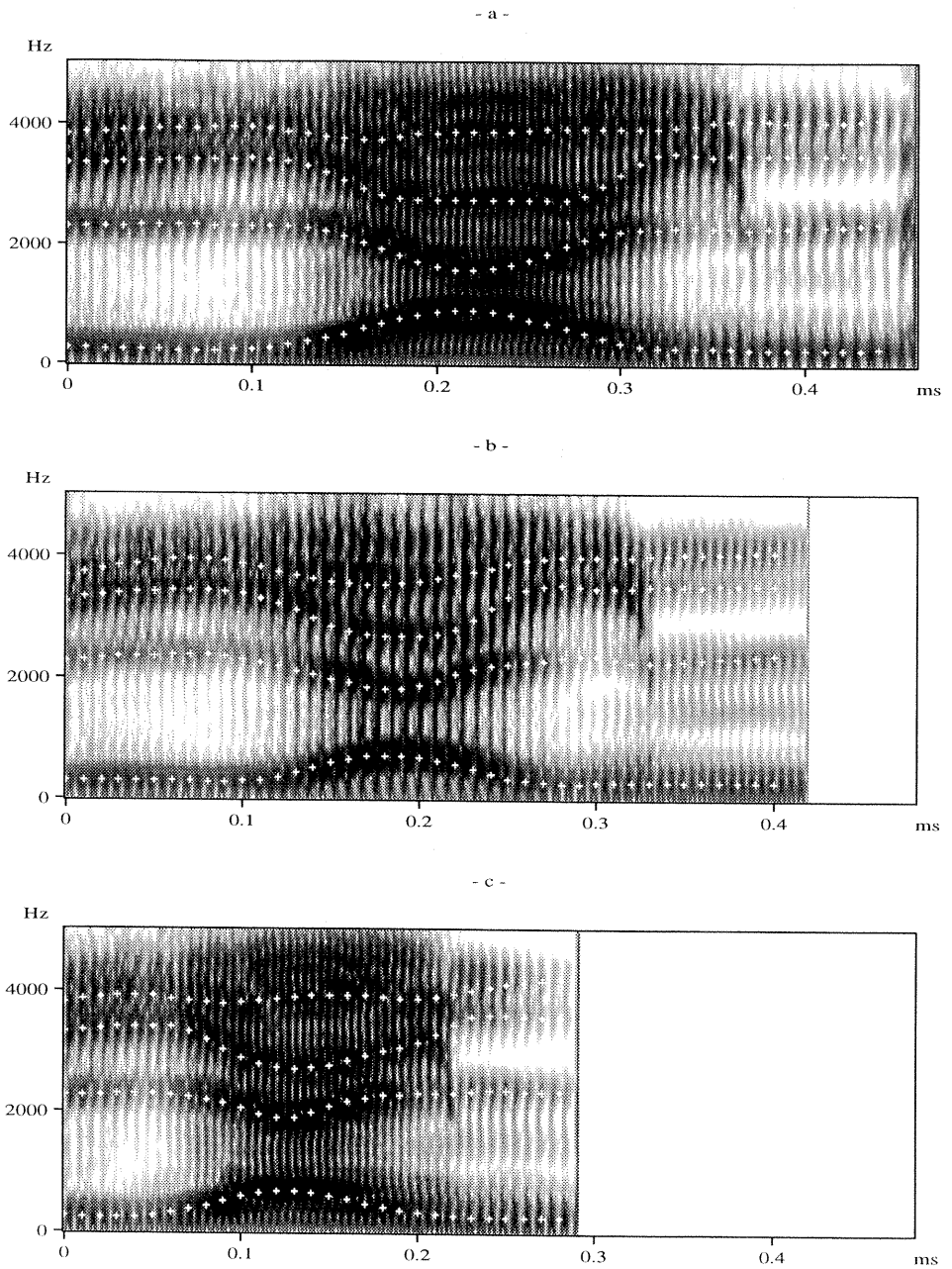



*Figure 3.3.* Sonagrammes de la séquence [iai] produite par le locuteur JLS dans trois conditions d'élocution. a : lent accentué, b : lent non-accentué, c : rapide accentué.

Le rapprochement F1/F2 caractéristique de la voyelle [a] est bien présent pour le cas lent accentué et l'on note bien la présence d'un palier pour cette voyelle : nous considérons donc ce cas comme "idéal", en ce sens que les trajectoires formantiques ne semblent pas souffrir de réduction vocalique. Dans les deux autres cas, le rapprochement est beaucoup moins prononcé : on est en présence d'un phénomène de réduction vocalique. Les valeurs formantiques du [a] dans les deux cas "non idéaux" se rapprochent en fait de celles d'un [ɛ]. Ceci est en accord avec les analyses présentées au paragraphe 3.2 sur l'effet

d'assimilation contextuelle ou coarticulation. La figure 3.4 qui donne les trajectoires de [iai] dans le plan F1/F2 pour les trois conditions d'élocution permet de mieux représenter cet effet. Pour le cas lent accentué (courbe en trait plein) la trajectoire fait une boucle passant par [i] puis [a] puis [i]. Dans les deux autres cas, les boucles (trait tireté pour lent non-accentué et pointillé pour rapide accentué) n'atteignent pas le [a], elles s'arrêtent en chemin, entre le [i] et le [a], c'est-à-dire près du [ɛ]. La même échelle logarithmique a été utilisée que pour le triangle vocalique classique (cf. paragraphe 3.2.1). De plus, les échantillons initiaux et finaux de chaque séquence sont supprimés. Les plateaux des /i/ entraînent en effet des stagnations dans le plan F1/F2 qui rendent le tracé confus.

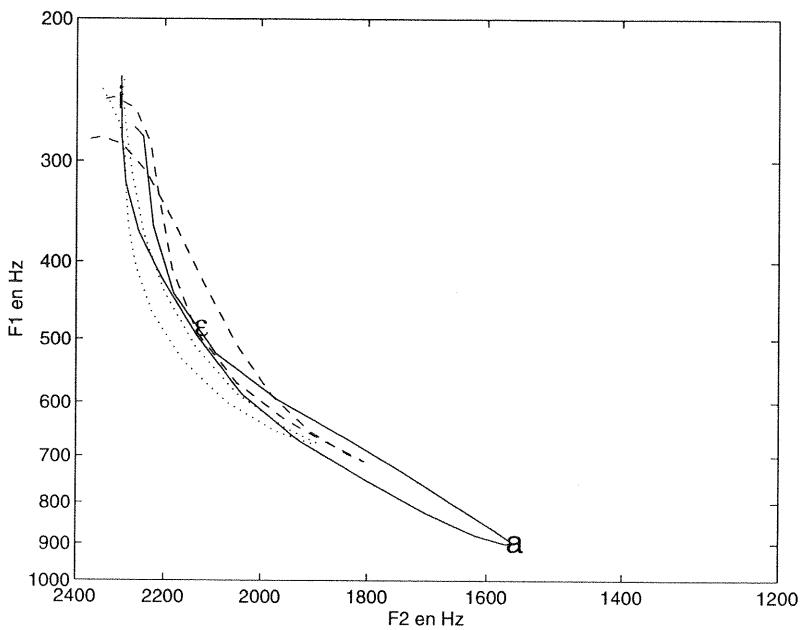



*Figure 3.4.* Séquences [iai] dans le plan traditionnel F1/F2 pour les trois conditions d'élocution. Lent accentué : trait plein, lent non-accentué : trait tireté, rapide accentué : trait pointillé.

# 3.4 Récupération des commandes centrales

Une étude préliminaire, dans notre tentative de simuler une des manifestations de variabilité intra-locuteur qu'est la réduction vocalique, consiste en la récupération des valeurs des paramètres de commande dans le cas prosodique idéal. Notre hypothèse est que la variation d'une partie de ces paramètres de commande, associée aux changements de style d'élocution, est responsable de la réduction vocalique observée. Ces paramètres sont recherchés parmi les commandes d'un modèle contrôlant les mouvements des articulateurs de la parole.

Il s'agit donc de récupérer les commandes à partir de grandeurs physiquement mesurables comme le signal acoustique ou les trajectoires articulatoires. C'est un problème typique d'inversion, comme celui posé en 1966 par le mathématicien américain M. Kac et résumé sous la formule : "peut-on entendre la forme d'un tambour?", c'est-à-dire, comme l'explicite Fabre [1988], peut-on, à partir de son spectre vibratoire, récupérer les caractéristiques géométriques d'un objet? Les travaux de Marr et Poggio sur la vision ont permis des avancées remarquables dans le domaine de l'inversion en général. Pour Poggio [1984], tout problème de perception visuelle bas-niveau est formellement un problème d'inversion :

*"Most of the goals of low-level vision can be seen as the solution to inverse problems. Consider, for instance, the problem of recovering the three-dimensional structure of a scene from the images of it. While in classical optics the problem is to determine the images given certain physical objects, we are confronted here with the inverse problem of finding their three-dimensional shape (and perhaps their physical properties) from the light intensity distribution in the image."*

Cependant, ces problèmes d'inversion sont **mal-posés** (au sens de Hadamard [1923]), c'est-à-dire que plusieurs objets tridimensionnels différents peuvent correspondre à une même image bidimensionnelle. En somme, comme est intitulé l'article de Cipra [1992], *"You can't hear the shape of the drum"*. Selon Poggio [1984], il est donc nécessaire d'introduire des contraintes et des hypothèses de régularité pour réduire l'espace fonctionnel des solutions acceptables.

Le second volet de ces avancées sur l'inversion est établi par Marr [1982] qui propose trois niveaux pour l'étude des systèmes humains complexes de traitement de l'information :

1. Le niveau de la théorie computationnelle, *i.e.* l'objectif des traitements.
2. Le niveau de l'algorithme, *i.e.* la façon dont s'effectuent les traitements.
3. Le niveau de l'implémentation, *i.e.* le substrat physiologique des traitements.

L'approche de Marr & Poggio, selon laquelle, pour comprendre le fonctionnement des processus humains de traitement de l'information, on doit s'intéresser à la fois aux aspects biophysiologiques et aux contraintes du monde physique, a été utilisée avec succès en perception visuelle —détection de contours (Torre & Poggio [1984], vision stéréoscopique (Marr & Poggio [1979], Marr [1982])— ainsi que dans de nombreux domaines, comme l'animation de visage avec la récupération des commandes faciales (Terzopoulos & Waters [1990]) et le contrôle des mouvements volontaires du bras avec le modèle computationnel de Kawato *et al.* [1987].

C'est dans ce cadre d'inversion que nous nous plaçons pour, en reprenant la métaphore de notre introduction, décoder la partition, qui a été composée en temps réel par

le Système Nerveux Central, à partir de son interprétation par le système musculo-squelettique. De façon plus précise, notre but est d'apporter quelques éléments de réponse à la question formulée lors du lancement du projet *Speech MAPS* (cf. Introduction), dont la forme est loin d'être innocente : "Un robot articulatoire peut-il apprendre à produire des gestes articulatoires à partir de sons?"

Il s'agit donc de récupérer les commandes qui assurent le contrôle du robot articulatoire à partir de signaux acoustiques émis par un locuteur humain. Pour cela, nous procédons en deux phases d'inversion successives. La première, traitée en détail au paragraphe 3.5, convertit les trajectoires acoustiques distales en trajectoires articulatoires proximales. La seconde, décrite au paragraphe 3.6, fournit les commandes motrices correspondant à la trajectoire d'un articulateur pertinent du point de vue acoustique.

La première phase utilise un modèle reliant positions des articulateurs et forme du conduit vocal, doublé d'un modèle harmonique acoustique associant fonction d'aire du conduit vocal et formants. Le modèle direct complet, permettant de passer des paramètres articulatoires aux formants, est purement géométrique : les concepts de force et de masse n'apparaissent nulle part. Nous qualifions ainsi cette première inversion de *cinématique* puisqu'elle vise à inférer les *mouvements* des articulateurs, c'est-à-dire la *géométrie* du conduit vocal, dans le temps. Cette inversion mal-posée est résolue par des contraintes de lissage sur les paramètres articulatoires de sortie (cf. 3.5.3).

La deuxième phase s'appuie sur un modèle (décrit au 3.6.2) représentant fonctionnellement la dynamique des articulateurs et reliant les mouvements aux *forces* qui les provoquent. On s'y réfère donc par le terme d'inversion *dynamique*. Des contraintes d'ordonnancement temporel sont imposées sur les commandes motrices (cf. 3.6.3).

Nous espérons, à l'aide de la procédure d'inversion globale, donner une vue "parcimonieuse" du contrôle du mouvement en parole, en désignant un petit nombre de paramètres responsables de la multitude des formes observées, tant articulatoires qu'acoustiques.

## 3.5 Inversion cinématique : des formants aux trajectoires articulatoires

### 3.5.1 Le modèle direct : des paramètres articulatoires aux formants

La récupération des trajectoires articulatoires à partir du signal acoustique utilise un modèle articulatoire couplé à un modèle acoustique. Le modèle articulatoire fournit la coupe sagittale du conduit vocal à partir d'un certain nombre d'articulateurs élémentaires.

Le modèle acoustique génère des formants à partir de la fonction d'aire correspondant à la coupe sagittale proposée par le modèle articulatoire.

### 3.5.1.1 Le modèle de Maeda

Le modèle articulatoire utilisé est celui de Maeda ([1979], [1990]). Ce modèle est fondé sur une analyse factorielle de données sur le profil sagittal du conduit vocal ainsi que de données vidéo de face sur les lèvres (Bothorel, Simon, Wioland & Zerling [1986]). L'hypothèse sous-jacente à ce modèle est que les activités complexes des organes articulatoires de la parole sont organisées autour d'un nombre limité de blocs fonctionnels contrôlables indépendamment : les articulateurs élémentaires (Maeda [1990]). Une grille semi-polaire est utilisée pour décrire la géométrie des contours sagittaux, la forme de la langue est ainsi décrite par un vecteur à 30 dimensions. L'analyse statistique factorielle permet, à partir des coordonnées des points d'intersection entre la grille et les différents contours, d'extraire sept composantes principales représentant justement l'influence géométrique de ces articulateurs élémentaires : la mandibule, les lèvres (aperture et protrusion), le corps, le dos et l'apex de la langue et le larynx (mouvement vertical). Le panneau de gauche de la figure 3.5 montre le tracé d'un contour sagittal du conduit vocal sur lequel est superposée la grille de mesure semi-polaire, le panneau de droite donne un modèle de contour à partir du vecteur à 30 dimensions.

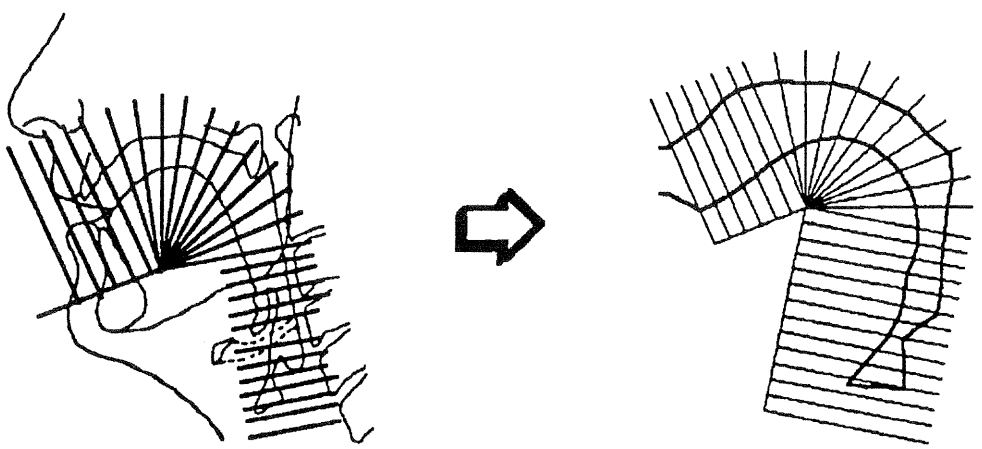


*Figure 3.5.* Analyse et restitution du conduit sagittal, d'après Boë [1993].

Les paramètres *mandibule*, *protrusion des lèvres* et *larynx* proviennent directement de cette analyse factorielle. Les paramètres de langue (*corps, dos, apex*) sont déterminés en soustrayant l'influence du paramètre *mandibule* sur la forme de la langue. Le paramètre

*corps* semble plutôt représenter la *position* (avant/haute ou arrière/basse) de la langue (influences horizontale et verticale), alors que le paramètre *dos* indique lui la *forme* (bombée ou plate) de la langue (influence verticale). L'influence de la mandibule sur l'ouverture des lèvres n'a pas été soustraite du paramètre *hauteur des lèvres*. Les paramètres de langue et la mandibule expliquent 88% de la variance des contours observés. La distribution de cette variance est la suivante : 15% pour la mandibule, 43% pour le corps, 23% pour le dos, 7% pour l'apex. L'influence du larynx est de 4%.

Un produit matriciel permet, à partir du vecteur articulatoire à 7 dimensions, de générer le vecteur géométrique à 30 dimensions et donc le contour sagittal du conduit vocal. Chaque paramètre a été normalisé et centré par Maeda en se fondant sur son écart-type et sa position moyenne. Ces paramètres sont alors exprimés en ce que nous choisissons de nommer *Unités Maeda* (UM). Précisons en termes de déplacements articulatoires la signification de ces Unités Maeda. Si l'on suppose que la distribution est normale, les paramètres articulatoires varient de -3UM à +3UM. La correspondance entre unités Maeda et unités de déplacement effectifs dépend de la ligne que l'on considère dans la grille semi-polaire. Pour le paramètre *corps* (qui sera celui auquel nous nous intéresserons par la suite, cf. paragraphe 3.6), une variation de -3UM à +3UM autour de la position neutre (moyenne) correspond ainsi à des déplacements maximum à l'horizontale de 2.7cm et à la verticale de 2.5cm. Étant donnée la structure linéaire du modèle, ces intervalles de variation sont valables pour toutes les configurations linguales tant qu'il n'y a pas de contact entre les lèvres ou entre la langue et les parois du conduit vocal.

La figure 3.6 présente les coupes sagittales du conduit vocal ainsi que les valeurs des paramètres articulatoires pour les voyelles [i] et [a] standards du modèle.

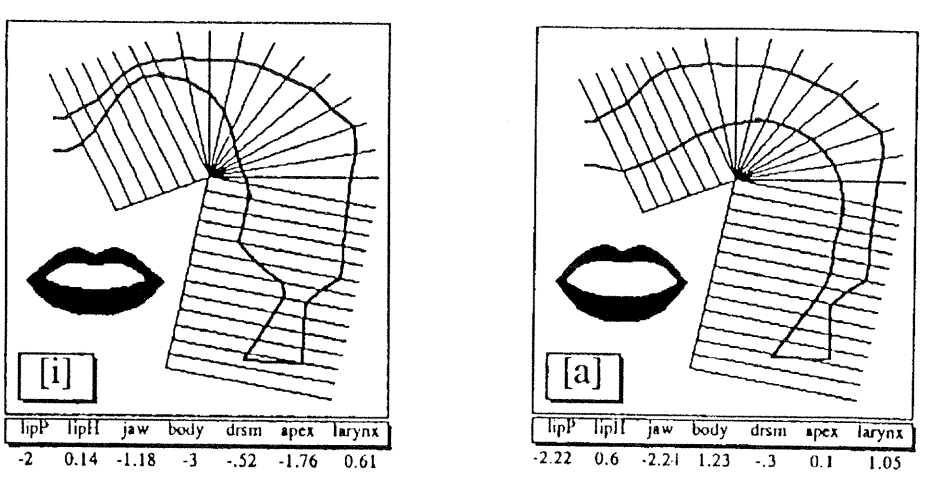



*Figure 3.6.* Coupes sagittales du [i] et du [a] standards du modèle de Maeda. D'après Boë [1994].

### 3.5.1.2 Le couplage avec un modèle acoustique

En couplant ce modèle articulatoire des profils du conduit vocal à un modèle acoustique, il est alors possible d'obtenir des valeurs formantiques à partir des sept commandes articulatoires. La fonction d'aire du conduit vocal (aires des différentes sections du conduit), depuis la glotte jusqu'aux lèvres, est d'abord calculée à partir de la distance sagittale, en utilisant un modèle géométrique de type αβ (Heinz & Stevens [1965]) fondé sur des mesures *scanner* et des moulages du conduit vocal (Perrier, Boë & Sock [1992]). On passe ainsi d'une coupe sagittale bidimensionnelle à un modèle tridimensionnel. La fonction de transfert du "tuyau vocal" ainsi obtenu est calculée à l'aide d'un modèle acoustique harmonique de production des voyelles, les pôles de cette fonction donnent les valeurs des formants (Badin & Fant [1984]). La figure 3.7 représente le passage des commandes articulatoires aux formants.

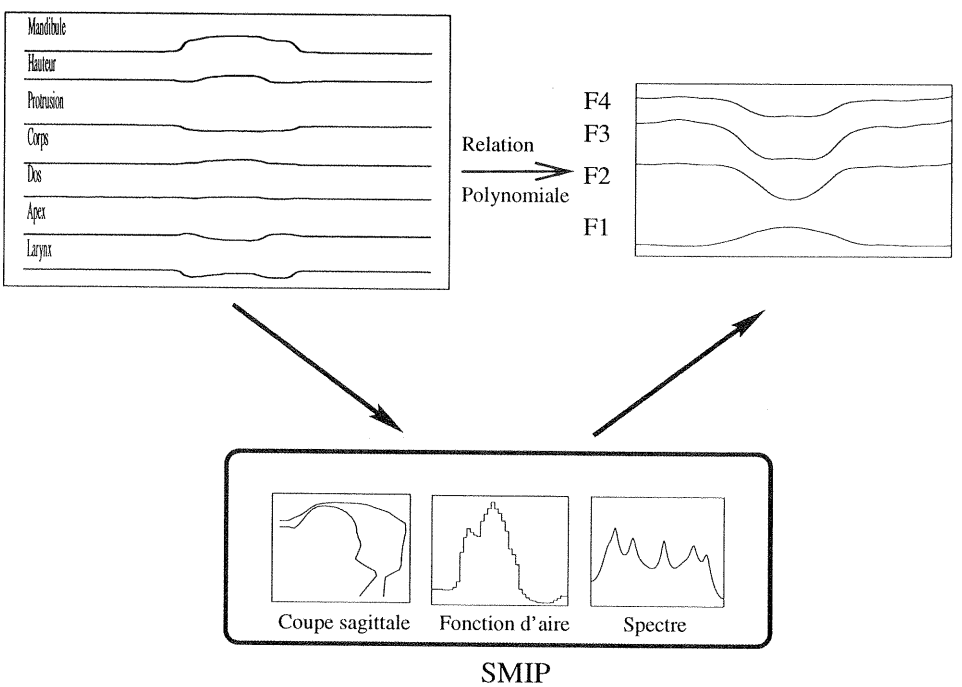



*Figure 3.7.* Le modèle direct de passage des positions des articulateurs aux formants.

Ce couplage entre le modèle articulatoire et le modèle acoustique est représenté approximativement par une relation analytique qui permet de passer directement des valeurs des paramètres articulatoires aux valeurs formantiques. Des études préliminaires (Bailly, Jordan, Mantakas, Schwartz, Bach & Olesen [1990]) ont montré qu'un polynôme du second ordre représente un compromis satisfaisant entre la simplicité d'implémentation, la stabilité et la précision requise. Une relation polynomiale du deuxième ordre a donc été

recherchée par une technique d'optimisation à partir d'un dictionnaire de correspondances articulateurs/formants (Morris [1990]) :

$$F^{(k)} = \sum_{i \leq j} a_{ij}^{(k)} A_i A_j$$

Les variables $A_i$, $A_j$ sont les paramètres articulatoires et $F^{(k)}$ est le k-ième formant. Nous prenons uniquement en compte ici les quatre premiers formants. Les coefficients $a_{ij}$ ont été optimisés pour rendre compte au mieux du dictionnaire de correspondances. Nous ferons par la suite référence à ce modèle articulatoire complet, qui a fait l'objet de mises au point détaillées lors du projet Européen ESPRIT *Speech MAPS*, sous son acronyme *SMIP*, pour *Speech Maps Interactive Plant* (Boë [1993], Boë, Gabioud & Perrier [1996]).

## 3.5.2 La normalisation du locuteur

### 3.5.2.1 La nécessité de normalisation

Un problème fréquent dans l'utilisation de modèles articulatoires est la normalisation du locuteur au modèle (ou l'inverse). En effet, dans certains cas, les transitions formantiques du locuteur peuvent sortir de l'espace acoustique du modèle et correspondre alors à des trajectoires articulatoires irréalistes. Aussi, afin que les trajectoires formantiques soient réalisables par le modèle, est-il nécessaire que les données acoustiques soient projetées dans l'espace acoustique de référence du modèle de Maeda, *i.e.* normalisées. La transformation que nous avons utilisée s'appuie sur la notion d'affiliation entre formants et cavités du conduit vocal (Payan & Perrier [1993], Perrier, Apostol & Payan [1995]). L'hypothèse fondamentale est qu'en exploitant les affiliations formants/cavités, la variabilité inter-locuteur peut être interprétée en termes de différences de longueurs de conduit vocal. Ainsi en rendant la géométrie du conduit vocal du locuteur semblable à celle du modèle de Maeda, nous nous assurons que les trajectoires formantiques générées par ce conduit vocal transformé, appartiennent bien à l'espace acoustique du modèle.

Présentons maintenant notre démarche pour la normalisation des formants du locuteur JLS obtenus avec le corpus présenté au paragraphe 3.3.

### 3.5.2.2 La notion d'affiliation

**Le modèle à quatre tubes, Fant [1960]**

Le modèle à quatre tubes proposé par Fant [1960] permet de rendre compte simplement des comportements acoustiques fondamentaux du conduit vocal. On peut en effet séparer le conduit vocal en quatre régions ou tubes, comme le montre le schéma 3.8. La constriction principale de la langue est représentée par le tube n°3 ; elle sépare le

conduit en deux cavités couplées qui constituent les tubes n°2 et n°4 : la cavité avant (des lèvres à la constriction) et la cavité arrière (de la constriction au larynx). La région correspondant aux lèvres est représentée par le tube n°1.

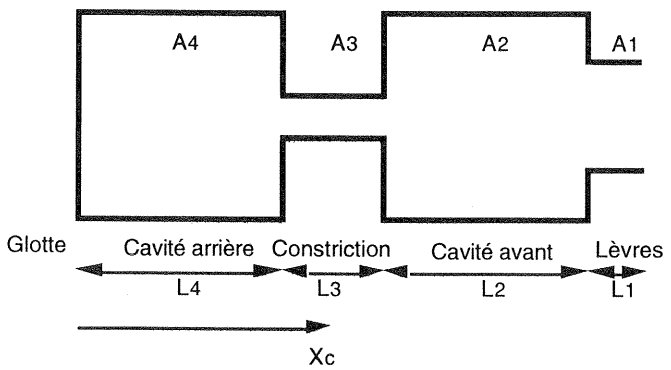



*Figure 3.8.* Le modèle à quatre tubes.

À l'aide de ce modèle, Fant interprète le comportement des formants lorsque la position de la constriction est déplacée depuis la glotte jusqu'aux lèvres. Il obtient ainsi, pour des aires de constriction et des ouvertures aux lèvres données, des *nomogrammes* du conduit vocal, *i.e.* des abaques des variations formantiques en fonction de la position de la constriction (pour plus de détails sur l'acoustique des voyelles, voir annexe A). La figure 3.9, extraite de Badin *et al.* [1988], donne un exemple de nomogramme pour une ouverture moyenne des lèvres. L'axe des abscisses est la position de la constriction en cm (Xc), de la glotte à gauche, aux lèvres à droite. Les fréquences en kHz sont en ordonnées.

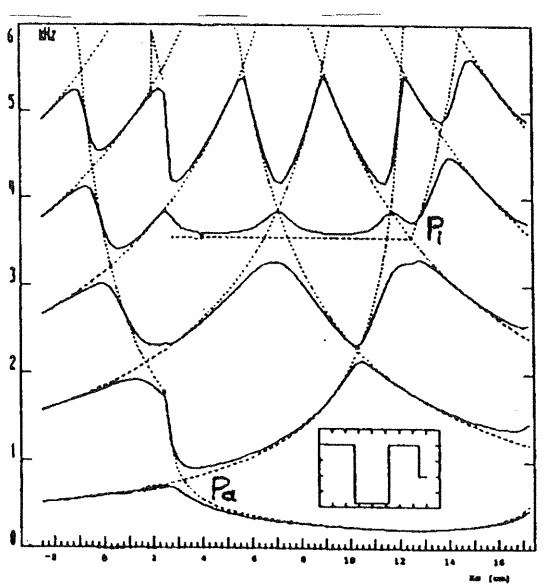



*Figure 3.9.* Nomogramme pour une ouverture moyenne des lèvres. D'après Badin *et al.* [1988]. Les points focaux $P_i$ et $P_a$ sont indiqués.

**La notion d'affiliation**

La notion d'*affiliation* d'un formant à une cavité du conduit vocal a été proposée par Fant [1960]. Badin *et al.* [1988] en ont précisé certains aspects à partir des nomogrammes eux-mêmes. Lorsque l'on associe les différents résonateurs que constituent chacun des tubes du modèle, des phénomènes de couplage apparaissent et les résonances qui en résultent s'écartent des résonances propres à chaque cavité isolée. Lorsqu'une résonance reste proche de celle propre à une cavité donnée, on parle d'*affiliation* de cette résonance (ou de ce formant) à la cavité en question. Sur le nomogramme de la figure 3.9, les affiliations correspondent aux régions où les lignes pointillées (résonances propres à une cavité isolée) sont proches des lignes continues (résonances des quatre tubes couplés).

### 3.5.2.3 Recherche des affiliations des formants pour la séquence [iai]

**Affiliation des formants du [i]**

**a. Locuteur JLS**

Les valeurs formantiques moyennes pour le [i] de JLS dans la séquence [iai] sont :

F1 = 250 Hz
F2 = 2305 Hz
F3 = 3480 Hz
F4 = 4010 Hz

On retrouve chez notre locuteur les caractéristiques formantiques habituellement observées pour un [i], *i.e.* le fort écart F1/F2, F1 étant faible et F2 élevé, et le rapprochement F3/F4. Sur le nomogramme précédent, les formants du [i] peuvent donc se trouver en avant ou en arrière du *point focal* (point de convergence formantique, cf. Boë & Abry [1986]) noté $P_i$. Pour les formants F1 et F2 les affiliations sont claires : F1 est la résonance Helmholtz de l'ensemble cavité arrière + constriction et F2 est affilié à la cavité arrière en tant que résonance demi-onde. Pour les formants F3 et F4, deux possibilités sont offertes : soit F3 est affilié à la cavité avant et F4 à la cavité arrière, si la voyelle [i] est située en arrière (à gauche) du point focal, soit l'inverse.

L'étude la transition [iyi] chez le locuteur JLS nous a permis de préciser ces affiliations. En effet le passage du [i] au [y] correspond à une fermeture des lèvres et donc à une modification de la cavité avant. Nous avons donc extrait les formants du signal acoustique correspondant à la production du logatome [iyi] au sein de la phrase porteuse "c'est iui ça" prononcée à débit "normal" et sans contrainte d'accentuation particulière. La figure 3.10 donne l'allure de la transition [iyi].

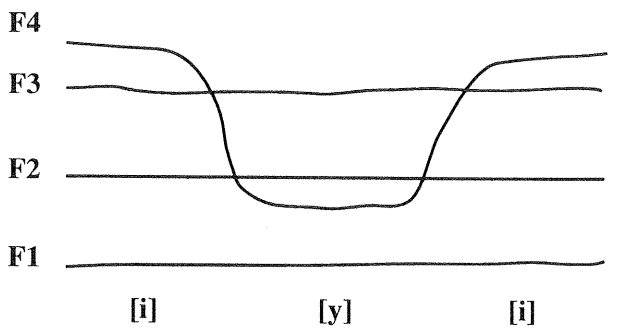



*Figure 3.10.* Allure de la transition [iyi] chez le locuteur JLS.

Au cours de cette transition, seul le formant F4 est modifié de façon significative, comme l'indique la figure 3.10, c'est donc lui qui est affilié à la cavité avant.

La cavité avant étant ouverte, les formants associés sont des résonances quart d'onde et harmoniques quart d'onde, alors que pour la cavité arrière (fermée), les formants affiliés sont une résonance de Helmholtz (pour la plus basse fréquence) et des résonances demi-onde (et harmoniques).

On en déduit donc finalement les affiliations suivantes pour le [i] du locuteur JLS :

F1: cavité arrière + constriction (Helmholtz)

F2: cavité arrière demi-onde

F3: cavité arrière onde

F4: cavité avant quart d'onde

**b. Modèle de Maeda**

Les valeurs formantiques moyennes pour le [i] du modèle de Maeda sont :

F1= 252 Hz

F2= 2191 Hz

F3= 3217 Hz

F4= 3840 Hz

Ces valeurs formantiques sont proches de celles du locuteur JLS. Pour préciser les affiliations de F3 et F4, il faut étudier la fonction de sensibilité[1] du modèle de Maeda pour la hauteur des lèvres (pour la cavité avant) et le corps de la langue (pour la cavité arrière). Lorsque l'on fait varier la hauteur des lèvres, seul le formant F4 est sensiblement affecté. F4 est donc clairement affilié à la cavité avant. Lorsque l'on fait varier le corps ou le larynx, F3 varie énormément. F3 variant avec la position du larynx, il faut déterminer si F3 est bien affilié à la cavité arrière ou plutôt à la cavité sub-épiglottique, car pour être affilié cavité arrière, F3 devrait être proche du double de F2 (F2 étant la résonance demi-onde de la

[1] Une fonction de sensibilité décrit l'évolution $\Delta Fi$ du formant Fi, lorsqu'un des paramètres de commande, Cj, varie de $\Delta Cj < 10\%$ autour de sa valeur standard pour la voyelle considérée.

cavité arrière et F3 l'harmonique suivante). Mais le formant suivant (F5) est beaucoup trop élevé pour valoir le double de F2. Le seul formant possible pour cette résonance onde est donc bien F3. Et la diminution de F3 par rapport au double de F2 est probablement due au couplage des deux cavités.

On a donc finalement pour la voyelle [i] du modèle de Maeda exactement les *mêmes affiliations* que pour celle du locuteur JLS.

**Affiliation des formants du [a]**

**a. Locuteur JLS**

Les valeurs formantiques moyennes pour le [a] de JLS dans le logatome [iai] sont :

F1= 890 Hz

F2= 1570 Hz

F3= 2740 Hz

F4= 3840 Hz

Le formant F1 est élevé, F2 est moyen, F3 est élevé, on note une focalisation F1/F2. Sur le nomogramme présenté plus haut (figure 3.9), le [a] de JLS est donc situé en arrière ou en avant du point focal $P_a$. On a donc deux possibilités pour F1 et F2 : soit F1 est affilié cavité avant et F2 cavité arrière, soit l'inverse.

Or du point de vue perceptif, le [a] en contexte [i] du locuteur JLS est très avant (on entend plus /a/ que */a/*), ce qui est cohérent avec l'influence du contexte antérieur [iVi]. Pour cette raison, nous suggérons que le [a] est situé en avant du point focal $P_a$. Les formants affiliés à la cavité arrière sont une résonance de Helmholtz et une résonance demi-onde et les affiliations de la cavité avant sont de type quart d'onde.

On a donc les affiliations suivantes:

F1: cavité arrière + constriction (Helmholtz)

F2: cavité avant quart d'onde

F3: cavité avant trois-quart d'onde

F4: cavité arrière demi-onde

**b. Modèle de Maeda**

Les valeurs formantiques pour le [a] standard du modèle de Maeda sont :

F1= 748 Hz

F2= 1207 Hz

F3= 2309 Hz

F4= 3521 Hz

L'étude des fonctions de sensibilité du modèle pour divers paramètres articulatoires permet de mettre en évidence les affiliations pour le [a] du modèle. F1 étant le formant qui

varie le plus lorsque la hauteur des lèvres varie, il est affilié cavité avant. Lorsque l'on fait descendre le larynx (et donc augmenter le volume de la cavité arrière), F2 et F4 diminuent. On a donc les affiliations suivantes pour le [a] du modèle de Maeda :

F1: cavité avant

F2: cavité arrière + constriction

Par conséquent, pour effectuer une normalisation pertinente (*i.e.* qui tienne compte du contexte antérieur [iVi]) il faut trouver, pour le modèle de Maeda, un [a] qui soit *antérieur* au point focal $P_a$. Pour cela, on cherche donc à produire avec le modèle un [a] dont le lieu de constriction soit plus avant que celui du [a] standard et dont l'ouverture aux lèvres soit moins grande. En appliquant les valeurs suivantes :

| | |
|---|---|
| Aperture L. : | 0.6 |
| Protrusion L. : | -2.22 |
| Mandibule : | -1.8 |
| Corps : | 1 |
| Dos : | 1 |
| Apex : | 0.1 |
| Larynx : | 2.01 |

on obtient alors les valeurs formantiques suivantes :

F1 = 673 Hz

F2 = 1418 Hz

F3 = 2266 Hz

F4 = 3535 Hz

et les mêmes affiliations que celles du [a] en contexte [i] du locuteur JLS.

La figure 3.11 résume les affiliations pour la transition [iai].

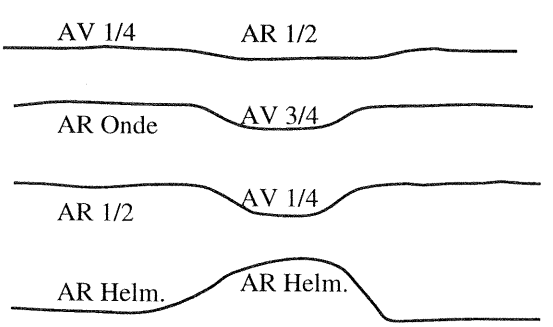



*Figure 3.11.* Affiliations pour la séquence [iai].

On notera que notre procédure de normalisation tient compte de la variabilité des allophones, liée au contexte phonétique.

### 3.5.2.4 Procédure de Normalisation

Les formants de notre locuteur et ceux du modèle de Maeda ayant mêmes affiliations, il est possible de procéder à une normalisation sur les résonances. Tout formant de JLS ($F_{jJLS}$) sera donc multiplié par un coefficient $\lambda_j$ fonction de son affiliation:

$F_j' = \lambda_j \, . \, F_{jJLS}$

La valeur formantique ainsi normalisée appartiendra à l'espace formantique du modèle de Maeda. On sépare les formants en deux groupes d'affiliation (avant et arrière) et l'on définit les coefficients suivants:

$\alpha_{AR} = L_{AR\,JLS} / L_{AR\,Maeda}$ où $L_{AR}$ est la longueur de la cavité arrière,

$\alpha_{AV} = L_{AV\,JLS} / L_{AV\,Maeda}$ où $L_{AV}$ est la longueur de la cavité avant.

Le résonateur de Helmholtz formé par une cavité de volume $V$ et une constriction de longueur $L_c$ et de faible section $A_c$, ayant pour fréquence de résonance :

$$F = \frac{c}{2\pi}\sqrt{\frac{A_c}{L_c V}},$$

la relation sur les longueurs fournit, pour les résonances Helmholtz de la cavité arrière :

$$F_{jJLS} = \frac{c}{2\pi}\sqrt{\frac{A_c}{L_c.(A_{AR}.L_{AR\_JLS})}} = \frac{c}{2\pi}\sqrt{\frac{A_c}{L_c.(A_{AR}.\alpha_{AR}.L_{AR\_Maeda})}}$$

Si l'on considère que l'aire à la constriction $A_c$, la longueur $L_c$ de la constriction et l'aire moyenne $A_{AR}$ de la cavité arrière, sont sensiblement équivalentes pour le locuteur JLS et pour le modèle de Maeda, on obtient alors une relation du type :

$$F_{jJLS} = \frac{1}{\sqrt{\alpha}} F_{jMaeda}$$

En négligeant les effets du couplage, on obtient de même pour les résonances demi-onde :

$$F_{jJLS} = \frac{1}{2.L_{JLS}} = \frac{1}{2.(\alpha.L_{Maeda})} = \frac{1}{\alpha} F_{jMaeda}$$

Et de même pour les résonances quart d'onde :

$$F_{jJLS} = \frac{1}{4.L_{JLS}} = \frac{1}{\alpha} F_{jMaeda}$$

Ainsi, compte-tenu de nos approximations, pour les résonances type Helmholtz, $\lambda$ vaut $\sqrt{\alpha}$ et pour les résonances demi-onde et quart d'onde, $\lambda$ vaut $\alpha$. Pour limiter les erreurs liées aux approximations faites dans l'évaluation théorique des coefficients, nous calculons à part les coefficients des résonances Helmholtz. D'autre part, lorsque l'on peut choisir, on calcule les coefficients sur les résonances de type onde ou trois-quart d'onde (plutôt que demi-onde et quart d'onde). En effet, l'erreur sur le coefficient $\alpha$ est proportionnelle à l'inverse de la valeur du formant F au carré :

$\alpha = F / F_{Maeda}$ => $\Delta\alpha = - \Delta F / ( F_{Maeda} . F^2 )$

Par conséquent, les coefficients choisis pour les résonances non-Helmholtz sont les suivants :

**Pour [i] :**

**Cavité arrière :** F3 (onde) : $\alpha_{[i]AR} = F3_{[i]Maeda} / F3_{[i]JLS}$

**Cavité avant :** F4 (quart d'onde) : $\alpha_{[i]AV} = F4_{[i]Maeda} / F4_{[i]JLS}$

**Pour [a] :**

**Cavité arrière :** F4 (demi-onde) : $\alpha_{[a]AR} = F4_{[a]Maeda} / F4_{[a]JLS}$

**Cavité avant :** F3 (trois-quart d'onde) : $\alpha_{[a]AV} = F3_{[a]Maeda} / F3_{[a]JLS}$

**Normalisation du formant F1**

Lors du passage de [i] à [a], les affiliations sont les mêmes. On choisit donc d'appliquer à $F1_{JLS}$ un coefficient $\lambda$ variant linéairement de [i] à [a]. On a donc :

$F1_{Maeda} = \lambda(F1_{JLS}) . F1_{JLS}$

avec : $$\lambda = \alpha_{[i]} + (\alpha_{[a]} - \alpha_{[i]}) \frac{F1_{JLS} - F1_{[i]JLS}}{F1_{[a]JLS} - F1_{[i]JLS}}$$

D'autre part, les affiliations étant de type Helmholtz, les coefficients $\lambda_{[v]}$ sont calculés ainsi :

$\alpha_{[i]} = F1_{[i]\,Maeda} / F1_{[i]\,JLS}$

$\alpha_{[a]} = F1_{[a]\,Maeda} / F1_{[a]\,JLS}$

**Normalisation du formant F2**

Lors de la transition [iai], F2 est affilié cavité arrière demi-onde, puis avant quart d'onde puis de nouveau arrière demi-onde. Pour les plateaux [i] puis [a] puis [i], on applique donc simplement les coefficients $\alpha_{[i]AR}$, puis $\alpha_{[a]AV}$ puis $\alpha_{[i]AR}$. Pour les deux transitions (de [i] à [a] et de [a] à [i]), on applique des coefficients variant linéairement de [i] à [a] pour la première transition et de [a] à [i] pour la seconde. Ainsi, les temps repérant les diverses portions du signal (plateaux et transitions) étant numérotés dans l'ordre $t_1$, ...$t_4$, on a :

si $t < t_1$ : $F2' = \alpha_{[i]AR}\, F2_{JLS}$

si $t_1 < t < t_2$ : $F2' = \alpha(F2_{JLS}) . F2_{JLS}$

où $\alpha(F2_{JLS}) = \alpha_{[i]AR} + (\alpha_{[a]AV} - \alpha_{[i]AR}).(F2_{JLS} - F2_{[i]JLS}) / (F2_{[a]JLS} - F2_{[i]JLS})$

si $t_2 < t < t_3$ : $F2' = \alpha_{[a]AV}\, F2_{JLS}$

si $t_3 < t < t_4$ : $F2' = \alpha(F2_{JLS}) . F2_{JLS}$

où $\alpha(F2_{JLS}) = \alpha_{[i]AR} + (\alpha_{[a]AV} - \alpha_{[i]AR}).(F2_{JLS} - F2_{[i]JLS}) / (F2_{[a]JLS} - F2_{[i]JLS})$

si $t > t_4$ : $F2' = \alpha_{[i]AR}\, F2_{JLS}$

**Normalisation du formant F3**

Lors de la transition [iai], F3 est affilié cavité arrière onde, puis avant trois-quart d'onde puis de nouveau arrière onde. On procède de même qu'avec F2 avec les coefficients idoines (les temps limites $t_1$, ...$t_4$, étant repérés à nouveau pour chaque formant).

**Normalisation du formant F4**

Lors de la transition [iai], F4 est affilié cavité avant quart d'onde, puis arrière demi-onde puis de nouveau avant quart d'onde. On procède de même qu'avec F2 avec les coefficients idoines.

### 3.5.2.5 Résultats

Les figures suivantes (3.12 et 3.13) présentent les résultats de la procédure de normalisation appliquée aux formants du locuteur JLS dans les trois conditions d'élocution.

La figure 3.12 donne les trajectoires formantiques dans le plan F1/F2 avant et après normalisation. Les positions des voyelles [i, ɛ, a] du locuteur sont indiquées, ainsi que les positions standard de ces voyelles pour le modèle de Maeda (lettres suivies d'un tiret). On remarque que les voyelles des séquences normalisées sont proches des positions standard.

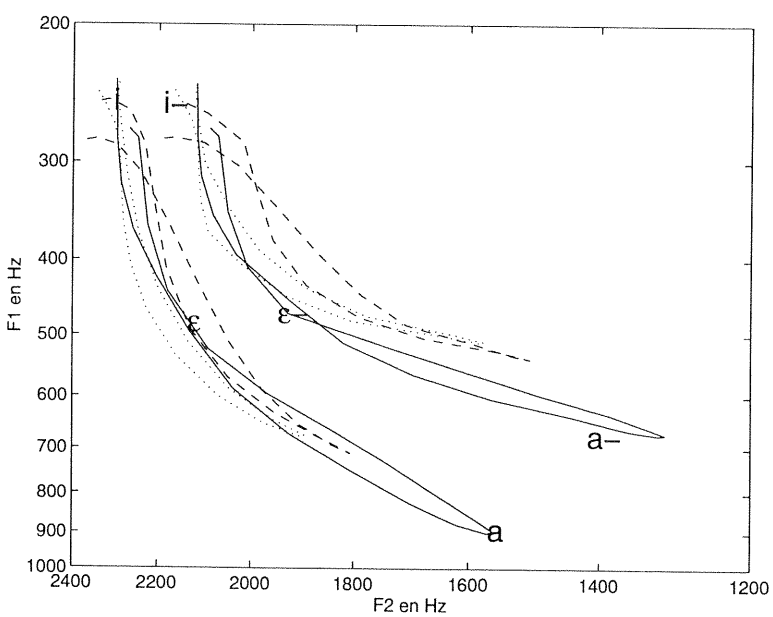



*Figure 3.12.* Séquences [iai] dans le plan traditionnel F1/F2 avant (en bas à gauche de la figure) et après (plus en haut et à droite) normalisation pour les trois conditions d'élocution. Lent accentué : trait plein, lent non accentué : trait tireté, rapide accentué : trait pointillé.

Les figures 3.13.a-c donnent les trajectoires temporelles des formants pour chaque condition d'élocution, avant (trait tireté) et après normalisation (trait plein). Les distances initiales entre F1 et F2 pour le [a] ne sont pas détériorées par cette procédure de normalisation.

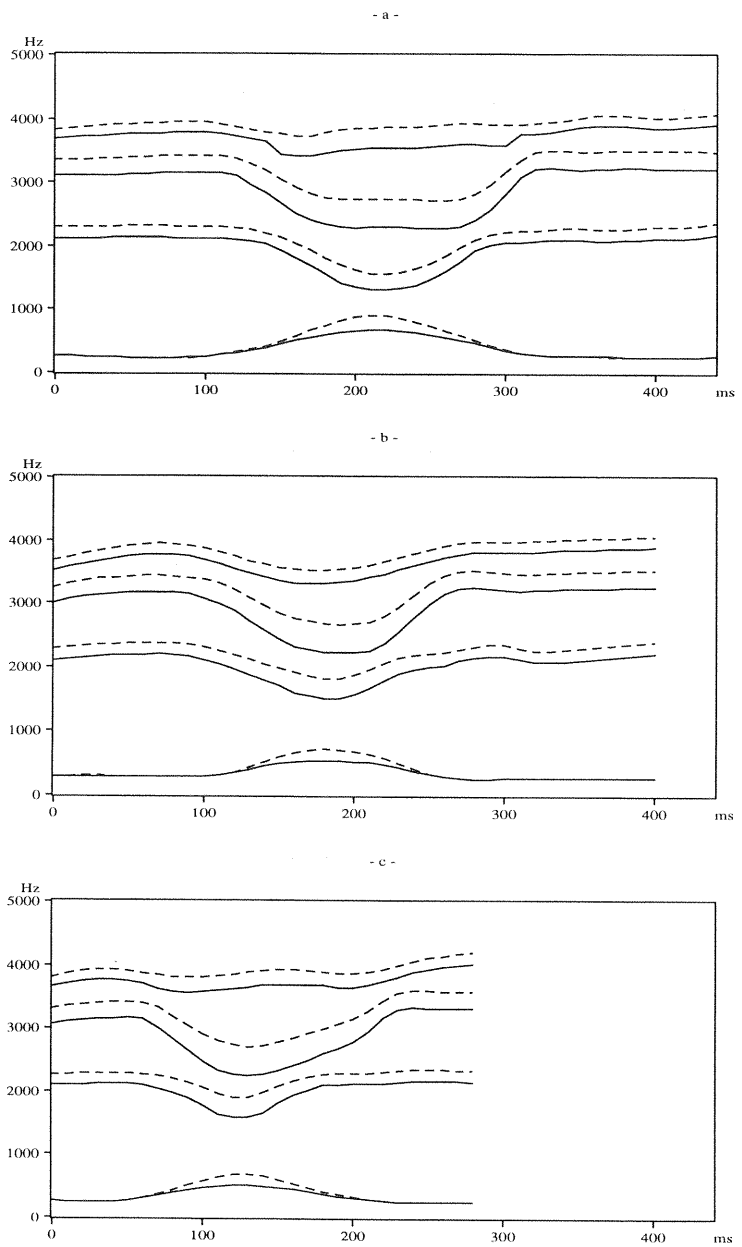



*Figure 3.13.* Trajectoires temporelles des formants pour les trois séquences [iai] avant et après normalisation. Formants normalisés : trait plein, formants originaux : trait tireté. a : lent accentué, b : lent non-accentué, c : rapide accentué.

### 3.5.3 La récupération des trajectoires articulatoires

**Le problème d'inversion**

Il s'agit de retrouver à partir du modèle articulatoire complet (le *SMIP*) les trajectoires articulatoires correspondant aux trajectoires formantiques normalisées. Le problème est donc d'inverser la relation polynomiale définie plus haut (cf. 3.5.1.2) pour déterminer les valeurs des sept paramètres articulatoires ($A_i$) rendant compte des trajectoires des quatre formants $F_d^{(k)}$ (l'indice $d$ signifie "donnée", l'exposant $k$ représente l'ordre du formant, de 1 à 4). Les données dont nous disposons sont les $N_e$ échantillons correspondant aux valeurs des quatre formants aux instants d'échantillonnage $t_1$, $t_2$, ..., $t_{Ne}$ : $F_d^{(k)}(1)$, $F_d^{(k)}(2)$, ..., $F_d^{(k)}(N_e)$. Il s'agit de déterminer les valeurs des $A_i(n)$ permettant de minimiser la distance entre les formants donnés, $F_d^{(k)}(n)$ et ceux obtenus par le modèle, $F^{(k)}(n)$. Ce problème d'inversion est typiquement un problème *mal-posé*. En effet de nombreuses configurations articulatoires peuvent être associées au même patron formantique (cf. par exemple Shroeder [1967] ou Mermelstein [1967]). Atal, Chang, Mathews & Tuckey [1978] ont représenté cette relation, dite *many-to-one*, à l'aide du concept de la *fibre acoustique*, sous-espace de l'espace des configurations du conduit vocal dans lequel le produit acoustique (les formants) est constant.

"*Large changes in the shape of the vocal tract can be made without changing the formant frequencies. These changes are consistent with the hypothesis that compensatory articulation is a possibility— that is, different people can produce the same sound with different vocal tract shapes. They are also consistent with the art of ventriloquism.*"

Pour résoudre ce problème d'inversion à excès de degrés de liberté, il est nécessaire d'introduire des contraintes (cf. paragraphe 3.4 et Poggio [1984]). Atal *et al.* proposent de réduire l'excès de degrés de liberté à l'aide de principes de minimisation de mouvement ou d'énergie pour passer d'un son à l'autre :

"*It seems worth investigating whether some minimum motion or minimum energy principle is applied in going from one sound to another.*"

Dans cette perspective, diverses études sur les trajectoires du bras ou des articulateurs de la parole (cf. Nelson [1983], pour les mouvements du bras et de la mandibule et Ostry, Keller & Parush [1983] pour les mouvements du dos de la langue) tendent à montrer que ces mouvements sont régis par des principes généraux d'*économie* d'effort (au sens large d'économie de l'énergie, de l'effort, de temps, etc.). Remarquant, dans la lignée de Nelson [1983], que les articulateurs de la parole se meuvent de façon régulière et lisse, nous ajoutons donc au *SMIP* une contrainte de lissage : minimisation de la distance entre deux points successifs de la trajectoire de tout articulateur. Le critère à minimiser devient donc la somme du coût de *lissage* (somme des coûts pour chaque articulateur) et du coût

*configurationnel* (erreur quadratique sur les formants donnés $F_d^{(k)}(n)$). Cependant les coûts configurationnel et de lissage sont contradictoires puisqu'un lissage parfait donnerait des paramètres articulatoires n'évoluant pas dans le temps. Pour remédier à ce problème, on donne des influences hiérarchisées aux différents coûts (cf. plus loin).

D'autre part une contrainte importante sur les paramètres articulatoires est qu'ils restent dans le domaine de variation prévu par le modèle de Maeda, sinon, les coupes sagittales du conduit vocal induites par ces paramètres "hors-norme" seraient erronées. Pour ce faire un coût supplémentaire est introduit, pénalisant les paramètres sortant de leur domaine de variations moyennes : le coût de *pénalisation*.

**L'approche connexionniste**

Notre problème étant de trouver *à chaque instant* la meilleure combinaison de paramètres articulatoires générant les formants donnés, l'approche connexionniste est assez séduisante. En effet, dans cette approche, les activités des divers composants (articulateurs ou formants ici) sont considérées en parallèle et en temps réel (Davalo & Naïm [1990]). De plus, l'organisation du réseau vers la production de sorties convenables se fait par coopération entre les cellules du réseau. La notion de coordination entre les articulateurs pour générer un produit acoustique est ainsi implicitement mise en œuvre.

Les principes de bases de l'approche connexionniste pour résoudre des problèmes de contrôle moteur ont été bien présentés par Jordan [1988, 1989, 1990]. Et c'est sous cet angle que Laboissière, Schwartz & Bailly [1990] ont choisi d'examiner quelques phénomènes spécifiques à la parole, comme la coarticulation. Ces auteurs ont ainsi eu recours à un réseau de neurones séquentiel utilisant un algorithme de rétropropagation du gradient. Leur réseau s'est révélé efficace pour la production de voyelles utilisant le *SMIP* présenté plus haut (cf. 3.5.1.2) et a donné des résultats semblables aux données expérimentales en ce qui concerne les phénomènes de coarticulation (Bailly & Laboissière [1993]). Ces résultats satisfaisants nous ont encouragés à utiliser cette technique pour notre problème d'inversion. Nous représentons donc les liens entre paramètres articulatoires et produits formantiques par un réseau de neurones dont certains poids de connexion sont optimisés afin de faire coïncider les trajectoires formantiques générées par le réseau avec celles données. On obtient alors les trajectoires articulatoires recherchées. L'optimisation progressive des poids se fait à l'aide d'une technique classique de *rétropropagation du gradient* de l'erreur totale à optimiser (Rumelhart, Hinton & Williams [1986]).

**Le réseau**

Le réseau est constitué de deux parties, du modèle direct (*SMIP*) sous la forme de sa représentation analytique du second ordre (décrite au paragraphe 3.5.1.2), ainsi que d'une structure permettant de représenter l'évolution temporelle des entrées du modèle direct, *i.e.* des paramètres articulatoires.

La partie dite *statique* du réseau, représentant le modèle direct, comprend sept cellules d'entrée, correspondant aux sept paramètres articulatoires et quatre cellules de sortie correspondant aux quatre formants. Chaque cellule de sortie reçoit une combinaison quadratique de tous les paramètres d'entrée. Les poids des connexions entre les sept cellules d'entrée et les quatre cellules de sortie sont les coefficients de la relation analytique du second ordre ; ils ne varient donc pas durant l'optimisation.

La partie dite *temporelle* du réseau, qui permet de simuler une trajectoire articulatoire de longueur $N_e$ (le modèle direct ne fournit qu'une relation *statique* entre entrées et sorties), est constituée d'une couche de $N_e$ cellules activées successivement, valant 1 lorsqu'elles sont activées et 0 sinon.

La partie *statique* du réseau est placée en aval de la partie *temporelle*, chacune des sept cellules articulatoires recevant une combinaison des $N_e$ unités de la couche *temporelle* (qui est donc la véritable couche d'entrée du réseau). Les poids des connexions, entre cellules articulatoires et cellules d'entrées, sont optimisés par la rétropropagation et représentent finalement la trajectoire articulatoire elle-même, ils seront nommés *poids articulatoires*.

L'estimation du coût de lissage nécessite le calcul de la distance entre deux points successifs d'une trajectoire articulatoire et donc la connaissance de l'activité de la cellule articulatoire aux temps $n$ et $n-1$. L'activité au temps $n-1$ est obtenue en introduisant un délai à la sortie de la cellule. Sept cellules de lissage, différence entre la sortie d'une cellule articulatoire et la sortie retardée de cette cellule, sont donc ajoutées aux cellules de sortie du réseau. On minimise le coût de lissage en spécifiant que les sorties désirées des cellules de lissage sont nulles. Le schéma 3.14 représente notre réseau.

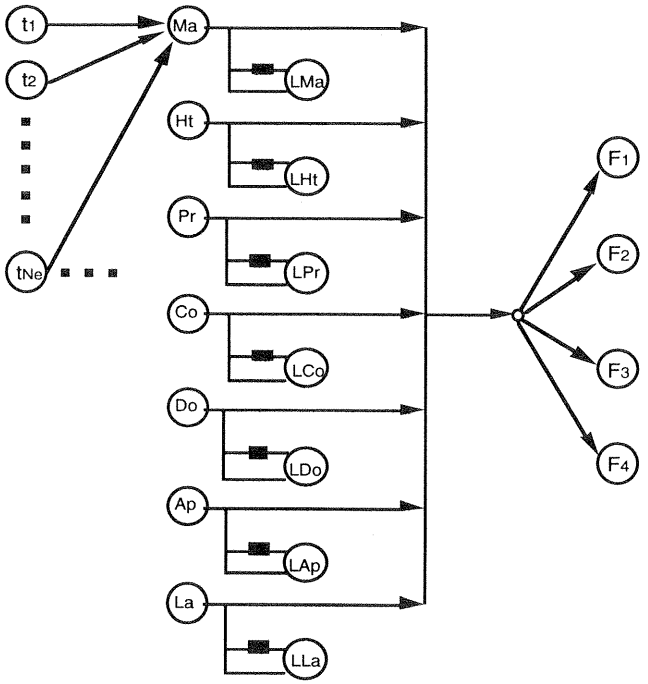



*Figure 3.14.* Schéma du réseau utilisé pour l'inversion cinématique. Ma..La : cellules articulatoires ; LMa..LLa : cellules de lissage ; F1..F4 : cellules de sortie. Les rectangles noirs représentent les délais introduits en sortie de cellule articulatoire.

**L'implémentation**

Le réseau ainsi défini est implémenté à l'aide d'un simulateur de réseaux de neurones formels : *XRBP* (pour *X terminal Recurrent Back Propagation*) (Laboissière [1992], Lœvenbruck [1992]). Fondé sur le logiciel original de Rumelhart, Hinton et Williams [1986], nommé *PDP* (*Parallel Distributed Processing*), qui implémente divers algorithmes pour les réseaux de neurones, et en particulier l'algorithme de rétropropagation du gradient (*BP*), *XRBP* permet en outre l'utilisation de réseaux récurrents et la visualisation des paramètres du réseau sous environnement *X windows*. Le simulateur est fondé sur l'apprentissage de patrons de sorties désirées : les poids des connexions entre les différents neurones (ou cellules) sont modifiés jusqu'à ce que les sorties du réseau soient égales à (ou très proche de) celles désirées. *XRBP* inclut d'autre part une notion d'*évolution temporelle* de l'état du réseau : il existe une variable nommée "instant" qui séquence le réseau et qui permet d'avancer (ou de reculer) dans le temps. *XRBP* permet de résoudre des problèmes d'inversion : pour connaître les valeurs de certains paramètres du réseau qui peuvent permettre d'obtenir un certain type de sorties, il suffit de faire apprendre au réseau le patron des sorties désirées et de lire ensuite les activités des cellules associées aux paramètres inconnus. Notons que les patrons formantiques désirés

sont entrés en barks[2] normalisés, afin que les quatre formants aient le même poids relatif dans le calcul de l'erreur.

La hiérarchisation des coûts configurationnel, de pénalisation et de lissage se fait en multipliant les coûts secondaires par des poids décroissant en fonction des itérations $t$ de l'algorithme. L'erreur totale à rétropropager, $E_T$, est donc (si $J_{conf}$ est le coût configurationnel, $J_{pén}$ le coût de pénalisation et $J_{liss}$ le coût de lissage) :

$$E_T = J_{conf} + \mu_1 J_{pén}\, e^{-\alpha 1 t} + \mu_2 J_{liss}\, e^{-\alpha 2 t}$$

Pour l'inversion présentée ici, nous avons fixé les coefficients μi et αi à des valeurs *ad hoc* :

$$\mu_1 = 0.1 \text{ et } \alpha_1 = 0.007$$
$$\mu_2 = 0.01 \text{ et } \alpha_2 = 0.007$$

### 3.5.4 Résultats

Afin de partir d'une position, dans l'espace des poids à optimiser, qui ne soit pas trop éloignée de la solution recherchée pour l'inversion de [iai], le réseau est d'abord entraîné avec une simple séquence [i]. Les cellules articulatoires du réseau sont initialisées avec les valeurs connues des paramètres pour un [i] standard (Boë [1993]) et on lance l'algorithme de rétropropagation pour optimiser les poids articulatoires. Lorsque l'on a obtenu des valeurs formantiques proches des valeurs désirées, on sauvegarde les valeurs des poids optimisés. On peut alors procéder à la procédure d'inversion proprement dite. On initialise cette fois cellules et poids articulatoires avec les valeurs connues et optimisées pour le [i].

Les premiers essais d'inversion sur la trajectoire [iai] dans le cas lent accentué, ont donné des trajectoires articulatoires correspondant à des coupes sagittales aberrantes. Ces résultats décevants étaient dus au décalage temporel entre les deux premiers formants (F1 et F2) et F3. Lorsque F1 et F2 entament leurs mouvements de transition du [a] au [i], F3 reste encore dans le plateau du [a] (cf. figure 3.3). Ce décalage peut être imputable à deux raisons :

- une mauvaise détection formantique due au croisement F2/F4 qui apparaît dans cette région de transition (cf. le paragraphe sur la normalisation 3.5.2).

- une action asynchrone des articulateurs dos et mandibule qui provoque des transitions formantiques en deux temps, d'abord F1 et F2 puis F3.

Le modèle articulatoire utilisé ne fournissant pas d'hypothèse sur la synchronisation des articulateurs, nous avons préféré nous en tenir à forcer l'inversion principalement sur les deux premiers formants : c'est justement la qualité du rapprochement F1/F2 qui permet

[2] La formule utilisée est celle de Chistoshiv : $F_{bark} = 6.7 \text{ argsh } ((F_{Hz} - 20)/600)$.

de juger de la réduction sur la voyelle [a]. Les résultats, présentés sur la figure 3.15, sont alors satisfaisants.

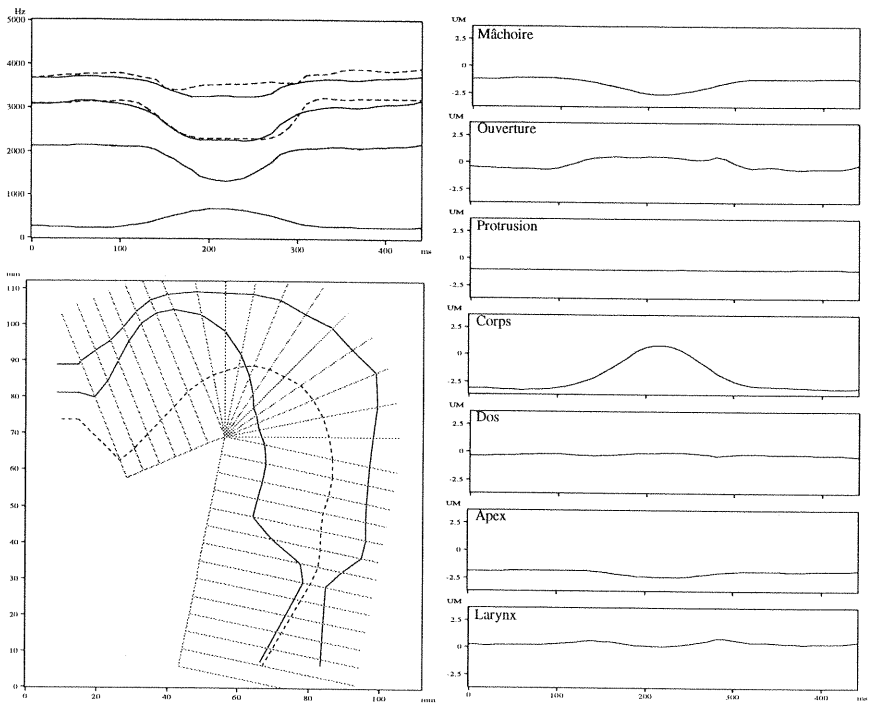



*Figure 3.15.* Résultats de l'inversion cinématique pour la séquence [iai]. Panneau supérieur gauche : données (trait tireté) et simulations (trait continu) formantiques. Panneau inférieur gauche : coupes sagittales du [i] (trait plein) et du [a] (trait tireté). Panneaux de droite : trajectoires articulatoires obtenues par l'inversion. Les valeurs des paramètres articulatoires sont exprimées en Unité Maeda (UM, cf. 3.5.1.1).

Les trajectoires articulatoires obtenues par cette inversion fournissent des patrons formantiques F1/F2 très proches des données. Le premier panneau de la figure 3.15 présente les trajectoires formantiques données (trait tireté) et calculées (trait continu). Le panneau en bas à gauche donne les coupes sagittales correspondant au premier [i] (trait continu) et au [a] (trait tireté). Les trajectoires des sept articulateurs sont fournies dans les panneaux de droite, dans l'ordre suivant : mandibule, hauteur et protrusion des lèvres, corps, dos et apex de la langue et hauteur du larynx.

**Test de synthèse**

Afin de vérifier que l'information essentielle n'a pas été perdue dans cette inversion privilégiant F1 et F2, nous avons procédé à un test de synthèse acoustique. Le sonagramme est obtenu à partir des configurations articulatoires fournies par l'inversion et en utilisant cette fois un simulateur temporel de l'appareil vocal complet (Scully, Castelli, Brearley, Shirt [1992]) et non plus l'approximation polynomiale du second ordre mise en œuvre pour l'inversion. Ce simulateur des phénomènes acoustiques, aérodynamiques et

mécaniques intervenant dans le production du signal de parole s'appuie sur les principes développés par Kelly & Lochbaum [1962] pour leur modèle analogue temporel du conduit vocal. Le conduit vocal complet est divisé en trois parties recevant chacune des paramètres de commande propres. L'appareil subglottique (trachée artère, bronches, bronchioles et poumons), aux dérivations multiples, est remplacé par un seul tube à section variable, équivalent du point de vue acoustique (Weiberl [1963]). La glotte qui représente la source vocale (ou l'excitateur) est décrite par un modèle mécanique à deux masses des cordes vocales (Pelorson, Hirshberg, Van Hassel, Wijnands & Auregan [1994]). Ce modèle des cordes vocales dépend essentiellement de deux paramètres indépendants (lors de la production des voyelles) : la pression subglottique et le facteur masse-tension regroupant les masses et les tensions des cordes vocales. Pour notre test de synthèse, ces deux paramètres sont fixés à des valeurs neutres et constantes, le signal acoustique ne contient ainsi pas d'effet d'accentuation lié aux cavités glottique et subglottique, seuls les effets des articulateurs du conduit vocal sont mis en œuvre. Enfin le conduit vocal (le filtre ou résonateur) est représenté par un tube dont la fonction d'aire est celle des coupes sagittales inférées par inversion (cf. 3.5.1.2 pour le passage des coupes sagittales à la fonction d'aire).

La figure 3.16 présente le sonagramme du signal synthétisé pour [iai]. Le rapprochement F1/F2 caractéristique du [a] est bien préservé. Un test perceptif informel indique que la qualité acoustique du [a] synthétique est tout-à-fait acceptable, en comparaison avec la qualité originale. Rappelons toutefois que les effets d'accentuation induits par l'augmentation d'énergie ou le $F_0$, ne sont pas analysés ici. Nous examinons l'accentuation dans son action sur la préservation des configurations formantiques.

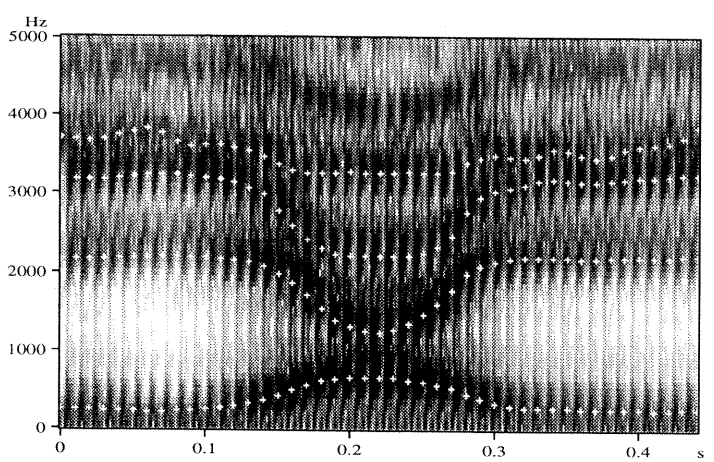



*Figure 3.16.* Sonagramme de la séquence [iai] obtenu à partir des trajectoires articulatoires inférées par inversion cinématique. Les quatre premiers formants sont représentés par les croix blanches.

# 3.6 Inversion dynamique : depuis la trajectoire articulatoire jusqu'aux commandes motrices

## 3.6.1 Le choix du paramètre articulatoire

**Les paramètres articulatoires liés à la langue**

Maeda, Honda et Kusawaka [1993] affirment que le système musculaire de la langue, quoique complexe, est organisé autour d'un nombre limité de blocs fonctionnels :

"*[...]although the tongue muscular system is anatomically complex, it is organised into a small number of functional blocks for speech production.*"

Ils montrent que la position de la langue pour les voyelles est déterminée par deux ensembles de muscles antagonistes du point de vue de leurs activités EMG. Le génioglosse postérieur (GGp) et l'hyoglosse (HG) fonctionnent symétriquement : l'activation de GGp correspond à un déplacement de la langue vers l'avant et vers le haut, tandis que l'activation de HG induit un déplacement vers le bas et l'arrière. De même, le styloglosse (SG) et le génioglosse antérieur (GGa) ont des actions antagonistes : vers l'arrière et le haut pour le premier et vers l'avant et le bas pour le dernier. La figure 3.17, d'après Maeda & Honda [1994], représente ces effets antagonistes.

Ces auteurs concluent que l'on peut considérer que la forme et la position du corps de la langue sont influencés par deux articulateurs linguaux indépendants et par la mandibule, chacun de ces articulateurs étant contrôlés par deux groupes de muscles antagonistes. Maeda et Honda [1994] proposent d'autre part un lien entre les deux principaux paramètres linguaux du modèle articulatoire de Maeda, présenté plus haut (cf. paragraphe 3.5.1.1), et ces deux ensembles de muscles antagonistes. Le paramètre *corps* de la langue, lié à la *position* de la langue rend compte de la synergie (HG, GGp) tandis que le paramètre *dos* de la langue, lié à la *forme* de la langue, est associé à la seconde paire (SG, GGa).

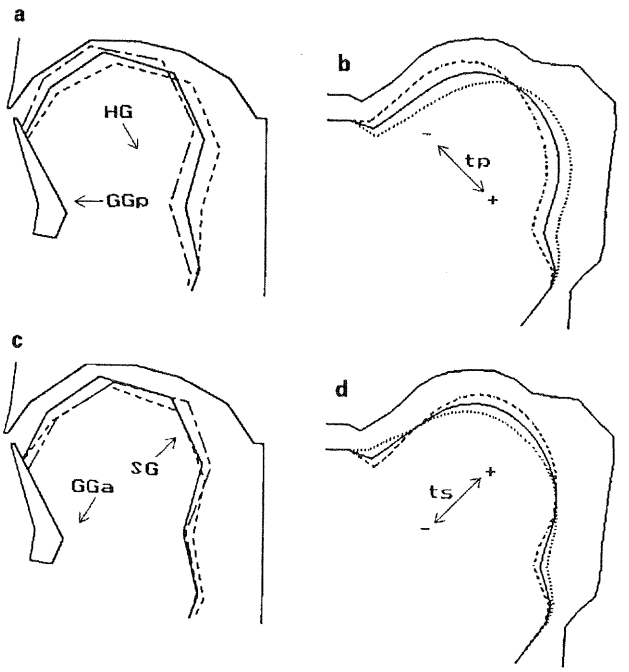



*Figure 3.17.* Les effets antagonistes de deux ensembles de muscles sur la langue. a : effets de HG (trait tireté) et GGp (trait mixte) sur les formes de la langue. La forme d'équilibre est représentée par le trait continu. Les flèches indiquent les orientations des forces de contraction des muscles. b : effets des modifications du paramètre corps de la langue du modèle de Maeda. c : effets des contractions de SG (trait mixte) et GGa (trait tireté). d : effets des modifications du paramètres dos de la langue. D'après Maeda & Honda [1994].

Les résultats de l'inversion cinématique montrent que des deux paramètres de langue qui nous intéressent, seul le paramètre *corps* présente une amplitude de mouvement significative. Nous le considérons donc comme le plus représentatif des mouvements de la langue pendant la séquence [iai]. Nous avons de surcroît vérifié cette hypothèse en calculant les formants lorsque le deuxième paramètre (le *dos*) est maintenu dans une position neutre (sa position moyenne), les autres paramètres conservant les valeurs déduites de l'inversion cinématique. La trajectoire formantique obtenue de cette façon reste très proche de l'originale. L'influence du paramètre *dos* est donc négligeable dans cette séquence et c'est pourquoi nous ne considérerons désormais que le paramètre *corps*.

## 3.6.2 Le modèle du second ordre

Pour générer les mouvements de l'articulateur *corps de la langue*, nous avons choisi une modélisation fonctionnelle simple des paires de muscles antagonistes. Une modélisation classique du second ordre présente un double avantage. D'une part, la simplicité du modèle

le rend inversible avec peu de contraintes (voir plus loin), d'autre part la dynamique des articulateurs est représentée, de façon fonctionnelle.

En effet, Cooke [1980] indique qu'une modélisation des membres reposant sur un modèle du second ordre permet de générer les mouvements simples observés empiriquement :

*"I would suggest that an adequate model for the generation of [simple skilled movements] is provided by considering the limb as a simple second order system. That is, one in which the limb behaves as if it were a damped spring having mass."*

Cette modélisation permet en effet de rendre compte d'un des aspects remarquables des mouvements simples : la relation linéaire entre le pic de vitesse et l'amplitude du mouvement.

Dans le même ordre d'idée, Nelson [1983] montre que, dans des conditions normales d'élocution, les déplacements de mandibule présentent des caractéristiques d'économie de mouvement :

*"[...]there is a consistent economy of movement in the way the jaw moves during speech, even though this is obviously not the primary objective of these movements."*

Selon lui un mouvement est économique s'il réalise un compromis entre les différents coûts physiques concurrents tout en respectant les conditions requises pour la tâche effectuée :

*"An 'economical' movement could be considered as one which is not optimal in any single criterion sense (not minimum-time, or maximum-distance, or minimum-energy, etc.) but rather one which represents a reasonable trade-off between the competing physical costs, while meeting the primary requirements of the movement task."*

Nelson remarque d'autre part que le profil de vitesse d'un système linéaire du second ordre non-amorti est remarquablement proche de celui obtenu avec un critère de minimisation du jerk (dérivée de l'accélération). Dans le cas du modèle du second ordre, il rappelle que le pic de vitesse du mouvement est lié à la distance parcourue (D) et à la durée du mouvement (T) par la relation :

$$V_{\max} \approx \frac{\pi}{2} \cdot \frac{D}{T}$$

On retrouve là la relation linéaire entre le pic de vitesse et l'amplitude du mouvement remarquée par Cooke.

Ce principe de minimisation du *jerk* est aussi utilisé par Hogan [1984]. Il montre que les profils de position, vitesse et accélération obtenus par minimisation du *jerk* sont proches des profils observés chez les singes lors de mouvements de pointage de cible avec l'avant-bras :

*"Within limits of experimental accuracy the minimum-jerk movement profile yields good qualitative and quantitative agreement with observed undisturbed movement profiles."*

Il propose aussi une interprétation physiologique de ce critère de minimisation du *jerk* : il permettrait de simplifier le contrôle moteur en réduisant la quantité d'information nécessaire pour déterminer la trajectoire:

*"[...] it seems that minimizing jerk may simplify the control of the system. Reducing the magnitude of the higher derivatives of the motion implies a reduction in the amount of information required to specify, store or predict the trajectory."*

Ostry, Keller & Parush [1983] obtiennent des profils de vitesse pour les mouvements du dos de la langue mesurés par échographie présentant la même relation linéaire entre amplitude et pic de vitesse.

S'appuyant sur les résultats de Nelson [1983] et Ostry *et al.* [1983], Perrier, Abry & Keller [1989] ont proposé, à l'instar de Cooke [1980], une modélisation des mouvements d'un articulateur à partir d'un modèle du second ordre. Afin de préserver la symétrie observée dans les profils de déplacement et de vitesse sur les mouvements du bras par exemple (Cooke [1980]), ils ont opté pour un modèle distribué, constitué de deux ressorts reliés par une masse. Le premier ressort représente ainsi le groupe des muscles agonistes et le second les muscles antagonistes. Cette modélisation leur a permis de simuler de façon correcte les trajectoires du dos de la langue obtenues par échographie dans différentes conditions d'élocution. Ce modèle a été repris par Abry, Perrier & Jomaa [1990] pour rendre compte des profils de vitesse de la mandibule recueillis par un kinésiographe mandibulaire dans diverses conditions d'élocution.

Cette modélisation ayant donc fait ses preuves pour divers articulateurs, nous avons choisi de l'utiliser ici pour générer les mouvements du corps de la langue.

**Le modèle distribué du second ordre**

Pour chaque degré de liberté articulatoire $i$, le modèle du second ordre d'un articulateur quelconque de la langue est décrit par l'équation suivante, normalisée par la masse :

$$\ddot{y}_i + b_i\dot{y} + K_i(y_i - y_{ei}) = 0$$

La figure 3.18 donne une représentation du modèle masse-ressort distribué (Perrier, Abry & Keller [1989]).

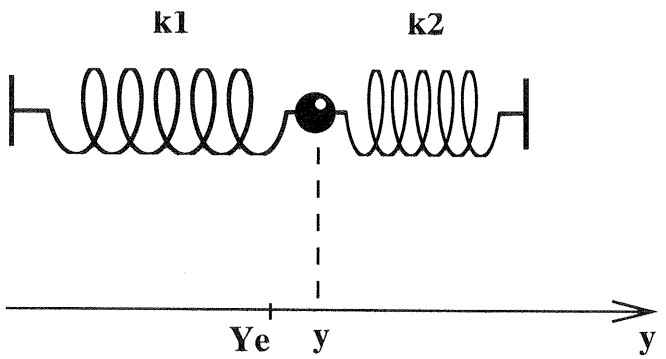



*Figure 3.18.* Le modèle masse-ressort distribué.

$K_i$ est la somme, normalisée par la masse, des raideurs $k_1$ et $k_2$ des deux ressorts et correspond à la cocontraction globale des muscles. Le coefficient de frottement $b_i$ est choisi de façon à ce que le système soit en régime pseudo-périodique amorti, *i.e.* : $b_i < 2\sqrt{K_i}$. La sensibilité du système à ce paramètre $b_i$ étant faible dans l'intervalle $]-1.6\sqrt{K_i}, 2\sqrt{K_i}[$, nous avons choisi la valeur *ad hoc* : $b_i = 1.89\sqrt{K_i}$ afin d'approcher le régime critique. La variable $y_{ei}$ correspond à la position d'équilibre spatial du système. Le choix d'un système sous-amorti est discutable (Kröger [1993]), toutefois c'est celui qui est fait le plus souvent en parole (cf. le modèle *Task Dynamics*, 2.2.1) et qui est justifié pour les mouvements du bras par Lacquaniti, Licata, Soeching [1982] ou Mac Kay, Crammond, Kwan, Murphy [1986].

**Le principe du contrôle**

Rappelons (cf. paragraphe 2.3) que notre hypothèse sur le contrôle est que, pour chaque degré de liberté, le système nerveux central associe à chaque phonème d'une séquence donnée une position d'équilibre spécifique, définissant ainsi les *cibles* successives du mouvement. Ces cibles posturales, vers lesquelles le mouvement est planifié, sont représentées par la variable $y_{ei}$ du modèle du second ordre.

D'autre part, le changement de position d'équilibre associé au passage d'une cible à une autre ne se fait pas de façon abrupte. Les temps de transition de cible à cible sont aussi contrôlés.

Pour chaque séquence, les *commandes centrales* correspondent donc à la spécification de points d'équilibre successifs, d'un niveau de cocontraction ainsi que des temps de transition et de maintien successifs de la trajectoire temporelle du point d'équilibre. La figure 3.19 donne un exemple de l'évolution temporelle de la position d'équilibre pour la production de trois phonèmes.

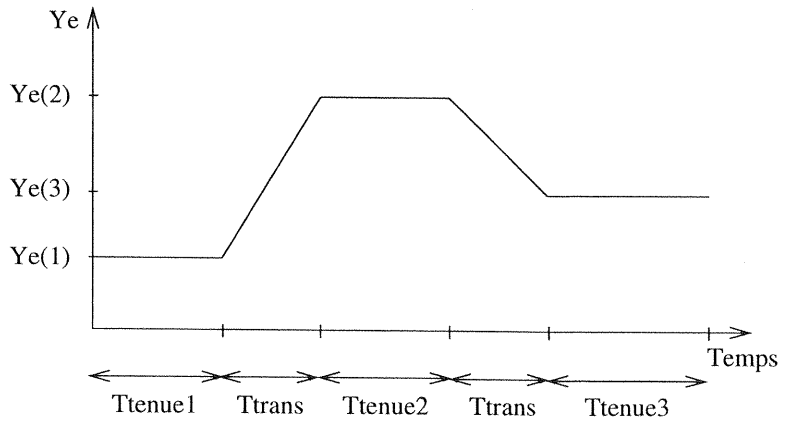



*Figure 3.19.* Évolution temporelle de la position d'équilibre pour la production de trois phonèmes successifs.

**L'hypothèse sur la variabilité prosodique**

Ainsi les cibles du mouvement sont spécifiées par $y_e$ tandis que la dynamique du mouvement est paramétrée par le niveau de cocontraction $K$, supposé constant le long de la séquence, et le *timing* de la commande centrale d'équilibre.

Dans ce cadre, notre hypothèse est que la variabilité associée aux effets prosodiques peut être simulée, avec la *même* séquence de points d'équilibre, en ajustant simplement le niveau de cocontraction et/ou le *timing* de la commande centrale d'équilibre.

## 3.6.3. La récupération des commandes centrales

### 3.6.3.1 La commande centrale d'équilibre (la commande posturale)

La figure 3.20 représente la trajectoire isolée du paramètre *corps* obtenue par inversion cinématique dans le cas lent accentué. Cette trajectoire est représentée en *Unités Maeda* (UM, cf. 3.5.1.1)[1].

[3]Afin de ne pas introduire de dissymétrie dans le modèle du second ordre, et pour des raisons historiques, le paramètre de longueur au repos ayant en effet été utilisé dans une version antérieure du modèle, nous avons en réalité recentré cette trajectoire et effectué un changement d'échelle pour que les déplacements aient un ordre de grandeur convenable. Le recentrage a été réalisé en soustrayant aux données leur valeur médiane (entre minimum et maximum). Les données ont été en outre multipliées par le facteur d'échelle 10/3, afin d'appartenir finalement à l'intervalle [-10, +10]. Cependant pour plus de clarté (*i.e.* pour que les déplacements soient lisibles en unités de grandeur habituelles), nous présenterons toujours les résultats en UM, sans tenir compte du recentrage et du changement d'échelle.

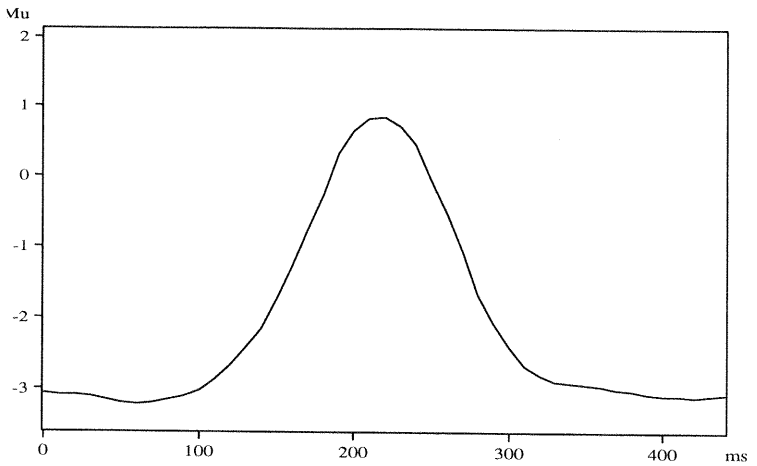



*Figure 3.20.* Trajectoire du corps de la langue pour la séquence [iai] dans le cas lent accentué.

Les positions d'équilibre successives sont directement lues sur la trajectoire du *corps* de la langue. Nous supposons en effet que dans le cas lent accentué, la trajectoire planifiée est sensiblement atteinte et qu'il y a peu d'*undershoot*. Nous avons d'abord pris pour les trois positions d'équilibre successives correspondant à [i], [a] et [i] les positions effectivement atteintes par l'articulateur pour chacune de ces voyelles. Quelques essais rapides nous ont finalement conduits à ajouter 2% de l'amplitude du mouvement à la position d'équilibre du [a] afin d'obtenir une meilleure adéquation entre les positions finales du [a] simulées et données. Les positions d'équilibre, ou les cibles ainsi définies, sont :

$y_e(i_1)$ = -3.223 UM[2]

$y_e(a)$ = 0.945 UM[3]

$y_e(i_2)$ = -3.113 UM[4]

et définissent la commande centrale dite "posturale".

### 3.6.3.2 Les commandes centrales dynamiques (commandes prosodiques)

**Le problème d'inversion**

Notre problème est de rechercher le niveau de cocontraction $K$ et les paramètres temporels $T_{hold1}$ (durée de la tenue de la première voyelle [i]), $T_{trans}$ (durée de la transition de [i] à [a])et $T_{hold2}$ (durée de la tenue de la voyelle centrale [a]), donnant la meilleure approximation de la trajectoire du *corps* de la langue obtenue par inversion acoustique.

[4]soit -6.812 en données corrigées
[5]soit 7.084 en données corrigées
[6]soit -6.447 en données corrigées

C'est un problème d'inversion depuis une trajectoire articulatoire vers une partie des commandes motrices : les commandes dynamiques (la commande posturale étant lue sur les données, cf. paragraphe 3.6.3.1). Nous disposons des données $y_d(1)$, $y_d(2)$, .., $y_d(Ne)$ correspondant aux $N_e$ échantillons de la trajectoire du corps aux instants $t_1$, $t_2$, ..., $t_{Ne}$. Il s'agit donc de déterminer les quatre paramètres $K$, $T_{hold1}$, $T_{trans}$ et $T_{hold2}$ minimisant l'erreur entre la trajectoire échantillonnée $y_d(n)$ et celle qui est obtenue par la résolution numérique de l'équation du second ordre (méthode de Runge-Kutta) et que nous nommerons $y(n)$. Plusieurs essais préliminaires ont montré que la minimisation de l'erreur moyenne sur la position, soit :

$$E = \frac{1}{N_e} \sum_{n=1}^{N_e} |y(n) - y_d(n)|$$

ne permettait pas d'approcher au mieux la courbe originale $y_d(n)$. En effet, l'erreur sur certains échantillons, comme par exemple les échantillons de début de transition, peut être importante numériquement alors que l'allure générale de la courbe est préservée. D'un autre côté, des erreurs, même faibles, sur les échantillons correspondant au plateau du [a] par exemple, peuvent provoquer des allures de courbe inacceptables. Les figures 3.21.a et 3.21.b montrent deux cas d'approximation de la courbe $y_d(n)$ représentée en trait tireté. Dans chaque figure, le premier panneau contient la position, le second la vitesse. Dans le premier cas (figure a), l'erreur $E$ est plus importante que dans le second ($E_1$ = 0.0406 et $E_2$ = 0.0367), alors qu'à l'évidence la première approximation est meilleure que la seconde, en position et en vitesse, si le critère est l'atteinte de la voyelle [a].

Aussi avons-nous choisi de tenir compte également de l'erreur sur la vitesse et de donner un poids double aux erreurs sur les positions correspondant au plateau du [a], qui est la zone sensible, dans le phénomène de réduction vocalique que nous analysons ici. Afin que les erreurs aient des poids équivalents, position et vitesse sont normalisées par les coefficients respectifs $Coef_y$ et $Coef_{\dot{y}}$ et appartiennent donc à l'intervalle [-1,1]. D'autre part, un poids moindre (*ad hoc*) est affecté à l'erreur sur la vitesse. L'erreur totale à minimiser devient alors :

$$E_T = \frac{1}{N_e . Coef_y} \sum_{n=1}^{N_e} |y(n) - y_d(n)| + 0.3 \frac{1}{N_e . Coef_{\dot{y}}} \sum_{n=1}^{N_e} |\dot{y}(n) - \dot{y}_d(n)| + E_{[a]}$$

où l'erreur sur le plateau du [a], calculée sur cinq échantillons (l'échantillon correspondant au maximum de déplacement associé au [a] et quatre échantillons pris de part et d'autre de celui-ci) est donnée par :

$$E_{[a]} = \frac{1}{5 . Coef_y} \left( |y(n_{a1}) - y_d(n_{a1})| + |y(n_{a2}) - y_d(n_{a2})| + \ldots + |y(n_{a5}) - y_d(n_{a5})| \right)$$

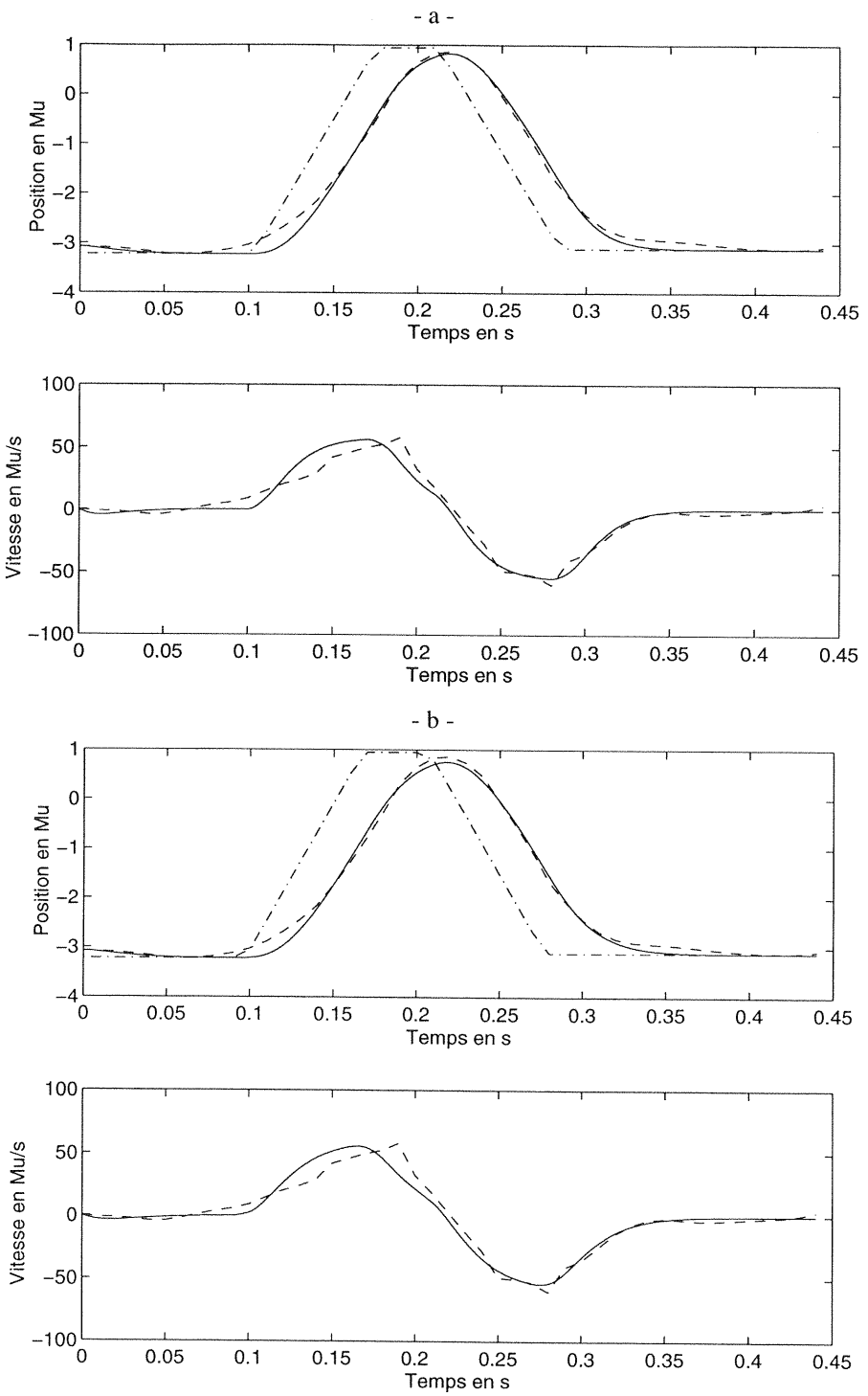



*Figure 3.21.* Deux exemples d'approximation des données articulatoires (voir texte pour les détails). L'erreur $E$ est plus importante dans le cas de la figure a, alors que cette approximation est meilleure. Traits tiretés : données, traits pleins : simulations, trait mixte : trajectoire de la commande centrale d'équilibre (commande posturale).

L'algorithme de rétropropagation du gradient utilisé pour l'inversion cinématique (cf. paragraphe 3.5) donne ici de piètres résultats dus au fait que, contrairement au cas précédent, nous ne disposons pas d'information *a priori* sur la position du point optimum dans l'espace à quatre dimensions des paramètres. L'algorithme s'arrête alors sur un minimum local insatisfaisant, ce qui est une faiblesse bien connue des descentes de gradient.

Les méthodes quasi-newtoniennes ont justement été introduites pour remédier à la lenteur des méthodes classiques de gradient. Présentons maintenant la méthode d'optimisation utilisée pour cette inversion, de la trajectoire articulatoire aux paramètres dynamiques.

**L'algorithme d'optimisation BFGS**

Cet algorithme fait partie de la classe des méthodes dites "quasi-newtoniennes" (Minoux [1983], Fletcher [1987]). Le problème étant de rechercher un minimum local $X^*$ d'une fonction réelle $f$ de $m$ variables réelles $(x_1, x_2, \ldots x_m)=X$, le principe général de ces méthodes est de générer une suite de points $X^{(0)}, X^{(1)}, \ldots, X^{(k)}$ convergeant vers ce minimum local. A chaque itération, le nouveau point $X^{(k+1)}$ est calculé de la façon suivante :

$$X^{(k+1)} = X^{(k)} - \lambda_k . H_k . \nabla f(X^{(k)})$$

$\lambda_k$ est le pas de déplacement choisi de façon à minimiser $f(X^{(k+1)}) = f(X^{(k)}+\lambda d_k)$ dans la direction de descente $d_k = -H_k . \nabla f(X^{(k)})$, c'est la *minimisation* (ou *recherche*) *unidimensionnelle*.

La matrice $H_k$ est une **approximation de l'inverse du Hessien de** $f$ (la formule itérative n'est donc autre qu'une généralisation de celle de Newton, qui ne converge pas lorsque le Hessien n'est pas défini-positif). Elle est modifiée à chaque itération en utilisant une formule de correction : $H_{k+1} = H_k + \Delta_k$ permettant par exemple de conserver la définie-positivité de la matrice $H_{k+1}$ (la définie-positivité du Hessien de $f$ étant une condition suffisante d'optimalité locale en un point stationnaire).

Parmi les diverses méthodes de correction existant, celle dite *BFGS* (car développée concurremment par Broyden [1970], Fletcher [1970], Goldfarb [1970] et Shanno [1970]) est probablement la plus utilisée. La formule de correction est la suivante :

$$H_{k+1} = H_k + \left[1 + \frac{\gamma_k{}^t H_k \gamma_k}{\delta_k{}^t \gamma_k}\right] . \frac{\delta_k \delta_k{}^t}{\delta_k{}^t \gamma_k} - \frac{\delta_k \gamma_k{}^t H_k + H_k \gamma_k \delta_k{}^t}{\delta_k{}^t \gamma_k}$$

Où:

$$\delta_k = X^{(k+1)} - X^{(k)}$$

$$\gamma_k = \nabla f(X^{(k+1)}) - \nabla f(X^{(k)})$$

Elle présente l'avantage de conserver la définie-positivité des matrices $H_k$ si le point $X^{(k+1)}$ est obtenu à partir de $X^{(k)}$ par minimisation unidimensionnelle dans la direction $d_k = -H_k . \nabla f(X^k)$.

L'algorithme BFGS, beaucoup moins sensible aux imprécisions dans la procédure de recherche unidimensionnelle que la plupart des algorithmes quasi-Newtoniens, tout en présentant une vitesse de convergence de même ordre, est considéré comme supérieur par de nombreux auteurs (Minoux [1983], Pierre [1986], Fletcher [1987]).

C'est cet algorithme que nous avons choisi pour optimiser la fonction de quatre variables $E_T(K, T_{hold1}, T_{trans}, T_{hold2})$ définie plus haut[5].

**Le problèmes des contraintes**

Les variables temporelles $(T_{hold1}, T_{trans}, T_{hold2})$ ne parcourent pas tout le domaine $[-\infty,+\infty]$, elles sont contraintes par la relation d'ordre suivante :

$$T_{début} < T_{hold1} < T_{hold1}+T_{trans} < T_{hold1}+T_{trans}+T_{hold2} < T_{hold1}+T_{trans}+T_{hold2}+T_{trans} < T_{fin} \quad (1)$$

où $T_{début}$ et $T_{fin}$ correspondent respectivement au début et à la fin du geste [iai].

Il faudrait donc employer plutôt un algorithme d'optimisation avec contraintes. Mais ceci impliquerait un nombre plus élevé de calculs à chaque itération. C'est pourquoi nous procédons à un changement de variables qui permet de contraindre artificiellement les variables temporelles $T_{hold1}$, $T_{trans}$, $T_{hold2}$ de façon à utiliser sans restriction l'algorithme BFGS. Pour ce faire, la relation d'ordre temporelle définie ci-dessus est transformée en plusieurs relations d'ordre pour chaque variable temporelle $T_i$ :

$$T_{i_{\inf}} < T_i < T_{i_{\sup}}$$

---

[7]Remarque sur la méthode des moindre carrés :
Une autre méthode couramment utilisée dans la recherche des paramètres $(x_1, x_2, ..., x_m)=X$, permettant d'ajuster une fonction $y(n,X)$ aux données $y_d(n)$ est la *méthode des moindre carrés* [Fletcher, 1987]. L'idée est de minimiser la fonction

$f(X) = \sum_{n=1}^{Ne} \left[y(n,X) - y_d(n)\right]^2 = \sum_{n=1}^{Ne} \left[r_n(X)\right]^2$ où les $r_n(X)$ sont appelés *résidus*.

Le gradient de $f$ en $X$ vaut donc :
$\nabla f(X) = 2A.R$ où $R$ est le vecteur des résidus et $A$ la matrice des dérivées partielles des résidus :
$A_{ij} = \partial r_j / \partial x_i$
On procède toujours par itérations successives :
$X^{(k+1)} = X^{(k)} - \lambda_k . d_k$
avec $d_k = H_k . \nabla f(X^{(k)})$ où $H_k$ est une approximation de l'inverse du Hessien de la fonction $f$ au point $X^{(k)}$.
Cette approximation minimise les résidus $r_n(X)$ et vaut :
$H_k = (2AA^t + \mu I)^{-1}$ où $\mu$ peut être positif ou nul.
(lorsque $\mu$ est nul c'est la méthode de *Gauss-Newton*, dans les autres cas, c'est la méthode de *Levenberg-Marquardt*).
Ces méthodes présentent des qualités de convergence similaires à celles des méthodes quasi-newtoniennes et en particulier à celles de l'algorithme BFGS [Fletcher, 1987]. Des méthodes hybrides ont même été retenues qui, selon les conditions, utilisent soit la formule BGFS, soit celle de Gauss-Newton.
Cependant, lorsque les dérivées partielles des résidus sont inconnues, la méthode de Gauss-Newton présente un coût computationnel beaucoup plus élevé qui la désavantage.
Après avoir effectué quelques essais avec une méthode des moindre carrés, nous sommes revenus à la méthode BFGS. En effet, cette méthode ne présentait pas de caractéristiques significativement meilleures dans notre cas.

Les différentes bornes inférieures et supérieures sont ajustées afin que la relation (1) soit vérifiée. La fonction à optimiser est donc finalement $E_T$ ($K$, $U_{hold1}$, $U_{trans}$, $U_{hold2}$) où les nouvelles variables temporelles $U_i$, définies par les relations suivantes, contraignent les variables $T_i$ à évoluer dans un intervalle limité $\left[T_{i_{\inf}}, T_{i_{\sup}}\right]$ :

$$U_i = \tan\left[\frac{\pi}{T_{i_{\sup}} - T_{i_{\inf}}}\left(T_i - \frac{T_{i_{\inf}} + T_{i_{\sup}}}{2}\right)\right] = \tan T'$$

$$T' \in \left[-\frac{\pi}{2}, +\frac{\pi}{2}\right]$$

Les différents paramètres $T_{inf}$ et $T_{sup}$ sont estimés tels que la relation d'ordre (1) soit bien vérifiée.

D'autre part, le paramètre de cocontraction $K$ est contraint à rester positif si le système a un quelconque sens physique. En outre, pour que $K$ soit du même ordre de grandeur que les paramètres temporels, nous divisons la variable à optimiser qui lui correspond par $10^5$. Finalement, nous entrons donc dans l'équation différentielle non pas directement $K$ mais la valeur $10^5\sqrt{U_K{}^2}$, $U_k$ étant la variable optimisée par l'algorithme.

**L'implémentation**

Cet algorithme a été mis en œuvre sous le logiciel MATLAB™ qui contient une boîte à outils d'optimisation fournissant diverses routines d'optimisation, dont l'algorithme BFGS. Le gradient de la fonction à minimiser ($\nabla E_T$) n'étant pas disponible de façon explicite, il est calculé avec une méthode aux différences finies. Comme nous l'avons vu précédemment, toute méthode d'optimisation met en œuvre une procédure de *minimisation unidimensionnelle* qui consiste à rechercher le pas de déplacement $\lambda$ qui minimise $f(X^{(k)}+\lambda d_k). = f(\lambda)$. Cette procédure met en œuvre une interpolation de la fonction $f$ pour obtenir une valeur de $\lambda$ dans un intervalle de valeurs convenables. L'algorithme que nous utilisons ici effectue une interpolation de la fonction $f$ avec un polynôme d'ordre 3.

La résolution de l'équation différentielle utilise, à chaque itération de l'algorithme d'optimisation, la méthode de Runge-Kutta. Remarquons que ces calculs supplémentaires ralentissent considérablement la méthode. Pour éviter de résoudre l'équation différentielle, une idée intéressante est de minimiser directement : $E = \ddot{y}_d + b\dot{y}_d + K(y_d - y_e)$. Cependant si $E$ ne tend pas vers 0 cela ne signifie pas nécessairement que les trajectoires simulées sont loin des données, mais tout simplement que la courbe des données n'est pas exactement du second ordre. Pour cette raison, nous avons préféré résoudre l'équation différentielle à chaque itération et calculer l'erreur à partir de la différence entre la solution de l'équation différentielle et la trajectoire donnée.

Aucune information n'étant disponible *a priori* sur la position du minimum, nous procédons en plusieurs passes jusqu'à ce que l'erreur $E_T$ soit satisfaisante. Des valeurs

initiales sont estimées "à vue d'œil" pour les paramètres temporels et l'on cherche d'abord à estimer le paramètre $K$. L'algorithme est ainsi lancé depuis plusieurs valeurs initiales de $K$ ($K$ variant de 1000 à 10000$s^{-2}$ par pas de 100)[6], les paramètres temporels prenant toujours comme valeurs initiales celles estimées. À la fin de la première passe, les valeurs des paramètres minimisant $E_T$ sont conservées et l'algorithme est relancé depuis plusieurs valeurs initiales du paramètre $T_{hold1}$ cette fois, les autres paramètres prenant pour valeur initiale la valeur obtenue en fin de première passe. On procède de même pour les deux autres paramètres. Si l'erreur augmente au cours d'une nouvelle passe, on revient sur les résultats de la passe précédente, et l'on choisit un point différent de l'espace des quatre paramètres comme point de départ de cette nouvelle passe. Cette procédure, assez longue reconnaissons-le, finit toutefois par fournir des paramètres dynamiques correspondant à une approximation acceptable de la courbe $y_d(n)$.

### 3.6.4 Résultats

Une approximation raisonnable des courbes de position et de vitesse pour le corps de la langue est obtenue pour les valeurs suivantes des paramètres dynamiques (prosodiques) :

$K$ = 7700 $s^{-2}$
$T_{hold1}$ = 103 ms
$T_{trans}$ = 73 ms
$T_{hold2}$ = 37 ms

Rappelons que les valeurs successives de la commande posturale (positions d'équilibre) pour [i], [a] et [i] ont été établies au paragraphe 3.6.3.1.

La figure 3.22 présente ces résultats. Le premier panneau contient les courbes de position estimée (trait plein) et "donnée" (trait tireté) du corps de la langue. La ligne brisée en trait mixte est la trajectoire de la commande d'équilibre. Le deuxième panneau contient les courbes de vitesse avec les mêmes conventions de traits.

Le niveau de cocontraction inféré par cette optimisation est relativement élevé. Appliqué à une masse de 50g sur un déplacement de 1cm (pour prendre un cas proche de la langue), il correspondrait à une force générée $F$ de 3.85N. Ce niveau est à comparer avec les niveaux de force observés par Dworkin, Aronson & Mulder [1980] : de 13 à 33N pour des locuteurs masculins à qui l'on demandait de maintenir la langue appuyée contre un capteur placé entre les dents pendant 7 secondes, la consigne étant de maintenir une contraction maximale.

[8]Pour une masse de 50g et un déplacement de 1cm, de telles valeurs du niveau de cocontraction correspondent à des forces de 0.5N à 5N.

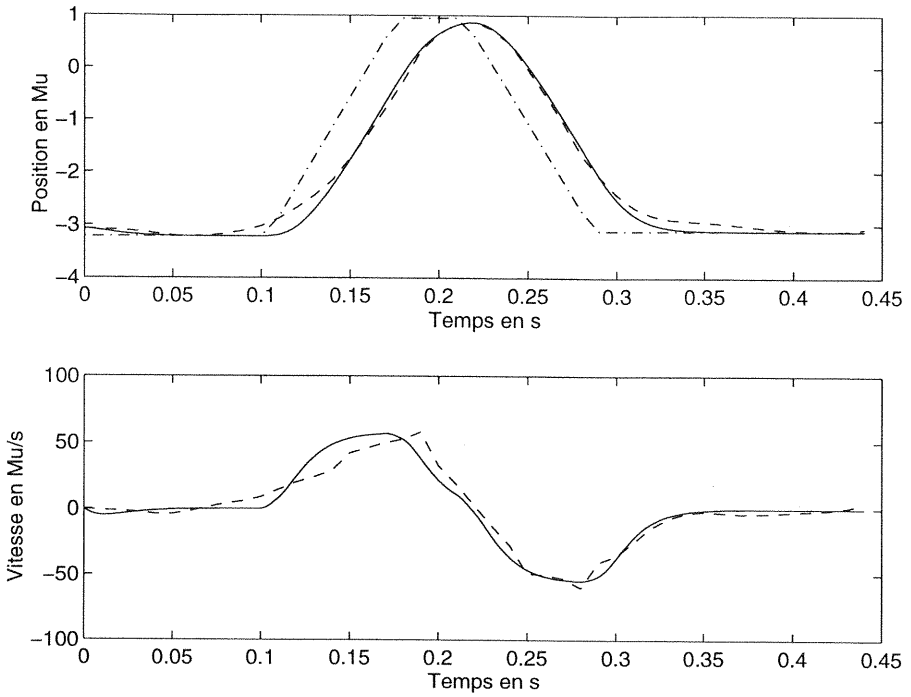



*Figure 3.22.* Résultats de l'inversion dynamique pour la séquence [iai].
Traits tiretés : données, traits pleins : simulations, trait mixte : trajectoire de la commande centrale d'équilibre (commande posturale).

Ce niveau de cocontraction élevé est en accord avec l'idée intuitive que l'on se fait de la cocontraction ou de la raideur. Rappelons les conclusions de Cooke [1980] au sujet d'une expérience de suivi de cibles visuelles avec l'avant-bras dans diverses conditions de précision et de vitesse. Dans cette expérience, la relation linéaire entre l'amplitude du mouvement et le pic de vitesse présente une pente plus élevée dans le cas des mouvements précis ou très rapides que dans le cas des mouvements moins rapides ou moins précis. Cette pente correspond dans le modèle du second ordre de Cooke à la raideur du système. Cooke montre qu'en augmentant la raideur du système il est possible d'augmenter la pente de la relation linéaire et il fait la relation avec la sensation de tension ou cocontraction concomitante à un mouvement très rapide :

"*[...] changing the initial stiffness changes the slope of the peak velocity-amplitude curve for movements in the model: an increased slope is produced by increasing the resting stiffness. This observation accords with the common experience of tensing or co-contracting in the expectation of performing a very rapid movement.*"

Nous reviendrons en détail sur la signification prosodique de la cocontraction au paragraphe 4.3.

En conclusion, le mouvement du corps de la langue étudié ici correspond à la production de la séquence [iai] pendant laquelle on a explicitement demandé au locuteur d'accentuer la voyelle [a]. Cette attention portée sur la voyelle centrale est liée dans l'espace

des commandes motrices à un niveau élevé de cocontraction. Cette variable cocontraction semble donc avoir une signification "physiologique" en accord avec les études présentées ci-dessus. Toutefois, il reste évidemment à vérifier que pour de plus faibles accentuations, le niveau de cocontraction diminue par rapport à celui obtenu ici. Nous nous y attacherons au chapitre 4.

## 3.7. Un test de synthèse à partir des commandes centrales

Pour vérifier que les commandes centrales obtenues par l'inversion chaînée sont pertinentes des points de vue articulatoire et acoustique, nous proposons un test de synthèse du signal acoustique. En premier lieu, les trajectoires articulatoires sont reconstruites : la trajectoire du corps de la langue est générée à partir des commandes centrales inférées ; le dos de la langue est maintenu constamment égal à sa valeur neutre de façon à s'assurer que les mouvements de la langue soient bien entièrement synthétiques ; les cinq autres paramètres articulatoires conservent les valeurs originales obtenues par l'inversion cinématique. Ces sept trajectoires articulatoires sont ensuite fournies au modèle de Maeda décrit au paragraphe 3.5.1.1. La fonction d'aire et les formants sont finalement obtenus par la méthode décrite au paragraphe 3.5.4. La figure 3.23 donne le sonagramme du signal synthétique pour la séquence [iai] lente et accentuée. Le rapprochement F1/F2 caractéristique du /a/ dans cette condition prosodique est bien généré. L'inversion globale, depuis le signal acoustique jusqu'aux commandes motrices est par conséquent acceptable.

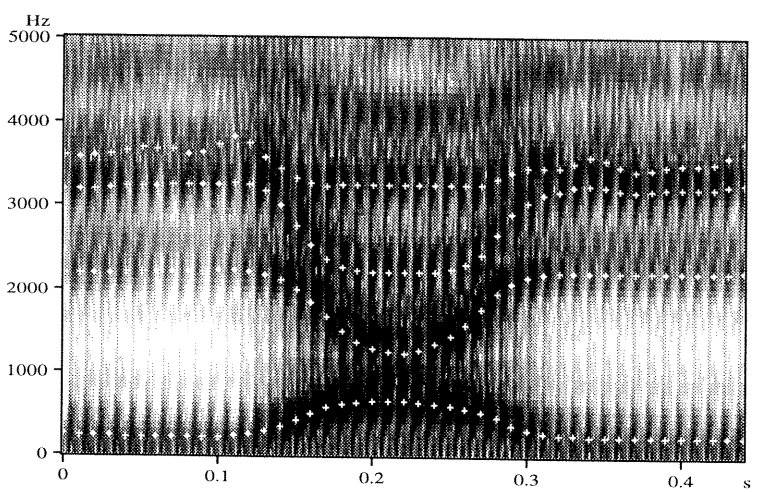



*Figure 3.23.* Sonagrammes synthétiques obtenus à partir des commandes centrales inférées par inversion globale, pour la séquence [iai].

# 3.8 Synthèse adaptative

## 3.8.1 Différents exemples de synthèse adaptative sur [iai]

Notre hypothèse est que la variabilité prosodique peut être générée, pour une séquence de parole donnée, en altérant les commandes centrales liées à la paramétrisation dynamique du mouvement, *i.e.* le niveau de cocontraction et le *timing* de la commande d'équilibre tout en gardant intacte la commande centrale liée aux phonèmes, *i.e.* les positions d'équilibre successives.

Pour tester cette hypothèse, une procédure de synthèse adaptative est mise en place, utilisant la même méthode que pour le test de synthèse décrit au paragraphe 3.7.

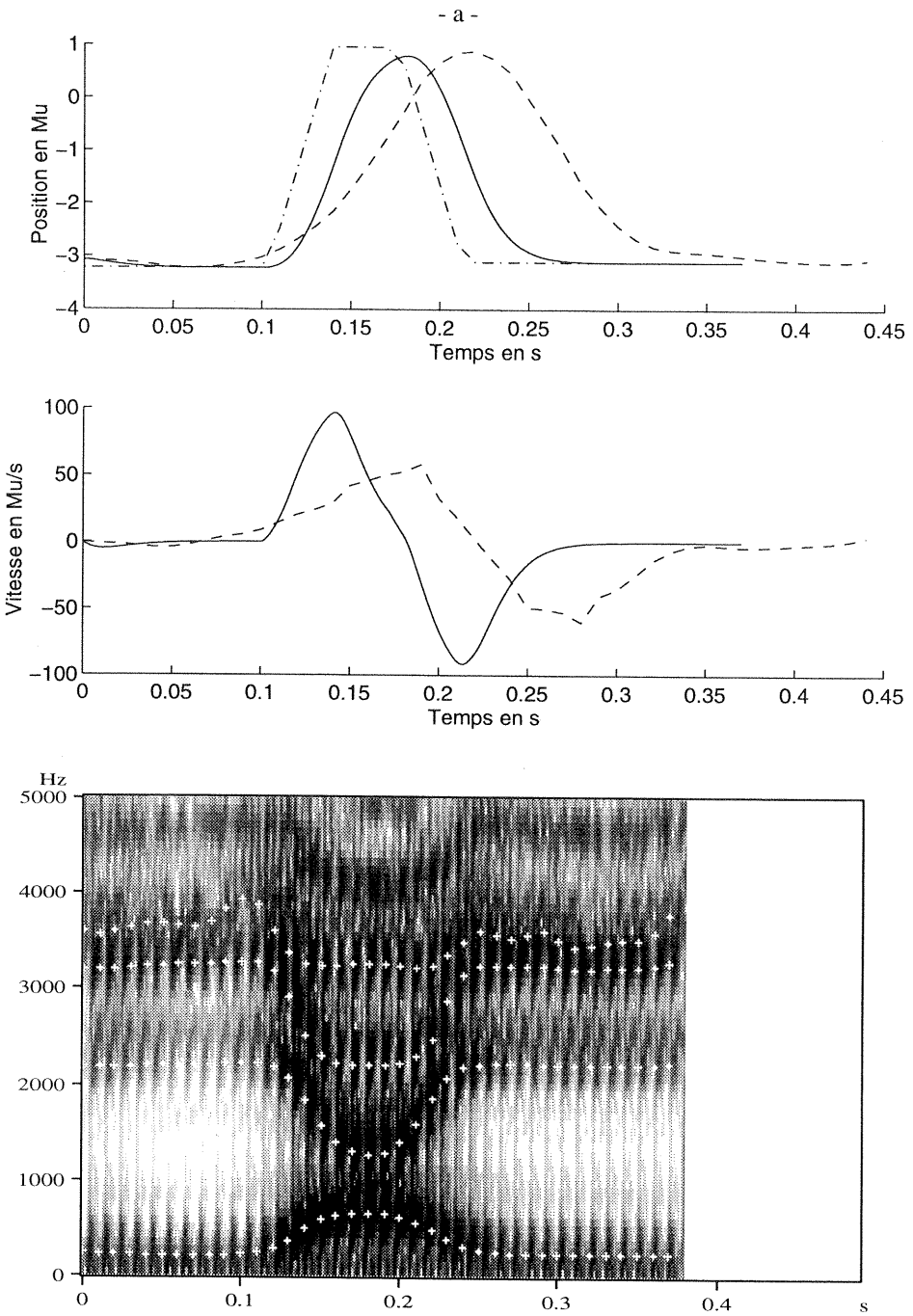
Tout d'abord, nous nous intéressons aux paramètres qui peuvent rendre compte du phénomène de réduction vocalique observé empiriquement (cf. paragraphe 3.3). Reprenant l'idée de Lindblom [1963], nous proposons un premier essai pour lequel la durée globale de la voyelle est réduite. Pour cela, on peut réduire soit le temps de maintien de la voyelle [a], soit le temps de transition, soit ces deux temps. La trajectoire du corps de la langue est donc calculée dans trois conditions :

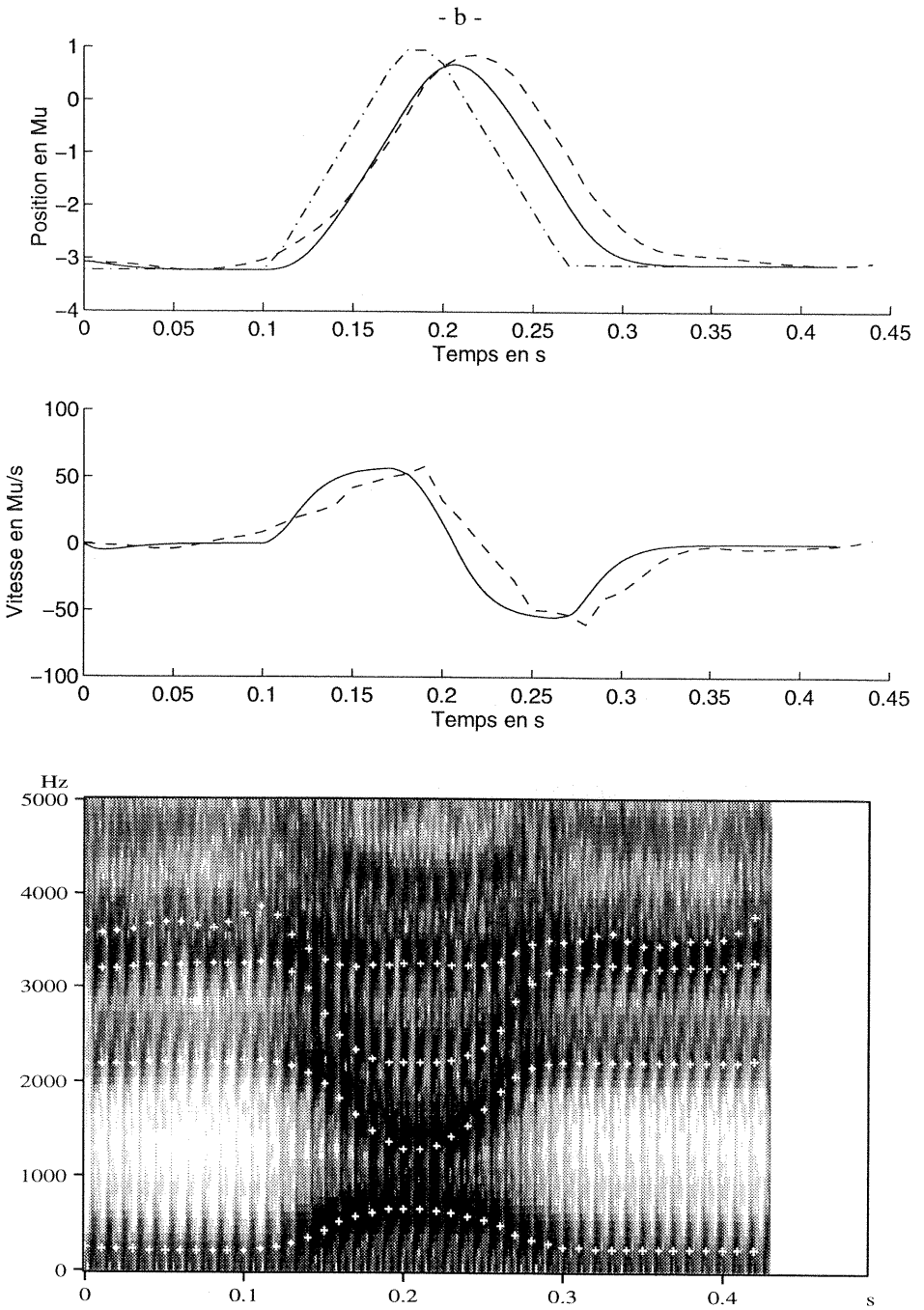
- condition a : $T_{trans}$ = 36 ms (*i.e.* réduit de 50% par rapport à l'original), les autres paramètres conservant leurs valeurs originales (celles obtenues par inversion globale dans le cas lent non-accentué).

- condition b : $T_{hold2}$ = 19 ms (*i.e.* réduit de 50%), les autres paramètres conservant leurs valeurs originales.

- condition c : $T_{trans}$ = 36 ms, $T_{hold2}$ = 19 ms, $T_{hold1}$, $K$ et $Y_e$ conservant leurs valeurs.

Dans ces trois cas (panneaux supérieurs des figures 3.24 a, b et c) on observe un *undershoot* articulatoire : la cible planifiée n'est pas atteinte. Les valeurs absolues des différences entre la position maximale du [a] atteinte dans le cas lent et accentué *(idéal)* et celles atteintes dans les conditions a, b et c correspondent respectivement à 2.1%, 4.9% et 9.8% de l'amplitude du mouvement *idéal*. L'augmentation typique de l'écart entre F1 et F2 (par rapport aux données originales, cf. paragraphe 3.3) est clairement observée sur les sonagrammes fournis dans les panneaux inférieurs des figures 3.24 et ceci alors que les cibles planifiées, codées au niveau de contrôle moteur en tant que positions d'équilibre, restent identiques. Les trajectoires formantiques obtenues dans ces trois conditions, présentent les caractéristiques observées empiriquement d'une augmentation du débit de parole, l'accentuation restant élevée (cf. figure 3.3.c). Notons toutefois que la diminution de la durée est relativement faible dans les conditions simulées (au maximum de 440ms à 366ms) par rapport à la diminution empiriquement observée (de 440ms à 280ms), ce qui explique que la réduction vocalique soit moins impressionnante que dans la réalité.


- a -
Position en Mu
1
0
-1
-2
-3
-4
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
Temps en s
Vitesse en Mu/s
100
50
0
-50
-100
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
Temps en s
Hz
5000
4000
3000
2000
1000
0
0
0.1
0.2
0.3
0.4
s


- b -
Position en Mu
1
0
−1
−2
−3
−4
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
Temps en s
Vitesse en Mu/s
100
50
0
−50
−100
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
Temps en s
Hz
5000
4000
3000
2000
1000
0
0
0.1
0.2
0.3
0.4
s

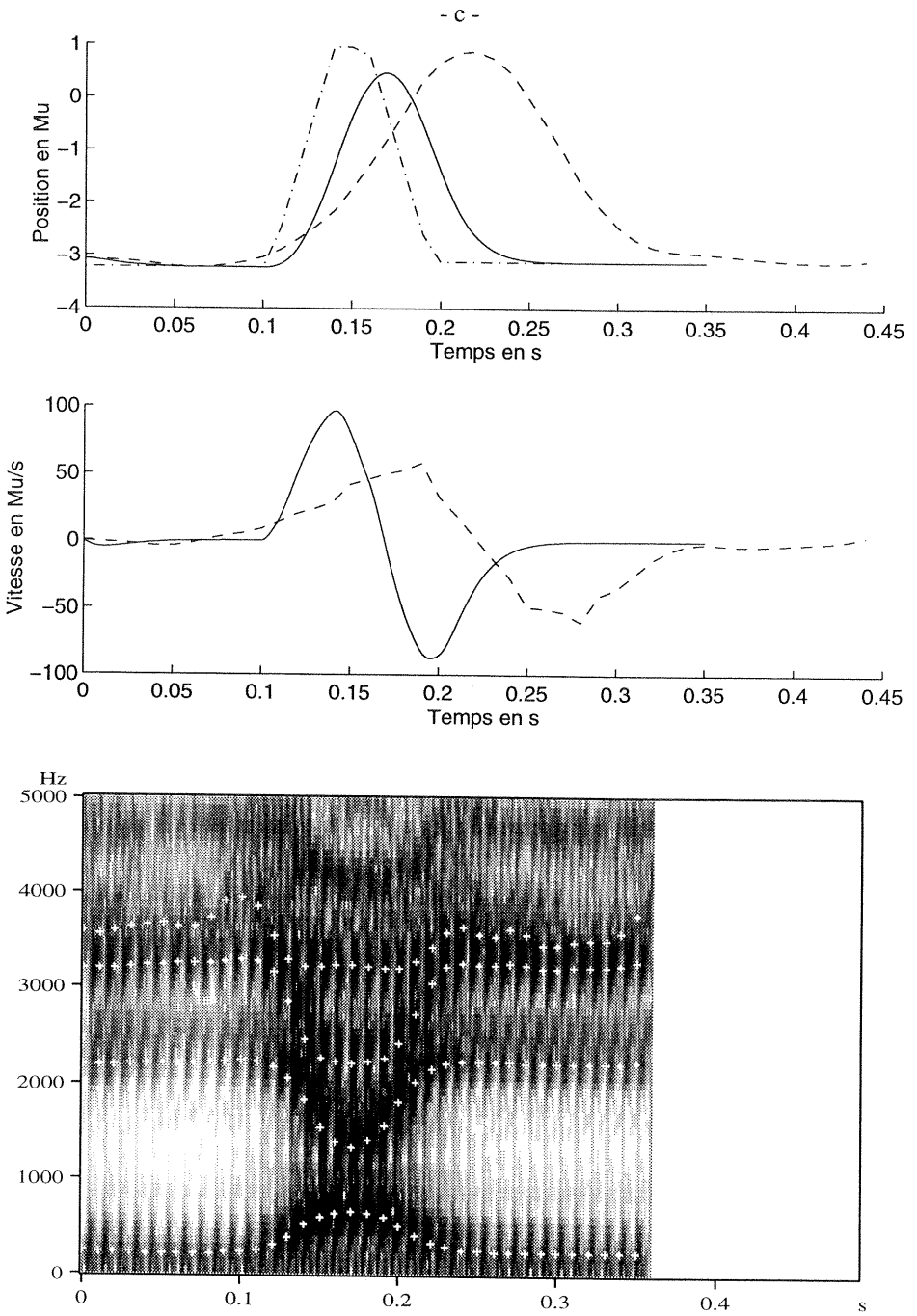
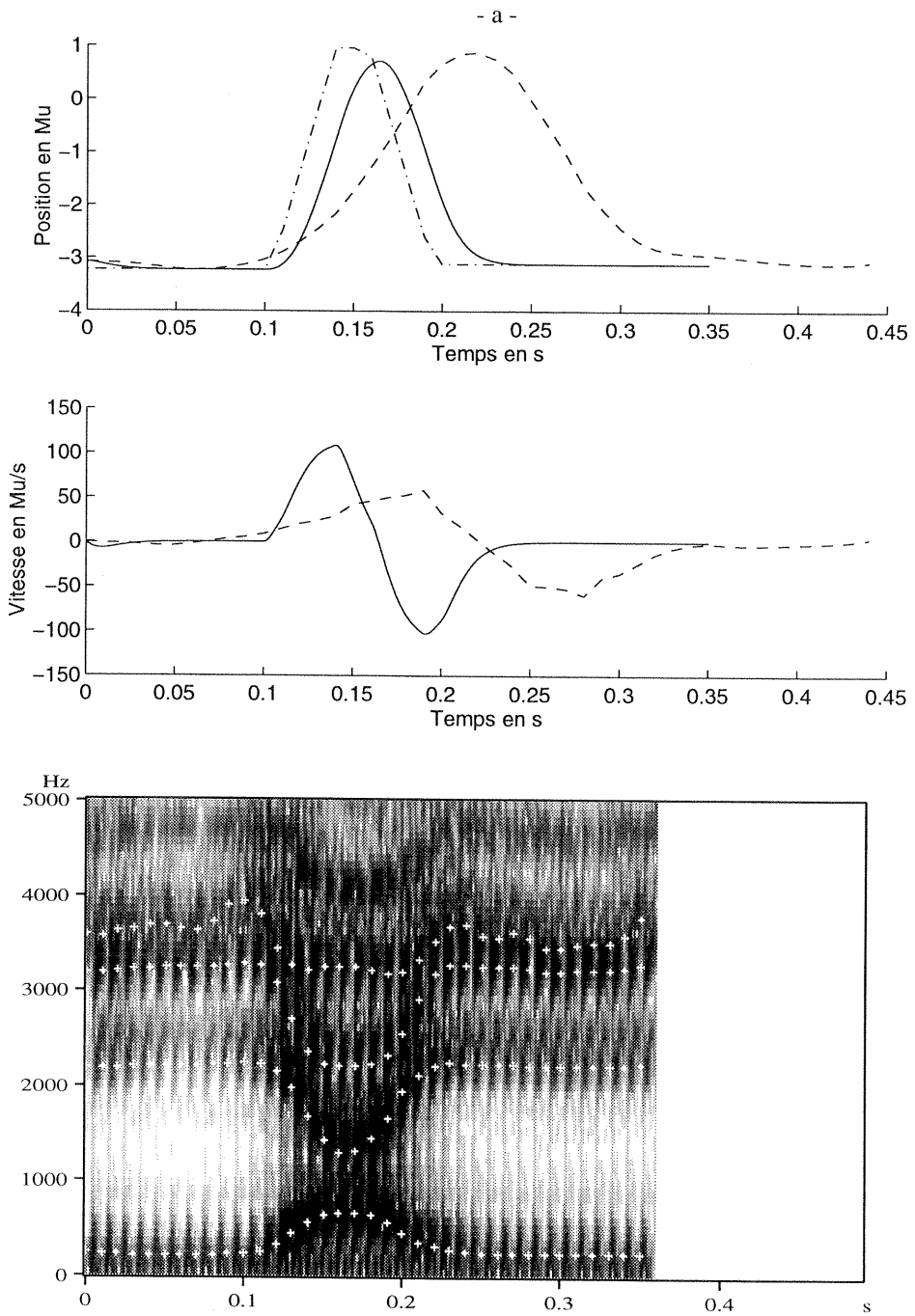



*Figure 3.24.* Synthèse adaptative : trois exemples de réduction à partir de diminutions du temps de transition ou du temps de maintien de la voyelle centrale. a : diminution du temps de transition ($T_{trans}$ = 36ms) ; b : diminution du temps de maintien ($T_{hold2}$ = 19ms) ; c : diminutions des temps de transition et de maintien ($T_{trans}$ = 36ms et $T_{hold2}$ = 19ms). Premiers panneaux : position du corps de la langue ; deuxièmes panneaux : vitesse du corps de la langue ; troisièmes panneaux : sonagramme. Trait continu : simulation, trait tireté : données, trait mixte : commande d'équilibre.

Cependant, comme il a été discuté au paragraphe 3.2, la durée n'est pas systématiquement liée à un *undershoot*. Par exemple, une augmentation suffisante du niveau de cocontraction peut contrecarrer l'effet d'une réduction de la durée. La figure 3.25.a présente les résultats obtenus lorsque le niveau de cocontraction est fixé à une valeur très élevée : $K$ = 15000 $s^{-2}$, alors que les paramètres de durée sont réduits : $T_{trans}$ = 36 ms, $T_{hold2}$ = 19 ms. La position d'équilibre planifiée est alors mieux atteinte que pour les essais précédents : la déviation par rapport à la position idéale maximale du [a] est limitée à 5.3% de l'amplitude (au lieu de 9.8%). Cette simulation peut correspondre prosodiquement à un débit rapide avec un très fort effet d'accent d'emphase.

De façon symétrique, une réduction de la cocontraction, peut induire un *undershoot*. La figure 3.25.b présente un *undershoot* obtenu en fixant la cocontraction à une valeur faible : K = 1000 $s^{-2}$, tout en maintenant les autres paramètres à leurs valeurs originales. Cette simulation (déviation de 18.4%) peut être comparée à une condition lente et non-accentuée (cf. le sonagramme empirique correspondant, figure 3.3.b).


- a -
Position en Mu
Temps en s
Vitesse en Mu/s
Temps en s
Hz
s

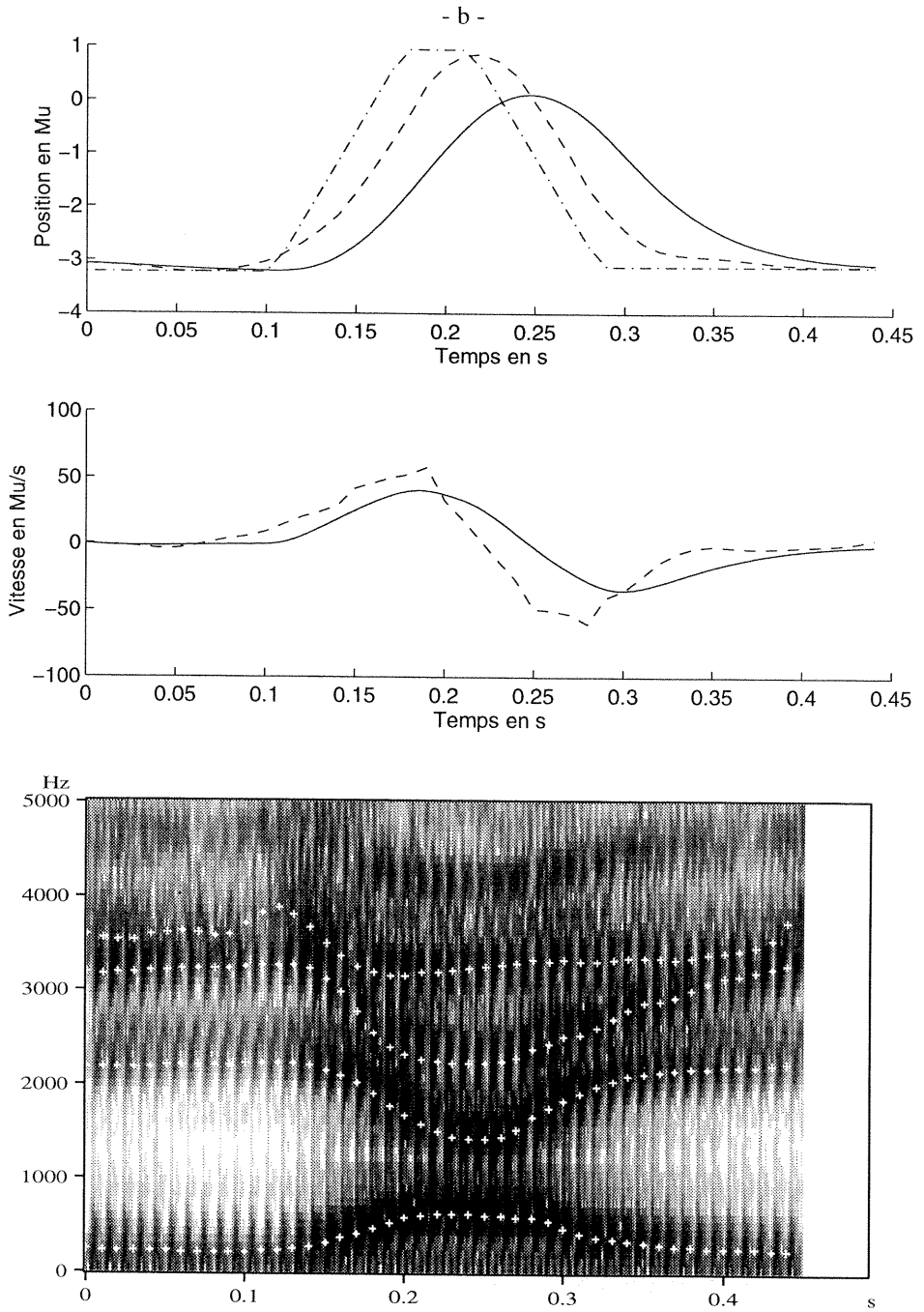



*Figure 3.25.* Synthèse adaptative : conjonctions des rôles de la cocontraction et des paramètres temporels. a : augmentation de la cocontraction et diminution des temps de transition et de maintien ($T_{trans}$ = 36ms, $T_{hold2}$ = 19ms et K = 15000 $s^{-2}$) ; b : simple diminution de la cocontraction (K = 1000 $s^{-2}$). Premiers panneaux : position du corps de la langue ; deuxièmes panneaux : vitesse du corps de la langue ; troisièmes panneaux : sonagramme. Trait continu : simulation, trait tireté : données, trait mixte : commande d'équilibre.

Dans un dernier essai, la conjonction des réductions du niveau de cocontraction et des paramètres de *timing* est étudiée. L'effet de réduction formantique observé est

encore plus fort que celui de chaque réduction séparée. Ce dernier cas (déviation de 43.0%) peut être considéré prosodiquement comme rapide et non-accentué, il est représenté sur la figure 3.26.

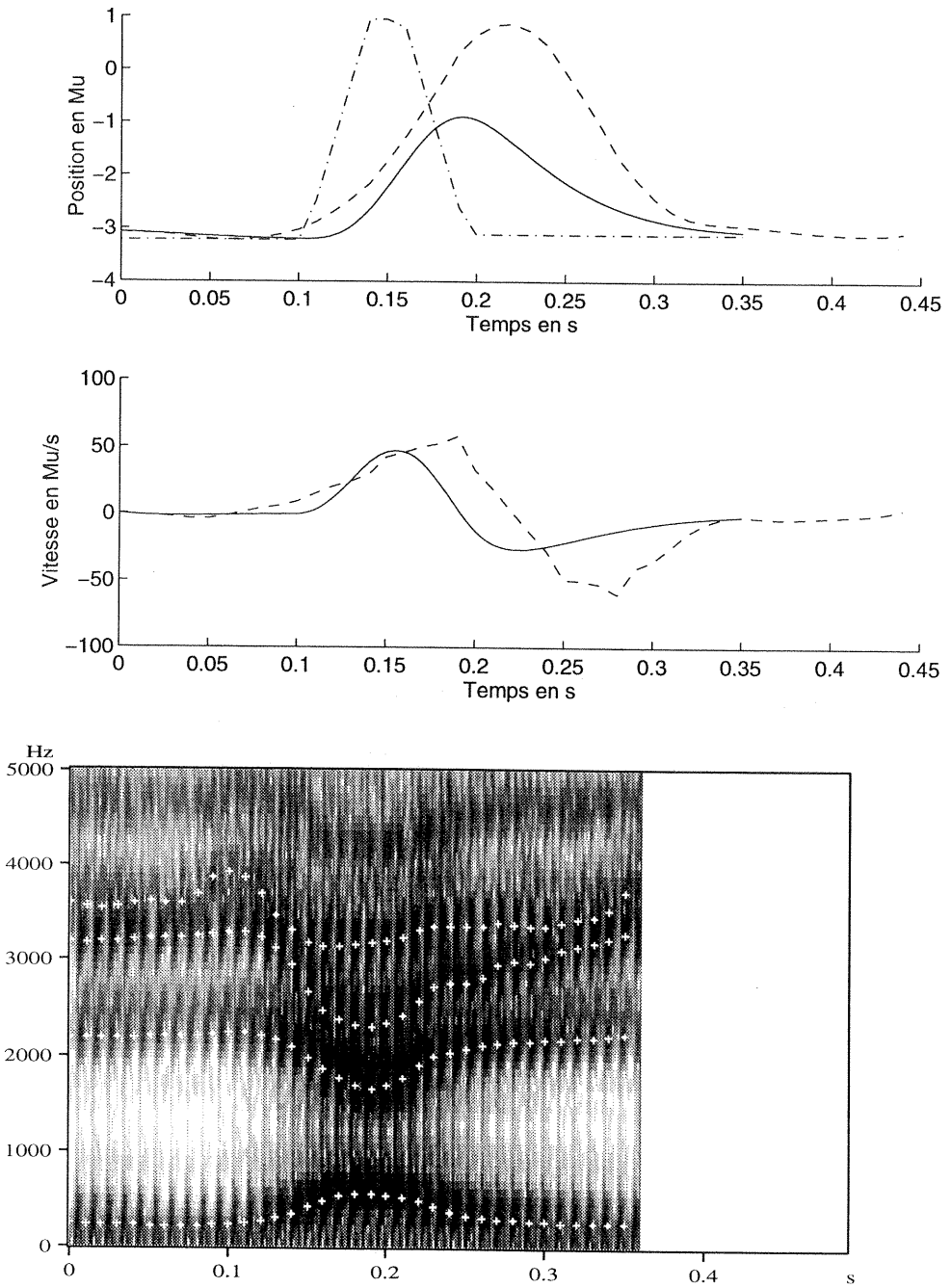



*Figure 3.26.* Synthèse adaptative : conjonctions des diminutions de la cocontraction et des paramètres temporels. Diminution de la cocontraction et des temps de transition et de maintien ($T_{trans}$ = 36ms, $T_{hold2}$ = 19ms et K = 1000 $s^{-2}$). Premiers panneaux : position du corps de la langue ; deuxièmes panneaux : vitesse du corps de la langue ; troisièmes panneaux : sonagramme. Trait continu : simulation, trait tireté : données, trait mixte : commande d'équilibre.

### 3.8.2 Discussion

Le *niveau de cocontraction* influence le niveau de force en jeu dans le mouvement : pour une trajectoire d'équilibre donnée, ainsi que des temps de transitions et de maintien donnés, un niveau de cocontraction minimum est donc requis pour atteindre les positions cibles ; pour des niveaux de cocontraction plus faibles, les cibles ne sont pas atteintes et un phénomène de réduction vocalique est observé.

Parallèlement, le *temps de transition* est corrélé avec la vitesse de la transition entre les différents points d'équilibre. Pour un niveau de cocontraction donné, au-delà d'un certain temps de transition, plus la pente de la transition est élevée, moins le mouvement suit la trajectoire planifiée. Par conséquent, la réduction du temps de transition peut elle aussi donner lieu à réduction vocalique, si le temps de transition est trop court et le niveau de cocontraction trop faible pour que l'articulateur atteigne la position d'équilibre spécifiée. Cependant, lorsque le niveau de cocontraction est suffisamment élevé (15000 $s^{-2}$ dans l'expérience précédente), la réduction du temps de transition peut au contraire aider à suivre le mouvement planifié. En effet, l'augmentation de la vitesse de transition augmente le niveau de force associé, puisque l'écart entre les trajectoires virtuelle et actuelle augmente plus rapidement pendant la première partie du mouvement. Il est par conséquent possible que cocontraction et temps de transition soient liés au contrôle de l'accent d'emphase (cf. 4.3.3 pour une analyse plus approfondie).

Le *temps de maintien* d'une voyelle influence sa durée ; il est donc lié au débit de parole. Une diminution de l'intervalle de temps prévu pour le maintien d'une voyelle entraîne naturellement un phénomène de réduction vocalique. Le débit de parole peut aussi influencer le *temps de transition* ; les modifications des temps de transition et de maintien peuvent coopérer pour suivre ou au contraire éviter la trajectoire d'équilibre planifiée.

Les résultats obtenus indiquent que la cocontraction et le *timing* agissent de paire et peuvent être complémentaires ou compensatoires suivant l'intention du locuteur. Un très haut niveau de cocontraction permet de compenser une réduction des paramètres temporels due à un débit de parole élevé. Ceci est cohérent avec la théorie H&H de Lindblom présentée au paragraphe 1.1 (cf. aussi 3.1) :

*"within limits speakers appear to have a choice whether to undershoot or not to undershoot."*[Lindblom, 1990].

D'autre part, l'idée que la cocontraction et le *timing* ont des rôles propres et distincts est cohérente avec l'hypothèse, proposée par Tuller, Harris & Kelso [1982], que les modulations de débit et d'accentuation ont des effets différents sur les activités

musculaires. À l'aide de mesures des activités EMG des muscles de la parole, ces auteurs montrent en effet que l'augmentation de l'accentuation est liée, chez tous les sujets et pour tous les muscles étudiés, à une augmentation de la durée et de l'amplitude de l'action musculaire. Par contre, à l'augmentation du débit correspondent des effets non uniformes suivant les sujets et les muscles étudiés. Lorsque le débit augmente, la durée peut diminuer, seule ou en parallèle avec une augmentation ou une diminution de l'amplitude qui peut elle aussi augmenter seule. Il semble donc, d'après leur étude, que les sujets modifient leur activité musculaire de façon similaire pour l'accentuation, mais ont recours à des stratégies différentes pour faire varier le débit. Les résultats de Kuehn & Moll [1976] à ce propos sont intéressants. Ils montrent que pour augmenter le débit, les locuteurs ont le choix entre augmenter la vitesse articulatoire, tout en maintenant le déplacement constant, ou conserver la même vitesse qu'en débit lent, en diminuant l'amplitude du mouvement. Par ailleurs Tuller *et al.* [1982], remarquent que la différence observée sur les effets de l'accentuation ou du débit se comprend intuitivement. D'une part, les buts recherchés sont distincts, puisque l'augmentation du débit correspond à une volonté d'accélérer le mouvement des articulateurs, alors que l'augmentation de l'accentuation vise à rendre certaines syllabes plus proéminentes. D'autre part, il est difficile pour un locuteur non-entraîné d'alterner syllabes rapides et syllabes lentes, alors qu'il est très facile d'alterner syllabes accentuées et non-accentuées. De même qu'il est facile de parler à débit constant, mais difficile de maintenir un niveau constant d'accentuation.

Le schéma de contrôle proposé ici, qui distingue paramètres temporels et paramètre de cocontraction, permet de rendre compte des effets différenciés du débit et de l'accentuation observés dans l'étude de Tuller *et al.* et ailleurs (Rakerd, Verbrugge & Shankweiler [1980]).

Notre modèle des effets prosodiques offre un cadre pour expliquer de façon quantitative comment l'*adaptativité* du *robot parlant* peut être mise en place. Les simulations montrent que le *robot* est capable de générer des formes acoustiques variées, incluant des phénomènes de réduction vocalique ou de compensation. De plus, il indique que la variabilité observée aux niveaux acoustique et articulatoire peut être synthétisée sans altérer toutes les commandes centrales : pour un même environnement phonémique, la commande centrale liée aux positions cibles des articulateurs reste invariante au cours des diverses stratégies prosodiques. Voici donc un premier pas dans notre tentative visant à réconcilier les descriptions physiques, phonétiques de la parole, qui révèlent la grande variabilité acoustique et articulatoire de la parole, avec les descriptions phonologiques, qui indiquent que le message linguistique est invariant.

# CHAPITRE IV

# Récupération de cibles :

## évaluation quantitative à partir de modèles dynamiques des articulateurs

Ce chapitre est une tentative de récupération des commandes motrices, cibles posturales et paramètres dynamiques, associées à deux séquences vocaliques [iai] et [iɛi] dans diverses conditions d'élocution. Il est dans la continuité du chapitre précédent et la plupart des principes qui y ont été mis en œuvre, sont réutilisés ici. Rappelons seulement que la récupération des commandes nécessite deux phases. La première est une inversion cinématique, depuis les formants acoustiques jusqu'aux trajectoires articulatoires. Les résultats de cette première inversion sont décrits au paragraphe 4.2. La deuxième phase consiste en l'inversion dynamique d'une trajectoire articulatoire et fournit les commandes centrales recherchées. Elle est présentée au paragraphe 4.3. Afin d'évaluer la pertinence des commandes centrales inférées, nous proposons au paragraphe 4.4 une synthèse du signal acoustique, à partir des trajectoires articulatoires qu'elles génèrent. Des tests perceptifs sont menés au paragraphe 4.5 sur ces signaux acoustiques synthétiques. Nous discutons de nos résultats, en les replaçant dans la perspective de l'existence de cibles vocaliques, au paragraphe 4.6.

# 4.1 Corpus

Le corpus utilisé est celui qui a été présenté au chapitre III, complété de la séquence [iɛi]. Il est donc composé des deux séquences [iai] et [iɛi], issues des phrase porteuse "Il **y a** immédiatement éternué" et "Il **y est** immédiatement retourné" prononcées par un locuteur Français masculin. Les mêmes conditions d'énonciation qu'au chapitre précédent sont considérées : lente et accentuée, lente et non-accentuée rapide et accentuée. La deuxième séquence [iɛ] a été choisie car elle met en œuvre un geste articulatoire s'effectuant dans la même direction que la séquence [ia], *i.e.* geste d'abaissement et de recul de la langue depuis le /i/ vers le /a/, mais de moindre amplitude. Les consignes données au locuteur ainsi que la méthode d'acquisition des données ont été décrites au chapitre III. Les sonagrammes des séquences [iai] ayant été fournis au chapitre III, la figure 4.1 ne présente que ceux des séquences [iɛi], dans les trois conditions d'élocution. Les formants, obtenus après correction et lissage des résultats d'une analyse LPC, sont représentés par des croix blanches.

On peut remarquer sur ces sonagrammes que le rapprochement F1/F2 du [ɛ] est nettement moins marqué que celui observé pour le [a] de la séquence [iai]. L'écart augmente encore pour les cas non-idéaux, *i.e.* lent non-accentué et rapide accentué. Les valeurs formantiques pour les conditions réduites tendent vers [e], ce qui confirme les hypothèses de coarticulation, avec le contexte antérieur du [i], évoquées au paragraphe 3.2.

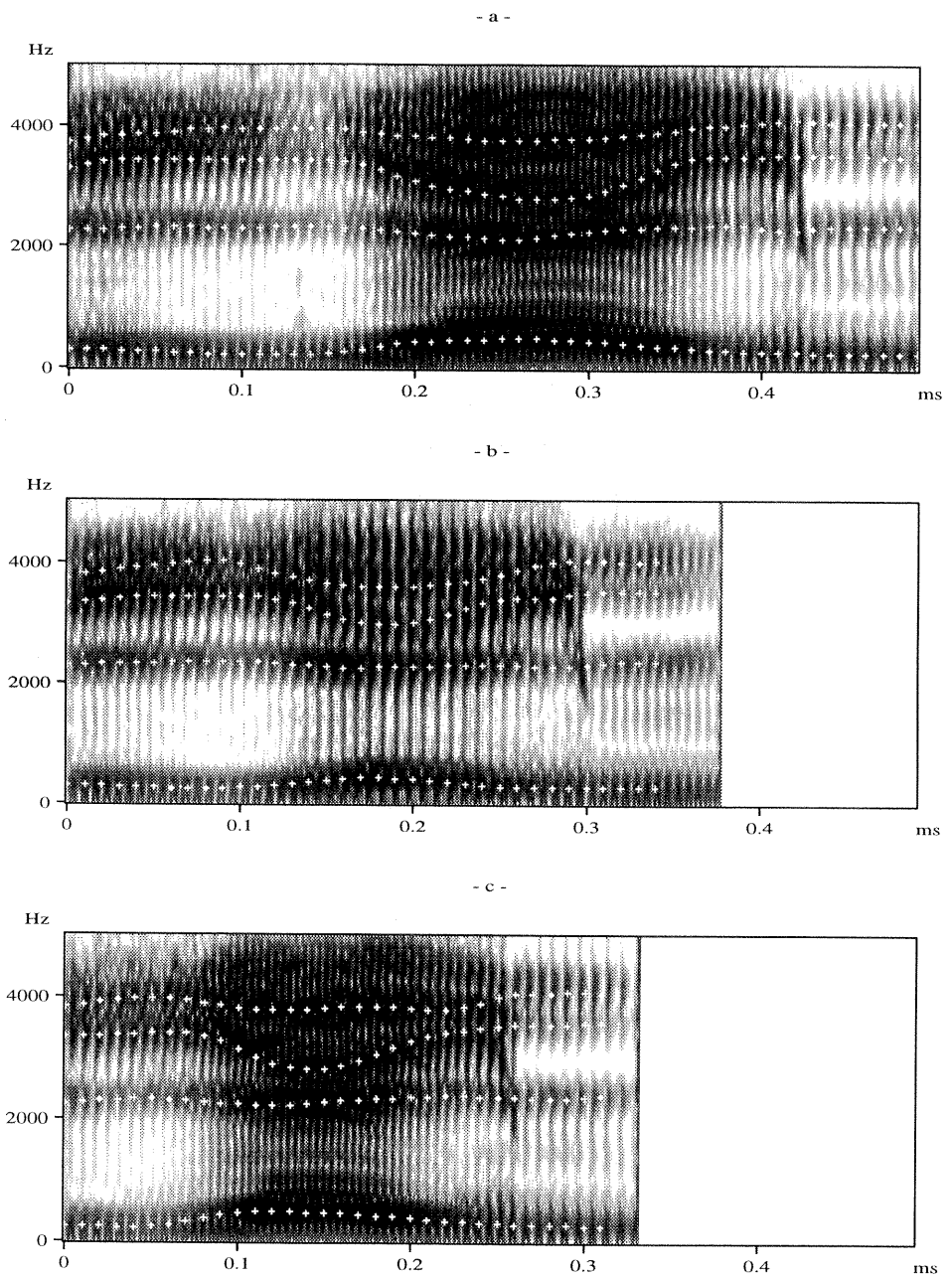



*Figure 4.1.* Sonagrammes de la séquence [iɛi] prononcée par le locuteur JLS dans trois conditions d'élocution. a : lent accentué, b : lent non-accentué, c : rapide accentué.

La figure 4.2 donne les trajectoires [iɛi] dans le plan F1/F2, pour les trois conditions d'élocution. Afin de permettre la comparaison avec les tracés équivalents pour la séquence [iai], les mêmes échelles logarithmiques sont utilisées. On a indiqué sur la figure la position "typique" de la voyelle [e] pour notre locuteur, mesurée au préalable, ainsi que la position du [a] de la séquence [iai]. On note que, comme pour les tracés de [iai], les premiers et les derniers échantillons ont été supprimés, pour plus de clarté.

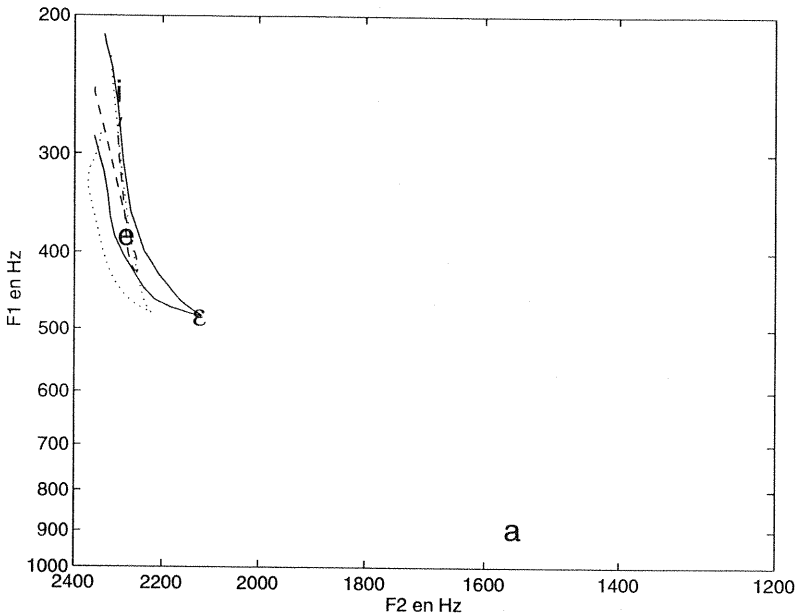



*Figure 4.2.* Séquences [iɛi] dans le plan traditionnel F1/F2 pour les trois conditions d'élocution. Lent accentué : trait plein, lent non-accentué : trait tireté, rapide accentué : trait pointillé.

# 4.2 Inversion cinématique : des formants aux trajectoires articulatoires

## 4.2.1 Inversions de [iai]

L'inversion de [iai] dans le cas idéal ayant été décrite au chapitre III, nous ne présentons ici que les résultats qui concernent les cas lent non accentué et rapide accentué. Les résultats de la normalisation des formants du locuteur JLS ont été présentés au paragraphe 3.5.2.5. Le même réseau de neurones est utilisé, avec la même erreur à rétropropager. Pour les mêmes raisons qu'au chapitre III, l'inversion est forcée principalement sur les deux premiers formants F1 et F2.

### 4.2.1.1 Résultats de l'inversion cinématique

Les résultats de l'inversion pour [iai] en conditions lente non-accentuée et rapide accentuée, sont présentés dans les figures 4.3 a et b. Les trajectoires des deux premiers formants, correspondant aux trajectoires articulatoires obtenues par inversion, sont très proches des données.

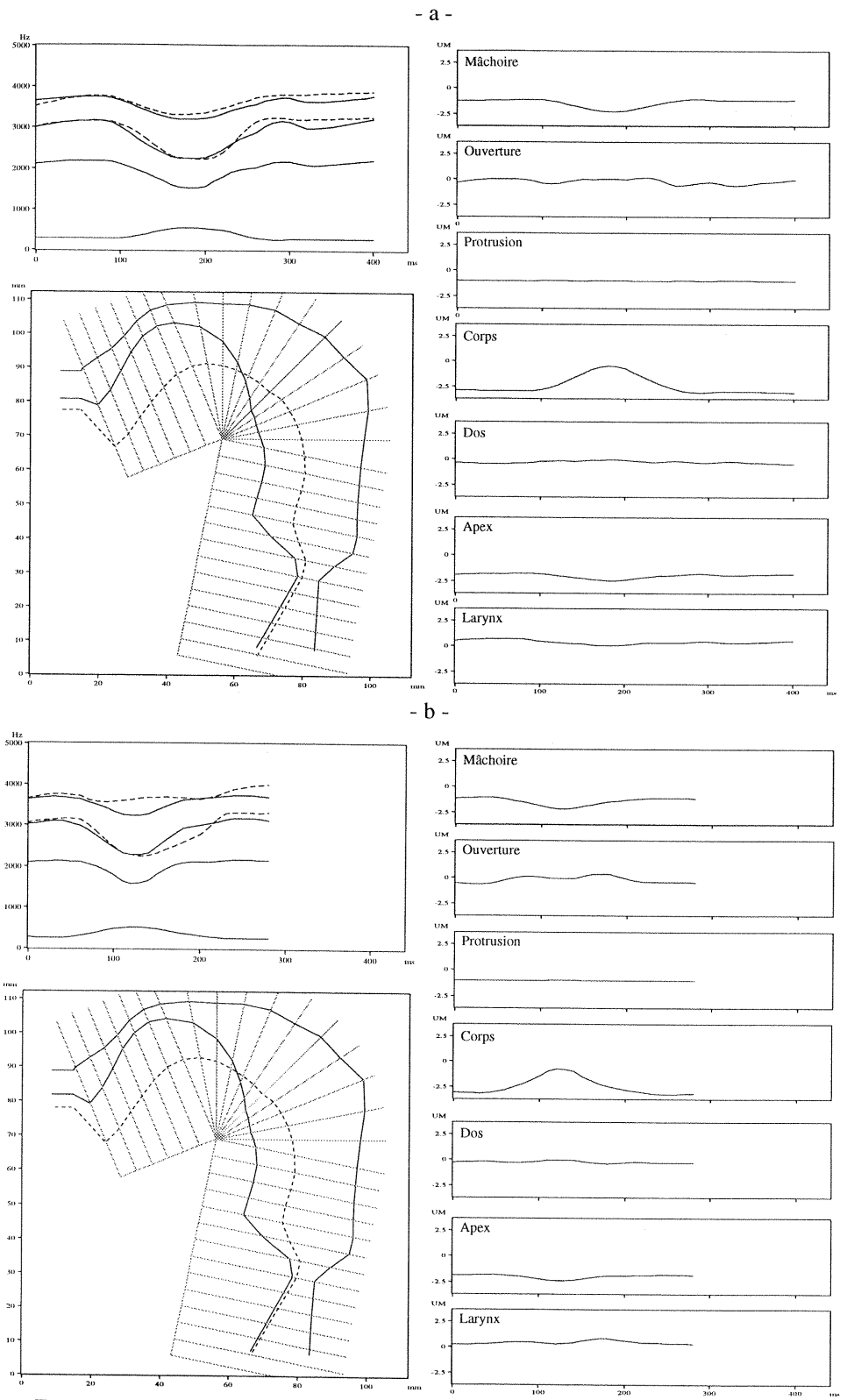



*Figure 4.3.* Résultats de l'inversion cinématique pour la séquence [iai]. Panneau supérieur gauche : données (trait tireté) et simulations (trait continu) formantiques. Panneau inférieur gauche : coupes sagittales du [i] (trait plein) et du [a] (trait tireté). Panneaux de droite : trajectoires articulatoires obtenues par l'inversion. a : lent non-accentué, b : rapide accentué.

### 4.2.1.2 Test de synthèse

Un test de synthèse permet de vérifier que l'information essentielle, sur la qualité phonétique des séquences vocaliques, n'est pas endommagée par l'inversion cinématique. Le même procédé qu'au paragraphe 3.5.4 est mis en œuvre. Les sonagrammes des signaux synthétisés sont présentés sur la figure 4.4. Un test perceptif informel indique que la qualité des [a] synthétiques est similaire à celle des originaux.

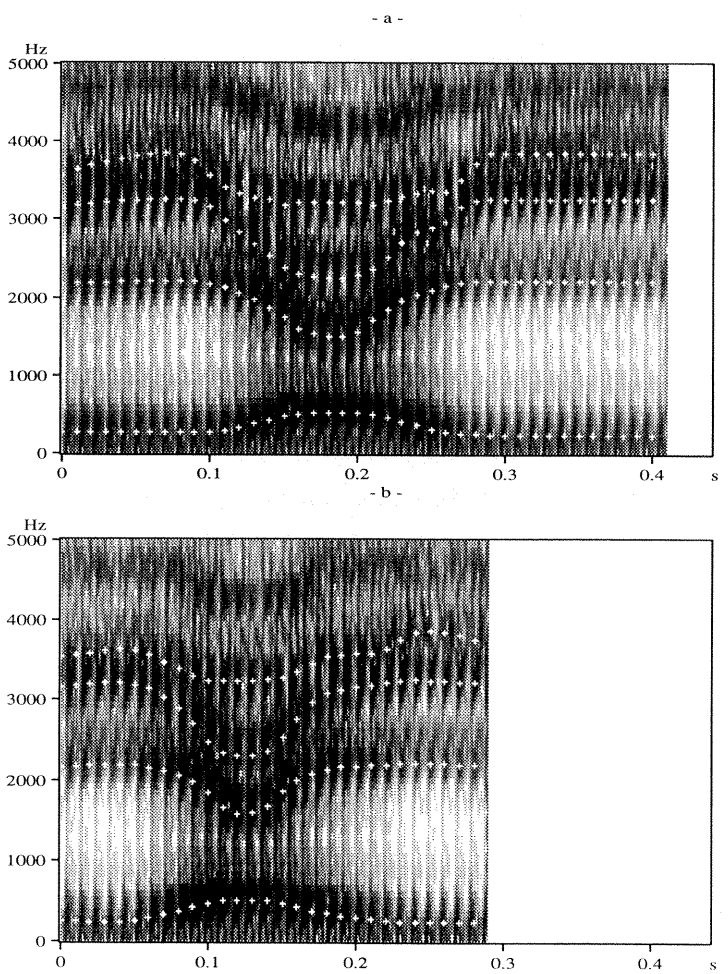



*Figure 4.4.* Sonagrammes de la séquence [iai] obtenus à partir des trajectoires articulatoires inférées par inversion cinématique. Les quatre premiers formants sont représentés par les croix blanches. a : lent non-accentué, b : rapide accentué.

### 4.2.1.3 Comparaison des trois conditions

L'inversion cinématique fournit sept trajectoires articulatoires, dont une, la trajectoire du corps de la langue, est utilisée pour l'inversion dynamique. Il est intéressant de comparer les trajectoires du corps de la langue obtenues dans les trois conditions

d'élocution. La figure 4.5 est une superposition des trois trajectoires, le cas lent accentué étant représenté en trait plein, le cas lent non-accentué en trait tireté et le cas rapide accentué en trait pointillé. Les trajectoires sont exprimées en unités UM.

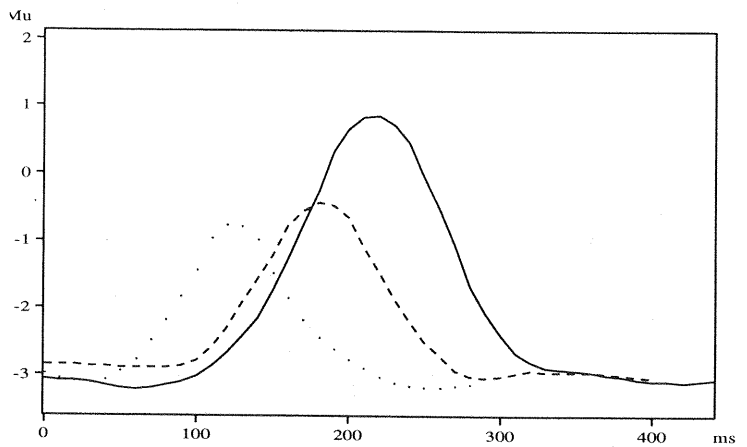



*Figure 4.5.* Trajectoires du corps de la langue pour la séquence [iai], inférées par inversion cinématique dans les trois conditions d'élocution : lente accentuée (trait plein), lente non-accentuée (trait tireté), rapide accentuée (trait pointillé).

On remarque que les trajectoires du corps sont réduites, en amplitude et en durée, dans les cas non-idéaux, par rapport à la trajectoire du cas idéal. D'autre part, l'asymétrie des trajectoires reflète l'asymétrie du contexte phonétique. Rappelons que la séquence [iai] est issue de la phrase porteuse "**il y a im**médiatement...", le premier [i] succède donc à un [l] alors que le second [i] précède un [m]. Le mouvement articulatoire du [l] au [i] n'est pas le symétrique du mouvement du [i] au [m]. Deux phénomènes de coarticulation différents sont en jeu dont les conséquences sont des positions différentes du corps de la langue pour les deux voyelles [i].

## 4.2.2 Inversions de [iɛi]

### 4.2.2.1 Normalisation

**Recherche des affiliations**

Les affiliations du [i] sont les mêmes que pour la séquence [iai] (cf. paragraphe 3.5.2.3). Les formants typiques pour le [ɛ] du locuteur JLS dans la séquence [iɛi] sont :

F1 = 480 Hz

F2 = 2135 Hz

F3 = 2785 Hz

F4 = 3755 Hz

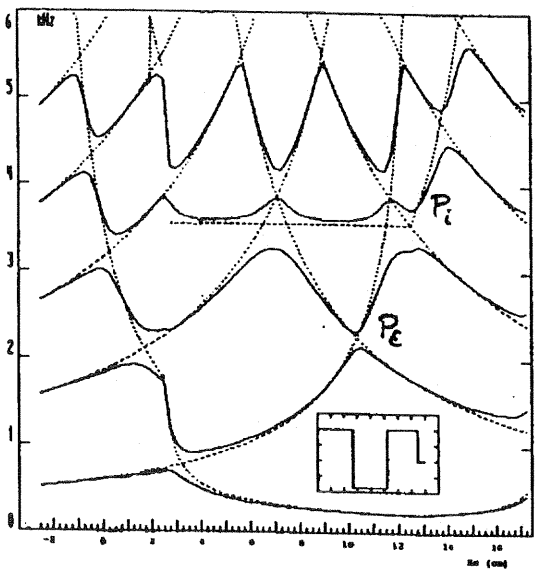



Figure 4.6. Nomogramme pour une ouverture moyenne des lèvres. D'après Badin *et al.* [1988]. Les points focaux $P_i$ et $P_\varepsilon$ sont indiqués.

On observe un écart F1/F2 et un rapprochement F2/F3. Sur le nomogramme ci-dessus, (figure 4.6) la voyelle [ɛ] de la séquence [iɛi] est donc située en arrière ou en avant du point focal $P_\varepsilon$. Or à l'écoute, le [ɛ] tend vers [e], il est donc très antérieur, ce qui est cohérent avec le contexte antérieur /iVi/. On remarque d'ailleurs, sur le tracé de la figure 4.2, que le [ɛ] du locuteur est très proche du [i]. Par conséquent, le [ɛ] est situé en avant du point focal $P_\varepsilon$. Les affiliations du [ɛ] du locuteur JLS sont donc :

F1: cavité arrière + constriction (Helmholtz)

F2: cavité arrière demi-onde

F3: cavité avant quart d'onde

F4: cavité arrière onde

Les valeurs formantiques d'un [ɛ] relativement antérieur standard, pour le modèle de Maeda sont :

F1 = 468 Hz

F2 = 1947 Hz

F3 = 2611 Hz

F4 = 3557 Hz

Les fonctions de sensibilité du modèle de Maeda indiquent que les affiliations sont identiques à celles de notre locuteur.

**Procédure de normalisation**

La procédure de normalisation est la même que celle qui a été utilisée pour la séquence [iai] (cf. 3.5.2.4). Pour normaliser le formant F1, les affiliations du [i] et du [ɛ] étant identiques, on applique un coefficient variant linéairement de $\alpha_{[i]}$ à $\alpha_{[\varepsilon]}$ (les coefficients $\alpha_{[voyelle]}$ sont les rapports des valeurs typiques de F1, du modèle de Maeda et du locuteur, pour la voyelle considérée). Pour les normalisations des formants F2, F3, et F4, on divise la séquence vocalique en cinq régions (plateaux des trois voyelles, et transitions de [i] à [ɛ] et de [ɛ] à [i]). Pour les trois plateaux, on applique respectivement les coefficients $\alpha_{[i]\ cavité}$, $\alpha_{[\varepsilon]cavité}$ et $\alpha_{[i]cavité}$. Les coefficients sont calculés ainsi :

Pour F2 : $\alpha_{[i]\ cavité} = \alpha_{[i]\ AR} = F2_{[i]Maeda} / F2_{[i]JLS}$

$\alpha_{[\varepsilon]\ cavité} = \alpha_{[\varepsilon]\ AR} = F2_{[\varepsilon]Maeda} / F2_{[\varepsilon]JLS}$

Pour F3 : $\alpha_{[i]\ cavité} = \alpha_{[i]\ AR} = F2_{[i]Maeda} / F2_{[i]JLS}$

$\alpha_{[\varepsilon]\ cavité} = \alpha_{[\varepsilon]\ AV} = F3_{[\varepsilon]Maeda} / F3_{[\varepsilon]JLS}$

Pour F4 : $\alpha_{[i]\ cavité} = \alpha_{[i]\ AV} = F4_{[i]Maeda} / F4_{[i]JLS}$

$\alpha_{[\varepsilon]\ cavité} = \alpha_{[\varepsilon]\ AR} = F2_{[\varepsilon]Maeda} / F2_{[\varepsilon]JLS}$

Remarquons que, contrairement au critère utilisé pour la normalisation de [iai], qui consistait à toujours choisir le formant d'ordre le plus élevé, nous avons calculé $\alpha_{[\varepsilon]\ AR}$ sur le formant F2 au lieu de F4. En effet les effets de couplage de cavité induisent des incertitudes sur l'affiliation du formant F4 pour le [ɛ] et il est par conséquent plus sûr d'utiliser un formant d'ordre moins élevé. D'autre part, le formant F2 étant affilié à la cavité arrière sur toute la séquence, il est choisi pour le calcul des coefficients de la cavité arrière, pour les deux voyelles. Ainsi $\alpha_{[i]\ AR}$ est aussi calculé sur un formant d'ordre moins élevé que pour la normalisation de [iai].

Pour les transitions, on applique un coefficient variant linéairement du coefficient de la voyelle initiale à celui de la voyelle finale.

**Résultats**

La figure 4.7 représente les résultats de la normalisation des formants du locuteur JLS dans les trois conditions d'élocution. Les tracés en tireté correspondent aux formants originaux du locuteur (après correction et lissage des résultats de l'analyse LPC), et les formants normalisés sont représentés en trait continu.

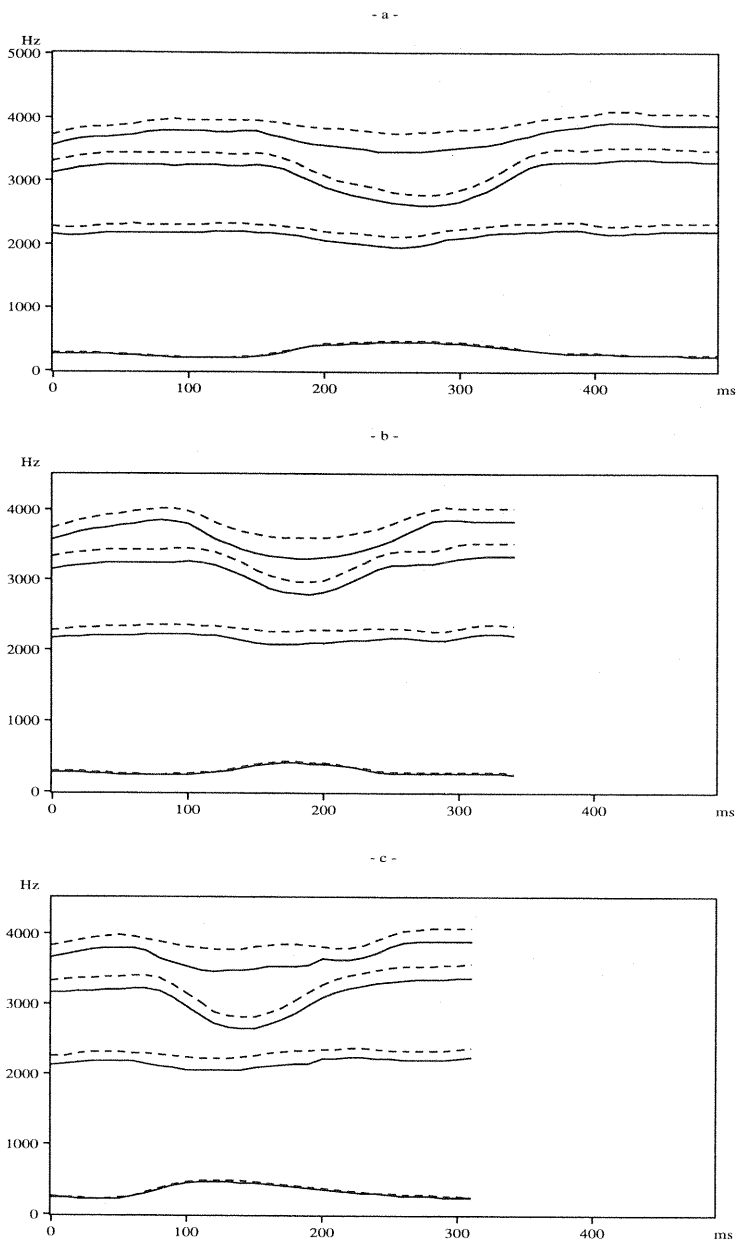



*Figure 4.7.* Résultats de la normalisation des formants du locuteur pour la séquence [iɛi]. Formants normalisés : trait plein, formants originaux : trait pointillé. a : lent accentué, b : lent non-accentué, c : rapide accentué.

L'analyse de ces trajectoires formantiques normalisées révèle l'existence de deux plateaux pour la voyelle centrale. Sur la figure 4.7a (cas lent accentué), le premier plateau apparaît comme un premier creux ou une première aspérité, vers t = 210ms pour F1, F2 et F3 ; le deuxième plateau se trouve vers t = 250ms pour F1 et F2 et vers t = 280ms pour F3. Une première inversion cinématique rapide répercute ces deux plateaux sur les trajectoires des articulateurs et fournit des configurations articulatoires insatisfaisantes. Tentons de

cerner l'origine de ces deux plateaux afin de décider si un lissage plus important, permettant de gommer ces irrégularités, est envisageable.

#### 4.2.2.2 Origine des deux plateaux de la voyelle [ɛ]

L'apparition de deux plateaux pour la voyelle [ɛ] pourrait être due à une désynchronisation des mouvements d'abaissement et de recul de la langue, liée à la coordination mandibule/langue pour l'abaissement.

Examinons l'effet d'un pur mouvement d'abaissement du corps de la langue à partir de la position correspondant au [i]. Sur le nomogramme de la figure 4.6, la position de la constriction serait inchangée, mais l'aire à la constriction augmenterait, ce qui aurait pour effet d'accroître le couplage entre les cavités. Les conséquences d'une légère augmentation de l'aire à la constriction, à partir d'une constriction relativement marquée, comme celle du [i], sont étudiées par Badin *et al.* [1988] (cf. aussi Fant [1960]). Ces auteurs indiquent que, dans ces conditions, F1 augmenterait, F2 et F3 diminueraient et F4 serait quasiment inchangé (ces changements sont représentés par la courbe en trait tireté sur le nomogramme, au niveau du point focal $P_i$).

Maintenant considérons un mouvement pur de recul à partir du [i]. Cette fois l'aire à la constriction demeure constante, mais la position de la constriction est reculée. Sur le nomogramme, on se déplace, depuis la région repérée par le point $P_i$, vers celle repérée par $P_\varepsilon$. Alors le formant F1 reste affilié Helmholtz, mais sa valeur augmente légèrement, F2 reste affilié à la cavité arrière et sa valeur augmente légèrement, F3 passe de cavité arrière à cavité avant et sa valeur diminue, enfin F4 passe de cavité avant à cavité arrière et diminue.

Supposons que, lors de la transition [iɛ], un pur mouvement d'abaissement de la langue ait lieu tout d'abord (sans recul concomitant). On a alors une assez forte augmentation de F1, une diminution de F2 et de F3 et peu d'effet sur F4. Supposons en outre que le mouvement de recul vienne s'ajouter ultérieurement au mouvement d'abaissement et que l'abaissement s'achève alors que le recul n'est pas terminé. Après une première augmentation de F1, induite par le mouvement d'abaissement seul, va s'ajouter l'augmentation due au mouvement de recul, puis F1 ne subira plus que la légère augmentation due au recul. Ceci crée l'aspérité que l'on peut observer sur F1 dans la transition [iɛ], avant le véritable plateau du [ɛ]. Pour F2 et F3, les mouvements d'abaissement et de recul ont des effets contradictoires. Ces formants entament donc une descente, puis une remontée vers le plateau du [ɛ], ce sont les creux observés sur les données. Pour F4, il semble que seul le recul ait un effet visible, F4 diminue donc vers le plateau du [ɛ]. En appliquant le schéma inverse pour la transition [ɛi], on arrive finalement à expliquer l'allure des trajectoires formantiques de toute la séquence [iɛi].

Une analyse rapide de trajectoires de la mandibule et du corps de la langue obtenues récemment, pour le même locuteur, par des chercheurs de l'Institut de la Communication Parlée, à l'aide d'un électro-magnétomètre (EMMA), confirme que pour [iɛi], ces mouvements sont effectivement désynchronisés, l'abaissement de la mandibule démarrant (et terminant) avant le recul du corps de la langue.

Il semble donc que les irrégularités formantiques observées sur [iɛi] proviennent bien de la coordination temporelle de l'ensemble mandibule-langue.

Revenons donc maintenant sur l'étude de la séquence [iai]. Cette désynchronisation devrait se manifester aussi. En considérant de nouveau le nomogramme, mais en s'intéressant cette fois à la région du [a], on voit que l'abaissement et le recul produisent les mêmes effets : forte augmentation de F1 et forte diminution de F2. Ainsi F1 et F2 ne sont pas affectés par la désynchronisation. Pour F3, l'effet dû à l'augmentation de l'aire à la constriction est une diminution importante, mais l'effet dû au recul est bien moindre. On s'attend donc à un creux sur la transition [ia] du formant F3. Si l'on considère en détail les courbes de la figure 3.13, on s'aperçoit que ce creux est bien présent. Cependant comme l'inversion a été forcée sur F1 et F2 uniquement, l'effet de la désynchronisation ne s'est pas fait remarquer.

Le modèle articulatoire que nous utilisons ne permettant pas de mettre en œuvre des hypothèses sur la coordination temporelle des articulateurs, nous devons nous résoudre à négliger ces effets. Nous imposons donc un lissage plus efficace sur les données (avant la normalisation) pour obtenir des transitions plus régulières. Nous devons reconnaître que de telles retouches sont susceptibles de modifier certaines caractéristiques cinématiques des trajectoires articulatoires. Les propriétés du contrôle moteur, que nous inférons à partir de ces données, pourraient donc être en partie l'image de ce lissage. Nous notons cependant que les différences capitales entre les trois conditions d'élocution sont préservées, à savoir, les différences de durée et d'amplitude formantique. Par ailleurs les tests perceptifs que nous présentons à la fin du chapitre incitent à penser que les conséquences de ce lissage sur l'inférence du contrôle sont négligeables.

La figure 4.8 représente les résultats d'une deuxième normalisation à partir de trajectoires mieux lissées.

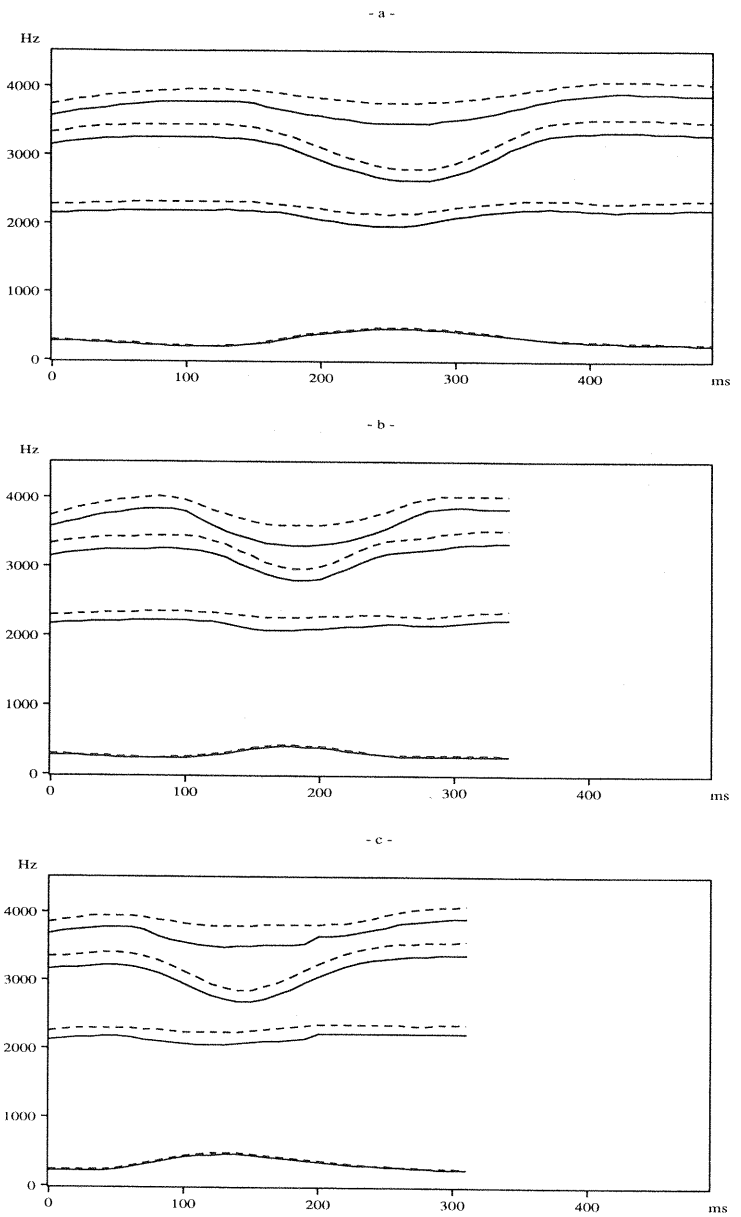



*Figure 4.8.* Résultats de la normalisation des formants du locuteur pour la séquence [iɛi], à partir d'un lissage plus efficace. Formants normalisés : trait plein, formants originaux mieux lissés : trait pointillé. a : lent accentué, b : lent non-accentué, c : rapide accentué.

La figure 4.9 représente les trois trajectoires formantiques dans le plan F1/F2, avant et après normalisation, avec le lissage plus efficace. Les positions "typiques" des voyelles [a, ɛ, e] du locuteur sont indiquées, ainsi que les positions standard de ces mêmes voyelles pour le modèle de Maeda (lettres suivies d'un tiret).

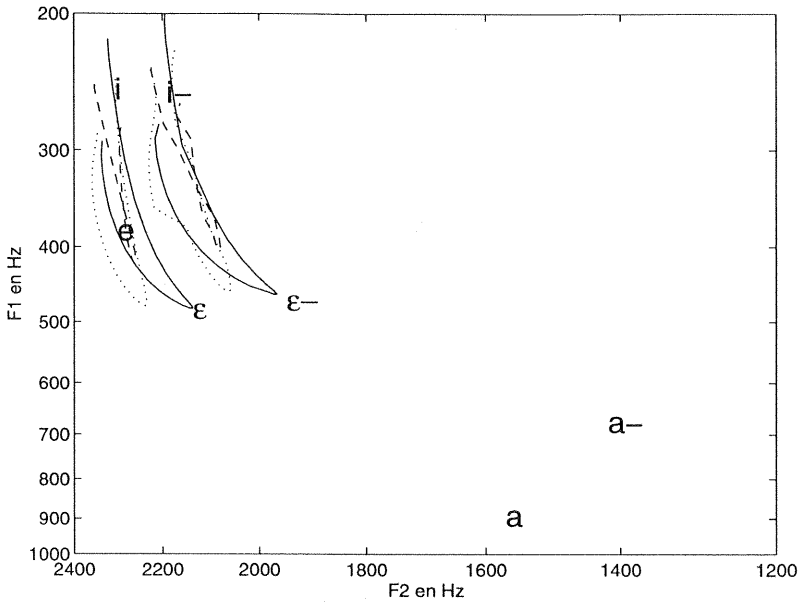



*Figure 4.9.* Séquences [iɛi] dans le plan traditionnel F1/F2, avant et après normalisation, pour les trois conditions d'élocution. Lent accentué : trait plein, lent non-accentué : trait tireté, rapide accentué : trait pointillé.

### 4.2.2.3 Résultats de l'inversion cinématique

Le même réseau de neurones est utilisé que pour la séquence [iai]. Les trajectoires formantiques étant mieux lissées que pour la séquence [iai], le problème du décalage du formant F3 n'apparaît plus. L'inversion est donc forcée cette fois sur les trois premiers formants. L'erreur à rétropropager est similaire à celle que nous avons utilisée pour la séquence [iai] :

$$E_T = J_{conf} + \mu_1 J_{pén}\, e^{-\alpha 1 t} + \mu_2 J_{liss}\, e^{-\alpha 2 t}$$

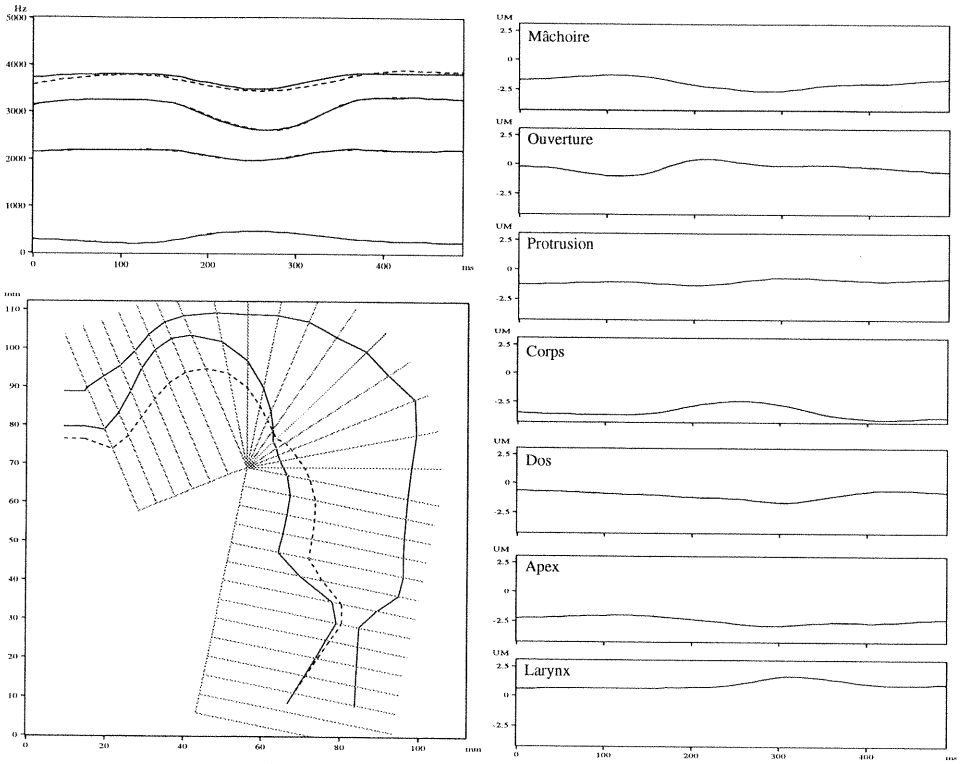
avec les coefficients *ad hoc* :

$\mu_1 = 0.1$ et $\alpha_1 = 0.001$

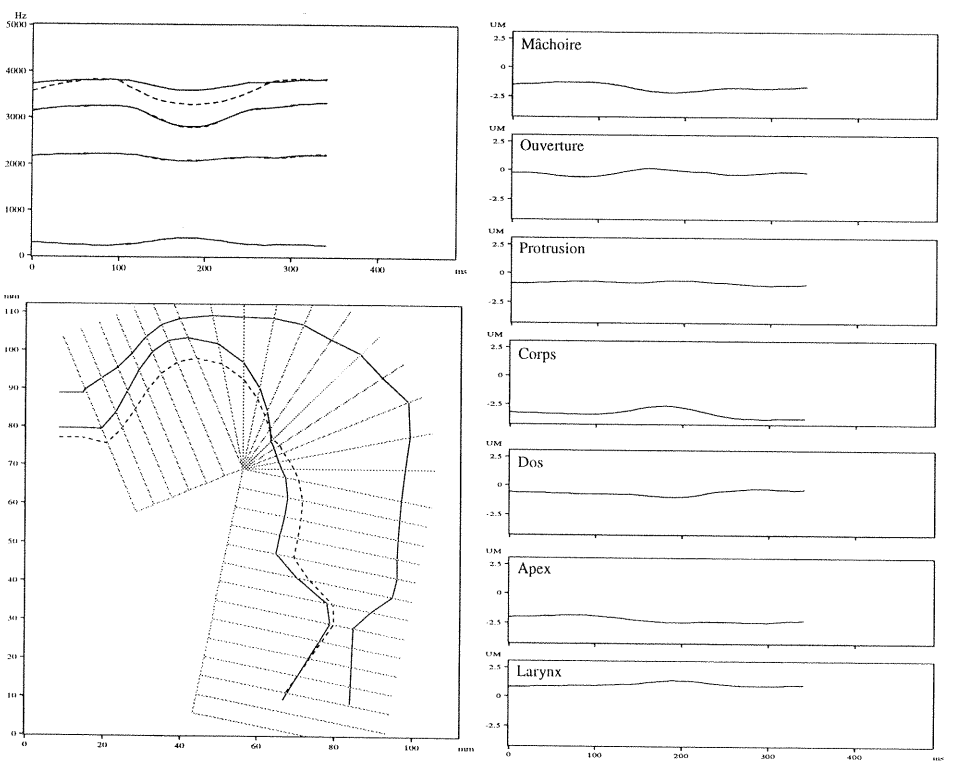
$\mu_2 = 0.1$ et $\alpha_2 = 0.005$

Les figures 4.10 a, b et c fournissent les résultats de l'inversion de [iɛi] pour les trois conditions d'élocution. Les trajectoires formantiques, associées aux trajectoires articulatoires inférées par l'inversion, sont proches des données originales.

- a -


Hz
ms
mm
UM
Mâchoire
Ouverture
Protrusion
Corps
Dos
Apex
Larynx

- b -


Hz
ms
mm
UM
Mâchoire
Ouverture
Protrusion
Corps
Dos
Apex
Larynx

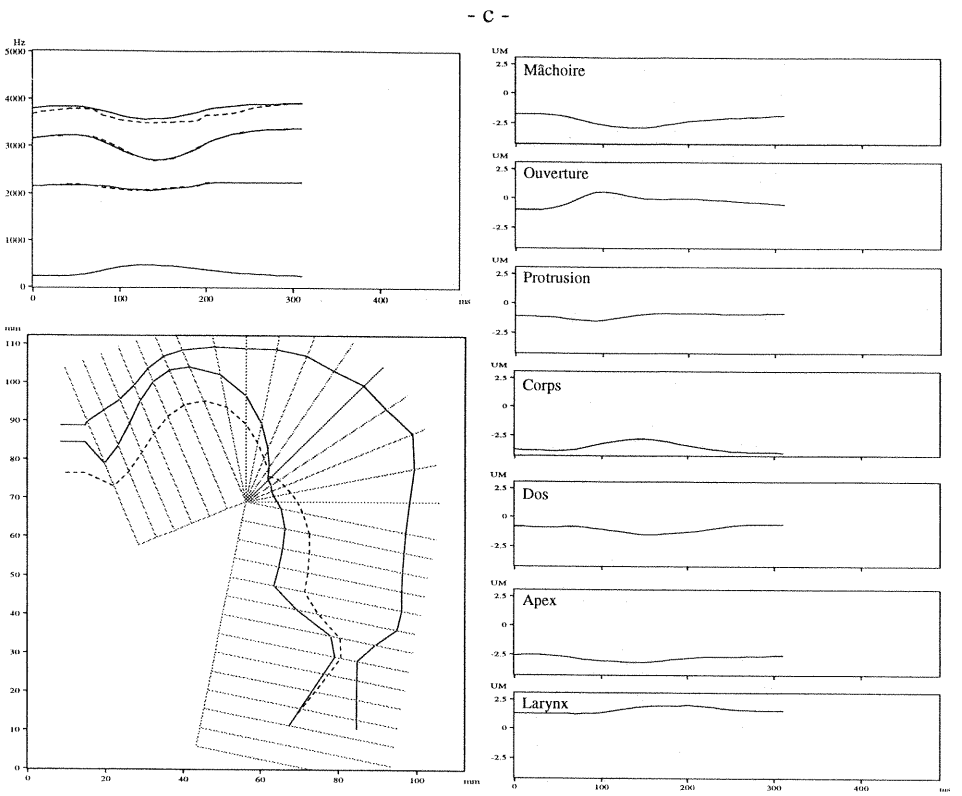
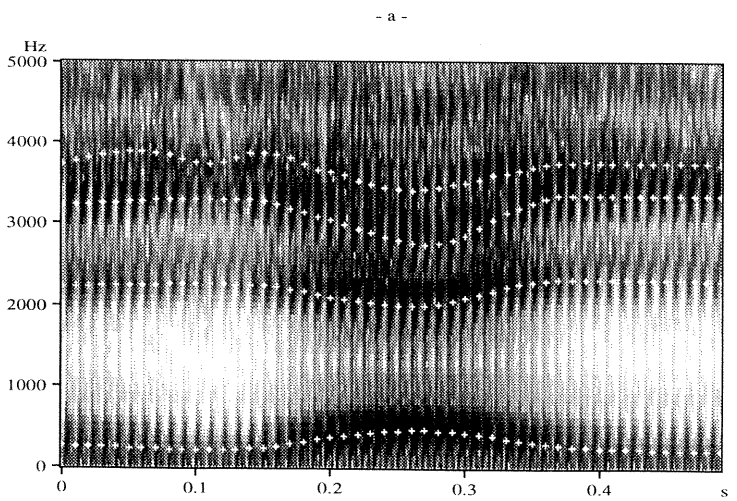



*Figure 4.10.* Résultats de l'inversion cinématique pour la séquence [iɛi]. Panneau supérieur gauche : données (trait tireté) et simulations (trait continu) formantiques. Panneau inférieur gauche : coupes sagittales du [i] (trait plein) et du [ɛ] (trait tireté). Panneaux de droite : trajectoires articulatoires obtenues par l'inversion. a : lent accentué, b : lent non-accentué, c : rapide accentué.

### 4.2.2.4 Test de synthèse

Les trois figures 4.11 a, b et c présentent les sonagrammes des signaux synthétisés. Un test perceptif informel indique que la qualité phonétique des voyelles [ɛ] est préservée.



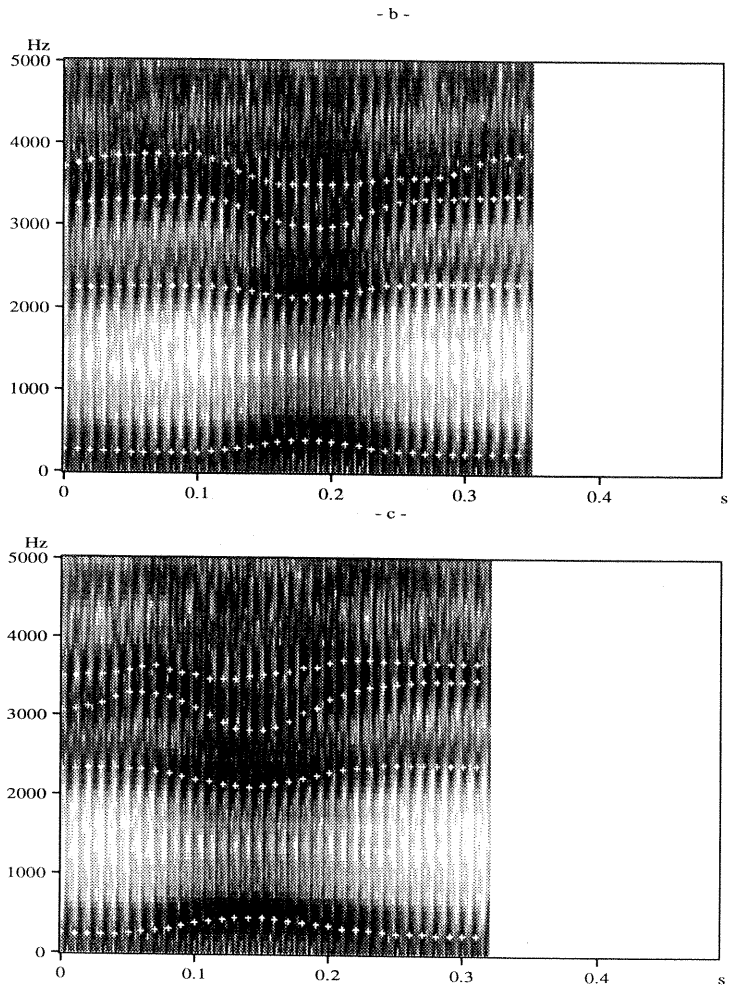



*Figure 4.11.* Sonagrammes de la séquence [iɛi] obtenus à partir des trajectoires articulatoires inférées par inversion cinématique. Les quatre premiers formants sont représentés par les croix blanches. a : lent accentué, b : lent non-accentué, c : rapide accentué.

### 4.2.2.5 Comparaison des trois conditions

La figure 4.12 présente les trois trajectoires du corps de la langue dans les trois conditions d'élocution, avec les mêmes conventions de trait que pour [iai] (cf. 4.2.1.3). Pour permettre une comparaison des amplitudes, la même échelle (à un décalage près), sur l'axe des ordonnées, a été employée que pour [iai].

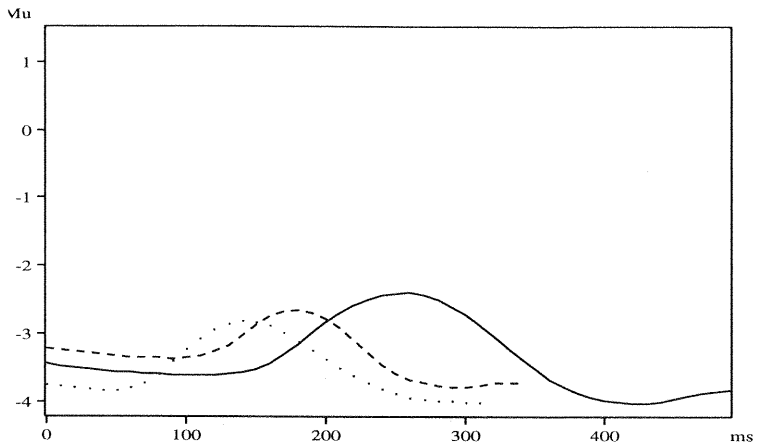



*Figure 4.12.* Trajectoires du corps de la langue pour la séquence [iɛi], inférées par inversion cinématique dans les trois conditions d'élocution : lente accentuée (trait plein), lente non-accentuée (trait tireté), rapide accentuée (trait pointillé).

L'amplitude du mouvement du corps de la langue pour les trois réalisations de [iɛi] est beaucoup plus faible que pour la séquence [iai]. Le mouvement du corps n'est plus nettement, comme pour la séquence [iai], le plus développé des sept mouvements articulatoires. Il est en concurrence notamment avec les mouvements de la mâchoire et du dos. La trajectoire lente accentuée présente la plus large amplitude et la plus longue durée, les deux autres sont réduites en amplitude et en durée. Les positions du corps pour les voyelles [i] varient selon les conditions d'élocution. On note une fois de plus que le choix de ne pas tenir compte des mouvements précédant et suivant la séquence [iɛi] présente des inconvénients. Il sera difficile, dans les cas réduits, de simuler précisément les parties de trajectoire correspondant aux [i], à partir des cibles du [i] lues sur la trajectoire lente accentuée.

# 4.3 Inversion dynamique : depuis la trajectoire articulatoire jusqu'aux commandes motrices

## 4.3.1 Résultats de l'inversion dynamique

### 4.3.1.1 Inversions de [iai]

Les résultats de l'inversion dynamique dans le cas lent accentué ont été présentés au paragraphe 3.6. Rappelons notre hypothèse : la variabilité associée aux effets prosodiques peut être simulée, avec la même séquence de points d'équilibre, en ajustant simplement le

niveau de cocontraction et/ou le *timing*. Nous supposons donc que la commande posturale reste la même dans les trois conditions d'élocution. Les positions d'équilibres inférées au 3.6.3.1, pour le cas lent accentué, sont donc réutilisées pour les cas lent non-accentué et rapide accentué. La même procédure d'optimisation qu'au paragraphe 3.6 est mise en œuvre. Notons que si l'utilisation d'un modèle fonctionnel très approximatif simplifie le contrôle du mouvement, elle complique par contre la tâche d'optimisation. Les propriétés dynamiques des articulateurs étant plus complexes, on sait *a priori* qu'aucune des solutions inférées par la procédure d'optimisation ne collera parfaitement aux données. La procédure d'optimisation est donc loin d'être automatique. Au contraire nous respectons volontairement un certain nombre d'étapes successives qui contraignent la convergence. Comme nous l'avons indiqué au paragraphe 3.6.3, nous estimons d'abord un intervalle de valeur convenable pour le niveau de cocontraction, en lançant la procédure d'optimisation depuis plusieurs valeurs initiales de $K$, puis nous affinons la recherche sur chacun des paramètres dynamiques. Et afin de s'assurer qu'un minimum local dans une autre région de l'espace des paramètres n'est pas ignoré, il est nécessaire de recommencer à partir de plusieurs points initiaux dans cet espace à quatre dimensions.

Les paramètres prosodiques inférés sont, pour le cas lent non-accentué :

$K$ = 1400 $s^{-2}$
$T_{hold1}$ = 67 ms
$T_{trans}$ = 73 ms
$T_{hold2}$ = $10^{-5}$ ms

et pour le cas rapide accentué :

$K$ = 2300 $s^{-2}$
$T_{hold1}$ = 46 ms
$T_{trans}$ = 46 ms
$T_{hold2}$ = $10^{-5}$ ms

Rappelons, pour comparaison, les résultats obtenues dans le cas lent accentué :

$K$ = 7700 $s^{-2}$
$T_{hold1}$ = 103 ms
$T_{trans}$ = 73 ms
$T_{hold2}$ = 37 ms

Les deux figures suivantes (4.13a,b) donnent les résultats de l'optimisation pour les deux conditions d'élocution. Pour chaque figure, le premier panneau comporte les positions estimée (trait plein) et "donnée" (trait tireté) du corps de la langue. La ligne brisée en trait mixte est la trajectoire du point d'équilibre. Le deuxième panneau contient les courbes de vitesse avec les mêmes conventions de traits.

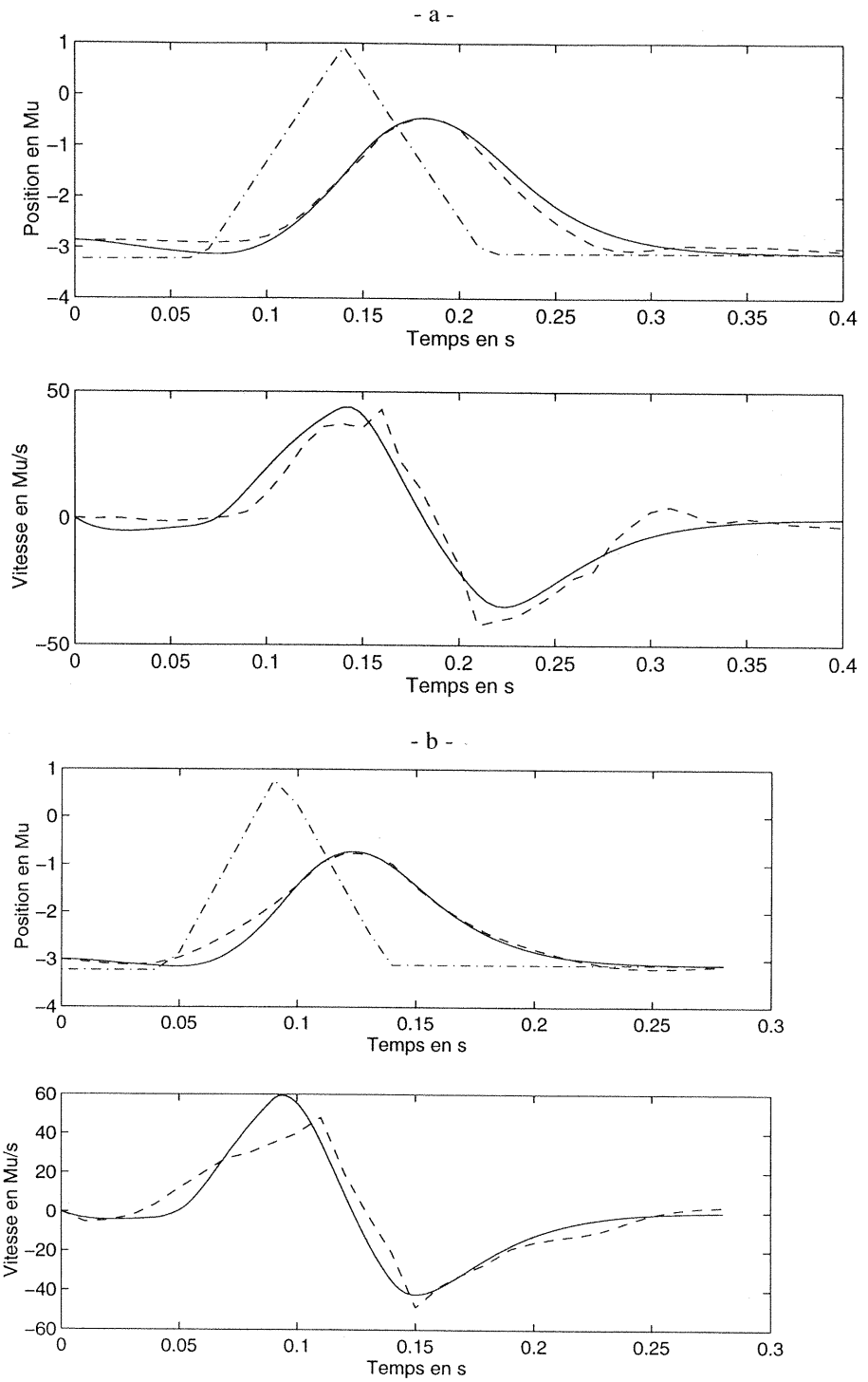



*Figure 4.13.* Résultats de l'inversion dynamique pour la séquence [iai]. a : lent non accentué, b : rapide accentué.

Les trajectoires articulatoires simulées sont relativement bien ajustées aux données. La partie correspondant à la voyelle [a] est particulièrement proche des données, ce que nous recherchions en augmentant le poids des échantillons correspondant au [a] dans le critère

d'erreur (cf. 3.6.3.2). Les transitions sont moins bien suivies, mais compte tenu des lissages effectués sur les données, ces erreurs ne sont peut-être pas rédhibitoires. Nous vérifierons au paragraphe 4.4 que les conséquences acoustiques de ces aléas sur les transitions sont faibles.

Les erreurs sur les plateaux des voyelles [i] sont probablement dues au fait que nous ne tenons pas compte des cibles précédant et suivant la séquence [iai]. Les plateaux des voyelles [i] correspondent en fait, pour le premier, à la fin du mouvement du [l] au [i], et, pour le second, au début du mouvement du [i] au [m]. Nous avons choisi, pour la simplicité du contrôle, de ne pas modéliser ces effets de conditions initiales et finales, mais il faudra évidemment vérifier que la qualité phonétique des [i] n'est pas affectée par cette approximation.

D'après ces résultats, il semble que, pour le locuteur JLS, les conditions plus accentuées soient caractérisées par des *niveaux de cocontraction* plus élevés et que la différence entre débit lent et débit rapide se fasse nettement sur le *timing* de la trajectoire d'équilibre : tous les paramètres temporels sont diminués des conditions lentes à la condition rapide. La différence d'accentuation affecte aussi les paramètres temporels.

Les deux conditions accentuées correspondent à des niveaux de cocontraction respectifs de $7700s^{-2}$ et $2300s^{-2}$, alors que la condition non-accentuée est associée à un niveau de $1400s^{-2}$.

*Les temps de transition* des deux conditions lentes sont similaires. Ce temps est plus faible pour le cas rapide accentué (diminution de 37% par rapport au cas idéal).

*Les temps de maintien du premier [i]* sont plus élevés pour les conditions lentes que la condition rapide, et ce temps diminue de la condition lente accentuée à lente non-accentuée (35% de diminution pour le cas lent non-accentué et 55% de réduction pour le cas rapide accentué).

*Le temps de maintien du [a]* diminue, et même disparaît, dans les deux conditions réduites, par rapport à la condition idéale. À débit lent, la voyelle [a] est donc tenue plus longtemps lorsqu'elle est accentuée que lorsqu'elle ne l'est pas. Il semble par conséquent que la tenue soit donc liée à l'accentuation, rendant disponible plus longtemps l'information véhiculée dans la partie stable de la voyelle. Cependant à débit rapide, la tenue de la voyelle est aussi fortement diminuée. Il semble donc que ce locuteur puisse accentuer sans nécessairement tenir la voyelle. Notons, à ce propos, que malgré la consigne d'accentuation, le [a] de la séquence rapide est le plus réduit des trois réalisations de [a], dans les espaces acoustique et articulatoire. La forte diminution de la cocontraction (de $7700s^{-2}$ à $2300s^{-2}$), de la condition idéale à la condition rapide accentuée (la cocontraction restant malgré tout supérieure à celle que l'on observe pour le cas lent non-accentué), ainsi que la quasi disparition du temps de maintien de la voyelle accentuée, sont donc cohérentes avec les observations acoustiques et articulatoires, qui indiquent que le locuteur a

probablement porté son attention prioritairement sur la consigne de rapidité, tout en respectant toutefois la consigne d'accentuation, puisque le [a] de la séquence rapide accentuée n'est pas dramatiquement réduit.

### 4.3.1.2 Inversions de [iɛi]

La même méthode que pour la séquence [iai] est utilisée pour l'inversion de [iɛi] dans les trois conditions d'élocution. La commande centrale posturale est inférée directement sur la trajectoire du corps dans le cas lent non accentué. Elle sera reprise pour les deux autres cas. Les commandes centrales prosodiques sont ensuite recherchées à l'aide d'une technique d'optimisation.

**Recherche de la commande centrale posturale**

Les positions d'équilibre successives sont estimées à partir de la trajectoire du corps de la langue, obtenue par inversion cinématique dans le cas lent accentué et représentée sur la figure 4.14.

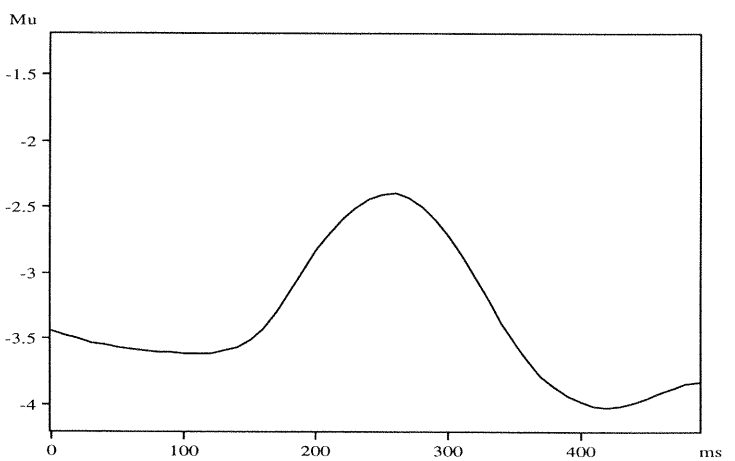



*Figure 4.14.* Trajectoire du corps de la langue pour la séquence [iɛi] dans le cas lent accentué.

Les positions d'équilibre des deux [i] sont lues directement sur la courbe. Comme pour la séquence [iai], nous ajoutons 2% de l'amplitude du mouvement à la position maximale du [ɛ] pour obtenir la position d'équilibre correspondante. Un changement d'échelle est effectué pour que le deuxième [i] de la séquence [iɛi] coïncide avec celui de la séquence [iai].

Les positions d'équilibre successives, qui définissent la commande centrale posturale, sont ainsi (avant changement d'échelle) :

$y_e(i_1)$ = -3.598 UM[1]

$y_e(\varepsilon)$ = -2.389 UM[2]

$y_e(i_2)$ = -4.027 UM[3]

**Recherche des commandes centrales prosodiques**

La technique d'optimisation est appliquée dans chacune des conditions d'élocution. Nous procédons en plusieurs étapes, de la même façon que pour la séquence [iai]. On obtient une approximation satisfaisante des courbes de position et de vitesse, avec les valeurs suivantes des paramètres prosodiques, pour le cas lent accentué :

$K$ = 5600 $s^{-2}$

$T_{hold1}$ = 118 ms

$T_{trans}$ = 92 ms

$T_{hold2}$ = 46 ms

le cas lent non-accentué :

$K$ = 1300 $s^{-2}$

$T_{hold1}$ = 61 ms

$T_{trans}$ = 57 ms

$T_{hold2}$ = 37 ms

et pour le cas rapide accentué :

$K$ = 1300 $s^{-2}$

$T_{hold1}$ = 41 ms

$T_{trans}$ = 57 ms

$T_{hold2}$ = 17 ms

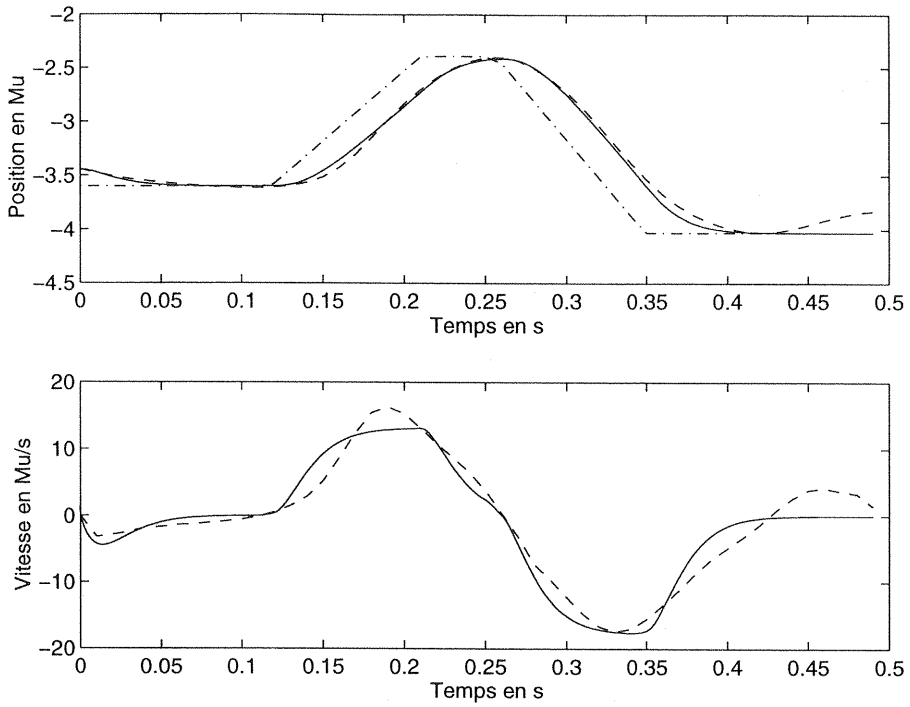
Les résultats de l'optimisation, pour la séquence [iɛi] dans les trois conditions d'élocution, sont représentées sur les figures 4.15 a, b et c. Les mêmes conventions de représentation sont utilisées que pour la séquence [iai] (cf. 4.3.1.1).

[1]soit -5.017 en données corrigées

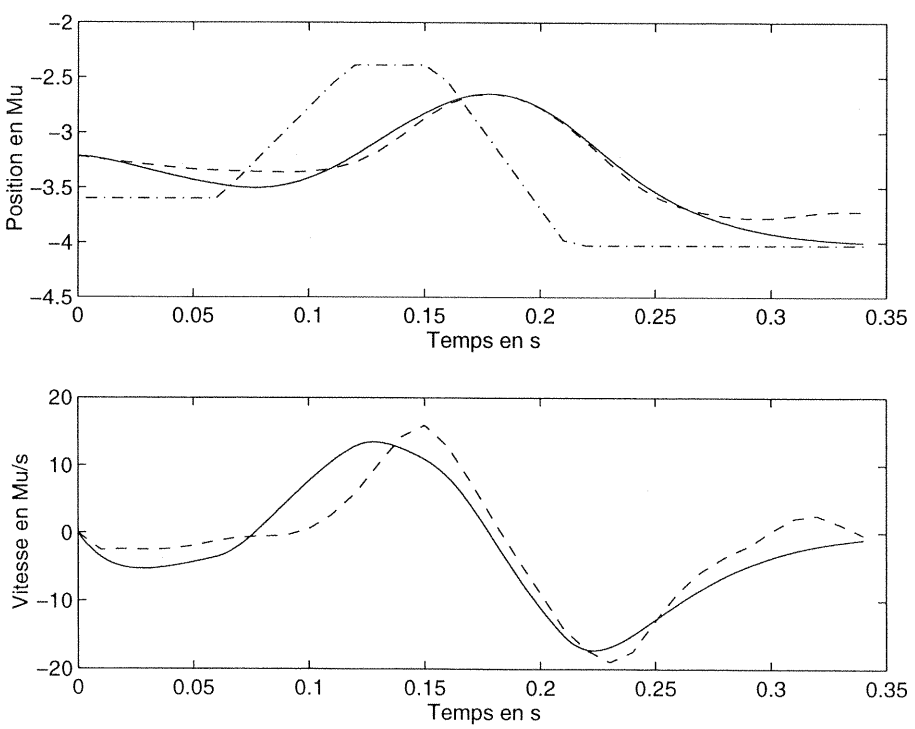
[2]soit -0.985 en données corrigées

[3]soit -6.447 en données corrigées (valeur identique à celle de $y_e(i_2)$ pour [iai])

- a -


Position en Mu
Temps en s
Vitesse en Mu/s
Temps en s

- b -


Position en Mu
Temps en s
Vitesse en Mu/s
Temps en s

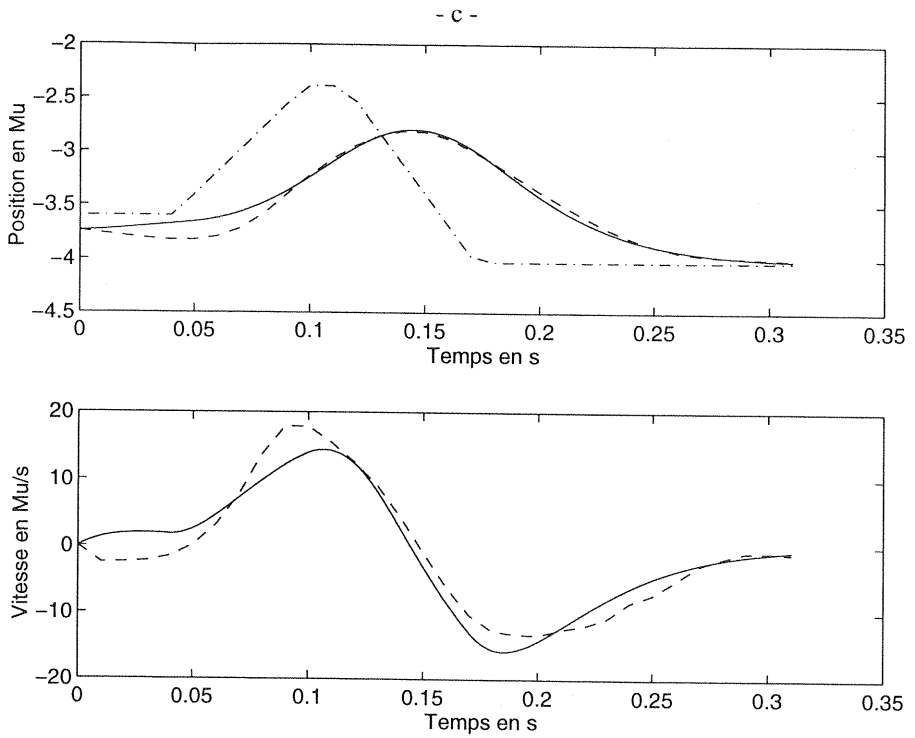



*Figure 4.15.* Résultats de l'inversion dynamique pour la séquence [iɛi]. a : lent accentué, b : lent non accentué, c : rapide accentué.

Les trajectoires simulées à partir des paramètres inférés par l'optimisation sont relativement proches des données. Les remarques formulées pour la séquence [iai], sur les parties transitoires et sur les plateaux des [i], s'appliquent également ici.

*Le niveau de cocontraction* est dans l'ensemble plus faible pour la séquence [iɛi] que la séquence [iai]. Dans le cas idéal, la cocontraction diminue en effet de $7700s^{-2}$ pour [iai] à $5600s^{-2}$ pour [iɛi], dans le cas lent accentué, elle conserve une valeur faible et dans le cas rapide accentué, elle diminue de $2300s^{-2}$ à $1300s^{-2}$. Cette variabilité des niveaux de cocontraction entre les deux séquences pour les mêmes conditions de production éclaire le rôle de la cocontraction dans la production de l'accentuation : il permet au locuteur de contrôler l'inertie du système articulatoire et donc d'agir sur la proximité entre la trajectoire virtuelle, telle qu'elle est donnée par l'évolution temporelle de la commande spécifiant le point d'équilibre, et la trajectoire effectivement suivie par l'articulateur. Ainsi on peut interpréter le plus faible niveau de cocontraction observé pour la séquence [iɛi] : l'amplitude du mouvement du corps de la langue étant plus faible pour [iɛ] que pour [ia], un niveau de cocontraction moindre est nécessaire dans la séquence [iɛi] pour maintenir un niveau de proximité équivalent entre la trajectoire virtuelle et la trajectoire effective. *Le temps de transition* dans le cas lent accentué est plus long pour la séquence [iɛi] que la séquence [iai]. Cette augmentation semble liée à l'augmentation générale de la durée de la

séquence [iɛi] (490 ms) par rapport celle de la séquence [iai] (440ms). On remarque par ailleurs que *tous* les paramètres temporels de la séquence [iɛi] idéale sont plus élevés que ceux de la séquence [iai]. Il n'a pas été imposé de tempo-référence absolu au locuteur, comme on aurait pu le faire à l'aide d'un métronome, par exemple. Par conséquent les différences temporelles observées sont vraisemblablement imputables à une légère différence de débit de parole d'une séquence à l'autre, même s'il s'agit des mêmes conditions de production. Il est intéressant de noter alors que la diminution du temps de transition dans le cas rapide accentué pour [iɛi] est exactement du même ordre que pour la séquence [iai] (38% de diminution). Il semble par conséquent que, pour ce locuteur, débit lent et débit rapide soient différenciés très précisément, même si à l'intérieur de chaque débit, on observe une variabilité des durées. La diminution du temps de transition de la condition lente accentuée à la condition rapide accentuée doit donc être interprétée comme une conséquence de la variation de débit. Nous observons en outre pour la séquence [iɛi] une diminution du temps de transition de la condition lente accentuée à la condition lente non-accentuée. Ceci n'a pas été observé pour la condition [iai]. Ce résultat illustre bien les degrés de liberté dont dispose le locuteur pour manipuler la proximité entre trajectoire virtuelle et trajectoire effective. Nous avons vu dans le chapitre III que temps de transition et cocontraction peuvent être manipulés conjointement à cet effet. Plus le temps de transition est court, plus les effets de l'inertie sont importants ou, en d'autres termes, plus la force de rappel est élevée ; de même plus la cocontraction est grande, plus la force de rappel est importante. En augmentant simultanément vitesse de transition et cocontraction, on diminue finalement le niveau total de cocontraction requis ; c'est la stratégie que l'on infère des données avec notre modélisation, dans le cas de la séquence [iɛi] lente non accentuée ; dans le cas de la séquence [iai], la stratégie impliquerait essentiellement la cocontraction. On peut mettre cette différence sur le compte de la variabilité intra-locuteur et nous n'avons pas assez de données pour infirmer cette hypothèse. Mais on peut aussi remarquer que l'amplitude du mouvement de [i] vers [ɛ] étant plus faible que de [i] vers [a] un temps de transition trop long pour le cas [iɛ] lent non accentué pourrait induire une transition peu naturelle.

*Les temps de maintien du premier [i]* évoluent dans le même sens que pour la séquence [iai] (48% de diminution pour le cas lent non-accentué et 65% pour le cas rapide accentué). Cette diminution est clairement liée à l'augmentation du débit pour le cas rapide accentué et à une moindre tenue des voyelles pour le cas lent non-accentué. Notons toutefois que, dans la mesure où nous ne prenons pas les contextes gauche et droite en considération dans nos simulations, la pertinence de ces temps de maintien est sujette à caution.

*Le temps de maintien du [ɛ]* diminue de la condition idéale aux conditions réduites, mais il ne disparaît pas comme pour la séquence [iai] (20% de réduction pour le cas lent non accentué et 63% pour le cas rapide accentué). Cette augmentation du temps de maintien dans les deux conditions réduites par rapport à la séquence [iai] peut s'interpréter dans une perspective d'isochronie moyenne des séquences [iVi] à débit donné : étant donnée l'amplitude plus faible du mouvement pour [iɛi], une tenue plus longue de la voyelle V est possible. Les diminutions de ce temps de maintien dans les conditions réduites par rapport à la condition idéale s'interprètent de la même façon que celles du temps de maintien du premier [i].

Une conclusion préliminaire semble donc ressortir de ces inversions : dans le cas de l'accentuation, le locuteur utiliserait ses possibilités de contrôle de la cocontraction, du temps de transition et du temps de maintien, pour assurer une meilleure réalisation des cibles planifiées, en s'efforçant de respecter une proximité suffisante entre la trajectoire virtuelle, qui va précisément aux cibles, et la trajectoire articulatoire effective. Dans le cas de l'accentuation, à débit donné, et par rapport à la condition non-accentuée, on peut donc observer, éventuellement simultanément, une augmentation de la cocontraction, du temps de transition et de la tenue de la voyelle centrale V.

### 4.3.2 Sensibilité à deux paramètres dynamiques

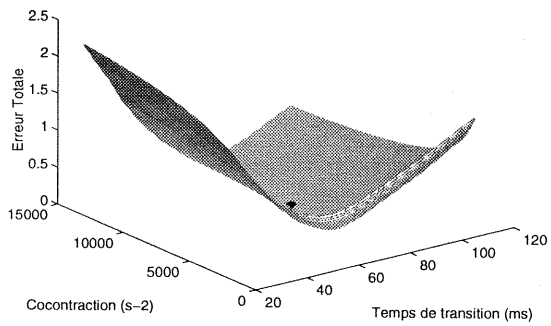
Afin d'évaluer la sensibilité des "solutions" obtenues aux paramètres variables, *i.e.* afin d'estimer la qualité des minima locaux obtenus, nous étudions comment évolue la courbe d'erreur, lorsque divers paramètres sont modifiés. Afin d'obtenir une courbe d'erreur lisible, donc au plus à trois dimensions, nous nous bornons à deux paramètres dynamiques. Les paramètres choisis sont la cocontraction et le temps de transition. Il nous semble en effet, ainsi que nous l'avons évoqué au paragraphe 3.8.2, que ces deux paramètres jouent des rôles complémentaires. Si le temps de transition est trop court, il ne permet pas que la trajectoire d'équilibre planifiée soit convenablement suivie. Mais si le niveau de cocontraction est suffisamment élevé, alors un faible temps de transition, ou une forte vitesse de transition, augmente le niveau de force de rappel et favorise le suivi de la trajectoire planifiée. Il est donc possible que deux solutions équivalentes et "réciproques" existent dans l'espace à quatre dimensions des paramètres dynamiques : l'une avec un fort niveau de cocontraction et un court temps de transition et l'autre avec un faible niveau de cocontraction et un long temps de transition.

Nous donnons dans les figures suivantes (4.16a-f), les courbes d'erreur à trois dimensions, en fonction du temps de transition et du niveau de cocontraction, dans les trois conditions d'élocution et pour les deux séquences [iai] et [iɛi]. L'erreur représentée est

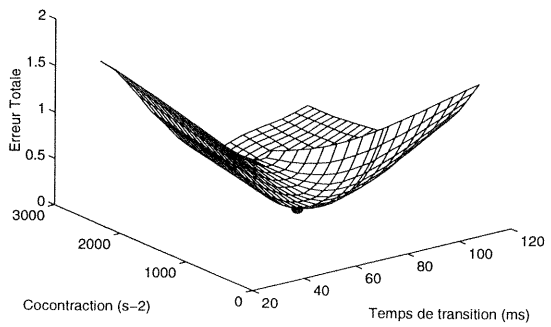
l'erreur totale que nous avons utilisée pour l'optimisation, *i.e.* la somme pondérée de l'erreur sur la position, la vitesse et le plateau de la voyelle centrale (cf. 3.6.3.2). Le minimum (correspondant à la "solution" de l'optimisation) est représenté par un point sur ces courbes.

Les figures 4.17a-l représentent les coupes de ces courbes selon les deux plans correspondant à la valeur optimale de chaque paramètre étudié, l'autre paramètre variant de part et d'autre de sa valeur optimale.

- a -




- b -




- c -

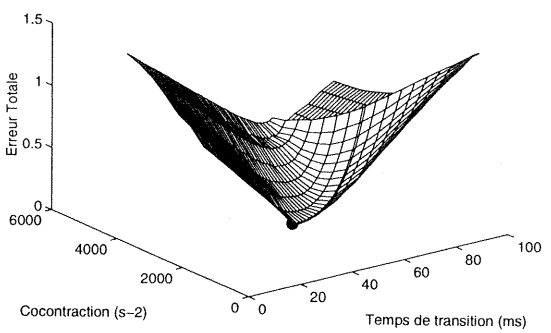
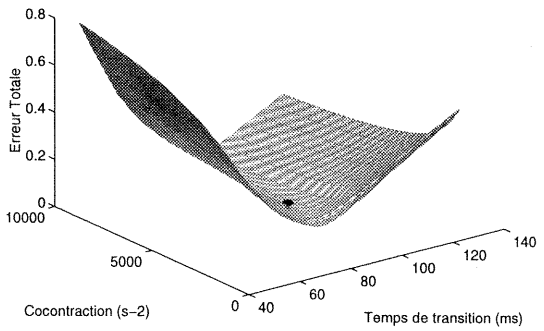


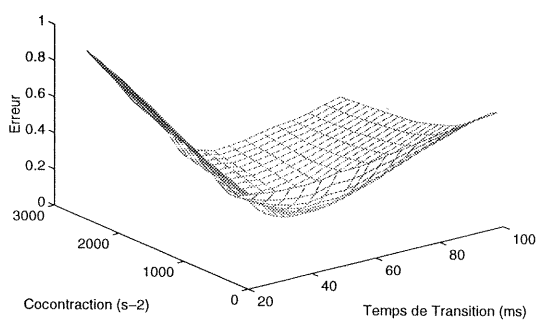

*Figure 4.16.* Erreur totale en fonction de la cocontraction et du temps de transition. a : [iai] lent accentué, b : [iai] lent non-accentué, c : [iai] rapide accentué.

- d -




- e -

Erreur d'optimisation




- f -

Erreur d'optimisation

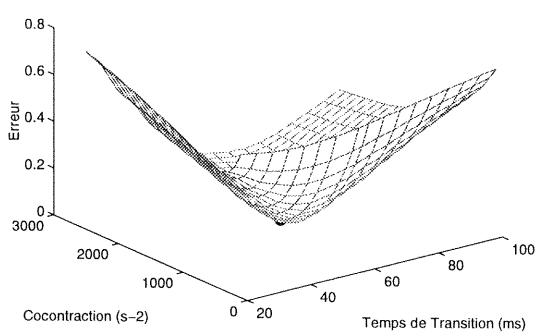
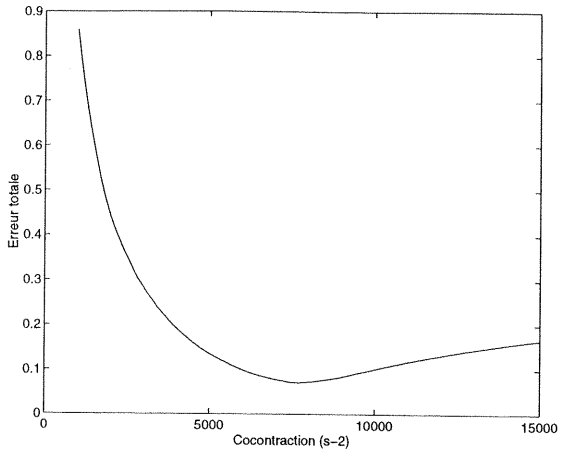
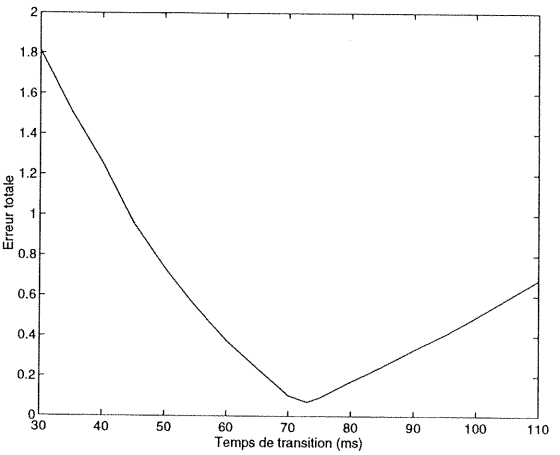



*Figure 4.16.* Erreur totale en fonction de la cocontraction et du temps de transition. d : [iɛi] lent accentué, e : [iɛi] lent non-accentué, f : [iɛi] rapide accentué.

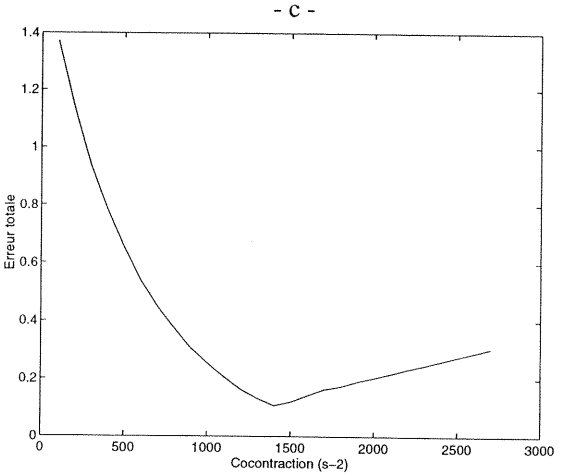
- a -


Erreur totale
Cocontraction (s-2)

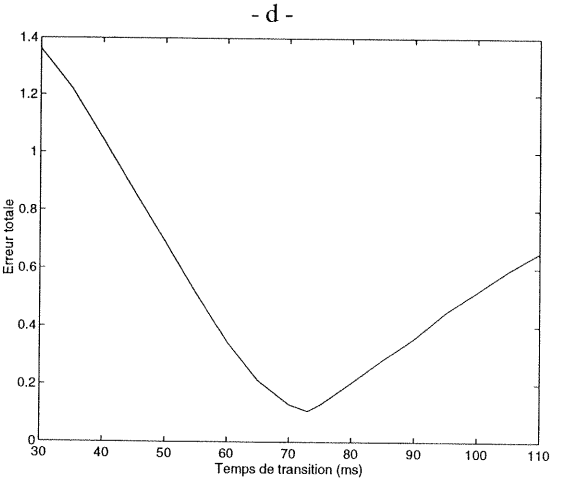
- b -


Erreur totale
Temps de transition (ms)

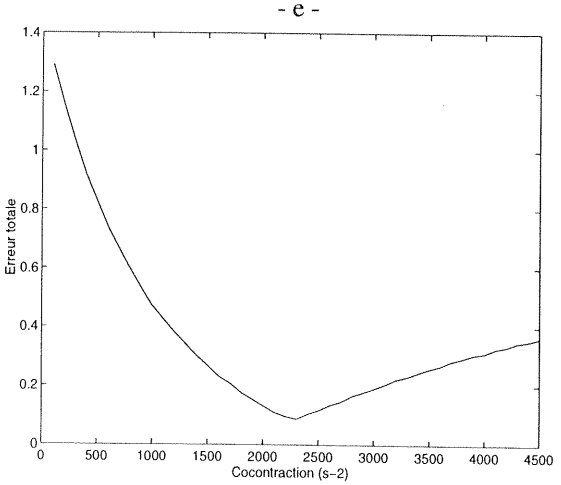
- c -


Erreur totale
Cocontraction (s-2)

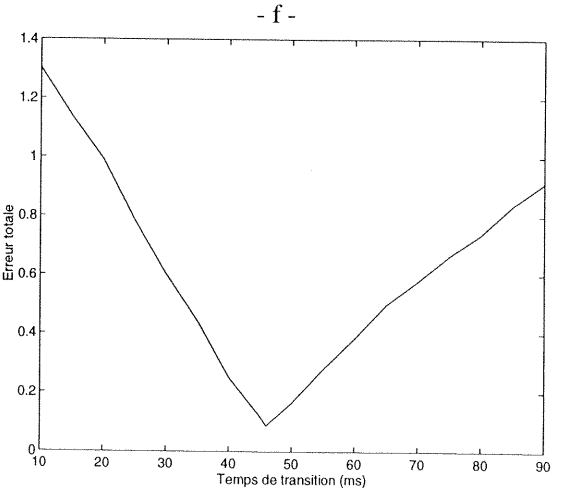
- d -


Erreur totale
Temps de transition (ms)

- e -


Erreur totale
Cocontraction (s-2)

- f -


Erreur totale
Temps de transition (ms)

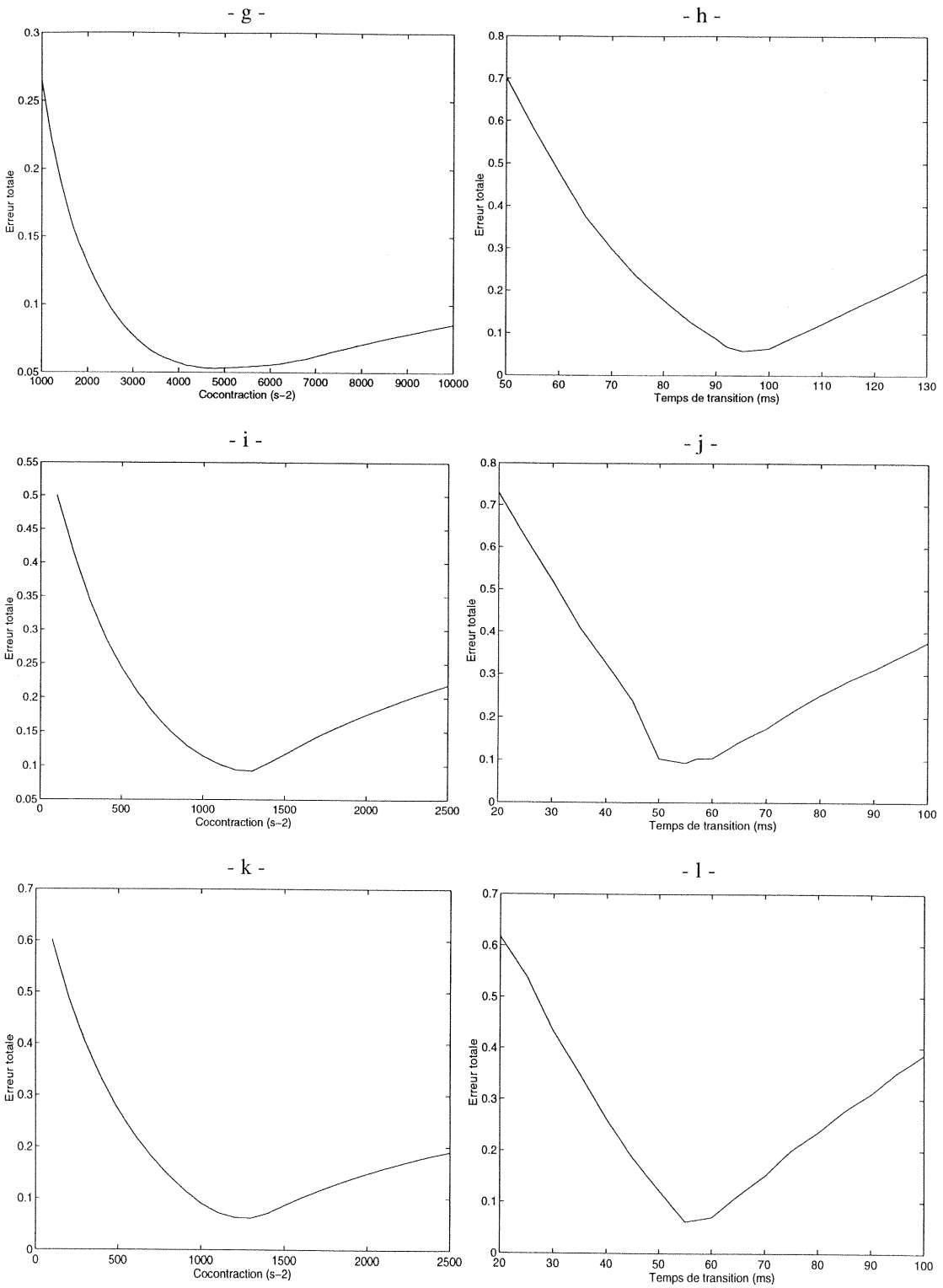



*Figure 4.17.* Erreur totale en fonction d'un seul paramètre dynamique, l'autre étant optimal. a, b : [iai] lent accentué; c, d : [iai] lent non-accentué; e, f : [iai] rapide accentué; g, h : [iɛi] lent accentué; i, j : [iɛi] lent non-accentué; k,l : [iɛi] rapide accentué.

En condition idéale, les courbes d'erreur pour les deux séquences vocaliques [iai] et [iɛi], présentent un minimum unique et sont relativement évasées. Il semble donc que la configuration du système, autour du point inféré par l'optimisation, soit relativement stable. Il existe ainsi une large région de l'espace à deux dimensions de ces paramètres dynamiques dans laquelle les données sont bien représentées. Ce résultat est intéressant en termes de contrôle : en jouant sur les paramètres de cocontraction et temporels, on modifie certes la trajectoire, mais d'une façon telle que la production de la voyelle ne soit pas gravement altérée. La stratégie de production que nous proposons pour l'accentuation est donc très efficace : elle ne nécessite pas un contrôle excessivement fin des paramètres pour maintenir la production de la voyelle quasiment constante. Ceci est cohérent avec l'idée qu'une voyelle accentuée est moins variable qu'une voyelle non-accentuée (cf. Lindblom *et al.* [1992]). Pour les conditions réduites, les courbes d'erreur forment des cuvettes plus étroites, indiquant que les optima inférés sont plus sensibles aux variations des paramètres dynamiques. Ici encore ces résultats sont cohérents avec les données observées sur des locuteurs.

Afin de donner une mesure numérique de la sensibilité des résultats de l'optimisation aux paramètres dynamiques étudiés, nous proposons de calculer les écarts d'erreur en fonction des écarts de valeur pour chaque paramètre. Nous définissons ainsi les sensibilités locales au point optimum :

$S_{opt}(K) = \frac{\Delta E_T}{\Delta K}$ , pour une variation relative de $K$ de 50% (arrondie à la centaine inférieure) et

$S_{opt}(T_{trans}) = \frac{\Delta E_T}{\Delta T_{trans}}$, pour une variation relative de $T_{trans}$ de 50% (arrondie à l'unité).

Le tableau 4.1 fournit les sensibilités pour [iai] et [iɛi] dans les trois conditions.

*Tableau 4.1.* Mesures des sensibilités locales aux paramètres K et Ttrans, pour [iai] et [iɛi] dans les trois conditions prosodiques.

| | [iai] | | | | [iɛi] | | | |
|---|---|---|---|---|---|---|---|---|
| | $S_{opt}(K).10^4$ | | $S_{opt}(T_{trans})$ | | $S_{opt}(K).10^4$ | | $S_{opt}(T_{trans})$ | |
| Condition | ΔK<0 | ΔK>0 | ΔTtr<0 | ΔTtr>0 | ΔK<0 | ΔK>0 | ΔTtr<0 | ΔTtr>0 |
| Lent acc | 0.340 | 0.145 | 0.038 | 0.016 | 0.107 | 0.068 | 0.014 | 0.005 |
| Lent non acc | 4.899 | 1.580 | 0.028 | 0.014 | 1.393 | 1.216 | 0.016 | 0.006 |
| Rap. acc | 2.673 | 1.432 | 0.034 | 0.020 | 1.990 | 1.269 | 0.014 | 0.007 |

Ces mesures numériques ne font que confirmer ce qu'indiquent les courbes à deux ou trois dimensions : en général, la sensibilité des résultats de l'optimisation aux paramètres dynamiques augmente en conditions réduites. Elle est relativement faible pour la condition idéale de chaque séquence vocalique. Par conséquent, les solutions obtenues par l'inversion

dynamique dans chaque cas idéal nous semblent acceptables. Les solutions inférées dans les cas réduits sont plus sujettes à caution. Dans la perspective d'une perception de la parole qui impliquerait une inversion (Marr [1982], Poggio [1984]), ces résultats se révèlent aussi intéressants : la récupération des paramètres dynamiques sous-jacents à la production des conditions réduites est plus délicate, car la zone de l'espace moteur correspondant est plus étroite. Les risques d'erreur et d'éventuelles confusions avec des solutions différentes sont donc plus grands. Nous confirmerons, au paragraphe 4.4, ces observations sur les sensibilités par un test perceptif sur l'identification et la qualité des séquences vocaliques selon les conditions prosodiques.

### 4.3.3 Interprétation prosodique des paramètres dynamiques

Nous venons de suggérer que si les paramètres temporels, temps de transition et de tenue, permettent de jouer sur le débit, la cocontraction joue un rôle majeur pour la production de l'accentuation : les voyelles accentuées seraient produites à raideur plus élevée, afin de permettre un meilleur suivi de la trajectoire virtuelle.

Comme nous l'avons indiqué brièvement au chapitre II, ce résultat est au centre d'un débat, puisque divers auteurs donnent une signification différente de la nôtre au paramètre de raideur ou de cocontraction.

En premier lieu, les résultats de Ostry *et al.* [1983] sur la relation entre le pic de vitesse et l'amplitude pour les mouvements du dos de la langue, semblent indiquer que les mouvements accentués présentent une pente (pour la relation vitesse maximale/amplitude) plus *faible* que les mouvements non accentués. Toutefois, comme le précisent Ostry & Munhall [1985], les différences de pente observées seraient liées à des durées relatives différentes. Ainsi, la voyelle non-accentuée étant produite à un débit plus rapide que la voyelle accentuée, la pente observée est plus élevée. Les résultats de Kelso, Bateson, Saltzman & Kay [1985] associant, dans certains cas, accentuation et faible pente, peuvent s'interpréter de même. La durée du mouvement doit être prise en compte dans l'estimation de la pente. Les résultats de Bateson & Kelso [1993], présentés au 2.2.1, indiquent aussi que les gestes accentués et par essence *plus lents* présentent une pente *Vm/Amp* plus faible que les gestes non-accentués et plus rapides. Ce malentendu sur la cause de la variation du rapport *Vm/Amp* a conduit Smith, Browman, McGowan et Kay [1993] à reproduire ce schéma accentuation = faible pente, dans leur estimation de paramètres dynamiques par une technique d'optimisation. Ils trouvent ainsi que la fréquence (ou plutôt pulsation) propre $\omega_0$, qui est proportionnelle à la racine carrée de la raideur du système du second ordre, donc au rapport *Vm/Amp*, est en général plus faible pour les gestes accentués que les gestes non-accentués. Toutefois cette conclusion est tempérée par la remarque que pour les syllabes en fin de mot, la différence d'accentuation ne joue pas sur la pulsation propre.

D'autre part, la relation entre raideur du système et pente de la courbe *Vm/Amp* n'est pas si simple dès qu'un amortissement non négligeable est impliqué.

Afin de clarifier les relations entre accentuation, débit et paramètres cinématiques, nous résumons dans les quatre tableaux suivants (4.2 a-d) les résultats obtenus par tous ces auteurs sur divers articulateurs de la parole.

*Tableau 4.2.a.* Ostry *et al.* 1983. Mouvements du dos de la langue obtenu par ultrason.

| Variable | Accentué (comparé à non-accentué) | Débit rapide (comparé à lent) |
|---|---|---|
| Durée | augmente | diminue |
| Amp | augmente | diminue |
| Vm | augmente | - augmente pour sujet DO<br>- inchangée pour sujet KM<br>- diminue pour sujet RF |
| Vm/Amp | - **diminue** pour abaissement du dos<br>- **pas significatif** pour élévation | **pas d'effet significatif** |

*Tableau 4.2.b.* Ostry & Munhall 1985. Dos de la langue

| Variable | Débit rapide (comparé à lent) |
|---|---|
| Durée | diminue |
| Amp | - diminue pour sujets SG et AD<br>- inchangée pour sujet CB |
| Vm | - augmente pour sujet CB<br>- inchangée pour sujets SG et AD |
| Vm/Amp | **augmente** (car la durée diminue) |

*Tableau 4.2.c.* Kelso *et al.* 1985. Mouvements de la lèvre inférieure.

| Variable | Accentué (comparé à non-accentué) | Débit rapide (comparé à lent) |
|---|---|---|
| Durée | augmente | diminue |
| Amp | augmente | diminue |
| Vm | augmente | - augmente pour sujet DW<br>- diminue pour sujet SK |
| Vm/Amp | - **diminue** pour sujet SK<br>- **diminue** pour DW, débit rapide<br>- **augmente** pour DW, débit normal | - **augmente** pour sujet SK<br>- **augmente** pour sujet DW si geste de fermeture non-accentué<br>- **diminue** pour DW dans autres cas |

*Tableau 4.2.d.* Bateson & Kelso 1993. Mouvements de la lèvre inférieure dans trois langues.

| Variable | Accentué (comparé à non-accentué) | Débit rapide (comparé à lent) |
|---|---|---|
| Durée | augmente | diminue |
| Amp | augmente | diminue |
| Vm | augmente | diminue |
| Vm/Amp | **pas significatif**<br>**diminue pour durées plus longues** | **augmente** |

Ces tableaux récapitulatifs montrent que les seules variables cinématiques qui présentent des variations systématiques sont la durée et l'amplitude. Les gestes accentués durent plus longtemps et ont des amplitudes plus élevées que les gestes non-accentués. Les gestes effectués à débit rapide ont des durées et des amplitudes plus faibles que les gestes effectués à débit lent. La vitesse maximale augmente systématiquement pour les gestes accentués par rapport aux gestes non-accentués, par contre ses variations ne sont pas systématiques pour les modifications du débit. Enfin, contrairement aux conclusions de Ostry *et al.* [1983] ou Kelso *et al.* [1985], le rapport *Vm/Amp* ne caractérise pas de manière claire les catégories d'accentuation alors qu'il est plus pertinent pour les catégories de durée (comme le suggèrent d'ailleurs Ostry & Munhall [1985]).

Dans notre étude, il semble aussi que le rapport *Vm/Amp* caractérise plutôt la différence entre gestes lents et rapide : lorsque la durée du mouvement diminue, le rapport *Vm/Amp* augmente. Nous présentons, dans les tableaux 4.3 a et b, les mêmes variables cinématiques, calculées dans les trois conditions prosodiques pour les mouvements du corps de la langue, inférés des données acoustiques et simulés à l'aide du modèle de la dynamique des articulateurs, pour la séquence [iai]. Les variables sont considérées pour les deux transitions /ia/ et /ai/. Le début et la fin de chaque transition correspondent aux passages par 0 de la vitesse.

*Tableau 4.3.a.* Mesures effectuées sur les trajectoires articulatoires inférées par inversion cinématique des données acoustiques.

| | Lent accentué | | Lent non-accentué | | Rapide accentué | |
|---|---|---|---|---|---|---|
| Variable | /ia/ | /ai/ | /ia/ | /ai/ | /ia/ | /ai/ |
| *Durée (s)* | 0.154 | 0.127 | 0.115 | 0.105 | 0.094 | 0.126 |
| *Amp (UM)* | 4.06 | 3.78 | 2.43 | 2.56 | 2.32 | 2.40 |
| *Vm (UM.s$^{-1}$)* | 57.49 | 59.77 | 42.73 | 41.91 | 47.80 | 47.70 |
| *Vm/Amp(s$^{-1}$)* | 14.14 | 15.79 | 17.56 | 16.31 | 20.56 | 19.82 |

*Tableau 4.3.b.* Mesures effectuées sur les simulations à l'aide du modèle de la dynamique des articulateurs.

| | Lent accentué | | Lent non-accentué | | Rapide accentué | |
|---|---|---|---|---|---|---|
| Variable | /ia/ | /ai/ | /ia/ | /ai/ | /ia/ | /ai/ |
| *Durée (s)* | 0.118 | 0.148 | 0.108 | 0.217 | 0.07 | 0.161 |
| *Amp (UM)* | 4.08 | 3.97 | 2.68 | 2.65 | 2.12 | 2.10 |
| *Vm (UM.s$^{-1}$)* | 56.57 | 54.96 | 43.90 | 34.79 | 55.50 | 35.18 |
| *Vm/Amp(s$^{-1}$)* | 13.85 | 13.83 | 16.37 | 13.09 | 26.05 | 16.73 |
| *$\sqrt{K}$ (s$^{-1}$)* | 87.74 | 87.74 | 37.41 | 37.41 | 47.95 | 47.95 |

Remarquons que les valeurs cinématiques obtenues pour les transitions /ai/ simulées reflètent la mauvaise adéquation avec les données articulatoires. Comme nous l'avons indiqué au 4.3.1, la séquence /iai/ est insérée dans une phrase porteuse et le mouvement continue après le deuxième /i/. Dans la simulation, nous ne rendons pas compte de cet aspect, puisqu'aucune cible n'est prévue après celle du /i/. Le système tend donc vers le point attracteur du deuxième /i/ et la vitesse tend asymptotiquement vers 0. Les durées des transitions (repérées par le passage par 0 de la vitesse) paraissent alors artificiellement plus longues.

S'il semble clair maintenant que le rapport *Vm/Amp* est fonction de la durée du mouvement, et donc aussi du débit, qu'en est-il de son lien avec la cocontraction (ou raideur) du système?

Pour un système du second ordre non-amorti, ou d'amortissement négligeable, on montre que le rapport *Vm/Amp* vaut la pulsation propre, c'est-à-dire la racine carrée de la raideur. Par contre, dès que l'amortissement est non négligeable —et c'est le cas ici, puisque nous l'avons choisi très proche de la valeur critique (cf. 3.6.2)— *Vm/Amp* dépend aussi de l'amortissement et de la durée. On ne peut plus relier ce rapport uniquement à la cocontraction. Si le rapport *Vm/Amp* est fonction de la *durée* du mouvement, la cocontraction, elle, a une interprétation physique différente, liée à la *force* avec laquelle s'effectue le mouvement. En effet, lorsque la cocontraction $K$ augmente, la force de rappel $K(y - y_e)$ du système du second ordre, appelé parfois *point attracteur* (Abraham & Shaw, [1982]), augmente aussi.

Ainsi selon notre conception, et contrairement aux points de vue défendus par les auteurs sus-cités, la notion de cocontraction n'est pas associée à la notion de durée. La durée est réglée par les temps de transition et de tenue, qui sont explicitement contrôlés. Pour nous, la notion de cocontraction est à exploiter en termes de tonicité, de force. Comme les résultats exposés plus haut le suggèrent, l'accentuation est liée à une augmentation de l'amplitude du mouvement. Cette augmentation d'amplitude est rendue possible par une augmentation de la cocontraction induisant une augmentation de l'attraction vers le point d'équilibre.

Pour Bateson *et al.* [1993] (voir aussi les travaux plus anciens de Ostry *et al.* [1983, 1985]), les mouvements articulatoires sont représentés par un système du second ordre *linéaire*, pour lequel la position d'équilibre (ou de repos) est constante pour chaque oscillation. Nous proposons un schéma différent, pour lequel la vitesse de transition, d'une position d'équilibre à une autre, est contrôlable. Notre système n'est donc linéaire que pendant les phases de tenue de la position d'équilibre. La durée des tenues et la pente des transitions déterminent la durée et l'amplitude du mouvement tout autant que la cocontraction.

L'hypothèse que le contrôle de l'accentuation puisse s'effectuer soit par un allongement de la transition, soit par une augmentation de la cocontraction, les deux agissant éventuellement en synergie, pourrait trouver des éléments de corroboration dans les données, si dans la variabilité observée, on trouvait des séquences accentuées où la durée du mouvement est égale ou inférieure à celle des séquences non-accentuées.

Regardons donc de plus près les différences de durées observées entre les gestes accentués et non-accentués, à même débit. Les tableaux II et III de Kelso *et al.* [1985] montrent que ces différences ne sont pas toujours significatives. Pour le locuteur DW,

l'écart de durée entre gestes d'ouverture (pour /ba/ à débit normal) accentué et non-accentué est de 11.5 ms (123.9 ms (accentué) *vs* 112.4 ms). Le même sujet présente des écarts également faibles, jusqu'à moins de 2 ms, pour ses gestes de fermeture dans /ba/ (débit normal : 91.8 ms (accentué) *vs* 82.5 ms; rapide : 66.3 ms (accentué) *vs* 64.7 ms). De même, le tableau VII de Bateson & Kelso [1993] indique que l'augmentation de durée pour les gestes accentués n'est pas si évidente. Pour un locuteur Anglais parmi les trois étudiés, la différence de durée moyenne entre gestes accentués et non-accentués est dans trois cas inférieure à 10 ms (geste d'ouverture pour /ma/ à débit rapide : 89 ms (accentué) *vs* 79 ms; geste de fermeture pour /ma/ à débit rapide : 78 ms (accentué) *vs* 71 ms; geste de fermeture pour /ba/ à débit rapide : 75 ms (accentué) *vs* 71 ms). Pour les trois locuteurs Français, les différences de durées moyennes entre gestes accentués et non-accentués ne sont absolument pas significatives pour les gestes de fermeture. Les durées des gestes accentués sont égales voire inférieures aux durées des gestes non-accentués. Pour les gestes de fermeture, la différence de durée est inférieure à 6 ms dans deux cas, pour un des trois locuteurs Français (/ba/ à débit rapide : 73 ms (accentué) *vs* 67 ms; /ma/ à débit rapide : 72 ms (accentué) *vs* 69 ms).

Aussi proposons-nous que parallèlement à la cocontraction, la durée du mouvement soit un élément de contrôle explicite de l'accentuation. Cette hypothèse a été rejetée par les travaux émanant des laboratoires *Haskins*, parce qu'elle paraît ne pas entrer dans l'élégant cadre théorique proposé par Carol Fowler [1980] sur le *timing* (ou l'ordonnancement) *intrinsèque*. Selon Fowler, les segments phonologiques doivent être définis dans un espace à quatre dimensions, c'est-à-dire que la représentation que le locuteur a de la séquence phonologique à produire doit inclure le temps. Cependant, diverses données sur la coarticulation et en particulier sur les stratégies d'anticipation labiale dans les séquences VCV (Abry & Lallouache [1996], Perkell & Matthies [1992]) tendent à contredire l'hypothèse d'un temps qui ne serait pas explicitement contrôlé et ne serait que la conséquence du comportement dynamique du système périphérique de production de parole. Il faut donc incorporer le contrôle du temps dans la planification de toute séquence de parole. Nous proposons donc un schéma de contrôle dans lequel non seulement la position d'équilibre intrinsèque est spécifiée mais aussi la cocontraction intrinsèque du système ainsi que l'allure temporelle de la trajectoire du point d'équilibre.

# 4.4 Validation perceptive

## 4.4.1 Synthèse à partir des commandes centrales

Comme au chapitre III, nous proposons de vérifier, à l'aide d'une synthèse du signal acoustique, que les commandes centrales inférées, pour les séquences [iai] et [iɛi], par l'inversion globale sont appropriées. Nous procédons de la même façon qu'il a été indiqué au paragraphe 3.7. Les sept trajectoires articulatoires sont d'abord générées. Celle du corps de la langue est construite à partir des commandes centrales fournies par l'inversion. Le dos est maintenu à une valeur constante (valeur neutre pour [iai], moyenne du [i] et du [ɛ] pour [iɛi]), afin de réduire son influence lors de la synthèse. Les cinq paramètres restant prennent les valeurs inférées par l'inversion cinématique. Le modèle de Maeda transforme ces sept trajectoires en coupes sagittales qui permettent de calculer la fonction d'aire et les formants.

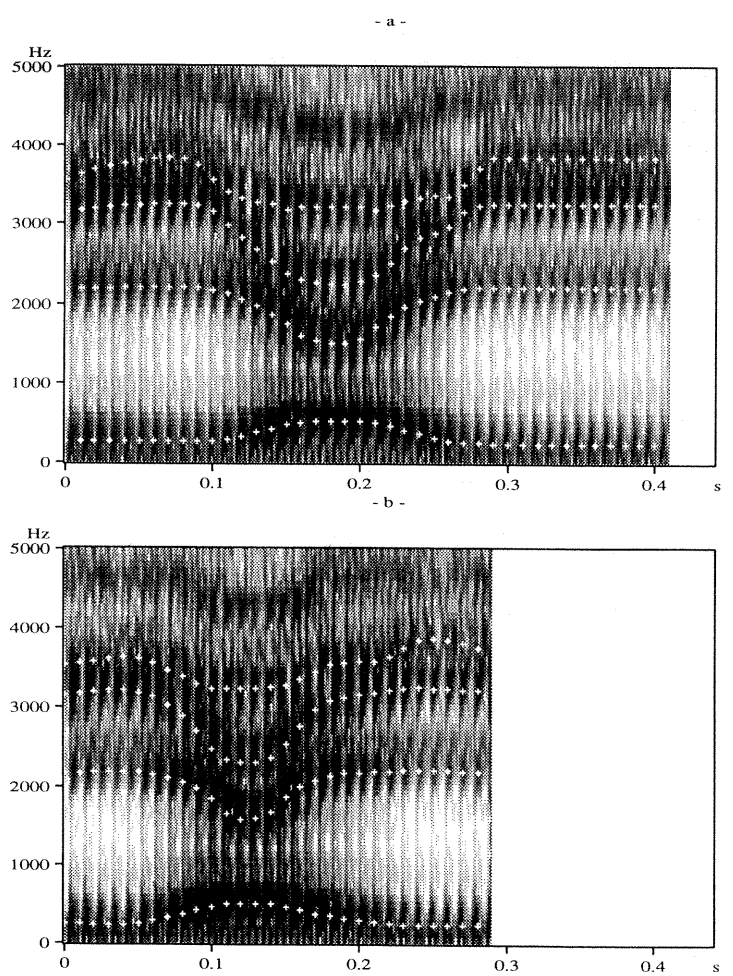



*Figure 4.18.* Sonagrammes synthétiques obtenus à partir des commandes centrales inférées par inversion globale, pour la séquence [iai]. a : lent non-accentué, b : rapide accentué.

Les figures 4.18 a-b présentent les signaux acoustiques synthétisés pour les séquences [iai] lente non-accentuée et rapide accentuée (cf. 3.7 pour le cas lent accentué). Les figures 4.19 a-c fournissent les mêmes signaux pour les trois séquences [iɛi].

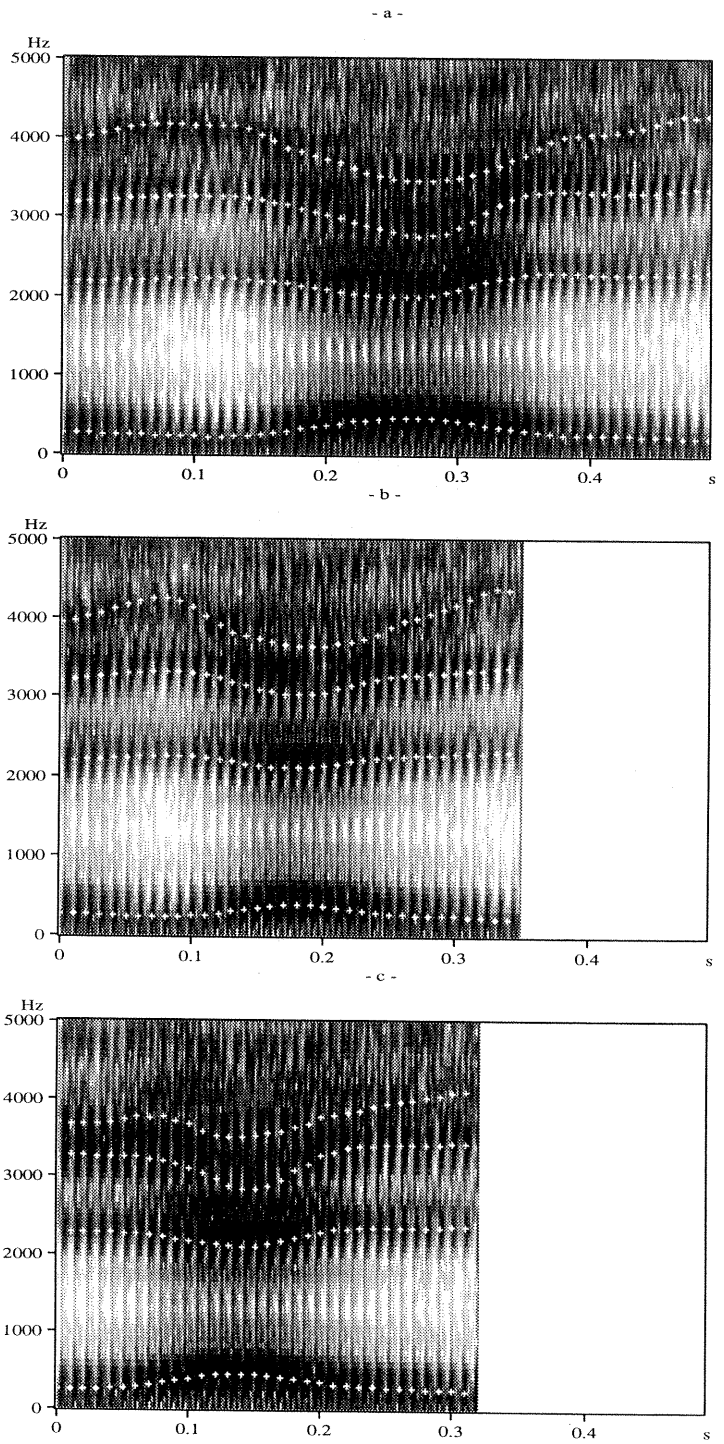



*Figure 4.19.* Sonagrammes synthétiques obtenus à partir des commandes centrales inférées par inversion globale, pour la séquence [iɛi]. a : lent accentué, b :lent non-accentué, c : rapide accentué.

On retrouve sur ces signaux synthétiques les caractéristiques essentielles des signaux originaux. Les tests perceptifs du paragraphe suivant valideront plus en détail la procédure d'inversion globale.

### 4.4.2 Tests Perceptifs

Des tests perceptifs sont menés pour évaluer la pertinence acoustique des résultats de l'inversion globale. Le principe est de vérifier, d'une part, que le robot parlant génère l'information suffisante pour que l'auditeur récupère la cible et, d'autre part, que l'ajustage de ses paramètres dynamiques permet de reproduire les effets de différentes conditions prosodiques.

Afin de vérifier que la cible est toujours récupérable, nous procédons à un test d'identification, dans lequel nous cherchons à vérifier que les séquences entières [iai] et [iɛi] restent distinguées, malgré des conditions prosodiques défavorables, et que la confusion entre les voyelles augmente lorsque l'information dynamique sur la cible est réduite. Des paires de séquences vocaliques ($iV_1i$ suivie de $iV_2i$, $iV_1$ suivie de $iV_2$, $V_1i$ suivie de $V_2i$ ou $V_1$ suivie de $V_2$) sont proposées aux auditeurs qui doivent décider à quelle catégorie appartient la deuxième voyelle inconnue, /a/ ou /ɛ/. La nature de la première voyelle n'est pas indiquée, elle peut être identique à celle de la seconde. Un test d'identification complémentaire est proposé pour déterminer si les voyelles sont identifiables, même sans référence. Les mêmes séquences vocaliques sont étudiées, présentées séparément cette fois (iVi , iV, Vi ou V).

Pour étudier la pertinence des modifications de paramètres dynamiques, nous proposons un test de jugement de qualité. Des paires de séquences vocaliques du même type que celle du premier test d'identification sont présentées aux auditeurs, mais la même voyelle est présente dans les deux séquences et les auditeurs sont informés de sa nature. Leur tâche est de déterminer dans quelle séquence la voyelle est le mieux identifiable.

#### Sujets

Sept sujets Français, âgés de 19 à 28 ans, ont participé aux expériences. Ils n'avaient aucune connaissance du problème considéré, ne présentaient pas de troubles de l'audition et savaient aisément manier la souris d'un ordinateur.

#### Conditions expérimentales

Les expériences ont eu lieu dans une chambre sourde. Les stimuli ont été présentés à travers un casque audio et des messages apparaissaient sur un moniteur couleur. Les auditeurs disposaient d'une souris pour sélectionner les réponses.

### 4.4.2.1 Première expérience d'identification

**Stimuli**

Les stimuli pour les transitions iV et Vi sont obtenus en tronquant simplement le signal acoustique après ou avant la voyelle V. Pour la transition iV, la frontière de la fin de la voyelle V est déterminée par la fin du plateau formantique de V (fin vocalique voisée, cf. Abry, Benoit, Boë, Sock [1985]), et pour la transition Vi, la frontière du début de V correspond au début de la partie stable de V (début vocalique voisé).

Les stimuli pour les voyelles isolées sont générés à partir d'une période excisée du signal acoustique et reproduite un grand nombre de fois, afin d'obtenir une voyelle stable et de longueur satisfaisante. On évite ainsi les effets de *gating*.

Chaque stimulus [iai], [ia], [ai] ou [a] est présenté en paire avec un stimulus équivalent, formé sur la voyelle [ɛ]. Les conditions prosodiques sont croisées ainsi que l'ordre des voyelles et l'on obtient ainsi 18 paires de stimulus par séquence (iVi, iV, Vi ou V), soit 72 paires pour l'ensemble des séquences. Un deuxième type de stimulus est proposé, pour étudier la récupération de la cible du [a] lorsque le contraste ne joue pas entre voyelles mais entre conditions prosodiques. Des paires de séquences vocaliques [iai/iai], [ia/ia], [ai/ai] ou [a/a] sont construites, dans lesquelles on présente d'abord la réalisation idéale puis une des deux réalisations réduites. On obtient ainsi 8 paires supplémentaires. Le premier test d'identification comporte donc 80 paires de stimuli, présentés 5 fois à chaque auditeur, dans un ordre aléatoire.

**Méthodes et consignes**

Les sujets étaient chargés d'identifier le deuxième stimulus de chaque paire présentée, selon une méthode à choix forcé. Ils avaient à "cliquer" avec la souris dans une des deux cases proposées, contenant, pour chaque séquence, les transcriptions orthographiques explicites des deux choix possibles ("i a i" ou "i è i", "i a" ou "i è", "a i" ou "è i", "a" ou "è"). Il a été en outre vérifié que les sujets comprenaient bien "è" comme la voyelle [ɛ] et non [e]. Les sujets étaient informés que les deux stimuli de chaque paire ne comportaient pas nécessairement des voyelles différentes. Deux essais de familiarisation, contenant respectivement des stimuli iVi et V, ont été proposés aux auditeurs avant de commencer l'expérience.

**Résultats**

Le tableau 4.4 (a-d) donne les résultats de ce premier test d'identification, pour l'ensemble des auditeurs (les résultats par sujet sont fournis en annexe B). Les pourcentages

d'identification correcte sont indiqués pour chacun des stimuli étudiés. Les erreurs de manipulation de la souris reportées par les sujets n'ont pas été corrigées.

*Tableau 4.4.a.* Scores d'identification (en pourcentage) des deuxièmes stimuli de chaque paire pour les séquences [iai] et [iɛi] pour 7 auditeurs et 5 répétitions. Les premiers stimuli sont indiqués en ligne. Lent accentué : "Lent acc", Lent non-accentué : "Lent Nacc", rapide accentué : "Rap. acc".

| | [iai] | | | | [iɛi] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [iɛi] Lent acc | 97 | 86 | 97 | [iai] Lent acc | 100 | 100 | 100 |
| [iɛi] Lent Nacc | 100 | 91 | 97 | [iai] Lent Nacc | 100 | 97 | 100 |
| [iɛi] Rap. acc | 100 | 86 | 97 | [iai] Rap. acc | 100 | 100 | 100 |
| [iai] Lent acc | | 69 | 80 | | | | |
| Total | 99 | 83 | 93 | Total | 100 | 99 | 100 |

*Tableau 4.4.b.* Scores d'identification (en pourcentage) des deuxièmes stimuli de chaque paire pour les séquences [ia] et [iɛ].

| | [ia] | | | | [iɛ] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [iɛ] Lent acc | 100 | 57 | 80 | [ia] Lent acc | 100 | 100 | 100 |
| [iɛ] Lent Nacc | 100 | 57 | 57 | [ia] Lent Nacc | 100 | 100 | 100 |
| [iɛ] Rap. acc | 100 | 63 | 63 | [ia] Rap. acc | 100 | 100 | 100 |
| [ia] Lent acc | | 29 | 34 | | | | |
| Total | 100 | 51 | 59 | Total | 100 | 100 | 100 |

*Tableau 4.4.c.* Scores d'identification (en pourcentage) des deuxièmes stimuli de chaque paire pour les séquences [ai] et [ɛi].

| | [ai] | | | | [ɛi] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [ɛi] Lent acc | 100 | 60 | 71 | [ai] Lent acc | 100 | 100 | 97 |
| [ɛi] Lent Nacc | 100 | 66 | 69 | [ai] Lent Nacc | 100 | 100 | 100 |
| [ɛi] Rap. acc | 100 | 77 | 69 | [ai] Rap. acc | 100 | 97 | 100 |
| [ai] Lent acc | | 54 | 57 | | | | |
| Total | 100 | 64 | 66 | Total | 100 | 99 | 99 |

*Tableau 4.4.d.* Scores d'identification (en pourcentage) des deuxièmes stimuli de chaque paire pour les séquences [a] et [ɛ].

| | [a] | | | | [ɛ] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [ɛ] Lent acc | 97 | 49 | 46 | [a] Lent acc | 100 | 100 | 100 |
| [ɛ] Lent Nacc | 100 | 57 | 46 | [a] Lent Nacc | 100 | 97 | 100 |
| [ɛ] Rap. acc | 100 | 54 | 49 | [a] Rap. acc | 100 | 100 | 100 |
| [a] Lent acc | | 20 | 17 | | | | |
| Total | 99 | 45 | 39 | Total | 100 | 100 | 100 |

Nous représentons dans la figure 4.20 les classifications de chaque voyelle, en fonction de la séquence dans laquelle elle a été présentée et selon les diverses conditions prosodiques.

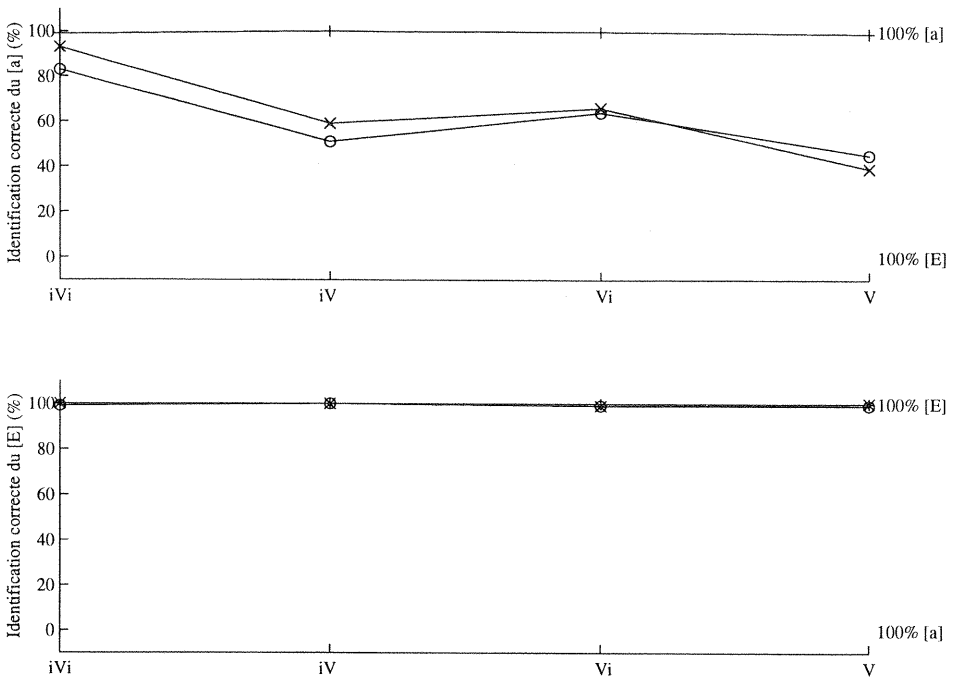



*Figure 4.20.* Pourcentage d'identification correcte du [a] et du [ɛ] selon les séquences vocaliques et dans les 3 conditions prosodiques ('+' : lent accentué, 'o' : lent non-accentué, 'x' : rapide accentué).

Ces résultats indiquent que les cibles du [ɛ] et du [a] sont bien récupérées lorsqu'elles sont présentées dans leur contexte iVi, quelle que soit la condition prosodique. La voyelle [ɛ] est remarquablement bien identifiée dans toutes les conditions prosodiques (99% d'identifications correctes), la voyelle [a] est mieux identifiée dans les conditions accentuées (99%, pour tous les stimuli précédents confondus, dans le cas lent accentué et 93% dans le cas rapide accentué) que dans la condition lente non accentuée (83%). Remarquons qu'un des sujets nous a signalé avoir omis l'information que les deux stimuli de la paire pouvait contenir des voyelles identiques. Les faibles pourcentages d'identification de la cible du [a] lorsque le stimulus précédent contient la même voyelle s'expliquent par cette erreur, les auditeurs plus attentifs ayant en effet des scores d'identification proches de 100% (cf. annexe B).

La voyelle [ɛ] est toujours bien identifiée, quelle que soit la séquence et la condition prosodique. Il semble donc qu'il soit impossible aux auditeurs de la classer dans l'unique autre catégorie proposée [a]. Ceci est cohérent avec le fait que la réduction du [ɛ] dans la séquence [iɛi] tend à déplacer la voyelle [ɛ] vers [e] et donc à l'éloigner encore du [a]. Nous vérifions ainsi que le changement de catégorie observé pour la voyelle [a] n'est pas dû à l'influence du premier stimulus de la paire.

La voyelle [a] en condition idéale est très bien perçue, quelle que soit la séquence (iV, Vi ou V). Les scores d'identifications se dégradent en conditions réduites lorsque le début ou la fin du signal sont tronqués ou lorsque la voyelle seule est présentée.

Ces résultats vont dans le même sens que les analyses de sensibilité présentées au paragraphe 4.3.2. Les séquences en conditions réduites sont plus sensibles aux paramètres dynamiques et les auditeurs commettent plus d'erreur de récupération de cibles.

Il apparaît donc que la réduction vocalique "synthétique" est conforme à celle que l'on observe chez le locuteur humain. La voyelle [a] est en effet toujours bien identifiée en condition idéale, alors qu'en conditions réduites, elle tend à être perçue comme un [ɛ]. D'autre part, lorsque le contexte iVi est intact, le robot fournit l'information utile aux auditeurs pour qu'ils récupèrent la cible, même en conditions prosodiques réduites. En l'absence d'une partie du contexte (séquences iV ou Vi), les auditeurs ne peuvent reconstruire que de façon imprécise la trajectoire des cibles planifiées par le robot et les scores d'identifications chutent aux alentours de 60%. Lorsque la voyelle isolée est présentée, les auditeurs sont en présence d'information statique et non plus dynamique, il n'est donc plus question de restaurer une trajectoire et la voyelle est classée dans la catégorie dont elle est la plus proche statiquement, *i.e.* le [ɛ].

### 4.4.2.2 Expérience d'identification complémentaire

**Stimuli**

Chaque séquence [iVi], [iV], [Vi] ou [V] est présentée de façon isolée, dans les différentes conditions prosodiques et pour les deux voyelles [a] ou [ɛ]. On dispose ainsi de 24 stimuli, présentés chacun 5 fois, dans un ordre aléatoire, aux auditeurs.

**Méthodes et consignes**

La même procédure que pour le premier test d'identification a été mise en place. Les sujets devaient "cliquer" avec la souris dans la case correspondant le mieux au stimulus présenté.

**Résultats**

Les résultats de ce test d'identification complémentaire sont fournis dans le tableau 4.5 suivant.

*Tableau 4.5* Scores d'identification (en pourcentage) des stimuli isolés pour 7 auditeurs et 5 répétitions.

| | a | | | ɛ | | |
|---|---|---|---|---|---|---|
| | Lent acc | Lent Nacc | Rap. acc | Lent acc | Lent Nacc | Rap. acc |
| iVi | 100 | 86 | 100 | 94 | 100 | 97 |
| iV | 100 | 46 | 83 | 100 | 100 | 100 |
| Vi | 97 | 60 | 86 | 100 | 100 | 100 |
| V | 100 | 40 | 34 | 100 | 100 | 100 |

Ces résultats sont très proches de ceux que l'on observe pour les stimuli par paire. Il semble que donc que la valeur contrastive du premier stimulus de chaque paire ne soit pas nécessaire pour que les auditeurs récupèrent la cible. Les phénomènes de changement de catégorie dus à la réduction vocalique sont aussi observés. On note que les piètres résultats du test précédent pour la voyelle [a] présentée deux fois de suite sont bien dus à une mauvaise compréhension des consignes par un des sujets. Cependant il faut noter que cette deuxième expérience a été mise en place en complément à la première. Les stimuli isolés ont donc été présentés APRÈS que les locuteurs se soient acclimatés aux contrastes [a]/[ɛ]. Il serait donc intéressant d'étudier ultérieurement, pour ce test d'écoute isolée, les réponses d'auditeurs n'ayant aucune connaissance des stimuli.

### 4.4.2.3 Expérience de jugement de qualité phonétique

#### Stimuli

Des paires de stimuli comprenant la même voyelle [a] ou [ɛ] sont construits à partir des séquences iVi, iV, Vi ou V. Les paires contiennent la condition idéale et une des deux conditions réduites. L'ordre de présentation des conditions prosodiques est aléatoire. On dispose de 32 stimuli en tout (16 par voyelles, 8 par ordre de présentation, 2 par séquence). Ces stimuli sont présentés 5 fois à chaque auditeur, dans un ordre aléatoire.

#### Méthodes et consignes

Pour chaque paire de stimuli, un message s'affichait à l'écran précisant la nature de la voyelle à juger ([a] ou [ɛ]). Les sujets devaient déterminer dans quel stimulus la voyelle leur semblait le mieux identifiable, et "cliquer" avec la souris sur la case appropriée ("premier" ou "deuxième").

**Résultats**

Nous présentons dans les tableaux 4.6a-d les résultats de l'expérience de jugement de qualité.

*Tableau 4.6.a.* Scores de qualité (en pourcentage) des stimuli lents accentués en séquence [iVi] pour 7 auditeurs et 5 répétitions.

| | [a] Lent acc | [ɛ] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 86 | 80 |
| suivi du cas rapide accentué | 91 | 57 |
| précédé du cas lent non accentué | 100 | 94 |
| précédé du cas rapide accentué | 97 | 86 |

*Tableau 4.6.b.* Scores de qualité (en pourcentage) des stimuli lents accentués en séquence [iV] pour 7 auditeurs et 5 répétitions.

| | [a] Lent acc | [ɛ] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 100 | 94 |
| suivi du cas rapide accentué | 100 | 66 |
| précédé du cas lent non accentué | 100 | 94 |
| précédé du cas rapide accentué | 100 | 91 |

*Tableau 4.6.c.* Scores de qualité (en pourcentage) des stimuli lents accentués en séquence [Vi] pour 7 auditeurs et 5 répétitions.

| | [a] Lent acc | [ɛ] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 100 | 91 |
| suivi du cas rapide accentué | 97 | 63 |
| précédé du cas lent non accentué | 94 | 91 |
| précédé du cas rapide accentué | 97 | 63 |

*Tableau 4.6.d.* Scores de qualité (en pourcentage) des stimuli lents accentués en séquence [V] pour 7 auditeurs et 5 répétitions.

| | [a] Lent acc | [ɛ] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 97 | 83 |
| suivi du cas rapide accentué | 100 | 69 |
| précédé du cas lent non accentué | 100 | 89 |
| précédé du cas rapide accentué | 100 | 66 |

La voyelle [a] est clairement jugée de meilleure qualité en condition idéale qu'en conditions réduites, quels que soient la séquence considérée (iVi, iV, Vi ou V) et l'ordre de présentation des conditions prosodiques. Il est de plus intéressant d'observer que la décision en faveur du [a] lent accentué par rapport au [a] rapide accentué, ou au [a] lent non accentué, est moins nette en contexte iVi complet que dans les autres cas, où les décisions sont quasi unanimes. Ceci va bien dans le sens d'une validation de notre stratégie de production de la séquence : il y a dans la dynamique des informations sur la cible intentionnelle que les auditeurs récupèrent, pour compenser les mauvaises réalisations de patron formantique. De plus, alors même qu'elles présentent, hors du contexte iVi complet, des évaluations perceptives très proches, les voyelles [a] lente non-accentuée et [a] rapide accentuée se différencient clairement lorsqu'elles sont présentées en contexte (cf. tableau 4.5). Ceci est une validation supplémentaire de notre stratégie de contrôle des effets prosodiques du type accentuation et débit d'élocution sur une séquence vocalique.

Pour la voyelle [ɛ], les choix sont moins tranchés, quels que soient les contextes dans lesquels la voyelle est présentée. C'est en particulier le cas pour la voyelle isolée : ce phénomène laisse penser que la réduction dans le domaine formantique est perceptivement moins importante que pour la voyelle [a]. Par ailleurs, les tendances observées pour la voyelle [a] se retrouvent pour [ɛ] : la décision en faveur du [ɛ] lent accentué est moins nette lorsque les voyelles sont présentées en contexte ; le cas rapide accentué est le concurrent le plus important du cas lent accentué, ceci dans tous les contextes, y compris le contexte isolé, et surtout dans le contexte complet. Cependant, dans le plan formantique (pour les données), le cas rapide accentué est moins réduit que le cas lent non-accentué (ceci se retrouve dans les évaluations perceptives en contexte isolé), alors que la trajectoire du corps de la langue est de plus faible amplitude. À l'évidence donc les autres paramètres articulatoires jouent un rôle important dans la réalisation du [ɛ]. Ceci doit pondérer la validation de nos hypothèses dans le cas du [ɛ].

## 4.5 Bilan

Cette première expérience d'application de la notion de cibles dans trois conditions prosodiques et pour deux séquences vocaliques [iai] et [iɛi] est encourageante. Il est en effet possible de simuler des trajectoires articulatoires, inférées à partir du signal acoustique enregistré dans diverses conditions prosodiques, en spécifiant, pour chaque chaîne phonémique, une seule et même commande posturale, en termes de positions d'équilibre cibles de l'articulateur et en déplaçant à vitesse contrôlée le point d'équilibre entre chacune des positions cibles. La variabilité acoustique et articulatoire, observée sur les données, est reproduite en ajustant des paramètres prosodiques liés à l'accentuation et au *timing*. Un

robot parlant qui serait contrôlé selon ces hypothèses serait donc capable, avec un nombre limité de paramètres variables de produire une large éventail de productions articulatoires et acoustiques des plus réduites au plus achevées. En augmentant la cocontraction et/ou le temps de maintien, le robot parlant peut mettre en œuvre l'accentuation, en diminuant tous les paramètres temporels, il peut simuler des effets dus à la rapidité du débit d'élocution. Diverses stratégies de contrôle sont possibles, qui diffèrent selon les contraintes imposées par la commande posturale. Si l'amplitude du mouvement de cible à cible est élevée (cf. [iai]), le robot est limité temporellement et doit ajuster précisément le niveau de cocontraction. Si l'amplitude du mouvement est faible, il peut plus à sa guise disposer des différents paramètres et ainsi, par exemple (cf. [iɛi]), réduire son effort, en diminuant le niveau de cocontraction tout en augmentant les diverses durées. La notion de cible, fort débattue, s'avère donc ici être la base d'un schéma de contrôle, parcimonieux mais efficace, de la production de voyelles. Dans le cadre d'une théorie de la perception qui mettrait en action des phénomènes d'inversion, nos résultats pourraient contribuer à expliquer l'intérêt perceptif de l'accentuation. Nous avons en effet montré que les configurations temporelles et dynamiques étaient beaucoup moins stables dans le cas de la réduction vocalique, suggérant une plus grande probabilité de confusion avec des séquences proches. Les tests perceptifs, menés sur les signaux acoustiques synthétiques, générés à partir des paramètres inférés par inversion globale, confirment la pertinence phonétique du petit nombre de paramètres extraits. Cependant cette première expérience de mise en œuvre des cibles s'est appuyée sur un modèle simplifié de la dynamique des articulateurs, qui représente mal la réalité neurophysiologique. La partie suivante présente une première tentative de validation sur un modèle biomécanique plus complexe.

## 4.6 Récupération de cibles à partir d'un modèle biomécanique

L'expérience d'inversion globale mise en œuvre précédemment semble conforter l'hypothèse de l'existence de cibles en production de parole. Cependant, les cibles sont inférées à partir d'un modèle fonctionnel très simple qui ne représente pas la réalité physique du contrôle. Un des inconvénients majeurs du modèle utilisé est qu'il ne distingue pas les aspects liés au contrôle du système des aspects purement biomécaniques. Dans ce chapitre, nous proposons de tester à nouveau l'hypothèse des cibles en utilisant le modèle biomécanique de l'ensemble mandibule/os hyoïde, proposé par Laboissière, Ostry & Feldman [1996]. Ce modèle représente en effet de façon sophistiquée et réaliste la biomécanique de la mandibule et décrit séparément les aspects spécifiques au contrôle neurophysiologique, qui mettent en œuvre la notion de Point d'Équilibre.

### 4.6.1 Le modèle biomécanique de la mandibule (Laboissière *et al.* [1996])

Le modèle proposé par Laboissière, Ostry & Feldman [1996] décrit les mouvements de la mandibule et de l'os hyoïde dans le plan sagittal. Il distingue explicitement les aspects biomécaniques et les aspects de contrôle neurophysiologiques de l'ensemble mandibule/os hyoïde.

Du point de vue biomécanique, le modèle comporte sept muscles (ou groupes de muscle) et quatre degrés de liberté cinématiques. La figure 4.21 représente les groupes de muscles en jeu.

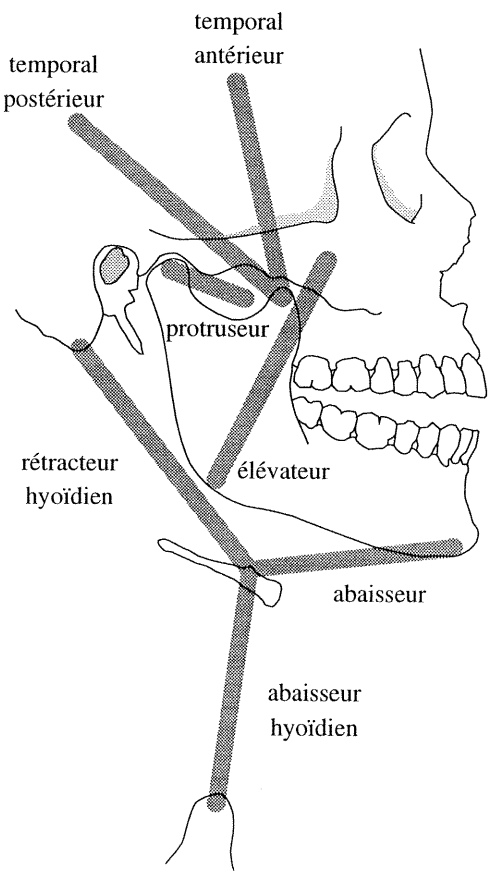



*Figure 4.21.* Représentation schématique des muscles du modèle et de leurs points d'insertion sur la mandibule et l'os hyoïde. D'après Laboissière, Ostry & Feldman [1996].

Le contrôle du système biomécanique met en œuvre l'hypothèse du Point d'Équilibre telle qu'elle a été explicitement décrite par Feldman [1966] (cf. 2.3.1) : les mouvements résultent de modifications des variables de contrôle neurophysiologique qui déplacent le point d'équilibre du système moteur. Les principales variables de contrôle

central agissent sur les potentiels de membrane des motoneurones, qui établissent une longueur de muscle seuil ($\lambda$) à partir de laquelle les motoneurones sont recrutés. L'activation musculaire, et par conséquent la force, varie en fonction de la différence entre le seuil spécifié et la longueur effective du muscle, ainsi qu'en fonction de la vitesse de changement de la longueur musculaire. Ainsi, un changement de seuil $\lambda$, résultant de modifications centrales des potentiels des motoneurones, fait passer le système d'une position d'équilibre à une autre et génère le mouvement. Les mécanismes élémentaires du contrôle selon l'hypothèse du Point d'Équilibre sont décrits plus en détail au 2.3.1.

Dans le modèle de l'ensemble mandibule/os hyoïde développé par Laboissière *et al.*, les mouvements ne sont pas contrôlés directement en termes de commandes centrales spécifiques à chacun des muscles. Les signaux de contrôle, correspondant à diverses combinaisons de $\lambda$s, sont spécifiés dans l'espace des degrés de liberté cinématiques du système. Ils permettent ainsi de générer des mouvements de rotation et de translation horizontale de la mandibule ainsi que des mouvements de translation horizontale et verticale de l'os hyoïde. Le niveau de cocontraction est également contrôlé, en spécifiant le niveau de force globale en jeu. Ces différents signaux de contrôle peuvent agir de façon isolée ou groupée.

La figure 4.22 décrit l'organisation du modèle. Les signaux de contrôle, correspondant à des spécifications dans l'espace des degrés de liberté du système, sont transformés en signaux de contrôle individuels pour chaque muscle, en termes de seuils d'activation des motoneurones. Une information afférente fournit la longueur musculaire et la vitesse.

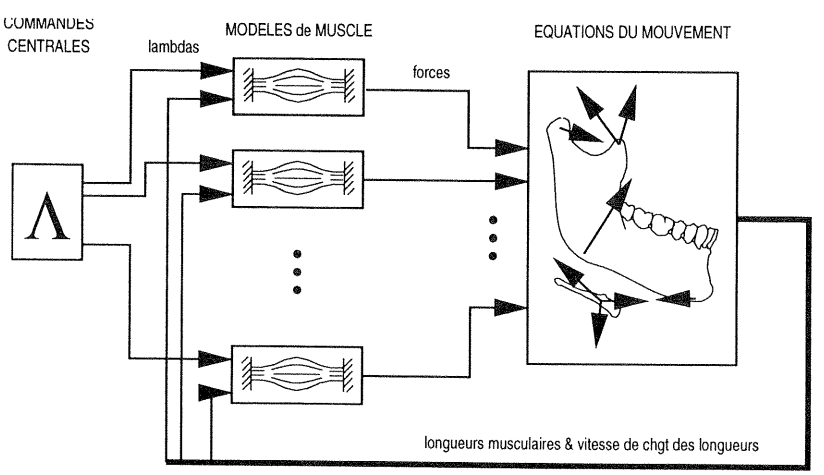



*Figure 4.22.* Représentation schématique de l'organisation du modèle de mandibule, incluant les signaux de contrôle central, les muscles modélisés, la dynamique et le *feedback* véhiculant l'information afférente sur les longueurs musculaires et la vitesse de changement des longueurs. D'après Laboissière *et al.* [1996].

Les quatre degrés de liberté du système dépendant de l'état de sept muscles, il existe une infinité de combinaisons de seuils musculaires $\lambda$ correspondant à toute configuration

géométrique statique. L'ensemble des points de l'espace des λ qui correspondent à une même position de l'ensemble mandibule/os hyoïde est appelé le *no-motion manifold* (la variété linéaire statique, VLS par la suite). Chaque configuration du système est ainsi associée à une VLS donnée et les mouvements correspondent à des changements de VLS. La figure 4.23 illustre la notion de VLS à l'aide d'une projection, calculée à l'aide du modèle, dans un espace à deux dimensions défini par les seuils λ de deux muscles (un abaisseur, un élévateur).

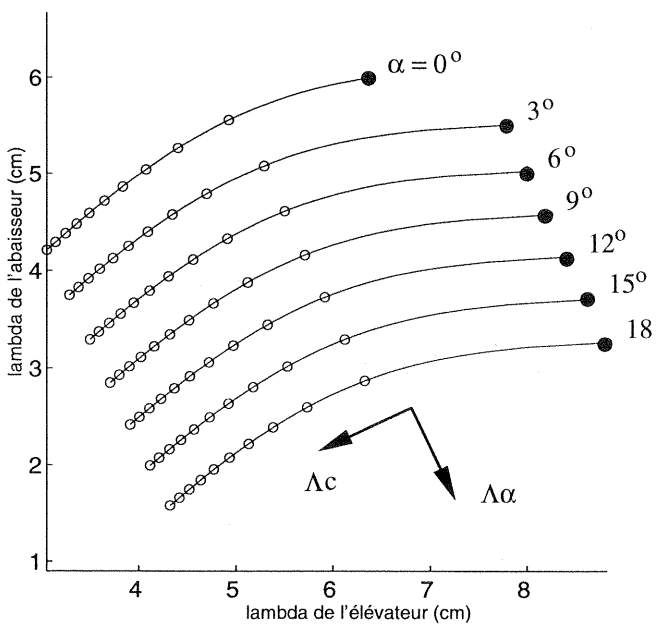



*Figure 4.23.* Projections de *Variétés Linéaires Statiques* dans un espace à deux dimensions, correspondant à un muscle abaisseur et un muscle élévateur. D'après Perrier, Ostry & Laboissière [1996].

Les différentes VLS qui composent cette figure correspondent à diverses configurations statiques de l'ensemble mandibule/os hyoïde (divers angles d'ouverture, dans ce cas simplifié). Chaque VLS est définie par les diverses combinaisons de λ pour les muscles abaisseur et élévateur qui correspondent à un angle donné. Les différentes combinaisons de λ, représentées par des cercles sur chaque VLS, sont associées à des forces musculaires totales allant de 10N à 100N, par pas de 10N. Les cercles pleins à l'extrémité de chaque VLS correspondent aux combinaisons où la force totale est minimale (compensant simplement le poids de la mandibule). Dans ce schéma, le mouvement est défini en spécifiant des vecteurs ($\Lambda_\alpha$ sur la figure) qui produisent des déplacements de λ d'une VLS à une autre. Parallèlement, il est possible de définir la cocontraction musculaire en définissant des déplacements de λ à l'intérieur d'une VLS (vecteur $\Lambda_c$ sur le schéma).

Dans le modèle complet, le mouvement est défini par des vecteurs à 4 dimensions (correspondant aux quatre degrés de liberté) et la cocontraction par des vecteurs à 3 dimensions (correspondant aux bases de chaque VLS à 3 dimensions).

Ainsi, ce modèle, qui sépare le contrôle des configurations d'équilibre et de la cocontraction, permet de mettre en œuvre nos hypothèses sur les cibles, exprimées en termes de positions d'équilibre du système, et la variabilité, paramétrée par la cocontraction, la durée des maintiens des positions d'équilibre et la vitesse de transition d'une position d'équilibre à une autre.

## 4.6.2 Acquisition des données

### 4.6.2.1 Corpus

Le corpus est similaire à celui décrit au paragraphe 3.3. Il a été recueilli au sein du Laboratoire de Contrôle Moteur de l'Université Mc Gill à Montréal. Il s'agit de la séquence vocalique [iai] issue de la phrase porteuse "il **y a** immédiatement". Un locuteur Français (TGM) et une locutrice Française (HL) ont participé. Trois conditions d'élocution sont étudiées : lente et accentuée, lente et non-accentuée, rapide et accentuée, les mêmes consignes étant données aux locuteurs que pour le corpus du 3.3. Pour chaque locuteur, parmi les dix répétitions de chaque énonciation, deux sont sélectionnées.

### 4.6.2.2 Méthode

Les mouvements de translation horizontale et de rotation, dans le plan sagittal, de la mandibule sont suivis grâce au système OPTOTRAK (Northern Digital), qui capte la lumière infrarouge émise par des diodes IRED (Infra-Red Emitting Diods).

Un appareil dentaire en acrylique ajusté à la dentition inférieure a été construit pour chaque sujet, à partir de ses empreintes dentaires. Une tige métallique légère, mais suffisamment rigide, est fixée sur la partie frontale de l'appareil, ses extrémités ressortant de la bouche horizontalement, au niveau des commissures des lèvres. Puisque la précision des données est améliorée si le nombre des cellules est élevé, et puisqu'une augmentation de la distance entre les cellules réduit le niveau de bruit sur les données, nous avons choisi de ne pas fixer les cellules IREDs directement sur cette tige. Des baguettes en bambou, formant un losange d'environ 15cm de long, sont donc attachées à la tige, permettant l'utilisation de cinq cellules IREDs bien espacées, dont les mouvements suivent ceux de la mandibule. Six cellules supplémentaires sont fixées à un casque et sont utilisées pour soustraire les

mouvements de la tête des mouvements de la mandibule. La photographie de la figure 4.24 donne une idée de l'appareillage utilisé.

*Figure 4.24.* Appareillage permettant l'acquisition des mouvements de rotation et de translation horizontale de la mandibule.

Les positions des cellules sont échantillonnées à 100 Hz et filtrées à l'aide d'un filtre bidirectionnel de Butterworth d'ordre 2 et de fréquence de coupure de 10Hz. Des enregistrements statiques et des mesures de la distance du condyle aux incisives inférieures permettent de calibrer le logiciel d'acquisition, associé au système OPTOTRAK, et de récupérer les données articulatoires, en termes d'angles pour la rotation et de positions pour la translation horizontale (cf. Bateson & Ostry[1995] pour plus de précisions sur l'acquisition des données avec le système OPTOTRAK).

Le signal acoustique est enregistré simultanément et échantillonné à 10kHz. La séquence [iai] est extraite de la phrase porteuse en utilisant les critères acoustiques de début et fin vocaliques voisés (cf. Abry, Benoit, Boë, Sock [1985]).

### 4.6.3 Simulations

Les transitions [ia] sont essentiellement caractérisées par des mouvements de rotation de la mandibule. Les mouvements de translation horizontale ainsi que les mouvements de

l'os hyoïde ne sont donc pas étudiés ici, dans le but de faciliter les simulations. Nous utilisons en outre une version simplifiée du modèle qui ne tient pas compte des commandes centrales liées aux mouvements de l'os hyoïde. Seule la commande centrale liée à la rotation (R) est par conséquent manipulée. La commande liée à la translation horizontale (T) passe par trois paliers fixes correspondant à des valeurs moyennes pour les trois voyelles successives. On a ainsi, quels que soient le locuteur et la condition prosodique :

$T(i_1)$ = 1mm

T(a) = 4.5 mm

$T(i_2)$ = 1.3mm

Nous supposons que cette commande évolue de façon synchrone avec la commande de rotation.

Nous reprenons la stratégie utilisée au paragraphe 4.3. L'hypothèse testée est que les mouvements accentués sont associés à des niveaux de force totale (ou de cocontraction) élevés et que les changements de débit d'élocution correspondent à des modifications du *timing* des commandes. Les cibles équilibres planifiées, en termes d'angles de rotation, sont donc d'abord inférées, par essais successifs, à partir de la trajectoire articulatoire obtenue dans la condition lente accentuée. Les différents paramètres temporels, ajustés ensuite, sont les temps de maintien des deux [i], les temps de transition (du [i] au [a] et du [a au [i]) et le temps de maintien du [a].

Les résultats des simulations pour la condition lente accentuée sont présentés, pour les deux énonciations sélectionnées, sur les figures 4.25.a et b pour les locuteurs TGM et HL respectivement. Pour chaque locuteur, les *mêmes* angles cibles équilibres sont utilisés dans les deux énonciations. Pour le locuteur TGM, les trois cibles équilibres définissant la commande centrale de rotation R sont :

$R(i_1)$ = -2.8°

R(a) = -13.3°

$R(i_2)$ = -2.5°

Et pour le locuteur HL :

$R(i_1)$ = -2°

R(a) = -10.5°

$R(i_2)$ = -2°

Les valeurs des paramètres temporels sont indiquées sur les figures. Les temps de transition sont de l'ordre de 100ms, les temps de maintien de la voyelle centrale de l'ordre de 70ms. La force totale en jeu est élevée (78N).

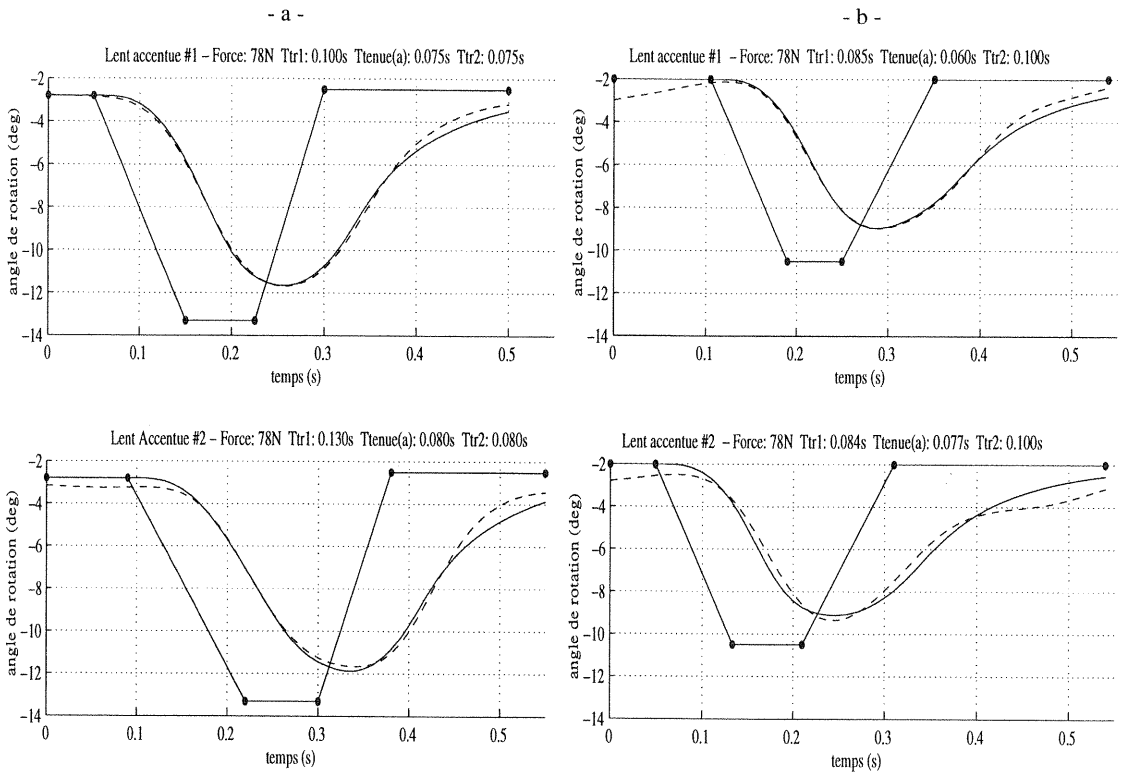



*Figure 4.25.* Simulation de la rotation de la mandibule pour deux répétitions de la séquence [iai] lente accentuée, pour chaque locuteur. Dans chaque panneau, la courbe en trait tireté correspond aux données expérimentales, celle en trait plein à la simulation. La commande centrale de rotation R est figurée par la ligne brisée joignant les points. a : locuteur TGM, b : locuteur HL.

Les courbes simulées (représentées en trait plein) sont assez proches des données (trait tireté). De la même façon que pour les simulations du 4.3, la non adéquation entre données et simulations sur les portions correspondant aux voyelles [i] est due au choix délibéré de ne pas tenir compte des contextes phonétiques précédant et suivant la séquence [iai]. Les légères différences entre données et simulations sur la voyelle [a] proviennent du couplage dynamique entre rotation et translation horizontale inhérent au modèle (cf. plus loin, paragraphe 4.6.4).

Pour la condition rapide accentuée, les mêmes plateaux de la commande centrale R sont utilisés que pour la condition idéale (lente accentuée). Les simulations pour les deux locuteurs et les deux énonciations sont représentées sur les figures 4.26.a et b. Le même niveau de force est impliqué (78N) que pour la condition idéale, mais les paramètres temporels varient. Le temps de maintien de la voyelle centrale est fortement diminué (de 70ms à 8ms), le temps de transition est du même ordre que dans la condition lente accentuée pour la descente du [i] au [a], mais diminue nettement pour la montée du [a] au [i], qui est d'ailleurs la transition la moins bien simulée. Il est probable que l'absence de cible après le deuxième [i] pénalise la simulation de la trajectoire du [a] au [i].

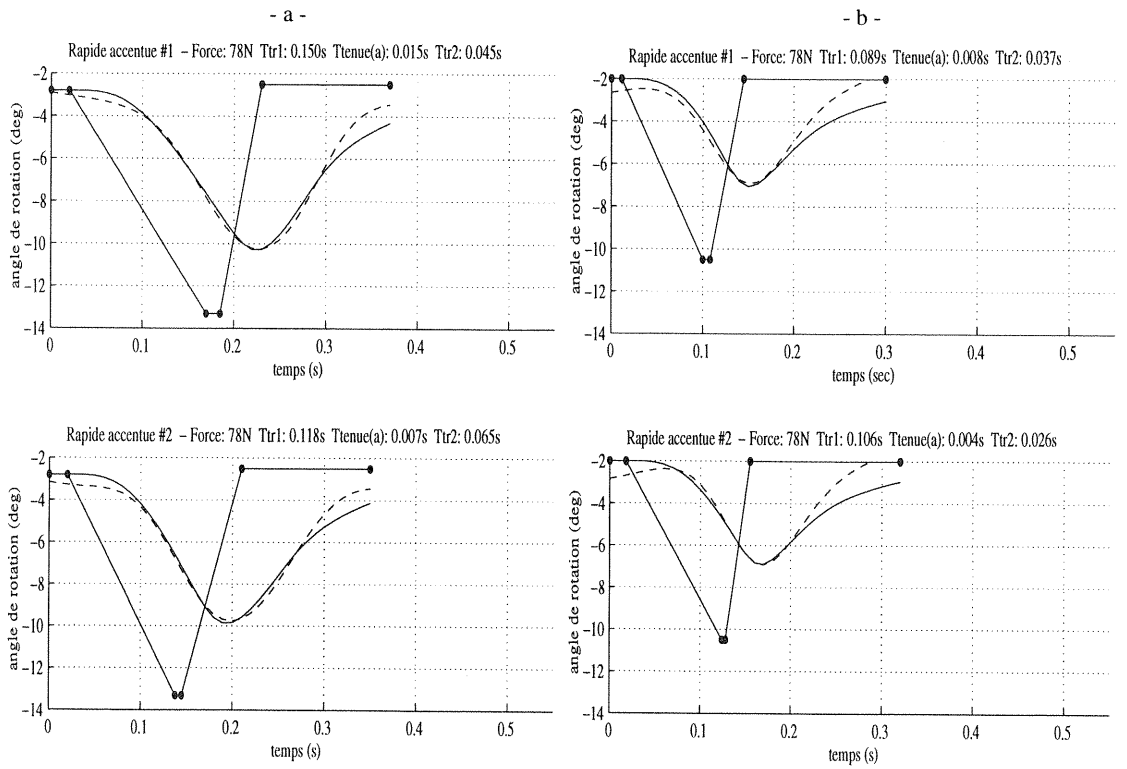



*Figure 4.26.* Simulation de la rotation de la mandibule pour deux répétitions de la séquence [iai] rapide accentuée, pour chaque locuteur. Les cibles équilibres sont les mêmes pour les deux répétitions de chaque locuteur. Les mêmes conventions de traits sont utilisées que précédemment. a : locuteur TGM, b : locuteur HL.

Une première simulation avec les cibles inférées pour la condition idéale, s'est avérée infructueuse pour la condition lente non-accentuée. Les résultats de cette simulation sont présentés sur les figures 4.27 a et b pour les deux répétitions des locuteurs TGM et HL. Même en réduisant le temps de maintien et la force totale au minimum (0.01ms et 10N) et en diminuant la vitesse de transition de cible à cible, il est impossible de simuler adéquatement les données articulatoires : bien que fortement réduite par rapport à la condition idéale, la trajectoire simulée dépasse en amplitude la trajectoire donnée.

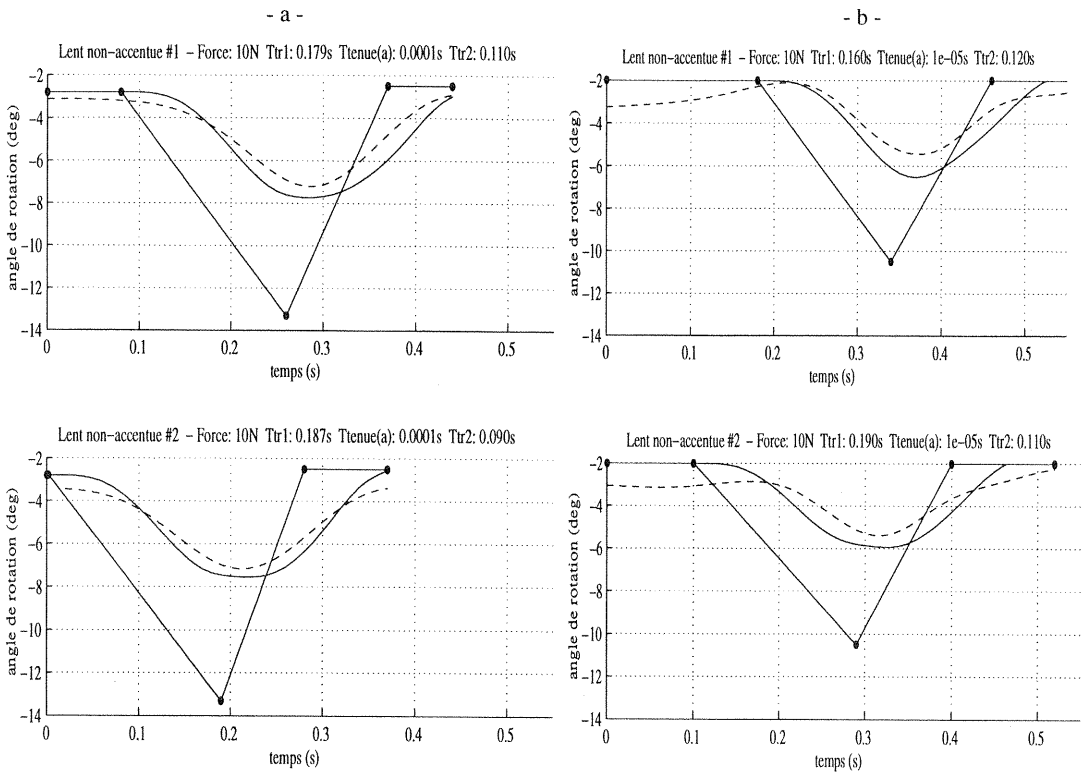



*Figure 4.27.* Tentative infructueuse de simulation de la rotation de la mandibule pour deux répétitions de la séquence [iai] lente non-accentuée, pour chaque locuteur. Les mêmes conventions de traits sont utilisées que précédemment. a : locuteur TGM, b : locuteur HL.

Une deuxième simulation pour laquelle l'amplitude du déplacement de cible est réduite par rapport à la condition idéale fournit une meilleure adéquation entre simulations et données (figures 4.28 a et b). Pour cette dernière simulation, le niveau de force totale est minimal (10N), les temps de transitions sont très élevés (de l'ordre de 170ms pour la transition du [i] au [a]), le maintien de la voyelle centrale a quasi disparu. Les nouvelles cibles définissant la commande centrale R sont, pour le locuteur TGM :

$R(i_1) = -2.8°$

$R(a) = -12°$

$R(i_2) = -2.5°$,

et pour le locuteur HL :

$R(i_1) = -2°$

$R(a) = -8.3°$

$R(i_2) = -2°$.

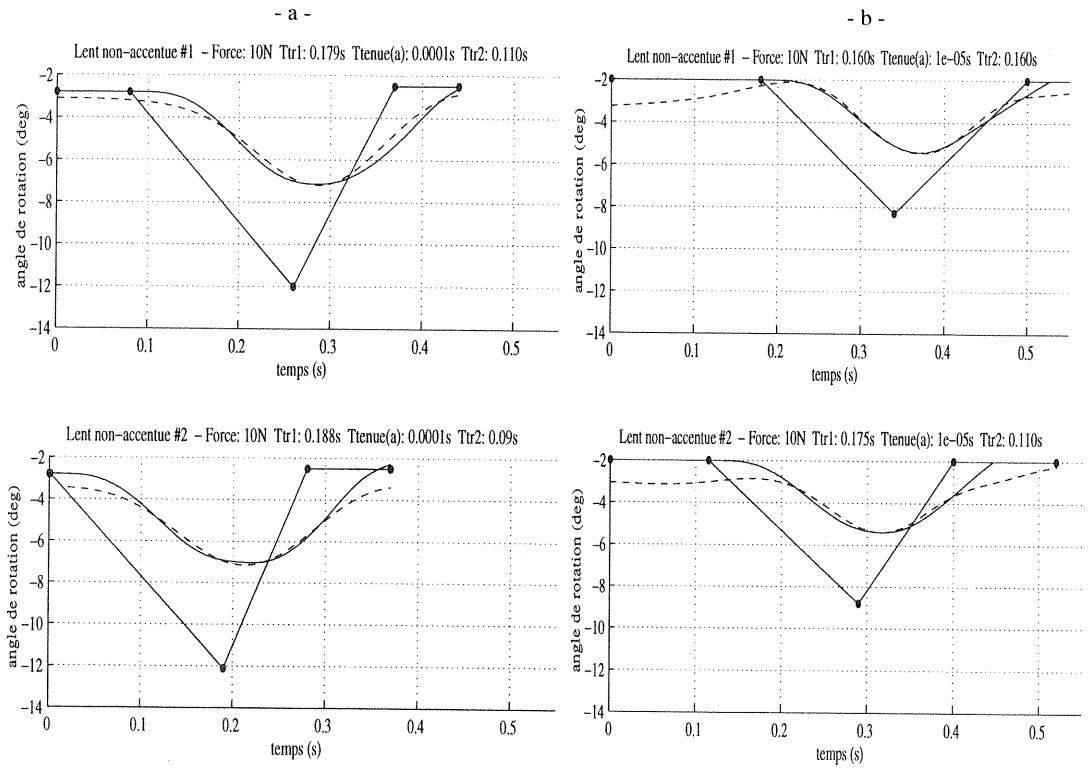



*Figure 4.28.* Simulation de la rotation de la mandibule pour deux répétitions de la séquence [iai] lente non-accentuée, pour chaque locuteur. Les mêmes conventions de traits sont utilisées que précédemment. a : locuteur TGM, b : locuteur HL.

### 4.6.4 Discussion

La simulation du cas rapide accentué fournit des résultats en accord avec notre hypothèse initiale. Une simple réduction du temps de maintien de la voyelle [a] engendre un *undershoot* de la position cible, comparable à celui que l'on observe sur les données empiriques. Le niveau élevé de force globale ainsi que la vitesse de transition relativement élevée permettent de simuler les transitions à fortes pentes observées.

La forte réduction de l'amplitude du mouvement obtenue en diminuant le niveau de force globale, pour la simulation du cas lent non-accentué, confirme le rôle important du niveau de cocontraction dans le contrôle de la dynamique articulatoire.

Cependant, la réduction du niveau de force et de la vitesse de transition ne permettent pas de produire un *undershoot* aussi important que celui observé empiriquement. La position d'équilibre correspondant à la voyelle [a] doit être modifiée pour obtenir une bonne adéquation entre simulations et données. Ceci contredit notre hypothèse de départ

selon laquelle les cibles articulatoires restent invariantes quel que soit le niveau d'accentuation. Ce résultat pourrait s'interpréter de la façon suivante :

Les cibles sont modifiées lorsque l'on passe d'une condition non-accentuée à une condition accentuée. L'accentuation d'emphase est une condition particulière de la parole qui demande des gestes particuliers, correspondant à des cibles phonémiques plus éloignées les unes des autres. On retrouve là la notion de *window model* de Keating [1988] (modèle à fenêtre), selon laquelle des fenêtres plus étroites correspondraient aux conditions plus accentuées. La figure 4.29 illustre cette notion pour la séquence [iai] dans deux conditions différant par leur niveau d'accentuation.

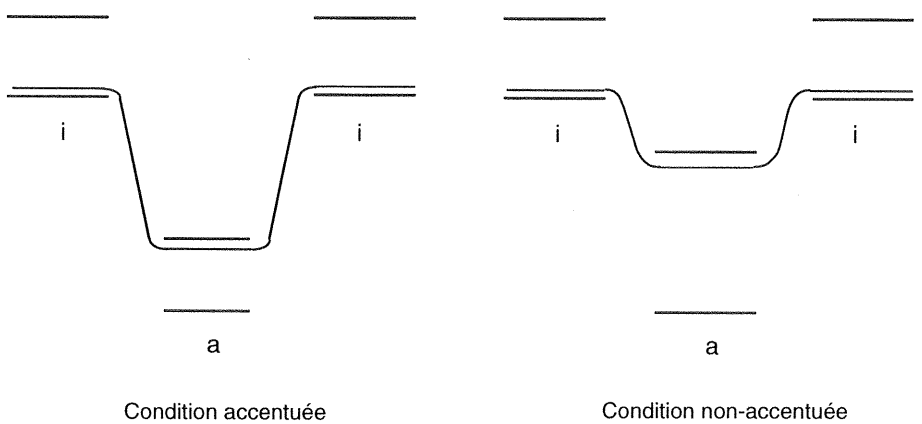



*Figure 4.29.* Une application du *window model* de Keating [1988] pour la séquence [iai] en conditions accentuée et non accentuée.

Cependant, nous nous garderons bien de conclure dans cette direction. Car les résultats que nous avons obtenus avec le modèle de la mandibule sont à considérer avec la plus grande précaution. En effet il existe, dans la version utilisée du modèle, de fortes interactions entre la dynamique de la rotation et de la translation de la mandibule. La figure 4.30, d'après Perrier *et al.* [1996], présente les résultats de simulations, mettant en œuvre des déplacements de configuration d'équilibre de 5mm, dans huit directions (indiquées sur le panneau de gauche).

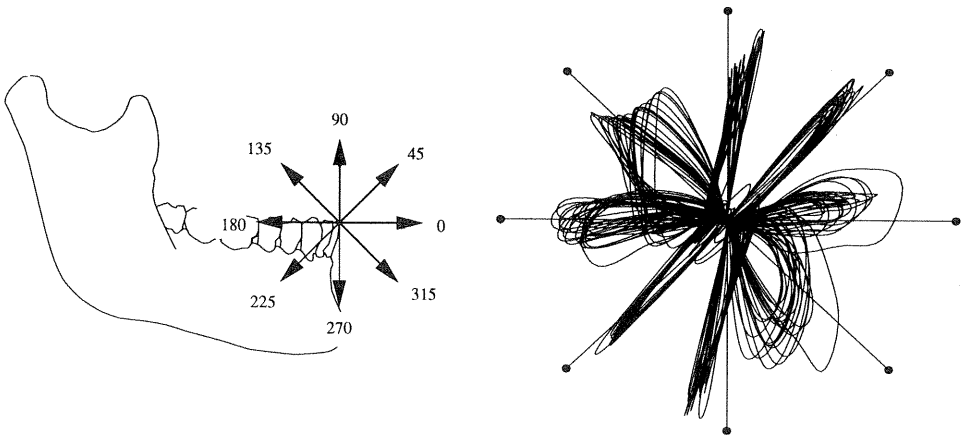



*Figure 4.30.* Simulations de déplacements de 5mm de la configuration d'équilibre dans huit directions. D'après Perrier *et al.* [1996].

Les configurations d'équilibre de départ et d'arrivée correspondent au centre du système de flèches, au niveau des incisives, et les positions intermédiaires sont écartées de 5mm dans la direction des flèches. Différents niveaux de cocontraction sont testés dans chaque direction. Les trajectoires simulées des incisives (représentées dans le panneau de droite) diffèrent selon la direction du mouvement et selon le niveau de cocontraction. À niveau de cocontraction faible, l'influence de la dynamique de la translation sur la rotation est relativement importante. Il apparaît donc qu'il n'est pas possible de simuler le mouvement selon un des deux degrés de liberté sans proposer une gestion explicite de la coordination des degrés de liberté. Comme nous l'avons souligné au chapitre II, l'hypothèse du contrôle par point d'équilibre ne propose rien sur les coordinations entre les degrés de liberté. Ces constats, tout en minimisant la portée des résultats que nous obtenons avec le modèle plus réaliste de la mandibule, soulignent la difficulté qu'il y avait pour nous à utiliser un tel modèle dans un processus complexe d'inversion des trajectoires formantiques vers les commandes motrices. Dans l'état actuel de nos connaissances sur les coordinations entre les différents degrés de liberté des articulateurs de la parole, notre choix d'un modèle articulatoire, rendant compte des coordinations cinématiques entre articulateurs est donc fondé.

Cependant, il est bien évident que les conclusions présentées dans les pages précédentes, doivent maintenant être validées sur un modèle plus complexe, du type de celui de la mandibule de Laboissière *et al.* ou de celui de la langue de Payan, Perrier & Laboissière [1995], pour lesquels les problèmes de coordination entre degrés de liberté doivent être maîtrisés.

# CONCLUSION

*Il suffit de passer le pont,*
*C'est tout de suite l'aventure !*

Georges Brassens

# 1. Cibles posturales + allure dynamique = communication parlée

Les résultats de notre étude indiquent que cibles abstraites invariantes et variabilités physiques peuvent être conciliées. Nous proposons que le locuteur vise des cibles articulatoires discrètes, reliées à des segments phonologiques, et projections dans l'espace proximal du locuteur de la cible ultime et essentielle de la parole, la cible perceptive. Nous avons montré aux chapitres III et IV qu'en mettant en œuvre des cibles invariantes, ou tout au moins des représentants de ces cibles, en termes de postures d'équilibre des articulateurs, légitimées par l'hypothèse neurophysiologique du Point d'Équilibre, on pouvait simuler des actes de production de la parole fort variés dans leurs réalisations articulatoires et acoustiques. Nous avons ainsi désigné un petit nombre de paramètres permettant d'ajuster ces réalisations au profit de l'auditeur ou du locuteur.

Pour confronter toujours plus nos hypothèses à la réalité de la production de la parole, nous avons effectué des tests préliminaires avec un modèle biomécanique de l'ensemble mandibule-os hyoïde, où propriétés dynamiques et contrôle moteur sont explicitement pris en compte séparément. Nous avons alors été confrontés au problème du contrôle coordonné des deux degrés de liberté en jeu, pour lequel l'absence actuelle de propositions enlève beaucoup de crédit aux conclusions que l'on pourrait tirer de nos simulations avec ce modèle. Ce semi-échec nous a confortés dans notre choix d'exploiter un modèle dynamique fonctionnel sur chacun des degrés de liberté du modèle articulatoire de Maeda.

La notion de cibles invariantes secondées par des paramètres variables s'inscrit dans un schéma de la communication parlée où l'invariance absolue est incongrue et la variabilité assumée.

Dans ce schéma, côté production, le message vocalique invariant serait codé par des cibles posturales abstraites et discrètes, images des segments phonologiques. Ces cibles abstraites seraient incarnées dans la mise en mouvement de l'appareil vocal, véhiculant une information concourante d'intonation, de flexion. Deux vecteurs de commandes régiraient ainsi la production de voyelle : *la posture* et *l'allure*.

Côté perception, si l'on se place dans un cadre d'inversion (Marr & Poggio), on peut imaginer que les cibles —en leur incarnation dynamique— que viserait le locuteur, seraient les objets perceptifs que récupérerait l'auditeur.

Dans la lignée de Lindblom, nous considérons que la communication parlée est négociation auditeur/locuteur, tractation sur l'objet produit/perçu. Le locuteur joue du vecteur d'allure pour façonner l'objet de la communication de la manière qu'il le souhaite, en fonction de son auditeur. Cet objet est une forme, au sens gestaltiste, à géométrie

variable. Le locuteur peut observer un principe d'économie de l'effort (la *lex facilitatis* de Hellwag), sacrifier à la bienséance (cf. le *stiff-upper-lip* très chic de l'aristocratie Britannique, qui trouble les canaux auditif et visuel), ou au contraire se contraindre à l'intelligibilité en affinant les contours de la forme transmise. L'auditeur, selon son degré d'attention (la *lex facilitatis* encore), d'intérêt ou de connivence, récupère plus ou moins complètement l'objet communiqué, dont il tire des informations, en termes de posture et d'allure, plus ou moins précises et exactes. La communication parlée est donc adaptative, flexible, régulée par le locuteur et l'auditeur.

# 2. Perspectives

Notre robot parlant est doué de flexibilité. Mais ses lacunes sont importantes et il convient maintenant d'esquisser quelques pistes pour les combler.

**La coordination articulatoire**

Dans le schéma de contrôle que nous avons proposé, le contrôle de la dynamique des articulateurs est considéré de façon individuelle. Dans la mise en œuvre des chapitres III et IV, nous avons délibérément choisi une séquence pour laquelle il est envisageable de privilégier le contrôle d'un seul articulateur. Mais il est bien évident que des synergies ont lieu, et que, puisque les locuteurs sont capables de réorganisation articulatoire (cf. les expériences de tube labial de Savariaux *et al.* [1995]), il existe un contrôle de la coopération inter-articulateurs. Les progrès en modélisation biomécanique de l'ensemble mandibulo-hyoïdien (Laboissière *et al.* [1996]) ou de la langue (Payan, Perrier & Laboissière [1995]), réalisés à l'Institut de la Communication Parlée en collaboration avec l'Université McGill de Montréal, vont permettre de mieux identifier les degrés de liberté de chaque articulateur, ainsi que le type de contrôle qui régit les mouvements selon ces degrés de liberté. Il faudra par ailleurs exploiter et développer des modèles de contrôle de la coordination temporelle selon ces degrés de liberté (cf. Morasso & Sanguinetti [1994], Laboissière, Sanguinetti & Payan [1995]).

**Stratégies des locuteurs**

Afin de clarifier les notions sur les tâches assignées à chaque paramètre et sur les types de stratégies utilisées par les locuteurs, il est essentiel de songer à étudier un plus grand nombre de locuteurs, sur une plus grande variété de conditions d'élocution. Il nous paraît plus efficace aussi, dans l'inférence de paramètres dynamiques, de travailler sur de véritables données articulatoires, et non pas sur des trajectoires inférées à partir de données acoustiques. Les travaux présentés au paragraphe 4.6 vont dans ce sens. L'électro-

magnétomètre, dont dispose maintenant l'ICP, devrait permettre d'étudier plus directement les paramètres dynamiques des mouvements du corps de la langue.

**L'optimisation**

Mais l'inférence de paramètres dynamiques, de relations temporelles inter-articulateurs, si elle doit être élargie à plus de locuteurs, de conditions d'élocution, d'articulateurs, va nécessiter le développement de procédures d'optimisation plus efficaces que celles que nous avons proposées ici. Dans le cas de l'utilisation du modèle simple du second ordre, il faudra, par exemple, songer à mieux contraindre le problème pour réduire le nombre de variables à optimiser ; dans le cas de la mise en œuvre d'un modèle biomécanique, il faudra commencer par rechercher au niveau du tronc cérébral les inhibitions ou les synergies entre les différents muscles contrôlant les degrés de liberté, pour contraindre l'inversion, rendue extrêmement délicate par l'ensemble des possibilités théoriques de compensation entre muscles et/ou degrés de liberté. Il sera alors possible d'envisager l'inférence des commandes centrales régissant le contrôle de la rotation *et* de la translation dans le modèle mandibulo-hyoïdien et de conclure plus nettement sur l'identité des cibles en contextes accentué *vs* non-accentué.

**Relations production/perception**

Afin de compléter la mise en œuvre du schéma de contrôle du robot parlant que nous avons proposé au chapitre II et de mieux décrire, par exemple, le *feedback* auditif, il faut prévoir d'explorer plus en profondeur les relations production/perception. Les tests perceptifs menés au chapitre IV nous ont confortés dans l'idée que certaines variables du contrôle de la production jouent un rôle perceptif bien précis. Des expériences perceptives plus systématiques pourraient compléter nos premières hypothèses sur les stratégies des locuteurs en désignant plus nettement les effets, chez l'auditeur, de certaines manipulations paramétriques. Le robot parlant n'a pas fini de balbutier...

# RÉFÉRENCES BIBLIOGRAPHIQUES

Abraham R.H. & Shaw C.D. (1982). *Dynamics—The Geometry of Behavior.* Aerial, Santa Cruz, CA.

Abry C. & Lallouache T.M. (1996). Le MEM: un modèle d'anticipation paramétrable par le locuteur. Données sur l'arrondissement du Français. *Bulletin de la Communication Parlée*, 3, Grenoble, France: Institut de la Communication Parlée.

Abry C., Benoit C., Boë L.J. & Sock R. (1985). Un choix d'événement pour l'organisation temporelle du signal de parole. *Actes des 14èmes Journées d'Étude sur la Parole,* Paris, 133-137.

Abry C., Perrier P. & Jomaa M. (1990). Premières modélisations sur le timing des pics de vitesse de la mandibule. *Actes des 18èmes Journées d'Étude sur la Parole.* Montréal (Québec), 99-102.

Akagi M. (1990). Evaluation of a spectrum target prediction model in speech perception. *J. Acoust. Soc. Am.*, 87, 2, 858-865.

Akagi M. (1993). Modeling of contextual effects based on spectral peak interaction. *J. Acoust. Soc. Am.*, 93, 2, 1076-1086.

Ainsworth W. (1975). Intrinsic and extrinsic factors in vowel judgments. In *Auditory analysis and perception of speech*, 103-113. Fant G. & Tatham M. (Eds.). Academic London.

Andruski J.E. & Nearey T.M. (1992). On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables. *J. Acoust. Soc. Am.*, 91 (1), 890-410.

Apostol L. (1994). *Étude et validation d'une procédure de normalisation par prise en compte des contraintes de production.* Rapport de Stage Ingénieur. Institut de la Communication Parlée. INP Grenoble. France.

Asatryan D.G. & Feldman A.G. (1965). Functional tuning of the nervous system with control of movement or maintenance of a steady posture. 1. Mechanographic analysis of the work of the joint on execution of a postural task. *Biophysics*, 10, 925-935.

Assman P.F., Nearey T.M. & Hogan T. (1982). Vowel identification: orthographic, perceptual, and acoustic aspects. *J. Acoust. Soc. Am.*, 71 (4), 975-989.

Atal B.S., Chang J.J., Mathews M.V. & Tukey J.W. (1978). Inversion of Articulatory-to-Acoustic Transformation in the Vocal Tract by a Computer Sorting Technique. *J. Acoust. Soc. Am.*, 63, 1535-1555.

Atkeson C.G. & Hollerbach J.M. (1985). Kinematic features of unrestrained vertical arm movements. *J. Neuroscience*, 5, 2318-2330.

Badin P. & Fant G. (1984). Notes on vocal tract computations, *STL QPSR, 2-3*, 53-108.

Badin P., Boë L.J., Perrier P. & Abry C. (1988). Vocalic nomograms: acoustic considerations upon formant convergence. *Bull. L.C.P. (2)*, 65-94.

Bailey P.J., Bevan K. & Burr T. (1995). Effects of formant frequency modulation on vowel identification. *Actes du 13ème Congrès International des Sciences Phonétiques*, Vol. 2, 682-685. Stockholm, Suède.

Bailly G. & Laboissière R. (1993). Learning coarticulation and compensation for selected VV with a control model. *2ème rapport annuel du projet ESPRIT-BR n°6975 SPEECH MAPS* (WP3. Deliverable 9).

Bailly G., Jordan M., Mantakas M., Schwartz J.L., Bach M. & Olesen M., (1990). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique. *J. Acoust. Soc. Am.* 87, S1, S105.

Beautemps D. (1993). *Récupération des gestes de la parole à partir de trajectoires formantiques: identification de cibles vocaliques non-atteintes et modèles pour les profils sagittaux des consonnes fricatives.* Thèse de l'Institut National Polytechnique de Grenoble, France.

Bell A.M. (1867). *Visible Speech: the Science of Universal Alphabetics*, London (Simpkin, Marshall & Co.).

Bennett S. & Weinberg B. (1979). Acoustic correlates of perceived sexual identity in preadolescents children's voices. *J. Acoust. Soc. Am.*, 66, 989-1000.

Bernstein N. (1967). *The co-ordination and regulation of movements.* Oxford: Pergamon Press.

Bizzi E., Hogan N., Mussa-Ivaldi F.A. & Giszter S. (1992). Does the nervous system use the equilibrium-point control to guide single and multiple joint movements? *Behavioral and brain sciences* 15, 603-613.

Bladon R.A.W., Henton C.G., Pickering J.B. (1984). Outline of an auditory theory of speaker normalization: In M.P.R. Van den Broecke & A. Cohen (Eds.), *Proceedings of the Tenth International Congress of Phonetic Sciences.* Dordrecht, Holland: Foris Publications.

Blumstein S.E. (1986). On acoustic invariance in speech. In Perkell J.S. & Klatt D.H. (Eds.), *Invariance and Variability in Speech Processes,* Ch.9, 178-193. Hillsdale N.J.: Lawrence Erlbaum.

Blumstein S.E. (1989). Theoretical implications of the quantal nature of speech: a commentary. *J. Phonetics*, 17 (1), 55-61.

Blumstein S.E. & Stevens K.N. (1979). Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants. *J. Acoust. Soc. Am.*, 66 (4), 1001-1017.

Blumstein S.E. & Stevens K.N. (1980). Perceptual invariance and onset spectra for stop consonants in various vowel environments. *J. Acoust. Soc. Am.*, 67, 648-662.

Blumstein S.E. & Stevens K.N. (1981). Phonetic features and acoustic invariance in speech. *Cognition*, 10, 25-32.

Boë L.J. (1993). Speech Maps Interactive Plant "SMIP". *1er rapport annuel du projet ESPRIT-BR n°6975 SPEECH MAPS* (Volume III: From speech signal to vocal tract geometry).

Boë L.J. (1994). Codebook and vowel prototypes of the vocal tract plant. *2ème rapport annuel du projet ESPRIT-BR n°6975 SPEECH MAPS* (Volume III: From speech signal to vocal tract geometry).

Boë L.J. & Abry C. (1986). Nomogrammes et systèmes vocaliques. *Actes des 15èmes Journées d'Étude sur la Parole*, 303-306. Aix en Provence, France.

Boë L.J. & Perrier P. (1988). C.F. Hellwag 200 ans après ou les éléments d'une fibre conductrice. *Actes des 17èmes Journées d'Étude sur la Parole*, 200-205. Nancy, France.

Boë L.J., Gabioud B. & Perrier P. (1996). SMIP : Speech Maps Interactive Plant. *Bulletin de la Communication Parlée*, 3, Grenoble, France: Institut de la Communication Parlée.

Boë L.J., Gabioud B., Perrier P., Schwartz J.-L. & Vallée N. (1995). Vers une unification des espaces vocaliques. In *Levels in Speech Communication: Relations and Interactions,* 63-71. C. Sorin, J. Mariani, H. Meloni & J. Schoentgen (Eds). Elsevier Science B.V..

Bothorel A., Simon P., Wioland F. & Zerling J.-P. (1986). *Cinéradiographie des voyelles et des consonnes du français.* Recueil de documents synchronisés pour quatre sujets: vues latérales du conduit vocal, vues frontales de l'orifice labial, données acoustiques. Institut de Phonétique, Strasbourg.

Broad D.J. & Clermont F. (1987). A methodology for modeling vowel formant contours in CVC context. *J. Acoust. Soc. Am.*, 81 (1), 155-165.

Browman C.P. & Goldstein L.M. (1985). Dynamic modeling of phonetic structure. In V. Fromkin (Ed.). *Phonetic linguistics.* New York: Academic.

Browman C.P. & Goldstein L.M. (1986). Towards an articulatory phonology. *Phonology Yearbook*, 3, 219-252.

Browman C.P. & Goldstein L.M. (1987). Tiers in articulatory phonology, with some implications for casual speech. *Haskins Laboratories Status report on Speech Research.* SR-92, 1-30.

Browman C.P. & Goldstein L.M. (1990). Gestural Specification Using Dynamically-Defined Articulatory Structures. *J. Phonetics, 18*, 299-320.

Browman C.P. & Goldstein L.M. (1992). Articulatory phonology: an overview. *Haskins Lab. Status report.* SR 111/112, 23-42.

Broyden C.G. (1970). The convergence of a class of double-rank minimization algorithms. *J. Inst. of Mathematics and its applic.*, 6, 76-90.

Catford J.C. (1981). Observations on the recent history of vowel classification. In R.E. Asher & E.J.A. Henderson: *Towards a history of phonetics*. University press, Edinburgh, 19-31.

Chistovich L.A. & Lublinskaya V.V. (1979). The 'Center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research*, 1, 185-195.

Cipra B. (1992). You can't hear the shape of the drum. *Science*, 255, 1642-1643.

Cooke J.D. (1980). The Organization of Simple, Skilled Movements. In G.E. Stelmach & J. Requin (Eds.), *Tutorials in Motor Behavior*, 199-212. Amsterdam, The Netherlands: Elsevier Science Publishers B.V. (North-Holland).

Cooper F.S., Liberman A.M. & Borst J.M. (1951). The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proc. Natl. Acad. Sci.* 37, 318-325.

Cooper F.S., Delattre P.C., Liberman A.M., Borst J.M. & Gerstman L.J. (1952). Some experiments on the perception of synthetic speech sounds. *J. Acoust. Soc. Am.*, 24 (6), 597-606.

Daniloff R.G. & Hammarberg R.E. (1973). On defining coarticulation. *J. Phonetics*, 1, 239-248.

Davalo E. & Naïm P. (1990). *Des réseaux de neurones*. Eyrolles. Paris.

Delattre P. (1948). Un triangle acoustique des voyelles orales du français. *The French Review*, XXI, 6.

Delattre P. (1969). The general phonetic characterisitics of languages. An acoustic and articulatory study of vowel reduction in four languages. *Final report, University of California*, Santa Barbara, CA, USA.

Delattre P.C., Liberman A.M. & Cooper F.S. (1955). Acoustic loci and transitional cues for consonants. J. *Acoust. Soc. Am.*, 27,4, 769-773.

Di Benedetto M.-G. (1989a). Frequency and time variations of the first formant: properties relevant to the perception of vowel height. *J. Acoust. Soc. Am.*, 86 (1), 67-77.

Di Benedetto M.-G. (1989b). Vowel representation: some observations on temporal and spectral properties of the first formant frequency. *J. Acoust. Soc. Am.*, 86 (1), 55-66.

Diehl R.L. & Kluender K.R. (1989). On the objects of speech perception. *Ecol. Psychol.*, 1, 121-144.

Diehl R.L. & Walsh M.A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *J. Acoust. Soc. Am.*, 85, 2154-2164.

Disner S.F. (1980). Evaluation of vowel normalisation procedures. *J. Acoust. Soc. Am.*, 67, 253-261.

Dworkin J.P., Aronson A.E. & Mulder D.W. (1980). Tongue force in normals and in dysarthric patients with amyotrophic lateral sclerosis. *J. Speech and Hearing Research,* 828-837.

Easton T.A. (1972). On the normal use of reflexes, *American Scientist*, 60, 591-599.

Engstrand O. (1988). Articulatory correlates of stress and speaking rate in swedish VCV utterances. *J. Acoust. Soc. Am.* 83, 1863-1875.

Fabre J.P. (1988). Peut-on entendre la forme d'un tambour? *La Recherche*, 202, 1104-1106.

Fairbanks G. & Grubb P. (1961). A psychophysical investigation of vowel formants. *J. Speech and Hearing Research*, 4, 203-219.

Fant G. (1960). On the predictability of formant levels and spectrum envelope from formant frequencies. *Acoustic Theory of Speech Production* . 216-228. The Hague: Mouton.

Feldman A.G. (1966). Functional Tuning of The Nervous System with Control of Movement or Maintenance of a Steady Posture – II Controllable Parameters of the Muscles. *Biophysics, 11*, 565-578.

Feldman A.G. (1981). The composition of central programs subserving horizontal eye movements in man. *Biological Cybernetics*, 42, 107-116.

Feldman A.G. (1986). Once more on the Equilibrium-Point hypothesis ($\lambda$ model) for motor control. *Journal of Motor Behavior, Vol. 18, 1*, 17-54.

Feldman A.G. & Levin M.F. (1995). The origin and use of positional frame of reference in motor control. *Behavioral & Brain Sciences*, 723-806.

Feldman A.G., Adamovich S.V., Ostry D.J. & Flanagan J.R. (1990). The Origins of Electromyograms - Explanations Based on the Equilibrium Point Hypothesis. In J.W. Winters & Woo S.L.Y. (Eds.), *Multiple Muscle Systems: Biomechanics and Movement Organization* (Section III - Chapter 10). Berlin, Germany: Springer Verlag.

Flanagan J.R., Feldman A.G. & Ostry D.J. (1992). Equilibrium trajectories underlying rapid target-directed arm movements. In G.E. Stelmach & J.R. Requin (Eds.), *Tutorials in motor behavior II.* Amsterdam, The Netherlands: North Holland.

Flanagan J.R., Ostry D.J. & Feldman A.G. (1990). Control of human jaw and multi-joint arm movements. In G.E. Hammond (Ed.), *Cerebral control of speech and limb*

*movements*, 29-58. Amsterdam, The Netherlands: Elsevier Science Publishers B.V. (North-Holland).

Flanagan J.R., Ostry D.J. & Feldman A.G. (1993). Control of trajectory modifications in target-directed reaching. *J. Motor Behavior*, Vol. 25, 3, 140-152.

Flash T. (1987). The control of hand equilibrium trajectories in multi-joint arm movements. *Biol. Cybern.*, 57, 257-274.

Flash T. & Hogan N. (1985). The coordination of arm movements; an experimentally confirmed mathematical model. *J. Neurosci* 5: 1688-1703.

Fletcher R. (1970). A New Approach to Variable Metric Algorithms. *TheComputer Journal, 13*, 317-322.

Fletcher R. (1987). *Practical methods of optimization.* John Wiley & Sons. Chichester.

Fodor J. (1983). *The modularity of mind.* Cambridge, MA: MIT Press.

Fowler C.A. (1980). Coarticulation and theories of extrinsic timing. *J. Phonetics*, 8, 113-133.

Fowler C.A. (1986). An event approach of the study of speech perception from a direct-realist perspective. *J. Phonetics*, *14*, 3-28.

Fujimura O. (1986). Relative Invariance of Articulatory Movements : An Iceberg Model. In J.S. Perkell & D.H. Klatt (Eds), *Invariance & Variability in speech processes*, 226-234. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Fujimura O. (1987). Fundamentals and applications in speech production research. *Actes du 11ème Congrès International des Sciences Phonétiques*, 10-27. Tallinn: Academy of Sciences of the Estonian SSR.

Fujimura O. (1991). Beyond the segment. In Mattingly I.G. & Studdert-Kennedy M. (Eds.), *Modularity and the Motor Theory of Speech Perception*, 25-31. Hillsdale, NJ: Lawrence Erlbaum.

Fujimura O. & Ochiai K. (1963). Vowel identification and phonetic contexts. *J. Acoust. Soc. Am.*, 35, 1889 (A).

Fujisaki H. & Kawashima T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics*, AU-16, 73-77.

Gaines (1969). Adaptive control theory (the structural and behavioural properties of adaptive controllers). *Encyclopædia of linguistics. Information and control,* 1-9. Ed. : A.R. Meetham. Pergamon Press. Oxford.

Gay T. (1978). Effects of speaking rate on vowel formant movements. *J. Acoust. Soc. Am.* Vol. 63, N°1, 223-230.

Gerstman L. (1968). Classification of self-normalised vowels. *IEEE Trans. Audio Electroacoust.* AU-16, 78-80.

Gibson J.J. (1966). *The sense considered as perceptual systems.* Boston: Houghton Mifflin.

Gleason H. A. (1955). *An introduction to descriptive linguistics.* New-York : Holt, Rinehart & Winston eds.

Goldfarb D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24, 109, 23-26.

Gottfried T.L. & Strange W. (1978). Identification of vowels in velar consonant contexts. *J. Acoust. Soc. Am.*, 63, S1, S4.

Gracco V.L. (1994). Some organizational characteristics of speech movement control. *J. Speech and Hearing Research*, 37, 4-27.

Hadamard J. (1923). *Lectures on the Cauchy problem in linear partial differential equations.* Yale University Press, New Haven.

Halle M. & Stevens K.N. (1964). Speech recognition: a model and a program for research. In J.A. Fodor & J.J. Katz (Eds.), *The structure of language.* Englewood Cliffs, N.J.: Prentice-Hall, 1964.

Hardcastle W.J. (1985). Some phonetic and syntactic constraints on lingual coarticulation during /kl/ sequences. *Speech Comm.*, 4, 247-263.

Harrington J. & Cassidy S. (1994). Dynamic and target theories of vowel classification: evidence from monophtongs and diphtongs in Australian English. *Language & Speech*, 37, 4, 357-373.

Harris K.S. (1975). Mechanisms of duration change. In *Proceedings of Speech and Communication Seminar*, Fant G. (Ed.), Almqvist & Wiksells, Stockholm, 299-305.

Harshman R. (1970). Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis. *Working Papers in Phonetics* 16, Phonetics Lab., UCLA.

Heinz J.M. & Stevens K.N. (1965). On the relations between lateral cineradiographs, area functions, and acoustic spectra of speech. *Proceedings of the 5th Int. Congr. of Acoustics*, A44.

Hellwag C.F. (1991). De Formatione Loquelæ. *Bulletin de la Communication Parlée*, 1, 26-105, Grenoble, France: Institut de la Communication Parlée (fac-similé et traduction française de la dissertation soutenue en 1781 à la faculté de Tübingen).

Hillenbrand J., Getty L.A., Clark M.J., Wheeler K. (1995). Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.*, 97, 5, 3099-3111.

Hirayama M., Vatikiotis-Bateson E., Kawato M. & Honda K. (1992). Neural network modeling of speech motor control. In *Proceedings of the International Conference on Spoken Language Processing*, Banff, 883-886.

Hirayama M., Vatikiotis-Bateson E., Kawato M. & Jordan M.I. (1992). Forward dynamics modeling of speech motor control using physiological data. In R.P. Lippman, J.E. Moody & D.S. Touretsky (Eds.). *Advances in neural information processing systems 4*, 191-198. San Mateo, CA: Morgan Kauffmann Publishers.

Hockett C.F. (1965). *A course in modern linguistics.* The Macmillan Company, New York.

Hogan N. (1984). An organising principle for a class of voluntary movements. *J. Neuroscience*, 4, 11, 1745-2754.

Hollerbach J.M. & Flash T. (1982). Dynamic interactions between limb segments during planar arm movement. *Biol. Cybern.*, 44, 67-77.

Holmes J. (1986). Normalisation in vowel perception. In Perkell J.S. & Klatt D.H. (Eds.), *Invariance and Variability in Speech Processes,* 346-357. Hillsdale N.J.: Lawrence Erlbaum.

Houdas Y. (1990). *Physiologie cardio-vasculaire.* Ed. Vigot. Paris.

Houde R.A. (1968). A study of tongue body motion during selected speech sounds. *SCRL Monograph,* No. 2. Santa Barbara: Speech Communications Research Laboratory.

Huang C.B. (1985). *Perceptual correlates of the tense/lax distinction in general American English.* Master's Thesis, MIT, Cambridge, MA.

Huang C.B. (1992). Modelling human vowel identification using aspects of formant trajectory and context. In Y. Tohkura, E. Vatikiotis-Bateson & Y. Sagisaka (Eds.). *Speech Perception, Production and Linguistic Structure*, 43-61. Amsterdam: IOS Press.

Jakobson R., Fant G. & Halle M. (1963). *Preliminaries to Speech Analysis.* Cambrige, MA: MIT Press.

Jenkins J.J. & Strange W. (1987). Identification of 'hybrid' vowels in sentence context. *J. Acoust. Soc. Am.*, 82, S1, S82.

Jones D. (1922). *An outline of English Phonetics*, 2ème éd., Leipzig (Teubner). 6ème éd. New-York : E.P. Dutton (1940).

Joos M. (1948). *Acoustic Phonetics*, Language Monograph no. 23, Linguistic Society of America, Baltimore.

Jordan M.I. (1986). *Serial order: a parallel distributed processing approach.* (Tech. Rep. No. 8604). San Diego: University of California, Institute for Cognitive Science.

Jordan M.I. (1988). Supervised learning and systems with degrees of freedom. *COINS Technical Report 88-27*, 1-41. University of Massachussetts, Amherst, MA.

Jordan M.I. (1989). Indeterminate motor skill learning problems. In M. Jeannerod (Ed.). *Attention and Performance, XIII.* MIT Press.

Jordan M.I. (1990). Motor Learning and the Degrees of Freedom Problem. In M. Jeannerod (Ed.), *Attention and Performance* (Chapter XIII). Hillsdale, NJ: Erlbaum.

Kac M. (1966). Can one hear the shape of the drum? *American Math. Monthly*. 73 (4), part II, 1-23.

Katayama M. & Kawato M. (1993). Virtual trajectories and stiffness ellipse during multijoint arm movement predicted by neural inverse models. *Biol. Cybern.*, 69, 353-362.

Kawato M., Furukawa K. & Suzuki R. (1987). A hierarchical neural network model for control and learning of voluntary movement. *Biol. Cybern.*, 57, 169-185.

Kawato M., Maeda Y., Uno Y. & Suzuki R. (1990). Trajectory formation of arm movement by cascade neural network model based on minimum torque-change criterion. *Biol. Cybern.*, 62, 275-288.

Keating P.A. (1988). The window model of coarticulation: articulatory evidence. *UCLA Working Papers in Phonetics, 69*, 3-29. Los Angeles : University of California.

Kelly J.L. & Lochbaum C.L. (1962). Speech synthesis. *Proceedings Stockholm Speech Communication Seminar, RIT,* 127-130.

Kelso J.A.S., Saltzman E. & Tuller B. (1986). The Dynamical Perspective on Speech Production: Data and Theory. *J. Phonetics, 14*, 29-59.

Kelso J.A.S, Vatikiotis-Bateson E, Saltzman E.L. & Kay B. (1985). A qualitative dynamic analysis of reiterant speech production: phase portraits, kinematics, and dynamic modeling. *J. Acous. Soc. Am. 77 (1)*, 266-280.

Kluender K.R. (1991). Effects of first formant onset properties on voicing judgments result from processes not specific to humans. *J. Acoust. Soc. Am.* 90 (1), 83-96.

Kluender K.R., Diehl R.L. & Wright B.A. (1988). Vowel-length differences before voiced and voiceless consonants: an auditory explanation. *J. Phonetics*, 16, 153-169.

Koopmans-Van Beinum F.J. (1980). *Vowel contrast reduction: an acoustic and perceptual study of Dutch in various speech conditions*, Doctoral dissertation, University of Amsterdam.

Kröger B. J. (1993). A gestural production model and its application to reduction in German. *Phonetica 50*, 213-233.

Kuehn D.P & Moll K.L. (1976). A cineradiographic study of VC & CV articulatory velocities. *J. Phonetics*, Vol. 4, 303-320.

Kuhl P. & Padden D.M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *J. Acoust. Soc. Am.* 73 (3), 1003-1010.

Kuwabara H. (1985). An approach to normalization of coarticulation effects for vowels in connected speech. *J. Acoust. Soc. Am., 77 (2)*, 686-694.

Laboissière R. (1992). *Préliminaires pour une robotique de la communication parlée : Inversion et contrôle d'un modèle articulatoire.* Thèse de l'Institut National Polytechnique de Grenoble, France.

Laboissière R., Ostry D.J. & Feldman A.G. (1996). The control of multi-muscle systems: human jaw and hyoid movements. *Biol. Cybern.*, 74, 373-384.

Laboissière R., Sanguinetti V. & Payan Y. (1995). On the biomechanical control variables of the tongue during speech movements. In *Proceedings of the 4th European Conference on Speech Communication and Technology*, 1289-1292. Madrid, Espagne. ESCA.

Laboissière R., Schwartz J.L. & Bailly G. (1990). Motor Control for Speech Skills: A Connectionnist Approach. In D.E. Touretsky, J.L. Elman, T.J. Sejnowski & G.E. Hinton (Eds.), *Proceedings of 1990 Connectionnist Models SUMMER School,* 319-327. San Mateo, CA: Morgan Kaufmann.

Lacquaniti F., Licata F. & Soeching J.F. (1982). The mechanical behavior of the human forearm in response to transient perturbations. *Biol. Cybern.* 44, 35-46.

Ladefoged P. (1967). Linguistic phonetics. *Working papers in Phonetics*, 6, Phonetics Laboratory. Los Angeles: University of California.

Ladefoged P. & Broadbent D. (1957). Information conveyed by vowels. *J. Acoust. Soc. Am.*, 29, 98-104.

Lehiste I. & Peterson G.E. (1961). Transitions, glides, and diphtongs. *J. Acoust. Soc. Am.*, 33, 268-277.

Liberman A.M. & Mattingly I.G. (1985). The motor theory of speech production revised. *Cognition, 21*, 1-36.

Liberman A.M., Cooper F., Shankweiler D. & Studdert-Kennedy M. (1967). Perception of the Speech Code. *Psychol. Review*, 74, 431-461.

Lindblom B. (1963). Spectrographic study of vowel reduction. *J. Acoust. Soc. Am., 35,* 1773-1781.

Lindblom B. (1968). *On the production and recognition of vowels*, summary of Doct thesis, University of Lund.

Lindblom B. (1988). Phonetic Invariance and the Adaptive Nature of Speech. In *Working Models of Human Perception*. London, UK: Academic Press.

Lindblom B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W.J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling*, 403-439. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Lindblom B. (1995). Approche intégrée de la production et de la perception. *Actes de l'École thématique: Fondements et Perspectives en Traitement Automatique de la Parole.* Marseille-Luminy juillet 1995. Ed : H. Méloni.

Lindblom B. & Studdert-Kennedy M. (1967). On the role of formant transitions in vowel recognition. *J. Acoust. Soc. Am., 42 (4),* 830-843.

Lindblom B., Brownlee S., Davis B. & Moon S.-J. (1992). Speech transforms. *Speech Comm.*, 11, 357-368.

Lloyd R.J. (1890). *Some researches into the nature of vowel-sound.* Turner & Dunnett, Liverpool, England.

Lobanov B.M. (1971). Classification of Russian vowels spoken by different speakers. *J. Acoust. Soc. Am* 49, 606-608.

Lœvenbruck H. (1992). *Contrôle d'un robot parlant : réseaux neuromimétiques et modèles dynamiques.* Rapport de DEA. Institut de la Communication Parlée. INP Grenoble.

Macchi M.J. (1980). Identification of vowels spoken in isolation versus vowels spoken in consonantal context. *J. Acoust. Soc. Am* 68 (6), 1636-1642.

MacNeilage P. (1970). Motor control of serial ordering of speech. *Psychological Review, 77,* 182-196.

MacNeilage P. (1980). Distinctive properties of speech motor control. *Tutorials in Motor behavior.* GE Stelmachs and J. Requin (eds). North-Holland Publishing Company.

Mac Kay W.A., Crammond D.J., Kwan H.C. & Murphy J.T (1986). Measurements of human forearm viscosity. *J. Biomech.* 19 (3), 231-238.

Maeda S. (1979). Un modèle articulatoire de la langue avec composantes linéaires. *Actes des 10èmes Journées d'Étude sur la Parole*, GALF, 152-164.

Maeda S. (1990). Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W.J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling*, 131-149. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Maeda S. & Honda K. (1994). From EMG to formant patterns of vowels: the implication of vowel systems spaces. *Phonetica* 51, 17-29.

Maeda S., Honda K. & Kusawaka N. (1993). From EMG to Vowel Formant Patterns. *3rd Seminar on Speech Production: Models and Data.* Old Saybrook, CT., 11-13 May.

Marr D. (1982). *Vision.* San Francisco: W.H. Freeman and Company.

Marr D. & Poggio T. (1979). A computational theory of human stereo vision. In *Proceedings of the Royal Society of London*, B., 204, 301-328.

Martinet A. (1970). *La linguistique synchronique. Études et recherche.* Presses Universitaires de France. Paris.

Mermelstein P. (1967). Determination of vocal tract shape from measured formant frequencies. *J.Acoust. Soc. Am.*, 41 (5), 1283-1967.

Miller J.D. (1984). Auditory perceptual correlates of the vowel. *J. Acoust. Soc. Am.* Suppl. 1, 76, S79 (A).

Miller J.D. (1989). Auditory-perceptual interpretation of the vowel. *J. Acoust. Soc. Am.*, 85 (5), 2114-2134.

Miller J.L. & Liberman A.M. (1979). Some effects of later-occurring information on the perception of stop-consonant and semivowel. *Percept. Psychophys.*, 25, 457-465.

Miller R.L. (1953). Auditory tests with synthetic vowels. *J. Acoust. Soc. Am.*, 25, 114-121.

Minoux M. (1983). *Programmation mathématique. Théorie et algorithmes.* Tome 1, Dunod Collection Technique et Scientifique des Télécommunications.

Monin M.-P. (1991). Introduction à *De Formatione Loquelæ. Bulletin de la Communication Parlée*, 1, 15-25, Grenoble, France: Institut de la Communication Parlée.

Moon S.-J. (1991). *An acoustic and perceptual study of undershoot in clear and citation-form speech.* Doctoral dissertation, University of Texas at Austin.

Morasso P. & Sanguinetti V. (1994). Representation of space and time in motor control. *Rapport annuel du projet ESPRIT-BR n°6975 SPEECH MAPS* (WP3).

Morin E. (1984). *Sociologie.* Ed. Fayard. Paris.

Morris A.C. (1990). *The use of non-linear net-input function MLP for learning the Maeda model function,* Internal technical report. Grenoble, France: Institut de la Communication Parlée, Institut National Polytechnique.

Morrish E.C.E. (1988). Compensatory articulation in subject with total glossectomy. *British Journal of Disorders of Communication*, 23, 13-22.

Nearey T.M. (1977). *Phonetic feature systems for vowels.* Unpublished doctoral dissertation, University of Connecticut, Storrs, CT.

Nearey T.M. (1978). *Phonetic Feature Systems for Vowels.* Indiana University Linguistics Club, Bloomington, IN.

Nearey T.M. (1989). Static, dynamic, and relational properties in vowel perception. *J. Acoust. Soc. Am.*, 85, 5, 2088-2113.

Nearey T.M. (1995). Evidence for the perceptual relevance of vowel-inherent spectral change for front vowels in Canadian English. *Actes du 13ème Congrès International des Sciences Phonétiques*, Vol. 2, 678-681. Stockholm, Suède.

Nearey T.M. & Assmann P.F. (1986). Modeling the role of inherent spectral change in vowel identification. *J. Acoust. Soc. Am.*, 80, 5, 1297-1308.

Nelson W.L. (1983). Physical principles for economies of skilled movements. *Biological Cybernetics, 46,* 135-147.

Nordström P.-E. & Lindblom B. (1975). A normalization procedure for vowel formant data. *Proceedings of the 8th International Congress of Phonetic Sciences,* Leeds, England.

Nord L. (1975). Vowel reduction - centralization or contextual assimimlation? In *Proceedings of Speech and Communication Seminar*, Fant G. (Ed.), Almqvist & Wiksells, Stockholm, 149-154

Nord L. (1986). Acoustic studies of vowel reduction in Swedish. *STL-QPSR* 4, 19-36 (Department of Speech Communication, RIT, Stockholm).

Öhman S.E.G. (1967). Numerical Model of Coarticulation. *J. Acoust. Soc. Am., 41,* 310-320.

Ostry D.J. & Munhall K.G. (1985). Control of rate and duration of speech movements. *J. Acoust. Soc. Am., 77,* 640-648.

Ostry D.J., Keller E. & Parush A. (1983). Similarities in the control of the speech articulators and the limbs: Kinematics of the tongue dorsum movement in speech. *J. Experimental Psychology (Human perception and performance), 9,* 622-636.

Payan Y. & Perrier P. (1993). Vowel Normalization by Articulatory Normalization: First Attempt for Vocalic Transitions. In *Proceedings of the 3rd European Conference on Speech Communication and Technology*, 417-420. Berlin, Germany: ESCA.

Payan Y., Perrier P. & Laboissière R. (1995). Simulation of tongue shape variations in the sagittal plane based on a control by the Equilibrium-Point hypothesis. *Actes du 13ème Congrès International des Sciences Phonétiques* , 2, 474-477. Stockholm, Suède, Août 1995.

Pelorson X., Hirshberg A., Van Hassel R.R., Wijnands A.P.J., Auregan Y. (1994). Theoretical and experimental study of quasi-steady flow separation within glottis during phonation. Application to a modified two-mass model. *J. Acoust. Soc. Am.*, 96, 3416-3431.

Perkell J.S. & Klatt D.H. (1986). *Invariance & Variability in speech processes.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Perkell J.S. & Matthies M.L. (1992). Temporal measures of anticipatory labial coarticulation for the vowel [u]: within- and cross-subject variability. *J. Acoust. Soc. Am.*, 91, 2911-2925.

Perrier P. & Ostry D.J. (1994). Dynamic modelling and control of speech articulators: Application to vowel reduction. In *Fundamentals in speech synthesis and speech recognition.* E. Keller (ed.), 231-251, London UK: J. Wiley and Son.

Perrier P., Abry C. & Keller E. (1989). Vers une modélisation des mouvements du dos de la langue. *Journal d'Acoustique, 2,* 69-77.

Perrier P., Apostol L. & Payan Y. (1995). Evaluation of a vowel normalisation procedure based on speech production knowmledge. *Proceedings of the 4th European Conference on Speech Communication and Technology*, 1921-1924. Madrid, Spain : ESCA.

Perrier P., Boë, L.J. & Sock R. (1992). Vocal Tract Area Function Estimation From Midsagittal Dimensions With CT Scans and a Vocal Tract Cast: Modeling the Transition With Two Sets of Coefficients. *J. Speech and Hearing Research, 35,* 53-67.

Perrier P., Lœvenbruck H. & Payan Y. (1996). Control of tongue movements in speech: The Equilibrium Point hypothesis perspective. *J. Phonetics,* 24, 53-75.

Perrier P., Ostry D.J. & Laboissière R. (1996). The equilibrium Point Hypothesis and its application to speech motor control. *J. Speech and Hearing Research,* 39 (2), 365-377.

Peterson G.E. (1961). Parameters of vowel quality. *J. Speech and Hearing Research,* 4, 10-29.

Peterson G.E. & Barney H.L. (1952). Control methods used in a study of the vowels. *J. Acoust. Soc. Am.*, 24, 2, 175-184.

Pierre D.A. (1986). *Optimization Theory with Applications.* New-York: Dover Publications, Inc.

Pitermann M. (1996). *Évaluation expérimentale de la théorie des cibles formantiques dans le cadre de la production des voyelles. Comparaison entre la variabilité des formants et des cibles correspondantes en fonction de la vitesse d'élocution et de l'accent d'insistance.* Thèse de l'Université Libre de Bruxelles. Faculté des Sciences. Institut des langues vivantes et de phonétique.

Poggio T. (1984). Low-level vision as inverse optics. In *Computational models of Hearing and Vision*, Tallinn, 123-127.

Pols L.C.W. & Van Son R.J.J.H. (1993). Acoustics and perception of dynamic vowel segments. *Speech Comm.* 13, 135-147.

Potter R.K. & Steinberg J.C. (1950). Toward the specification of speech. J. *Acoust. Soc. Am.*, 22, 807-820.

Rakerd B., Verbrugge R.V. & Shankweiler D.P. (1980). Speaking rate, syllable stress and vowel identity. *Haskins Lab. Status Rep. Speech Res.*. SR-62, 149-160.

Rubin P., Baer T. & Mermelstein P. (1981). An articulatory synthesizer for perceptual research. *J. Acoust. Soc. Am.*, 70, 2, 321-328.

Rumelhart D.E., Hinton G.E. & Williams R.J. (1986). Learning Internal Representation by Error Propagation. In D.E. Rumelhart, J.L. McClelland (eds.), *Parallel Distributed Processing: Exploration in the Microstructure of Cognition*, 318-362. Cambridge, MA: MIT Press.

Ryalls J.H. & Lieberman P. (1982). Fundamental frequency and vowel perception. *J. Acoust. Soc. Am.*, 72, 1631-1634.

Saltzman E.L. (1986). Task dynamic coordination of the speech articulators: a preliminary model. *Experimental brain research, Series* 15, 129-144.

Saltzman E.L. & Kelso J.A.S. (1983). Toward a dynamical account of motor memory and control. In R. Magill (Eds.), *Memory and control of action*, 17-38. Amsterdam: North-Holland.

Saltzman E.L. & Munhall, K.G. (1989). A Dynamical Approach to Gesture Patterning in Speech Production. *Ecological Psychology, 1*, 1615-1623.

de Saussure F. (1916). *Cours de Linguistique Générale*, C. Bailly & A. Séchehaye (Eds.). Paris: Payot.

Savariaux C., Perrier P. & Orliaguet J.-P. (1995). Compensation strategies for the perturbation of the rounded vowel [u] using a lip-tube: a study of the control space in speech production. *J. Acoust. Soc. Am.*, 98, 5, 2428-2442.

Schmidt R.A. (1982). *Motor Control and learning. A behavioral emphasis.* Human Kinetics Publishers, Inc. Champaign, Illinois, Réed. 1988.

Schmidt R.A., Sherwood D.E., Zelaznik H.N. & Leikind B. (1985). Speed-accuracy trade-offs in motor behavior: Theories of impulse variability. In H. Heuer, U. Kleinbeck & K.H. Schmidt (Eds.), *Motor behavior: Programming, control and acquisition*, 79-123. Berlin: Springer-Verlag.

Schulman R. (1989). Articulatory dynamics of loud and normal speech. *J. Acoust. Soc. Am.*, 85, 1, 295-312.

Schwartz J.L., Beautemps D., Arrouas Y., Escudier P. (1992). Auditory analysis of speech gestures. In V.J. van Heuven & L.C.W. Pols (eds.) *The auditory processing of speech, from sounds to words,* 239-252. Berlin, New-York,

Scully C., Castelli E., Brearley E. & Shirt M. (1992). Analysis and simulation of a speaker's aerodynamic and acoustic pattern for fricatives. *J. Phonetics*, 20, 39-51.

Shankweiler D., Verbrugge R.R. & Studdert-Kennedy M. (1978). Insufficiency of the target for vowel perception. *J. Acoust. Soc. Am.*, 63, S1, S4.

Shanno D.F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24, 111, 647-656.

Shearme J.N. & Holmes J.N. (1962). An experimental study of the classification of sounds in continuous speech according to their distribution in the formant1-formant2 plane. In A. Sovijärvi & P. Aalto (Eds.), *Actes du 4ème Congrès International des Sciences Phonétiques*, 234-240, Mouton & Co., The Hague.

Shepard R.N. (1984). Ecological constraints on internal representation: resonant kinematics of perceiving, imagining, thinking and dreaming, *Psychol. Review*, 91, 4, 417-447.

Shroeder M.R. (1967). Determination of the geometry of the human vocal tract by acoustic measurements. *J. Acoust. Soc. Am.*, 41, 4, 1002-1010.

Smith C.L., Browman C.P., Mc Gowan R.S., Kay B. (1993). Extracting dynamic parameters from speech movement data. *J. Acous. Soc. Am. 93 (3)*, 1580-1588.

Sommerhof (1950). *Analytical Biology.* Oxford: The University Press.

Stålhammar U., Karlsson I. & Fant G. (1973). Contextual effects on vowel nuclei, *STL-QPSR*, 4, 1-18.

Stein R.B. (1982). What muscle variable(s) does the nervous system control in limb movements. *The Behavioral and Brain Sciences*, 5, 535-577.

Stetson R.H. (1928). *Motor Phonetics : a study of speech movements in action.* Archives néerlandaises de phonétique expérimentale, 3, 1-216. (2ème édition., 1951, North Holland: Amsterdam. Rééd. 1988, par Kelso J.A.S. et Munhall K.G., Boston).

Stevens K.N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In David Jr. E.E. & Denes P.B. (Eds;), *Human Communication: A unified view*, 51-66. New-York: Mc Graw Hill.

Stevens K.N. (1989). On the quantal nature of speech. *J. Phonetics*, 17, 3-45.

Stevens K.N. & Blumstein S.E. (1978). Invariant cues for place of articulation in stop consonants. *J. Acoust. Soc. Am., 64,* 1358-1368.

Stevens K.N. & Blumstein S.E. (1981). The search for invariant acoustic correlates of phonetic features. In P. Eimas & J. Miller (Eds.), *Perspectives on the study of speech.* Hillsdale N.J.: Lawrence Erlbaum Associates.

Stevens K.N. & House A.S. (1963). Perturbation of vowel articulations by consonantal context: an acoustical study, *J. Speech and Hearing Research*, 6, 111-128.

Strange W. (1989). Dynamic aspects of coarticulated vowels spoken in sentence context. *J. Acoust. Soc. Am.*, 85, 2135-2153.

Strange W., Jenkins J.J. & Edman T.R. (1978). Dynamic information specifies vowel identity. *J. Acoust. Soc. Am.*, 63, S1, S5.

Strange W., Jenkins J.J. & Johnson T.L. (1983). Dynamic specification of coarticulated vowels. *J. Acoust. Soc. Am.*, 74, 3, 695-705.

Strange W., Verbrugge R.R., Shankweiler D.P. & Edman T.R. (1976). Consonant environment specifies vowel identity. *J. Acoust. Soc. Am.*, 60, 213-224.

Studdert-Kennedy M. (1976). Speech Perception. In N.J. Lass (Ed.), *Contemporary issues in experimental phonetics.* New York: Academic Press.

Syrdal A.K. (1984). Aspects of a model of the auditory representation of American English vowels. *Speech Comm.*, 4, 121-135.

Sweet H. (1877). *Handbook of Phonetics,* Oxford: Henry Frowde.

Sweet H. (1899). *The practical study of languages. A guide for teachers and learners.* Réed. 1964: Oxford University Press, London.

Terzopoulos D. & Waters K. (1990). Physically-based facial modelling, analysis, and animation. *J. Visualization and Computer Animation*, 1 (2), 73-80.

Tiffany W.R. (1959). Non-random sources of variation in vowel quality. *J. Speech and Hearing Research*, 2, 305-317.

Torre V. & Poggio T. (1984). *MIT Artificial Intelligence Laboratory Memo* N°776.

Traunmüller H. (1981). Perceptual dimension of openness in vowels. *J. Acoust. Soc. Am.*, 69,5, 1465-1475.

Traunmüller H. (1985). The role of the fundamental and higher formants in the perception of speaker size, vocal effort and vowel openness, *Perilus IV*, Stockholm University, 92-102.

Tuller B., Harris K.S. & Kelso J.A.S. (1982). Stress and rate: differential transformations of articulation. *J. Acoust. Soc. Am.*, 71, 6, 1534, 1543.

Uno Y., Kawato M. & Suzuki R. (1987). Formation of optimum trajectory in control of arm movement - minimum torque-change model. *Japan IEICE Technical Report MBE*, 86-79: 9-16.

Van Bergem D.R. (1993). Acoustic vowel reduction as a function of sentence accent, word stress and word class. *Speech Comm.* 12, 1-23.

Van Son R.J.J.H. (1993). Vowel perception: a closer look at the literature. *IFA Proceedings 17*, 33-64.

Vatikiotis-Bateson E. & Kelso J.A.S. (1993). Rhythm type and articulatory dynamics in English, French and Japanese. *J. Phonetics 21*, 231-265.

Vatikiotis-Bateson E. & Ostry D.J. (1995). An analysis of the dimensionality of jaw motion in speech. *J. Phonetics*, 23, 101-117.

Vatikiotis-Bateson E., Hirayama M. & Kawato M. (1991). Neural network modeling of speech motor control using physiological data. *Perilus XIV* (Stockholm Univ. Linguistics Dept.), 63-67.

Vatikiotis-Bateson E., Hirayama M., Wada Y. & Kawato M. (1993). Generating articulator motion from muscle activity using artificial neural networks. In *Annual Bulletin Research Institute of logopedics & phoniatrics* (RILP), 27, 67-77.

Vatikiotis-Bateson E., Tiede M., Wada Y., Gracco V. & Kawato M. (1994). Phoneme extraction using via point estimation of real speech. In *Proceedings of the International Conference on Spoken Language Processing*, 2, 631-634, Yokohama.

Verbrugge R.R. & Isenberg D. (1978). Syllable timing and vowel perception. *J. Acoust. Soc. Am.*, 63, S1, S4.

Verbrugge R.R. & Rakerd B. (1986). Evidence of talker-independent information for vowels. *Language and Speech*, 29, 1, 39-57.

Weiberl E.R. (1963). *Morphometry of the human lung*. Springer-Verlag: Berlin.

# ANNEXE A

# Acoustique des voyelles

# ACOUSTIQUE DES VOYELLES

Cette annexe s'inspire du rapport de stage d'ingénieur effectué à l'Institut de la Communication Parlée par Lian Apostol [1994].

**Un peu d'anatomie**

On peut décrire le processus de parole de façon simplifiée comme une onde de vibrations acoustiques créée par le flux d'air à travers le conduit vocal. D'un point de vue fonctionnel on peut diviser l'appareil de production de la parole en trois parties : subglottique (poumons, trachée), glottique ou larynx et supraglottique (conduit vocal).

La source d'énergie de la parole est le mécanisme de la respiration, qui utilise les poumons et les muscles de la poitrine et de l'abdomen. Pendant l'expiration, à cause de la compression de la cage thoracique, l'air est envoyé vers *le larynx*, à une pression supérieure à la pression atmosphérique.

Le flux d'air produit par les poumons, avant d'arriver dans le conduit vocal, est modifié par la structure du larynx, situé entre le pharynx et la trachée, dans la région moyenne du cou. Au-dessus du larynx on trouve l'os hyoïde, en forme de fer à cheval, qui constitue en même temps la base de la langue. Chaque mouvement de cet os entraîne des déplacements des cartilages du larynx. La position du larynx varie avec l'âge du locuteur et aussi avec le sexe (les femmes ont un larynx plus élevé que les hommes, donc leur pharynx est plus court). Cette différence anatomique justifie les décalages qui existent entre les formants vocaliques des hommes et des femmes. La cavité interne du larynx peut être divisée en trois parties : le vestibule, l'étage moyen et l'étage inférieur. Les deux dernières parties sont délimitées par des replis, tendus horizontalement, les cordes vocales. Sous l'action de différents muscles, les cordes vocales modulent de façon pseudo-périodique le débit d'air issu des poumons. Elles sont à la base de la production des voyelles et des consonnes voisées. Elles sont ouvertes de façon à laisser passer l'air librement pendant la respiration et en parole lors de la production de sons non voisés (fricatives et consonnes non voisées). Pour la production de sons voisés, elles se rapprochent l'une de l'autre et vibrent sous l'action du flux d'air envoyé par les poumons et de la chute de pression qui s'établit entre les cavités sous-glottique et supra-glottiques. Le signal ainsi obtenu va exciter les cavités supraglottiques.

Le conduit vocal est formé de deux parties, orale et nasale. La partie orale est composée à son tour du pharynx et de la cavité buccale. Le pharynx est situé au-dessus du larynx, en face de la colonne vertébrale. Il communique, à la partie supérieure, avec les

cavités nasales et avec la bouche. À sa partie inférieure il se continue avec l'œsophage. La cavité buccale est délimitée par : le palais dur, le voile du palais, le corps de la langue, les joues, les dents et les lèvres. La dimension du conduit oral, mesurée des cordes vocales aux lèvres, varie d'un locuteur à un autre, la longueur moyenne étant de 17 cm environ. Les parties mobiles comme la langue, la mâchoire et les lèvres permettent des grandes variations dans la géométrie du conduit, formant ainsi un tube acoustique d'une forme complexe dont la section varie spatialement de la glotte aux lèvres et temporellement selon le son prononcé. Le couplage entre le conduit nasal et le conduit buccal est contrôlé par l'intermédiaire de la position du vélum (baissé ou relevé). La partie nasale est composée des fosses nasales qui se rejoignent avant d'être connectées au conduit buccal, et des sinus, cavités connectées aux fosses nasales par des étroits canaux. Les cavités nasales constituent des résonateurs pour les voyelles ou consonnes nasales, qui sont prononcés avec la valve vélopharyngeale ouverte.

**La production des voyelles**

En tenant compte des fonctions des différentes parties de l'appareil vocal, on peut le schématiser à l'aide de l'analyse des systèmes. Voici une représentation synthétique du système de production de la parole :

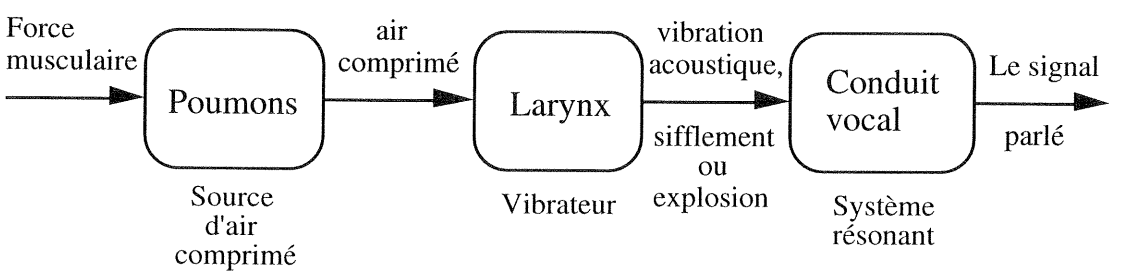



*Figure A.1.* Le système de production de la parole

Les voyelles sont des sons voisés, c'est-à-dire qu'elles sont produites avec la vibration des cordes vocales. Chaque fois que les cordes vocales s'ouvrent et se ferment, une pulsation d'air s'échappe des poumons, les cordes vocales jouant ainsi le rôle de générateur de débit. Cette pulsation met en vibration le conduit vocal d'une façon déterminée par ses dimensions et par sa forme. Pendant la prononciation d'une voyelle, l'air vibre dans le conduit plus particulièrement à quatre ou cinq fréquences privilégiées. Pour ces fréquences, les cavités supraglottiques présentent une faible impédance à la propagation de l'onde acoustique et laissent passer vers les lèvres un maximum d'énergie. Ces fréquences sont donc les fréquences de résonance de cette forme particulière du conduit vocal et sont appelées *formants* (voir figure A.2). On ne doit pas oublier ici la *fréquence fondamentale*

*de la voix* (notée F0) déterminée par la fréquence de vibration des cordes vocales, et qui donne au spectre une structure de raies.

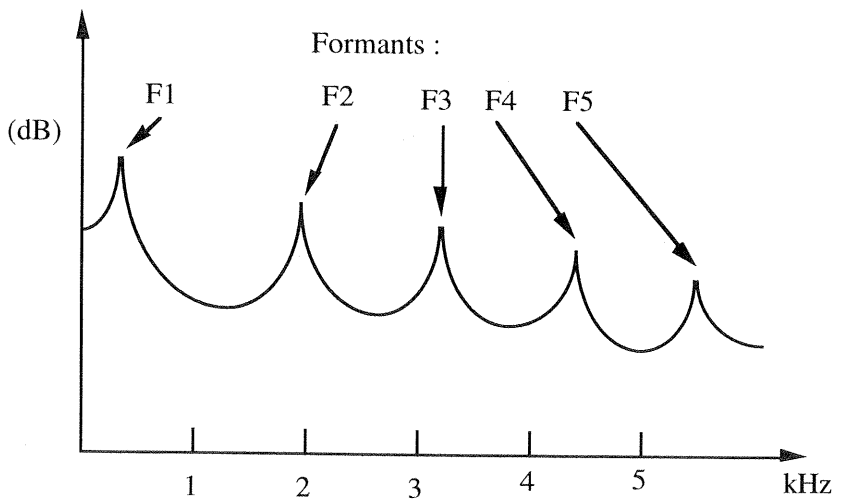



*Figure A.2* . L'enveloppe spectrale ; les formants.

Des expériences ont montré que l'oreille humaine est surtout sensible aux maxima de l'enveloppe spectrale, d'où l'importance des valeurs formantiques. Par ailleurs, la position sur l'axe de la fréquence des trois premiers formants caractérise parfaitement une voyelle. Mais on peut décrire suffisamment bien les voyelles en gardant seulement les deux premiers formants, d'où leur représentation courante sur un plan F1/F2.

**Modélisation acoustique de la production des voyelles**

On peut modéliser le processus de production de la parole en faisant une analogie avec les circuits électriques. Les cavités du conduit vocal peuvent être regardées comme un filtre qui est excité avec des impulsions quasi-périodiques (pour les sons voisés), fournies par une source qui dans notre cas est la glotte. Le schéma de ce système est :

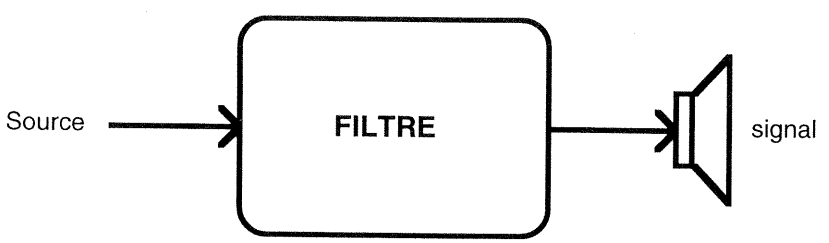



*Figure A.3.* La notion de source-filtre.

En première approximation, on peut considérer qu'il n'y a pas d'interaction entre la source et le filtre (c'est le concept source-filtre du système phonatoire, Fant [1960]). On

peut faire l'hypothèse que, lorsqu'on produit des sons voisés (et particulièrement des voyelles), la forme du conduit vocal évolue de la glotte vers les dents d'une manière suffisamment progressive pour que l'écoulement d'air soit considéré comme laminaire. Si les variations de la section transversale ne sont ainsi pas trop grandes et si on se situe dans un domaine de fréquence en deçà de 5000 Hz, la longueur d'onde correspondante est grande devant la dimension transversale maximale de ce tuyau acoustique et on peut supposer une propagation unidimensionnelle d'ondes planes, selon l'axe moyen du conduit. D'où l'idée de découper le conduit vocal en **n** tubes cylindriques élémentaires dont l'aire de la section suit l'évolution de la géométrie du conduit dans le plan frontal. Les paramètres d'entrée d'un modèle à **n** tubes consistent en une fonction d'aire, qui correspond à un "sample and hold" de la géométrie volumique continue du conduit vocal :

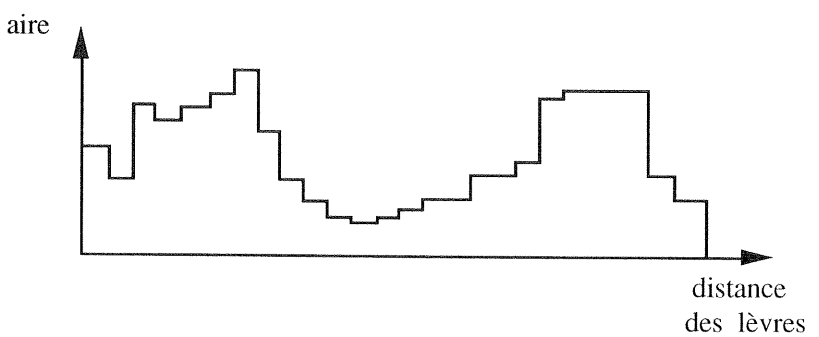



*Figure A.4.* Fonction d'aire.

On montre que le débit et la pression acoustique se comportent dans un tuyau unité respectivement comme le courant et la tension dans un quadripôle électrique. Du point de vue harmonique, le filtre, et donc le conduit vocal, peut être modélisé par un ensemble de circuits résonants, conçu à partir de la structure physique du conduit, et dont les éléments tendent à décrire aussi bien que possible le comportement spectral des cavités. Chaque tube élémentaire a un analogue électrique.

Même si la forme du tuyau acoustique que constitue le conduit vocal est complexe, on observe toujours, de manière plus ou moins prononcée, un rétrécissement et un seul entre la glotte et les dents, lors de la production des voyelles. Ce rétrécissement porte le nom de *constriction*. Celle-ci permet de distinguer, de manière générale, deux cavités en couplage : la cavité avant (des lèvres à la constriction), et la cavité arrière (de la constriction au larynx). Fant [1960] est parti de cette observation pour introduire un modèle simplifié du conduit vocal : le modèle à quatre tubes cylindriques.

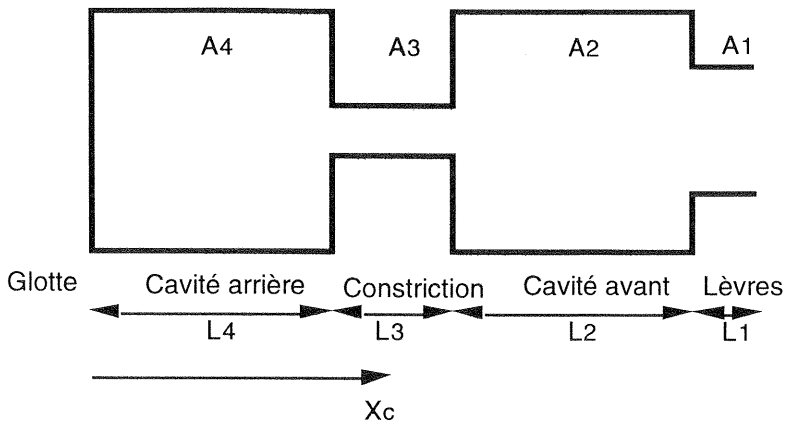



*Figure A.5.* Le modèle à quatre tubes

La constriction est représentée par le tube n° 3 (Aire A3, Longueur L3); les cavités avant et arrière, de part et d'autre de la constriction, sont schématisées par les tubes n° 2 et n° 4; la partie associée aux lèvres est simulée par le tube n° 1. Fant a étudié le comportement des formants lorsque la constriction se déplace de la glotte jusqu'aux lèvres; il obtient ainsi des *nomogrammes* du modèle à quatre tubes utilisés dans notre travail.

**Un peu d'acoustique**

L'air contenu dans un tube uniforme de longueur L peut vibrer de deux façons différentes, selon les conditions aux extrémités du tube :

- Si les extrémités sont toutes les deux soit ouvertes, soit fermées, la fréquence de résonance est un multiple de c/2L ( c étant la célérité du son dans l'air) : celle-ci est appelée "résonance demi-onde".

- Si une extrémité est ouverte et l'autre fermée, ou inversement, la fréquence de résonance est un multiple de c/4L : c'est une "résonance quart-d'onde".

La résonance associée à un système composé d'un résonateur de volume V comportant une embouchure de longueur L et de faible section A, est appelée "résonance de Helmholtz" et est donnée par la relation :

$$F = \frac{c}{2\pi}\sqrt{\frac{A}{LV}}$$

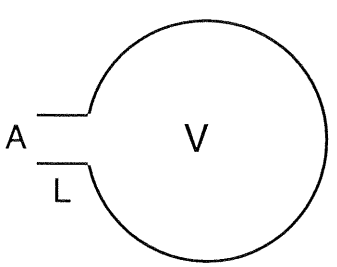



Figure A.6. Le résonateur de Helmholtz

# ANNEXE B

# Résultats des tests perceptifs par locuteur

# Sujet SF

- **Premier test d'identification**
- **Test d'identification complémentaire**
- **Test de jugement de qualité**

## Premier test d'identification

*Tableau A.1.a.* Nombre d'identifications correctes des deuxièmes stimuli de chaque paire pour les séquences [iai] et [iei] pour 5 répétitions. Lent accentué : "Lent acc", Lent non-accentué : "Lent Nacc", rapide accentué : "Rap. acc".

| | [iai] | | | | [iei] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [iei] Lent acc | 5 | 5 | 5 | [iai] Lent acc | 5 | 5 | 5 |
| [iei] Lent Nacc | 5 | 5 | 5 | [iai] Lent Nacc | 5 | 5 | 5 |
| [iei] Rap. acc | 5 | 5 | 5 | [iai] Rap. acc | 5 | 5 | 5 |
| [iai] Lent acc | | 5 | 5 | | | | |

*Tableau A.1.b.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [ia] et [ie].

| | [ia] | | | | [ie] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [ie] Lent acc | 5 | 5 | 5 | [ia] Lent acc | 5 | 5 | 5 |
| [ie] Lent Nacc | 5 | 5 | 5 | [ia] Lent Nacc | 5 | 5 | 5 |
| [ie] Rap. acc | 5 | 5 | 5 | [ia] Rap. acc | 5 | 5 | 5 |
| [ia] Lent acc | | 3 | 4 | | | | |

*Tableau A.1.c.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [ai] et [ei].

| | [ai] | | | | [ei] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [ei] Lent acc | 5 | 5 | 5 | [ai] Lent acc | 5 | 5 | 5 |
| [ei] Lent Nacc | 5 | 5 | 5 | [ai] Lent Nacc | 5 | 5 | 5 |
| [ei] Rap. acc | 5 | 5 | 5 | [ai] Rap. acc | 5 | 5 | 5 |
| [ai] Lent acc | | 5 | 5 | | | | |

*Tableau A.1.d.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [a] et [e].

| | [a] | | | | [e] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [e] Lent acc | 4 | 5 | 5 | [a] Lent acc | 5 | 5 | 5 |
| [e] Lent Nacc | 5 | 5 | 5 | [a] Lent Nacc | 5 | 4 | 5 |
| [e] Rap. acc | 5 | 5 | 5 | [a] Rap. acc | 5 | 5 | 5 |
| [a] Lent acc | | 3 | 3 | | | | |

## Test d'identification complémentaire

*Tableau A.2.a.* Nombre d'identifications correctes des stimuli pour 5 répétitions.

| | a | | | e | | |
|---|---|---|---|---|---|---|
| | Lent acc | Lent Nacc | Rap. acc | Lent acc | Lent Nacc | Rap. acc |
| iVi | 5 | 5 | 5 | 5 | 5 | 5 |
| iV | 5 | 5 | 5 | 5 | 5 | 5 |
| Vi | 5 | 5 | 5 | 5 | 5 | 5 |
| V | 5 | 5 | 5 | 5 | 5 | 5 |

## Test de jugement de qualité phonétique

*Tableau A.3.a.* Scores de qualité des stimuli lents accentués en séquence [iVi] pour 5 répétitions.

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 4 |
| suivi du cas rapide accentué | 5 | 3 |
| précédé du cas lent non accentué | 5 | 4 |
| précédé du cas rapide accentué | 5 | 3 |

*Tableau A.3.b.* Scores de qualité des stimuli lents accentués en séquence [iV]

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 5 |
| suivi du cas rapide accentué | 5 | 5 |
| précédé du cas lent non accentué | 5 | 4 |
| précédé du cas rapide accentué | 5 | 3 |

*Tableau A.3.c.* Scores de qualité des stimuli lents accentués en séquence [Vi].

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 5 |
| suivi du cas rapide accentué | 5 | 4 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 5 | 5 |

*Tableau A.3.d.* Scores de qualité des stimuli lents accentués en séquence [V].

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 5 |
| suivi du cas rapide accentué | 5 | 3 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 5 | 3 |

# Sujet PG

- **Premier test d'identification**
- **Test d'identification complémentaire**
- **Test de jugement de qualité**

## Premier test d'identification

*Tableau A.4.a.* Nombre d'identifications correctes des deuxièmes stimuli de chaque paire pour les séquences [iai] et [iei] pour 5 répétitions. Lent accentué : "Lent acc", Lent non-accentué : "Lent Nacc", rapide accentué : "Rap. acc".

| | [iai] | | | | [iei] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [iei] Lent acc | 5 | 0 | 5 | [iai] Lent acc | 5 | 5 | 5 |
| [iei] Lent Nacc | 5 | 2 | 4 | [iai] Lent Nacc | 5 | 5 | 5 |
| [iei] Rap. acc | 5 | 0 | 4 | [iai] Rap. acc | 5 | 5 | 5 |
| [iai] Lent acc | | 0 | 3 | | | | |

*Tableau A.4.b.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [ia] et [ie].

| | [ia] | | | | [ie] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [ie] Lent acc | 5 | 0 | 4 | [ia] Lent acc | 5 | 5 | 5 |
| [ie] Lent Nacc | 5 | 0 | 1 | [ia] Lent Nacc | 5 | 5 | 5 |
| [ie] Rap. acc | 5 | 0 | 0 | [ia] Rap. acc | 5 | 5 | 5 |
| [ia] Lent acc | | 0 | 0 | | | | |

*Tableau A.4.c.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [ai] et [ei].

| | [ai] | | | | [ei] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [ei] Lent acc | 5 | 2 | 3 | [ai] Lent acc | 5 | 5 | 5 |
| [ei] Lent Nacc | 5 | 2 | 5 | [ai] Lent Nacc | 5 | 5 | 5 |
| [ei] Rap. acc | 5 | 3 | 5 | [ai] Rap. acc | 5 | 5 | 5 |
| [ai] Lent acc | | 4 | 4 | | | | |

*Tableau A.4.d.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [a] et [e].

| | [a] | | | | [e] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [e] Lent acc | 5 | 0 | 0 | [a] Lent acc | 5 | 5 | 5 |
| [e] Lent Nacc | 5 | 0 | 0 | [a] Lent Nacc | 5 | 5 | 5 |
| [e] Rap. acc | 5 | 0 | 0 | [a] Rap. acc | 5 | 5 | 5 |
| [a] Lent acc | | 0 | 0 | | | | |

## Test d'identification complémentaire

*Tableau A.5.a.* Nombre d'identifications correctes des stimuli pour 5 répétitions.

| | a | | | e | | |
|---|---|---|---|---|---|---|
| | Lent acc | Lent Nacc | Rap. acc | Lent acc | Lent Nacc | Rap. acc |
| iVi | 5 | 1 | 5 | 5 | 5 | 5 |
| iV | 5 | 0 | 3 | 5 | 5 | 5 |
| Vi | 5 | 0 | 5 | 5 | 5 | 5 |
| V | 5 | 0 | 0 | 5 | 5 | 5 |

## Test de jugement de qualité phonétique

*Tableau A.6.a.* Scores de qualité des stimuli lents accentués en séquence [iVi] pour 5 répétitions.

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 5 |
| suivi du cas rapide accentué | 5 | 4 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 5 | 4 |

*Tableau A.6.b.* Scores de qualité des stimuli lents accentués en séquence [iV]

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 4 |
| suivi du cas rapide accentué | 5 | 3 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 5 | 5 |

*Tableau A.6.c.* Scores de qualité des stimuli lents accentués en séquence [Vi].

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 5 |
| suivi du cas rapide accentué | 5 | 3 |
| précédé du cas lent non accentué | 5 | 4 |
| précédé du cas rapide accentué | 5 | 2 |

*Tableau A.6.d.* Scores de qualité des stimuli lents accentués en séquence [V].

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 4 |
| suivi du cas rapide accentué | 5 | 2 |
| précédé du cas lent non accentué | 5 | 4 |
| précédé du cas rapide accentué | 5 | 3 |

# Sujet CC

- **Premier test d'identification**
- **Test d'identification complémentaire**
- **Test de jugement de qualité**

## Premier test d'identification

*Tableau A.7.a.* Nombre d'identifications correctes des deuxièmes stimuli de chaque paire pour les séquences [iai] et [iei] pour 5 répétitions. Lent accentué : "Lent acc", Lent non-accentué : "Lent Nacc", rapide accentué : "Rap. acc".

| | [iai] | | | | [iei] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [iei] Lent acc | 5 | 5 | 4 | [iai] Lent acc | 5 | 5 | 5 |
| [iei] Lent Nacc | 5 | 5 | 5 | [iai] Lent Nacc | 5 | 4 | 5 |
| [iei] Rap. acc | 5 | 5 | 5 | [iai] Rap. acc | 5 | 5 | 5 |
| [iai] Lent acc | | 5 | 5 | | | | |

*Tableau A.7.b.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [ia] et [ie].

| | [ia] | | | | [ie] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [ie] Lent acc | 5 | 3 | 5 | [ia] Lent acc | 5 | 5 | 5 |
| [ie] Lent Nacc | 5 | 3 | 5 | [ia] Lent Nacc | 5 | 5 | 5 |
| [ie] Rap. acc | 5 | 4 | 4 | [ia] Rap. acc | 5 | 5 | 5 |
| [ia] Lent acc | | 3 | 2 | | | | |

*Tableau A.7.c.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [ai] et [ei].

| | [ai] | | | | [ei] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [ei] Lent acc | 5 | 3 | 4 | [ai] Lent acc | 5 | 5 | 5 |
| [ei] Lent Nacc | 5 | 5 | 3 | [ai] Lent Nacc | 5 | 5 | 5 |
| [ei] Rap. acc | 5 | 5 | 3 | [ai] Rap. acc | 5 | 5 | 5 |
| [ai] Lent acc | | 3 | 4 | | | | |

*Tableau A.7.d.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [a] et [e].

| | [a] | | | | [e] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [e] Lent acc | 5 | 1 | 1 | [a] Lent acc | 5 | 5 | 5 |
| [e] Lent Nacc | 5 | 4 | 1 | [a] Lent Nacc | 5 | 5 | 5 |
| [e] Rap. acc | 5 | 3 | 2 | [a] Rap. acc | 5 | 5 | 5 |
| [a] Lent acc | | 1 | 1 | | | | |

## Test d'identification complémentaire

*Tableau A.8.a.* Nombre d'identifications correctes des stimuli pour 5 répétitions.

| | a | | | e | | |
|---|---|---|---|---|---|---|
| | Lent acc | Lent Nacc | Rap. acc | Lent acc | Lent Nacc | Rap. acc |
| iVi | 5 | 5 | 5 | 5 | 5 | 5 |
| iV | 5 | 4 | 5 | 5 | 5 | 5 |
| Vi | 4 | 4 | 5 | 5 | 5 | 5 |
| V | 5 | 1 | 3 | 5 | 5 | 5 |

## Test de jugement de qualité phonétique

*Tableau A.9.a.* Scores de qualité des stimuli lents accentués en séquence [iVi] pour 5 répétitions.

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 2 | 2 |
| suivi du cas rapide accentué | 3 | 3 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 4 | 5 |

*Tableau A.9.b.* Scores de qualité des stimuli lents accentués en séquence [iV]

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 5 |
| suivi du cas rapide accentué | 5 | 5 |
| précédé du cas lent non accentué | 5 | 4 |
| précédé du cas rapide accentué | 5 | 5 |

*Tableau A.9.c.* Scores de qualité des stimuli lents accentués en séquence [Vi].

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 3 |
| suivi du cas rapide accentué | 5 | 2 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 5 | 4 |

*Tableau A.9.d.* Scores de qualité des stimuli lents accentués en séquence [V].

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 0 |
| suivi du cas rapide accentué | 5 | 1 |
| précédé du cas lent non accentué | 5 | 2 |
| précédé du cas rapide accentué | 5 | 3 |

# Sujet PP

- **Premier test d'identification**
- **Test d'identification complémentaire**
- **Test de jugement de qualité**

### Premier test d'identification

*Tableau A.10.a.* Nombre d'identifications correctes des deuxièmes stimuli de chaque paire pour les séquences [iai] et [iei] pour 5 répétitions. Lent accentué : "Lent acc", Lent non-accentué : "Lent Nacc", rapide accentué : "Rap. acc".

| | [iai] | | | | [iei] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [iei] Lent acc | 4 | 5 | 5 | [iai] Lent acc | 5 | 5 | 5 |
| [iei] Lent Nacc | 5 | 5 | 5 | [iai] Lent Nacc | 5 | 5 | 5 |
| [iei] Rap. acc | 5 | 5 | 5 | [iai] Rap. acc | 5 | 5 | 5 |
| [iai] Lent acc | | 4 | 5 | | | | |

*Tableau A.10.b.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [ia] et [ie].

| | [ia] | | | | [ie] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [ie] Lent acc | 5 | 3 | 5 | [ia] Lent acc | 5 | 5 | 5 |
| [ie] Lent Nacc | 5 | 4 | 5 | [ia] Lent Nacc | 5 | 5 | 5 |
| [ie] Rap. acc | 5 | 4 | 5 | [ia] Rap. acc | 5 | 5 | 5 |
| [ia] Lent acc | | 3 | 5 | | | | |

*Tableau A.10.c.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [ai] et [ei].

| | [ai] | | | | [ei] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [ei] Lent acc | 5 | 1 | 2 | [ai] Lent acc | 5 | 5 | 5 |
| [ei] Lent Nacc | 5 | 1 | 1 | [ai] Lent Nacc | 5 | 5 | 5 |
| [ei] Rap. acc | 5 | 4 | 1 | [ai] Rap. acc | 5 | 5 | 5 |
| [ai] Lent acc | | 2 | 2 | | | | |

*Tableau A.10.d.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [a] et [e].

| | [a] | | | | [e] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [e] Lent acc | 5 | 1 | 0 | [a] Lent acc | 5 | 5 | 5 |
| [e] Lent Nacc | 5 | 1 | 1 | [a] Lent Nacc | 5 | 5 | 5 |
| [e] Rap. acc | 5 | 1 | 0 | [a] Rap. acc | 5 | 5 | 5 |
| [a] Lent acc | | 1 | 2 | | | | |

## Test d'identification complémentaire

*Tableau A.11.a.* Nombre d'identifications correctes des stimuli pour 5 répétitions.

| | a | | | e | | |
|---|---|---|---|---|---|---|
| | Lent acc | Lent Nacc | Rap. acc | Lent acc | Lent Nacc | Rap. acc |
| iVi | 5 | 5 | 5 | 5 | 5 | 5 |
| iV | 5 | 5 | 5 | 5 | 5 | 5 |
| Vi | 5 | 2 | 5 | 5 | 5 | 5 |
| V | 5 | 4 | 2 | 5 | 5 | 5 |

## Test de jugement de qualité phonétique

*Tableau A.12.a.* Scores de qualité des stimuli lents accentués en séquence [iVi] pour 5 répétitions.

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 5 |
| suivi du cas rapide accentué | 5 | 5 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 5 | 5 |

*Tableau A.12.b.* Scores de qualité des stimuli lents accentués en séquence [iV]

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 5 |
| suivi du cas rapide accentué | 5 | 5 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 5 | 5 |

*Tableau A.12.c.* Scores de qualité des stimuli lents accentués en séquence [Vi].

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 5 |
| suivi du cas rapide accentué | 5 | 4 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 5 | 2 |

*Tableau A.12.d.* Scores de qualité des stimuli lents accentués en séquence [V].

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 5 |
| suivi du cas rapide accentué | 5 | 3 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 5 | 2 |

# Sujet CV

- **Premier test d'identification**
- **Test d'identification complémentaire**
- **Test de jugement de qualité**

## Premier test d'identification

*Tableau A.13.a.* Nombre d'identifications correctes des deuxièmes stimuli de chaque paire pour les séquences [iai] et [iei] pour 5 répétitions. Lent accentué : "Lent acc", Lent non-accentué : "Lent Nacc", rapide accentué : "Rap. acc".

| | [iai] | | | | [iei] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [iei] Lent acc | 5 | 5 | 5 | [iai] Lent acc | 5 | 5 | 5 |
| [iei] Lent Nacc | 5 | 5 | 5 | [iai] Lent Nacc | 5 | 5 | 5 |
| [iei] Rap. acc | 5 | 5 | 5 | [iai] Rap. acc | 5 | 5 | 5 |
| [iai] Lent acc | | 0 | 0 | | | | |

*Tableau A.13.b.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [ia] et [ie].

| | [ia] | | | | [ie] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [ie] Lent acc | 5 | 5 | 5 | [ia] Lent acc | 5 | 5 | 5 |
| [ie] Lent Nacc | 5 | 5 | 5 | [ia] Lent Nacc | 5 | 5 | 5 |
| [ie] Rap. acc | 5 | 5 | 5 | [ia] Rap. acc | 5 | 5 | 5 |
| [ia] Lent acc | | 0 | 0 | | | | |

*Tableau A.13.c.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [ai] et [ei].

| | [ai] | | | | [ei] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [ei] Lent acc | 5 | 5 | 5 | [ai] Lent acc | 5 | 5 | 5 |
| [ei] Lent Nacc | 5 | 5 | 5 | [ai] Lent Nacc | 5 | 5 | 5 |
| [ei] Rap. acc | 5 | 5 | 5 | [ai] Rap. acc | 5 | 5 | 5 |
| [ai] Lent acc | | 1 | 0 | | | | |

*Tableau A.13.d.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [a] et [e].

| | [a] | | | | [e] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [e] Lent acc | 5 | 5 | 5 | [a] Lent acc | 5 | 5 | 5 |
| [e] Lent Nacc | 5 | 5 | 5 | [a] Lent Nacc | 5 | 5 | 5 |
| [e] Rap. acc | 5 | 5 | 5 | [a] Rap. acc | 5 | 5 | 5 |
| [a] Lent acc | | 0 | 0 | | | | |

## Test d'identification complémentaire

*Tableau A.14.a.* Nombre d'identifications correctes des stimuli pour 5 répétitions.

| | a | | | e | | |
|---|---|---|---|---|---|---|
| | Lent acc | Lent Nacc | Rap. acc | Lent acc | Lent Nacc | Rap. acc |
| iVi | 5 | 5 | 5 | 4 | 5 | 5 |
| iV | 5 | 2 | 5 | 5 | 5 | 5 |
| Vi | 5 | 5 | 5 | 5 | 5 | 5 |
| V | 5 | 0 | 0 | 5 | 5 | 5 |

## Test de jugement de qualité phonétique

*Tableau A.15.a.* Scores de qualité des stimuli lents accentués en séquence [iVi] pour 5 répétitions.

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 2 |
| suivi du cas rapide accentué | 4 | 1 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 5 | 4 |

*Tableau A.15.b.* Scores de qualité des stimuli lents accentués en séquence [iV]

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 5 |
| suivi du cas rapide accentué | 5 | 0 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 5 | 4 |

*Tableau A.15.c.* Scores de qualité des stimuli lents accentués en séquence [Vi].

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 5 |
| suivi du cas rapide accentué | 5 | 1 |
| précédé du cas lent non accentué | 4 | 5 |
| précédé du cas rapide accentué | 4 | 4 |

*Tableau A.15.d.* Scores de qualité des stimuli lents accentués en séquence [V].

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 4 | 5 |
| suivi du cas rapide accentué | 5 | 5 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 5 | 5 |

# Sujet JMB

- **Premier test d'identification**
- **Test d'identification complémentaire**
- **Test de jugement de qualité**

## Premier test d'identification

*Tableau A.16.a.* Nombre d'identifications correctes des deuxièmes stimuli de chaque paire pour les séquences [iai] et [iei] pour 5 répétitions. Lent accentué : "Lent acc", Lent non-accentué : "Lent Nacc", rapide accentué : "Rap. acc".

| | [iai] | | | | [iei] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [iei] Lent acc | 5 | 5 | 5 | [iai] Lent acc | 5 | 5 | 5 |
| [iei] Lent Nacc | 5 | 5 | 5 | [iai] Lent Nacc | 5 | 5 | 5 |
| [iei] Rap. acc | 5 | 5 | 5 | [iai] Rap. acc | 5 | 5 | 5 |
| [iai] Lent acc | | 5 | 5 | | | | |

*Tableau A.16.b.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [ia] et [ie].

| | [ia] | | | | [ie] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [ie] Lent acc | 5 | 1 | 0 | [ia] Lent acc | 5 | 5 | 5 |
| [ie] Lent Nacc | 5 | 0 | 1 | [ia] Lent Nacc | 5 | 5 | 5 |
| [ie] Rap. acc | 5 | 1 | 0 | [ia] Rap. acc | 5 | 5 | 5 |
| [ia] Lent acc | | 0 | 1 | | | | |

*Tableau A.16.c.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [ai] et [ei].

| | [ai] | | | | [ei] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [ei] Lent acc | 5 | 0 | 1 | [ai] Lent acc | 5 | 5 | 5 |
| [ei] Lent Nacc | 5 | 0 | 1 | [ai] Lent Nacc | 5 | 5 | 5 |
| [ei] Rap. acc | 5 | 0 | 0 | [ai] Rap. acc | 5 | 5 | 5 |
| [ai] Lent acc | | 0 | 1 | | | | |

*Tableau A.16.d.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [a] et [e].

| | [a] | | | | [e] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [e] Lent acc | 5 | 0 | 0 | [a] Lent acc | 5 | 5 | 5 |
| [e] Lent Nacc | 5 | 0 | 0 | [a] Lent Nacc | 5 | 5 | 5 |
| [e] Rap. acc | 5 | 0 | 0 | [a] Rap. acc | 5 | 5 | 5 |
| [a] Lent acc | | 0 | 0 | | | | |

## Test d'identification complémentaire

*Tableau A.17.a.* Nombre d'identifications correctes des stimuli pour 5 répétitions.

| | a | | | e | | |
|---|---|---|---|---|---|---|
| | Lent acc | Lent Nacc | Rap. acc | Lent acc | Lent Nacc | Rap. acc |
| iVi | 5 | 4 | 5 | 5 | 5 | 4 |
| iV | 5 | 0 | 1 | 5 | 5 | 5 |
| Vi | 5 | 0 | 0 | 5 | 5 | 5 |
| V | 5 | 0 | 0 | 5 | 5 | 5 |

## Test de jugement de qualité phonétique

*Tableau A.18.a.* Scores de qualité des stimuli lents accentués en séquence [iVi] pour 5 répétitions.

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 4 | 5 |
| suivi du cas rapide accentué | 5 | 1 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 5 | 5 |

*Tableau A.18.b.* Scores de qualité des stimuli lents accentués en séquence [iV]

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 4 |
| suivi du cas rapide accentué | 5 | 1 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 5 | 5 |

*Tableau A.18.c.* Scores de qualité des stimuli lents accentués en séquence [Vi].

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 5 |
| suivi du cas rapide accentué | 5 | 3 |
| précédé du cas lent non accentué | 4 | 5 |
| précédé du cas rapide accentué | 5 | 4 |

*Tableau A.18.d.* Scores de qualité des stimuli lents accentués en séquence [V].

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 5 |
| suivi du cas rapide accentué | 5 | 5 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 5 | 4 |

# Sujet EZ

- **Premier test d'identification**
- **Test d'identification complémentaire**
- **Test de jugement de qualité**

## Premier test d'identification

*Tableau A.19.a.* Nombre d'identifications correctes des deuxièmes stimuli de chaque paire pour les séquences [iai] et [iei] pour 5 répétitions. Lent accentué : "Lent acc", Lent non-accentué : "Lent Nacc", rapide accentué : "Rap. acc".

| | [iai] | | | | [iei] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [iei] Lent acc | 5 | 5 | 5 | [iai] Lent acc | 5 | 5 | 5 |
| [iei] Lent Nacc | 5 | 5 | 5 | [iai] Lent Nacc | 5 | 5 | 5 |
| [iei] Rap. acc | 5 | 5 | 5 | [iai] Rap. acc | 5 | 5 | 5 |
| [iai] Lent acc | | 5 | 5 | | | | |

*Tableau A.19.b.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [ia] et [ie].

| | [ia] | | | | [ie] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [ie] Lent acc | 5 | 3 | 4 | [ia] Lent acc | 5 | 5 | 5 |
| [ie] Lent Nacc | 5 | 3 | 1 | [ia] Lent Nacc | 5 | 5 | 5 |
| [ie] Rap. acc | 5 | 3 | 3 | [ia] Rap. acc | 5 | 5 | 5 |
| [ia] Lent acc | | 1 | 0 | | | | |

*Tableau A.19.c.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [ai] et [ei].

| | [ai] | | | | [ei] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [ei] Lent acc | 5 | 5 | 5 | [ai] Lent acc | 5 | 5 | 4 |
| [ei] Lent Nacc | 5 | 5 | 5 | [ai] Lent Nacc | 5 | 4 | 5 |
| [ei] Rap. acc | 5 | 5 | 5 | [ai] Rap. acc | 5 | 5 | 5 |
| [ai] Lent acc | | 4 | 4 | | | | |

*Tableau A.19.d.* Scores d'identification des deuxièmes stimuli de chaque paire pour les séquences [a] et [e].

| | [a] | | | | [e] | | |
|---|---|---|---|---|---|---|---|
| 1er stimulus | Lent acc | Lent Nacc | Rap. acc | 1er stimulus | Lent acc | Lent Nacc | Rap. acc |
| [e] Lent acc | 5 | 5 | 5 | [a] Lent acc | 5 | 5 | 5 |
| [e] Lent Nacc | 5 | 5 | 4 | [a] Lent Nacc | 5 | 5 | 5 |
| [e] Rap. acc | 5 | 5 | 5 | [a] Rap. acc | 5 | 5 | 5 |
| [a] Lent acc | | 2 | 0 | | | | |

## Test d'identification complémentaire

*Tableau A.20.a.* Nombre d'identifications correctes des stimuli pour 5 répétitions.

| | a | | | e | | |
|---|---|---|---|---|---|---|
| | Lent acc | Lent Nacc | Rap. acc | Lent acc | Lent Nacc | Rap. acc |
| iVi | 5 | 5 | 5 | 4 | 5 | 5 |
| iV | 5 | 0 | 5 | 5 | 5 | 5 |
| Vi | 5 | 5 | 5 | 5 | 5 | 5 |
| V | 5 | 4 | 2 | 5 | 5 | 5 |

## Test de jugement de qualité phonétique

*Tableau A.21.a.* Scores de qualité des stimuli lents accentués en séquence [iVi] pour 5 répétitions.

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 4 | 5 |
| suivi du cas rapide accentué | 5 | 3 |
| précédé du cas lent non accentué | 5 | 4 |
| précédé du cas rapide accentué | 5 | 4 |

*Tableau A.21.b.* Scores de qualité des stimuli lents accentués en séquence [iV]

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 5 |
| suivi du cas rapide accentué | 5 | 4 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 5 | 5 |

*Tableau A.21.c.* Scores de qualité des stimuli lents accentués en séquence [Vi].

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 4 |
| suivi du cas rapide accentué | 4 | 5 |
| précédé du cas lent non accentué | 5 | 3 |
| précédé du cas rapide accentué | 5 | 1 |

*Tableau A.21.d.* Scores de qualité des stimuli lents accentués en séquence [V].

| | [a] Lent acc | [e] Lent acc |
|---|---|---|
| suivi du cas lent non accentué | 5 | 5 |
| suivi du cas rapide accentué | 5 | 5 |
| précédé du cas lent non accentué | 5 | 5 |
| précédé du cas rapide accentué | 5 | 3 |

## PISTES POUR LE CONTRÔLE D'UN ROBOT PARLANT CAPABLE DE RÉDUCTION VOCALIQUE.

**Résumé**
L'objectif de cette thèse est d'étudier comment la commande phonémique, par essence abstraite et invariante, est codée en termes de contrôle moteur pour générer un signal acoustique physique et variable. Pour cela, un modèle de contrôle moteur est exploité, s'appuyant sur la Théorie "du Point d'Équilibre", proposée par A. Feldman, selon laquelle tout mouvement résulterait de changements des variables de contrôle des conditions d'équilibre mécanique du système moteur. L'hypothèse étudiée ici consiste à relier les commandes motrices discrètes, spécifiant des points d'équilibre cibles, aux phonèmes. Ce modèle permet de générer des trajectoires articulatoires et acoustiques réalistes, pour des séquences Voyelle-Voyelle-Voyelle, prononcées par un locuteur français dans diverses conditions d'élocution. En effet, partant du signal acoustique produit par le locuteur, une première étape d'inversion d'un modèle articulatoire du conduit vocal fournit les trajectoires des articulateurs du conduit vocal. Puis, pour l'articulateur corps de la langue, le plus pertinent dans les séquences vocaliques étudiées, un modèle dynamique fonctionnel des articulateurs de la parole est inversé (deuxième étape d'inversion), afin de trouver les commandes motrices sous-jacentes à la trajectoire observée dans une des conditions d'élocution, dite idéale. Puis à partir des commandes motrices ainsi inférées pour la condition idéale, il est possible, en ne faisant varier que les paramètres dynamiques appropriés, de générer des signaux articulatoires et acoustiques variables, similaires à ceux des autres conditions d'élocution. Ainsi une piste nouvelle pour la synthèse adaptative de la parole est ouverte. Des analyses de sensibilité et des tests perceptifs effectués sur les signaux de synthèse attestent de la capacité de ce modèle de contrôle à simuler les effets d'accent d'emphase et de débit à partir de commandes posturales constantes du type "Point d'Équilibre".

**Mots-clefs :** communication parlée, production, contrôle moteur, réduction vocalique, invariance et variabilité, notion de cible, inversion, hypothèse du point d'équilibre.

## PROPOSALS FOR THE CONTROL OF A SPEAKING ROBOT CAPABLE OF VOWEL REDUCTION.

**Abstract**
The purpose of this thesis is to study how the phonemic command, abstract and invariant in nature, is coded in terms of motor control, to generate a physical and variable acoustic signal. A motor control model is exploited, based on the Equilibrium Point Hypothesis, proposed by A. Feldman, which suggests that movement result from shifts in variables controlling the system's mechanical equilibrium. The hypothesis studied here consist in relating discrete motor commands, specifying target equilibrium points, to phonemes. Realistic articulatory and acoustic trajectories were generated with this model for Vowel-Vowel-Vowel sequences, uttered by a French speaker under several speech conditions. Articulatory trajectories were inferred from the acoustic signal by a first inversion involving an articulatory model of the vocal tract. A functional dynamic model of speech articulators, applied to the tongue body articulator, which is considered as the most relevant in the studied sequences, is then inverted to recover the motor commands underlying the trajectory observed in the "ideal" speech condition. Then from the motor commands thus inferred for the ideal condition, articulatory and acoustic signals, similar to the ones observed in other conditions, are generated by simply tuning appropriate dynamic parameters. A new track is thus open for adaptive speech synthesis. Sensitivity analysis and perceptual tests on synthetic signals attest for the control model's ability to simulate focus stress and rate effects using constant postural (equilibrium point) commands.

**Keywords :** speech communication, production, motor control, vowel reduction, invariance and variability, vowel target , inversion, equilibrium point hypothesis.