Thèse de l'Institut National Polytechnique de Grenoble
Spécialité Sciences Cognitives
Décembre 1997

présentée par
Yann MORLEC

Génération multiparamétrique de la prosodie du français par apprentissage automatique

Composition du jury :

Jean-Marc DOLMAZON
Jacqueline VAISSIÈRE
Daniel HIRST
Gérard BAILLY
Véronique AUBERGÉ

Institut de la Communication Parlée
46 av. Félix Viallet, 38031 Grenoble CEDEX 1 France

Résumé

Cette thèse a pour objet la conception, l'apprentissage automatique et l'évaluation d'un modèle de génération de la prosodie du français.

Ce modèle, d'inspiration cognitive, suppose un encodage direct des informations linguistiques et para-linguistiques véhiculées par un énoncé via des formes prosodiques prototypiques. Ces formes - ou mouvements - sont enchaînées au sein d'un même niveau linguistique (phrase, groupe, mot...) et se superposent à des enchaînements encodant les niveaux supérieurs. Cette morphologie superpositionnelle opère de manière multiparamétrique. Dans le cadre de cette thèse, nous avons étudié deux niveaux linguistiques - la phrase et le groupe - et deux dimensions prosodiques - la mélodie et le rythme. Pour ces deux niveaux, nous avons élaboré un système de génération conjoint des contours mélodiques et rythmiques, paramétré par un apprentissage automatique hiérarchique sur des corpus de parole naturelle. L'architecture du modèle est un réseau de neurones récurrent à connections partielles.

Les corpus de parole ont été conçus de façon à prendre en compte le phénomène linguistique étudié avec une représentativité statistique suffisante, et à faire varier systématiquement le contenu des niveaux linguistiques inférieurs. A partir d'un corpus de phrases isolées prononcées par un locuteur selon six consignes attitudinales, nous avons entrepris la génération des prototypes mélodiques et rythmiques du niveau de la phrase véhiculant l'attitude du locuteur vis-à-vis de son propos. Nous avons également étudié, sur la base d'un corpus de phrases déclaratives, les modulations du niveau porté de groupe prenant en charge la fonction de structuration de l'énoncé.

Chaque étape d'apprentissage de notre modèle a été sanctionnée par un test de perception destiné à évaluer sa capacité à capturer les informations nécessaires et suffisantes pour communiquer les fonctions linguistiques données comme consignes a priori dans chaque corpus.

Abstract

This thesis presents the design, automatic training and evaluation of a prosody generation system for French text-to-speech synthesis. The cognitive model that underlies this research assumes a direct encoding of linguistic and para-linguistic information carried by an utterance via prototypic prosodic contours. These contours - or movements - are chained within a given linguistic level (sentence, group, word...) and superposed with those that encode higher levels. Prosody is thus described as the superposition of independent multiparametric movements. We study two linguistic levels - sentence and group - and two prosodic dimensions - melody and rhythm. For these two levels we design a model generating both melodic and rhythmic contours. The parameters of the model are tuned on an ad hoc database of natural speech. This functional model is a recurrent neural network with partial connections.

Starting from a corpus of isolated utterances uttered with six different prosodic attitudes by a male speaker, we show that such a functional model can generate appropriate melodic and rhythmic prototypes. Using an additional corpus of declarative utterances we also study and generate prosodic contours at the group level. We show that these synthetic contours cue the syntactic structure of the sentence. At each step of the hierarchical training process, we perform a perceptual test in order to evaluate the ability of our model to generate the necessary and sufficient information conveying linguistic functions suggested a priori in the training corpus.

Remerciements

Parce qu'une thèse n'est pas le fruit d'un travail solitaire mais celui d'une communauté de têtes pensantes dans un environnement favorable, je tiens à remercier tous ceux qui m'ont permis de mener à bien et d'évaluer mes activités de recherche.

Ces personnes, qui font partie de ma vie professionnelle et de ma vie privée, j'ai préféré les remercier de vive voix plutôt que d'en faire une liste plus ou moins exhaustive et qui, de toutes les façons, ne saurait traduire la gratitude que je leur porte.

Je tiens simplement à dédicacer ce mémoire à ma moitié, Isabelle, qui a su me soutenir dans les moments douloureux de l'enfantement de ce mémoire.

Table des matières

Introduction

Ce mémoire présente la conception, l'apprentissage automatique et l'évaluation d'un modèle de génération multiparamétrique de la prosodie. Ce modèle est assis sur une approche cognitive des formes prosodiques proposée par Aubergé (1991), et qui suppose un encodage direct des informations linguistiques et paralinguistiques associées à un énoncé via des formes prosodiques prototypiques. Ces formes - ou mouvements - sont enchaînées au sein d'un même niveau (phrase, groupe, mot…) et se superposent à des enchaînements encodant les niveaux supérieurs. La thèse de Barbosa (1994) complète cette assise en mettant en évidence, pour le français, une structuration rythmique par contours dont l'empan temporel est de l'ordre du groupe.

Ces positions fournissent un paradigme théorique fort : une approche morphologique et superpositionnelle de la prosodie mettant en rapport direct forme et sens. Cette thèse confronte cette approche à de nouvelles données, enrichit le modèle initial par une négociation plus fine dans l'usage des paramètres prosodiques par les niveaux linguistiques et, surtout, propose un système de génération conjoint des durées segmentales et de la mélodie. Elle permet de plus de valider notre approche expérimentale fondée sur une démarche hypothético-déductive : corpus–validation–modélisation–évaluation. Ce cycle est générique et sera appliqué dans ce travail à deux fonctions langagières de la prosodie : attitudinale et hiérarchisante.

Ce manuscrit est organisé en trois parties :

- La première est un bref état de l'art qui nous permet, en nous appuyant sur un ensemble volontairement limité de références bibliographiques servant notre propos, de préciser notre position vis-à-vis de l'objet prosodique et les principes de modélisation que nous avons adoptés.
- La seconde présente les différents corpus de parole qui ont été nécessaires à l'apprentissage de notre modèle de génération automatique de la prosodie. Nous détaillons, en premier lieu la méthodologie de conception que nous avons employée. L'analyse acoustique des données nous permet de renforcer l'hypothèse selon laquelle l'information prosodique est répartie sur l'ensemble de l'unité qui la porte. Un test original de dévoilement progressif montre que cette information « distribuée » est bien perçue par l'auditeur. Un Modèle d'Expansion du Mouvement Prosodique (MEMP) met en

évidence les caractéristiques spatio-temporelles des contours révélés par l'analyse.

- La troisième partie retrace les différentes étapes de conception, d'apprentissage hiérarchique et d'évaluation de notre MEMP. Elle montre l'intérêt de l'utilisation d'outil connexionniste et d'une architecture modulaire qui permet une évaluation progressive à chaque stade d'apprentissage. Deux niveaux de représentation linguistique ont été étudiés et modélisés : le niveau de la *phrase* exprimant la position du locuteur vis-à-vis de l'objet dont il parle et de son interlocuteur ; le niveau du *groupe* véhiculant la structuration de l'énoncé.

Ces étapes successives nous conduiront à reformuler nos hypothèses de travail pour tenir compte des effets de négociation entre les différents niveaux considérés et pour modéliser les synergies développées par les deux paramètres prosodiques étudiés.

En annexe, vous trouverez d'une part le détail des corpus de phrases isolées qui ont servi à l'apprentissage des différents modules de notre modèle de génération et, d'autre part, un rapide inventaire des outils et des différentes étapes qui nous ont permis de mener à bien cette étude.

Chapitre 1

Représentations phonétiques et modèles de génération de la prosodie

1.1 Prosodie, accentuation, rythme et intonation

1.1.1 Quelques définitions

Cette section établit, en guise de préambule à l'étude qui va suivre, un ensemble de définitions prélevées dans la littérature de la prosodie, de l'accentuation, de l'intonation et du rythme. Nous tenterons ensuite de préciser notre propre point de vue en établissant une distinction claire entre le niveau acoustique, perceptif et fonctionnel.

1.1.1.1 La prosodie

Elle est née d'une opposition théorique entre ce qui distingue la segmentalité de la suprasegmentalité, cette dernière caractérisant les faits prosodiques. La notion de suprasegmentalité a été proposée par Hockett (1942) : « Features which clearly follow each other in the stream of speech are segmental. Those which clearly extend over a series of several segmental groupings are suprasegmental. ». Ce terme a eu beaucoup de succès auprès des linguistes traitant des faits prosodiques par opposition aux phonèmes. Martinet (1991) dit qu'est suprasegmental « tout ce qui n'entre pas dans le cadre phonématique », c'est-à-dire tout ce qui au niveau phonique de la langue ne se prête pas à l'étude phonologique. Ducrot & Todorov (1972) reprennent cette idée « Tous les faits phoniques qui échappent à l'analyse des phonèmes et des traits distinctifs » ont d'après eux un caractère suprasegmental. Dubois (1994) donne une définition de suprasegmentalité en associant le trait suprasegmental au trait prosodique. Selon cette dernière définition, c'est « une caractéristique phonique qui affecte un segment plus long que le phonème : l'accent, l'intonation, la durée sont des traits suprasegmentaux. » .

Il est aisé de surprendre dans cette constatation une contradiction : si l'on dit qu'un trait suprasegmental couvre un segment plus long, ce dernier ne reste qu'un segment. Il n'y a donc pas de raison de distinguer notionnellement un segment d'un autre. C'est pourquoi nous préférons à ces définitions qui proposent une dichotomie entre caractère segmental et suprasegmental celle de Vaissière (1980, p. 530) :

« Il est difficile de donner une définition simple du mot 'prosodie'. Pour notre part, la 'prosodie' recouvre les phénomènes de variations dans l'actualisation des phonèmes (qui sont des unités 'idéales') ; ces variations peuvent être décrites sur le plan acoustique (description de l'évolution de la courbe de fondamental, des durées des segments successifs et de l'intensité comparée des phonèmes), sur le plan perceptuel (perception du rythme des phrases, de leur 'mélodie', d'accent, d'intonation. . .) et sur le plan fonctionnel (fonctions linguistiques et para-linguistiques de ces variations. »

Les phénomènes de variations dans l'actualisation des phonèmes sont nombreux : la courbe mélodique, la durée, l'intensité, la dynamique spectrale de chaque son sont autant de degrés de liberté offerts à la structuration intonative et rythmique d'un énoncé.

Dans le cadre de cette thèse, nous nous intéresserons aux variations de **fréquence fondamentale** et de **durée** (cf. section 1.1.3) liées à deux fonctions communicatives essentielles pour les systèmes de génération de la prosodie en synthèse de la parole : véhiculer la modalité de la phrase et la structuration de l'énoncé. Nous avons étendu l'étude de la modalité, où l'on distingue classiquement *déclaration* et *question*, à quelques attitudes (cf. section 1.1.2.2).

Ces deux paramètres participent à la perception de la **hauteur** ("pitch") et de la **durée perçue** sans toutefois qu'on puisse établir une association directe entre paramètres physiques et corrélat perceptif (Botte, Canévet, Demany & Sorin 1989, Fraisse 1974, Fletcher & Munson 1933). Les notions d'**intonation** et de **rythme** désignerons respectivement les variations à long terme de hauteur et de durée perçue. Nous avons délibérément choisi cette définition très large de l'intonation afin de la caractériser non pas par ses fonctions, mais par sa substance. En effet, les auteurs sont souvent en désaccord sur les fonctions assumées par l'intonation - par exemple : Bolinger (1989, p.3) exclut l'étude du système tonal du cadre intonatif - et sur leur importance relative. Ils ont même douté de son statut linguistique.

Parmi les diverses fonctions intonatives et rythmiques, **l'accentuation** consiste à mettre en relief un certain nombre de syllabes par rapport à leur environnement immédiat. Étant donné l'importance de la proéminence dans les études prosodiques, nous allons détailler les diverses catégories qui ont été classiquement distinguées.

1.1.1.2 L'accentuation

Le terme *accentuation* regroupe les phénomènes de mise en relief dont les manifestations, les domaines et les fonctions sont différents selon la langue étudiée. Du point de vue de la substance, l'accent est à envisager comme un complexe de facteurs physiologiques (production) et acoustiques (transmission) qui confèrent à une syllabe donnée une proéminence (perception) plus grande qu'à celles qui l'entourent.

On distingue généralement deux catégories de langues :

- Les langues à **accent fixe** pour lesquelles la position accentuable du mot (accent d'insistance excepté) est toujours placée sur une syllabe déterminée (la syllabe finale du mot pour le français).
- Les langues à **accent libre** pour lesquelles la position accentuable dépend de la fonction lexicale et de la structure morphologique du mot.

La particularité de ces dernières est de posséder un accent lexical.

Accent lexical

Cet accent, comme son nom l'indique, fait partie intégrante de l'unité lexicale[1]. Sa non-réalisation et/ou un placement erroné peut conduire à une identification incorrecte du mot[2].

Ainsi, il permet de distinguer des couples homophones par sa seule position comme le montre les exemples qui suivent :

- Italien : an**co**ra : *« encore » vs* **an**kora : *« ancre »*
- Anglais : under**ta**king : *« entreprise » vs* **un**dertaking : *« pompes funèbres »*

Cependant, l'accent lexical n'est pas toujours accompagné d'une variation mélodique ("pitch accent") comme l'explique Bolinger en rappelant, pour l'anglais, sa conception du "stress" (1989, p. 214) :

« Since accent is thus often not realized at all, we speak of the *potential* for it to be realized on a given syllable, rather than the acual realization, as the distinguishing feature, and for that purpose we use the term of STRESS. Thus *úndertaking* is stressed on its first syllable, and also accented on that syllable *when* the speaker decide to accent it. »

[1]Comme le système tonal qui utilise un inventaire de mouvements mélodiques distinctifs (1989, p. 214) : « Accent, in its arbitrary guise, is like the arbitrary tone of a tone language : it can be used to distinguish words. ».

[2]L'identification résulte d'un large faisceau de contraintes linguistiques : l'accent peut donc parfois être omis si le contexte au sens large permet la désambiguïsation.

Pour Bolinger, le *stress* correspond donc à une syllabe qui a la potentialité d'être accentuée par un mouvement mélodique. Cette définition est d'ailleurs confirmée par Sluijter & van Heuven (1996, p. 2471) « Stress is a structural, linguistic property of a word that specifies which syllable in a word is the strongest. In our view "stressed" refers to syllables which are the potential docking sites for accent placement. » et Campbell & Beckman (1997) « ...the stress in English is nothing more than a structural marking of potential for a syllable to bear pitch accent ». De même, Kohler (1986, p.460) distingue les notions de *lexical stress* et de *sentence stress*.

Si la réalisation d'un accent sur une syllabe avec le trait *+stress* dépend de l'intention de communication du locuteur, ce trait est cependant toujours réalisé sur les noyaux vocaliques de cette syllabe par une mise en relief sur d'autres paramètres prosodiques tels que le timbre - " full" *vs* "reduced" -, la durée et l'intensité.

Cette position accentuelle constitue, en outre, un *attracteur* pour les proéminences d'intonation comme nous allons le voir par la suite.

Accent emphatique

Une syllabe peut être mise en relief pour mettre en valeur de façon subjective un mot. On parlera alors d'accent emphatique. Virtuellement, toute syllabe peut recevoir un tel accent comme, par exemple, lors d'une correction de prononciation. D'autre part, il y a de nombreux moyens syntaxiques, phonétiques (coups de glotte...) et prosodiques (pauses antéposées...) autres qu'un accent prosodique pour mettre en valeur un élément du discours.

Classiquement, on distingue l'**accent d'insistance** et l'**accent contrastif** :

- L'accent d'insistance n'altère pas le sens général de la séquence. Il crée dans la chaîne parlée un contraste d'ordre quantitatif. Pour le français, cet accent est principalement placé sur la première syllabe du mot.
- L'accent contrastif ne vise pas simplement à mettre en relief le mot mais également à le mettre en contraste avec d'autres items lexicaux présents dans la phrase ou implicites dans le discours. Par exemple : « Je veux **deux** baguettes. » (pas une ou trois) *vs* « je veux deux **baguettes**. » (pas deux flûtes).

Accent "d'intonation"

Accent primaire : L'accent "d'intonation" est également appelé accent primaire, accent principal ou accent de phrase. Il est l'outil privilégié de la structuration de l'énoncé. Comme nous l'avons vu ci-dessus, les langues imposent une place spécifique à cet accent :

– Dans les langues à accent libre, seules les syllabes fortes ("stressed") peuvent le recevoir.
– Dans les langues à accent fixe, l'accent "d'intonation" est toujours placé sur une syllabe déterminée, comptée à partir du commencement ou de la fin du groupe accentuel. Le groupe accentuel est défini comme un ensemble de mots recevant au moins un accent "d'intonation". Cette unité est constituée par un ou plusieurs **syntagmes accentogènes** (Garde 1968) formés d'un mot lexical éventuellement accompagné de **clitiques** qui ont la propriété de ne pas pouvoir recevoir l'accent "d'intonation". Le clitique est dit **proclitique** s'il s'intègre au le groupe accentuel qui le suit et **enclitique** s'il s'intègre au groupe accentuel qui le précède. Le regroupement des syntagmes accentogènes pour former un groupe accentuel est régi essentiellement par des contraintes rythmiques.

Accent secondaire : En français, le syntagme accentogène se termine par l'accent "d'intonation" mais il n'est pas forcément le seul accent de ce syntagme. La proportion de syllabes accentuées *vs* inaccentuées est de l'ordre de 30% (Pasdeloup 1990, Wenk & Wioland 1982). Le groupe accentuel serait donc d'environ de trois à quatre syllabes. Dans les mots polysyllabiques, certains accents dits **accents secondaires**, ne résultant pas d'une focalisation, apparaissent (voir revue de question dans Pasdeloup 1990, p.93–218). Leur position est très variable : majoritairement initiale chez Rossi (1985), plutôt localisée sur l'antépénultième chez Garde (1968). Pasdeloup, quant à elle, montre que sa position est régie par de nombreux facteurs : nombre de syllabes du mot, présence/absence d'une attaque consonantique à l'initiale du mot, contexte accentuel immédiat, morphologie interne... La dernière syllabe d'un morphème non final d'un mot polymorphémique est ainsi un fort attracteur pour cet accent qui joue ainsi un rôle de démarcation, comparable à celui de l'accent primaire pour le mot. Cette fonction est donc à rapprocher de la notion de "secondary stress" versus "primary stress" des langues à accent libre.

L'accent secondaire se caractérise généralement d'un point de vue acoustique par un ictus mélodique faiblement ancré sur la syllabe accentuée et sans allongement significatif. Nous reprendrons la conclusion de Pasdeloup sur la fonction présumée de cet accent secondaire qu'elle différencie de celle de l'accent primaire qui assume une fonction clairement linguistique : « L'accent secondaire aurait pour fonction de réguler la production accentuelle de telle sorte qu'elle soit conforme aux normes biologiques et psychologiques de production du rythme. » (Pasdeloup 1990, p.218). Ainsi, cet accent semble remplir des fonctions rythmiques, Sa position est régie par des contraintes essentiellement phonotactiques et morphologiques.

Cet accent secondaire apparaît rarement dans nos corpus. Nous en trouvons principalement sur les monomots en initiale ou en frontière morphologique interne (cf. figure 1.1). Nous avons pourtant veillé à varier de manière systématique la position de la frontière morphologique interne des mots polymorphémiques dans les monomots de nos corpus.

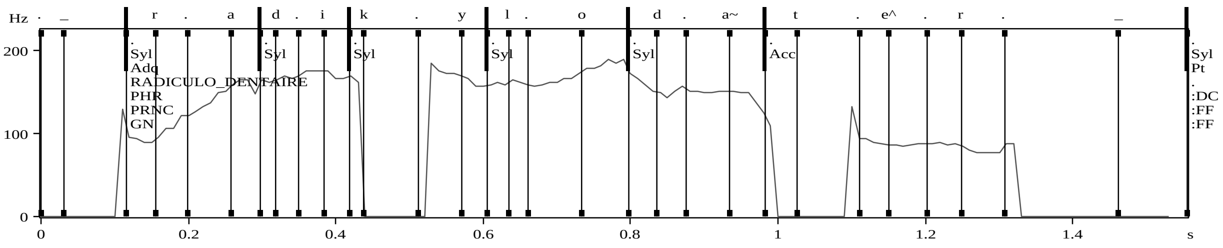


FIG. 1.1: Exemple de réalisation d'un accent en frontière morphologique interne sur la syllabe [lɔ] dans le monomot « radiculodentaire ». Notez le faible accroissement de la durée syllabique et le petit ictus mélodique réalisé.

Accentuation et intonation : Dans le cas du français, les auteurs s'accordent pour associer l'accent "d'intonation" à la fin d'un mouvement intonatif. Il ne peut donc pas se distinguer de l'intonation ni par sa nature ni par sa fonction (Rossi 1980, p.17-19). Notre position est plus contrastée : **Nous considérons l'accent "d'intonation" comme la partie émergée de l'iceberg *intonation***, c'est-à-dire comme une proéminence perceptive liée à l'évolution rapide et coordonnée des paramètres prosodiques dans un mouvement intonatif. Nous verrons que les accents conçus de cette manière ne sont pas forcément finaux (voir le mouvement mélodique associé au *doute-incrédulité* en section 2.2).

1.1.1.3 Le rythme

Pour certains auteurs, le rythme relève de l'étude distributionnelle des accents. Les corrélats perceptifs des paramètres prosodiques sont alors simplement réduits à l'étude de l'intonation et de l'accentuation (Rossi 1993). Nous allons détailler ce que nous entendons par rythme : en partant d'une définition de la durée phonémique, nous établirons la notion de rythme en insistant sur les éléments qui nous ont conduits à opter pour le Groupe Inter P-Center comme unité minimale macrorythmique.

La durée

La durée est le paramètre le plus difficile à préciser car rien n'indique comment le système de contrôle, de production ou de perception de parole mesure le temps. Pour pouvoir parler de durée, il faut fixer au moins deux événements entre lesquels le temps s'écoule. Lorsque le choix d'une paire d'événements représentatifs du phénomène observé est effectué, on peut faire une mesure objective du temps : on parlera alors de **durée objective**. A la composante perceptive, on associera le terme de **durée perçue**.

Mesures de la durée objective : Lorsque ce phénomène est un signal de parole, on peut retenir une dizaine d'événements acoustiques sur lesquels appuyer une segmentation (Abry, Benoît, Boë & Sock 1985). Si un certain nombre d'études se sont fondées sur des représentations de phasage entre ces divers événements (Abry, Orliaguet & Sock 1990, Löfqvist 1991, Sock & Löfqvist 1995), la plupart des études prosodiques partent d'une segmentation rudimentaire du signal de parole en durées phonémiques. Dans cette approche, le phonème est vu comme une suite de phones homogènes (Vigouroux & Caelen 1985) dont la structure séquentielle peut être facilement décrite par un automate de type chaîne de Markov (Wightman & Talkin 1997). Si cette technique de segmentation et d'alignement automatique est largement répandue, il faut, néanmoins, bien préciser quelle est l'appartenance de certains phones dits de transition. Ainsi, nous incluons toujours les segments d'explosion et d'écoulement glottique dans la réalisation d'une occlusive sourde. L'aspiration est, quant à elle, incluse dans la durée vocalique comme l'a suggéré Fant (1973, p.119) : « ...we must add to the length of the vowel, after a voiced stop, the duration of the voiceless or weakly voiced interval between release transient and visible onset of the following vowel. ». Dans l'exemple présenté en figure 1.2, le choix de la frontière *i2* conduirait aux durées de [t] et [i] respectivement égales à 140 et 30 ms, alors que notre choix de la frontière *i3* conduit à 70 et 100 ms. Notre segmentation est donc essentiellement guidée par des contraintes d'homogénéité spectrale maximale des sons. Les détails phonétiques tels que les dévoisements partiels ou complets non encodés phonologiquement sont traités pour l'instant dans le système de synthèse par l'emploi de polysons adéquats sans une gestion particulière de leur proportion.

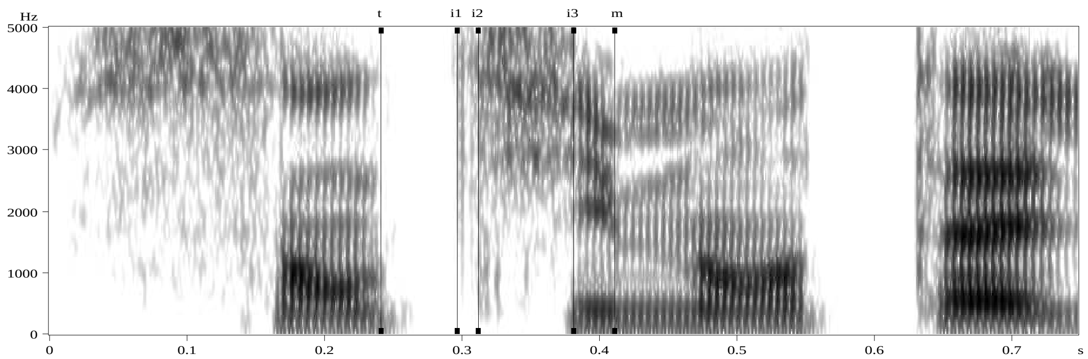


FIG. 1.2: Sonagramme de la séquence [sãtimã] dans le mot "sentimentaliser". Les étiquettes *t* et *m* indiquent le début des sons correspondants. Les étiquettes *i1*, *i2* et *i3* indiquent les débuts candidats de [i].

Niveaux d'observation : Le rythme en parole a été étudié à l'aide de mesures diverses : la suite des durées des phonèmes du discours incluant les silences, la durée des syllabes ou des pieds, l'intervalle entre syllabes accentuées... Nous reviendrons en section 1.3.2.1 sur les modèles de prédiction de la durée des phonèmes qui opèrent, sans supposer une régulation explicite du temps, sur des unités rythmiques plus larges que le phonème. Or la durée acoustique résulte des gestes articulatoires permettant de mettre en forme notre conduit vocal et ainsi de produire les sons distinctifs du langage articulé. Le système de production impose notamment des fréquences d'oscillation propres à chaque articulateur : globalement, plus un geste est ample et complexe, plus il dure longtemps (comparer, dans le tableau 1.1, les durées des voyelles ouvertes *vs* fermées, nasales *vs* orales, fricatives *vs* occlusives...). Ces limites expliquent d'ailleurs quelques asymétries entre les systèmes de production et de perception qui contredisent la théorie motrice de la perception (MacKay 1987) : la parole compressée temporellement reste intelligible jusqu'à 400 mots à la minute (MacKay 1987, p.352), alors que nous sommes incapables de la reproduire. De même, tous les sons ne peuvent être tenus pendant la même durée : le maintien du voisement dans une occlusive voisée en français nécessite une expansion du conduit bucal. Chaque son a donc un facteur de compressibilité/d'expansion qui lui est propre. Cependant malgré une grande variabilité de contenu phonématique, des expériences ont montré que l'humain est capable de synchroniser une activité langagière avec une autre activité : rythmique, langagière ou autre (Allen 1975, Berthier, Abry & Lalouache 1991, Fraisse 1974, Gérard & Auxiette 1992).

Rythme et unité rythmique

Le rythme : Dans l'acception la plus générale du terme, le rythme se définit par la **répétition** d'un **repère** à des intervalles réguliers et perceptibles. En parole, l'intervalle défini par deux repères consécutifs constituera l'unité rythmique minimale. Celle qui reste la plus couramment employée est la syllabe. Le mouvement quasi-périodique de la mâchoire a donné très tôt un rôle prépondérant à la syllabe dans l'organisation rythmique de la parole en production comme en perception(Fraisse 1974). Nous choisirons un autre candidat déjà proposé par Barbosa (1994) : le Groupe Inter P-Center (Marcus 1975).

L'isochronisme : La tendance à l'isochronie semble être l'une des caractéristiques rythmiques essentielles et a donné lieu à une large discussion sur l'**isochronisme** et l'**isosyllabisme** (Pike 1945, Lehiste 1977, Wenk & Wioland 1982). Dans les langues pour lesquelles le rythme possède une organisation temporelle syllabique, les syllabes sont produites à intervalle de temps régulier ; pour les langues à organisation temporelle fondée sur le pied, les syllabes accentuées sont isochrones.

Mais la notion d'isochronisme en production est largement controversée (Nooteboom 1991, Fant, Kruckenberg & Nord 1991). Elle semble être beaucoup plus liée à la perception comme l'écrivent Fant et ses collaborateurs (1991) :

« Stress timing is not a matter of physical isochrony of interstress intervalls but a perceptual dominance of heavy syllables, the succession of which is sensed as quasi-periodical. »

Le *Perceptual-center* : Marcus (1975) a proposé un modèle de perception qui tient compte de cette quasi-périodicité en définissant un nouveau paradigme en perception. L'auteur a montré que les auditeurs étaient capables d'ajuster l'intervalle de temps entre deux occurrences syllabiques (Voyelle *vs* Consonne-Voyelle) pour percevoir cette alternance comme isochrone. Ses expériences lui ont permis de définir le **Perceptual-center** (PC) ou Centre de Perception de l'isochronisme qui agit comme un point de repère psychoacoustique de la perception de l'isochronisme en parole.

On peut reprocher à Marcus de ne pas traiter de la parole continue et de forcer les sujets à se concentrer sur une synchronisation bien particulière. Mais des expériences à base de "tapping" et contenant des stimuli de parole plus "écologiques" ont conduit à des résultats similaires comme indiqué ci-après.

Corrélat acoustique du *P*-Center : Allen (1975) a montré que les sujets plaçaient leur coup ("beat") au voisinage de l'onset des syllabes *+stress*. Pour le français, Berthier et al. (1991) ont mis en évidence une relation temporelle entre la percussion d'un couteau sur une flûte et l'établissement vocalique qui la suit (entre 0 et 100 ms), mais jamais avec la fin de la voyelle qui la précède. Pompino-Marschall (1989) a réalisé une expérience pour calculer la position absolue de ces PC. L'auteur aboutit à la conclusion qu'en dépit de la variété des consonnes et des voyelles, les PC sont localisés au voisinage de l'établissement de la voyelle que ce soit dans des expériences de production ou de perception. Cependant, il montre qu'il est difficile de le repérer en parole continue (Pompino-Marschall 1992).

Scott (1993) a implémenté un modèle pour la localisation du PC sur la base de la dérivée première de la fonction d'intensité. D'après son modèle, un accroissement significatif de l'intensité dans la bande de fréquence $[196Hz, 1638Hz]$ détermine de façon optimale la position du PC.

Dans notre étude, nous faisons coïncider le PC avec l'établissement vocalique. Les unités phonémiques entres deux PC adjacents constitueront le **Groupe Inter P-Center** (GIPC). Nous montrerons, dans la suite de ce mémoire, que le GIPC constitue un candidat plus performant que la syllabe pour le contrôle macrorythmique.

1.1.1.4 L'intonation

De nombreux auteurs ont réduit l'étude de l'intonation à celle des fonctions de communication assumées par la mélodie : évolution de la fréquence fondamentale au cours du temps. Nous situons sa nature dans un domaine clairement perceptif[3]. L'étude de l'intonation inclut donc par définition les phénomènes accentuels liés à la perception de proémi-

[3]Par abus de langage, nous emploierons parfois dans ce mémoire le terme de contour intonatif pour décrire une courbe de fréquence fondamentale.

nences tonales. Ces phénomènes émergents sur des portions réduites de l'énoncé (la syllabe en général) sont alors inclus dans des mouvements plus vastes ou des structures phonologiques hiérarchiques. La première approche considère les accents comme des phénomènes émergents, comme les points d'ancrage de contours plus vastes avec la structure linguistique et phonologique de l'énoncé voire de la structure du discours. La seconde approche s'appuie sur ces événements saillants pour lever une structure phonologique plus profonde. Une comparaison plus fine de ces approches phonétiques et phonologiques de l'intonation sera proposée en section 1.2.1. Dans la littérature, l'intonation est cependant plus souvent définie par ses fonctions que par ses manifestations prosodiques.

1.1.2 Les fonctions de l'intonation

Cette section est destinée à donner un aperçu des différentes fonctions linguistiques et para-linguistiques ou expressives de l'intonation en mettant en évidence les problèmes que pose une analyse séparée de ces deux composantes.

En s'appuyant sur les propos de plusieurs auteurs (Di Cristo 1978, Ohala 1996, Quilis 1979, Bolinger 1972, Léon 1993), nous pouvons élaborer une classification des faits prosodiques en quatre catégories :

1. Un niveau fondamental de l'expression des signaux nécessaires à la survie de l'individu. On classera parmi ces signaux ceux qui expriment la sécurité ou au contraire la menace.
2. Un degré où l'intonation est spontanée, naturelle et stimulée psychophysiologiquement. Il exprime l'état du sujet sans que celui-ci veuille l'exprimer consciemment. A ce niveau, les réalisations prosodiques reflètent généralement l'état émotionnel du locuteur (joie, tristesse. . .).
3. Un niveau où ces formes naturelles d'intonation sont employées intentionnellement. C'est là que s'opère la reconnaissance de ce que le sujet pense ou interprète sur l'auditeur, le contenu de son message ou sur lui-même. Dans ce cas, la forme de l'expression fait probablement l'objet d'un apprentissage et peut donc varier selon la langue ou la culture de l'individu. On trouve à ce niveau l'expression des attitudes du locuteur.
4. Un niveau référentiel - où apparaîssent des unités prosodiques assumant des fonctions morphologiques, syntaxiques, ou plus généralement structuratrices de l'énonciation - qui entre dans la structure particulière de chaque langue.

Cette classification permet de distinguer au moins deux types de fonctions de l'intonation : la première que nous appellerons expressive (1, 2, 3) et la seconde que nous qualifierons de syntaxique (4).

1.1.2.1 Fonction syntaxique

L'analyse linguistique de l'intonation implique une segmentation en unités qui constituent un système d'oppositions. Les auteurs ne s'accordent généralement pas sur la taille et le nombre de ces unités mais tous signalent l'importance de la fin du contour mélodique avec ses mouvements ascendants ou descendants. Les fonctions syntaxiques se réalisent à travers ces contours intonatifs et peuvent être caractérisées comme intégrative, démarcative ou délimitative, hiérarchisante et modale (Rossi 1985).

Par son pouvoir intégratif, l'intonation transforme une chaîne de mots en unités communicatives. Elle permet d'accorder à une suite non grammaticalement correcte le statut d'énoncé intono-syntaxiquement bien formé. D'une manière complémentaire, l'intonation joue un rôle de délimitation : elle aide à segmenter le discours en un nombre déterminé d'unités délimitant ainsi les frontières de certains constituants syntaxiques. Par son pouvoir hiérarchisant, l'intonation contribue à séparer les informations importantes des informations secondaires. Les unités intonatives du français nécessaires à l'actualisation de ces fonctions sont définies par Rossi & al. (1981) comme des intonèmes de continuité et de finalité divisés en continuatif et conclusif majeurs ou mineurs.

Le mode est une catégorie grammaticale traduisant le type de communication institué par le locuteur entre lui et son interlocuteur (nous verrons plus loin qu'il est délicat d'instaurer une frontière entre fonction modale et attitudinale). Par le mouvement mélodique de la dernière partie de l'unité intonative sont signalées des valeurs de finalité (mouvement descendant) ou de continuité (mouvement ascendant) dont l'importance pour l'identification du mode phrastique (assertion *vs* question) a fait l'objet de nombreuses recherches. Ainsi, Delattre (1969) dégage la fonction modale significative de l'intonation au moyen d'oppositions de sens reposant uniquement sur la forme et le niveau des courbes mélodiques.

1.1.2.2 Fonction expressive

Par l'intonation, un locuteur ne transmet pas simplement des informations structurales et modales. L'intonation renseigne également sur l'état émotionnel (fonction émotionnelle) ou l'attitude du locuteur vis-à-vis de son discours ou de son interlocuteur (fonction attitudinale). Avant d'aller plus loin, nous nous proposons d'identifier les principales différences entre émotion, attitude et modalité. En effet, dans le cadre de notre recherche, nous nous intéressons aux faits prosodiques intentionnels associés à un code linguistique conventionnel appris par le locuteur et donc dépendant de sa langue maternelle. Nous n'étudierons donc pas ici tout ce qui ne fait pas partie de ce code et tout ce qui échappe au contrôle du locuteur.

Pour appuyer notre démonstration, nous nous référerons aux propos de deux auteurs

(Ohala 1996, Fónagy 1987).

Ohala aborde la question de l'universalité de l'expression des émotions en soulignant le rôle essentiel de l'éthologie [4] pour construire une taxonomie de ce type d'expressions. Il envisage les trois niveaux suivants :

- L'expression des signaux nécessaires à la survie.
- L'expression des signaux exprimant l'état psychophysiologique du sujet.
- L'expression de l'opinion du sujet sur son propos, sur l'auditeur ou sur lui-même.

Ohala positionne donc le code émotionnel dans une **histoire naturelle** où l'association entre la cooccurence d'un son (ou les caractéristiques globales d'une suite de sons), et les caractéristiques physiologiques de l'émetteur ou le contexte dans lequel ils sont émis, se "sédimentarise" et passe ainsi au niveau conventionnel. Ce code pourra ainsi évoquer la correspondance directe puis la situation par le son ou la séquence elle-même. Le niveau d'abstraction final permettra alors d'y associer non plus un concept situationnel mais la position du locuteur vis à vis de ce concept. On passe donc de manière continue du « savoir » au « faire-savoir » puis au « faire-croire » (Austin 1962, Searle 1969 - réédité en 1985) par altérations successives de l'impression vers l'expression. Les éléments de base notés comme universaux par Ohala incluent le lien entre taille, poids et F0, relation que les auditeurs savent remarquablement bien faire (Lass, Hendricks & Iturriaga 1980). L'une des altérations possibles consiste, par réflexivité, à associer ces invariants sonores avec la confiance ou la déférence… Cet ancrage biologique des émotions permet de comprendre pourquoi des auditeurs humains naïfs sont capables d'identifier les messages émotionnels dans les vocalisations de macaques (Leinonen, Linnankoski, Laakso & Aulanko 1991) et pourquoi ces performances au cours de l'enfance (Linnankoski, Laakso, Aulanko & Leinonen 1994) se développent de manière similaire pour la reconnaissance des émotions humaines (van Bezooyen 1984).

Fónagy, quant à lui, décompose en trois catégories les informations émotionnelles véhiculées par la prosodie :

- Les émotions primaires (« primary emotions ») comme la colère, la haine, la joie, la peur…
- L'expression des attitudes sociales (« social attitudes ») comme l'incrédulité, l'indignation…
- Les patrons intonatifs marquant la catégorie modale.

Il démontre le caractère universel des émotions par des tests d'identification multilingues sur des stimuli délexicalisés. Par contre, les attitudes du français ne sont pas identifiées correctement par des sujets hongrois. Les patrons modaux représentent, selon l'auteur, le niveau le plus élevé d'organisation linguistique de l'énoncé. Cependant, l'auteur insiste sur les difficultés à effectuer une séparation claire entre attitudes et modalités qu'il considère comme les attitudes "les plus générales".

[4] Science qui a pour objet l'étude du comportement animal.

A partir de ces deux répartitions, nous proposons de représenter, sur un axe parcourant le degré d'universalité et l'intentionnalité, les expressions d'émotions, d'attitudes et de modalités (cf. figure 1.3).

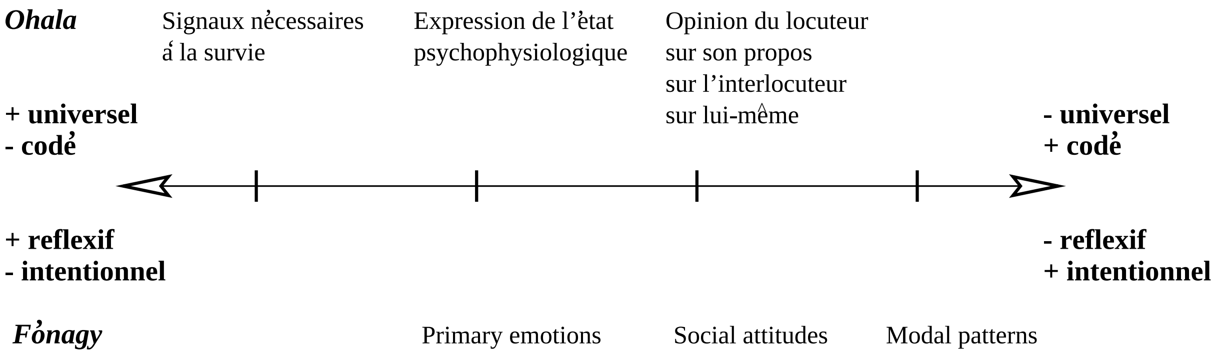


FIG. 1.3: Continuum émotions-attitudes-modalités (selon Alt 1997, p.13).

1.1.2.3 Discussion

Delattre (1972, p. 159), dans l'introduction de son article sur la fonction distinctive de l'intonation sépare deux écoles :

- La première considère l'intonation comme un phénomène linguistique : les différences de niveau et de forme terminale dans les inflexions de la voix sont porteuses de sens et participe à la structuration du discours.
- La seconde estime que les contours intonatifs ne sont pas vraiment distinctifs : ils sont principalement le reflet de l'attitude du locuteur.

Delattre pose ici le problème de la séparation théorique entre l'analyse de la fonction linguistique et de la fonction para-linguistique de l'intonation.

Pour les adeptes de la première école, la fonction essentielle de l'intonation est celle qui permet la structuration du discours. Pour Lieberman (1965), la fonction primaire de de l'intonation est celle qui fournit des indices acoustiques permettant à l'auditeur de segmenter la parole en blocs pour le processus syntaxique. Plus récemment, A. Cutler note dans son introduction au symposium *Prosody in situations of communication* : « Two of the major uses of prosodic information [. . .] are to encode salience and segmentation » (Cutler & Norris 1991). Les adeptes de la seconde école, quant à eux, privilégient la fonction expressive de l'intonation et s'intéressent principalement au caractère universel de celle-ci.

La raison majeure de cette séparation théorique pourrait provenir de la discussion sur le statut linguistique, c'est-à-dire le caractère conventionnel de l'expressivité de l'intonation. Les résultats des tests de perception de Fónagy (1987) montre bien que contrairement aux émotions, les formes attitudinales du français ne sont pas reconnues par des sujets étrangers et feraient donc partie des signes conventionnels codés linguistiquement.

Si l'on ajoute qu'il n'est pas possible dans la pratique d'observer une intonation purement grammaticale (sans attitude) (Bolinger 1951, p. 205), nous adopterons une position selon laquelle l'intonation n'est jamais purement grammaticale ou expressive. C'est pourquoi nous étudierons ensemble ces deux fonctions en essayant d'abord, de démontrer le caractère linguistique des attitudes et, de révéler à partir d'un corpus de phrases, produites selon différentes consignes d'expressivité, les interactions entre les dimensions structurales et les informations attitudinales de l'intonation.

1.1.3 Caractérisation de la substance prosodique

1.1.3.1 Les paramètres acoustiques

La substance acoustique de la prosodie est définie classiquement par quatre paramètres :

La fréquence fondamentale (notée F0) : corrélat acoustique de la fréquence laryngienne. Elle est essentiellement la conséquence de la tension des cordes vocales et de la pression trans-glottique.

La durée : mesure d'un intervalle de temps donné, elle est inhérente à tout processus physique. Cette notion comprend le débit de parole, la durée et la répartition des pauses, les allongements syllabiques...

L'intensité : elle correspond à l'énergie contenue dans le signal au cours d'un intervalle de temps donné.

Le timbre : il est le corrélat essentiel de la distinction entre syllabe forte ("stressed") et faible ("unstressed"). Un des phénomènes le plus étudié est la réduction vocalique, c'est-à-dire la déviation entre la cible acoustique atteinte au centre de la voyelle et une cible idéale lorsque la voyelle est produite hors-contexte ou accentuée. Cette réduction est d'ailleurs l'un des principaux enjeux de la prosodie articulatoire (Beckman & Edwards 1992, Loevenbruck & Perrier 1993, Fougeron & Keating 1997).

Il convient de distinguer plusieurs causes de variations de ces quatre paramètres. On peut établir schématiquement une distinction entre les variables contrôlées *i.e.* les variations régies par le contenu linguistique du message ; et les variables non contrôlées inhérentes au mécanisme de production. Les premières sont corrélées aux changements de valeurs intonatives, rythmiques ou accentuelles. Les secondes sont liées à des contraintes **intrinsèques** et **co-intrinsèques** dues au mode de production de chaque phonème et de son contexte.

1.1.3.2 La microprosodie

Les contraintes intrinsèques et co-intrinsèques ont été regroupées sous le terme de microprosodie (Di Cristo 1985) et entraînent les effets suivants :

La micromélodie concerne les variations locales de la fréquence fondamentale et peut avoir deux sources. L'une est liée à la F0 intrinsèque des voyelles - plus haute pour les voyelles fermées que pour les voyelles ouvertes (Rossi, Di Cristo, Hirst, Martin & Nishinuma 1981). La seconde est relative aux effets co-intrinsèques c'est-à-dire à l'influence de certains segments sur les segments adjacents.

La durée segmentale intrinsèque est liée principalement à l'amplitude et la complexité des gestes articulatoires sous-jacents (cf. page 9). Du fait de la coarticulation - fortement anticipatrice - la durée de chaque son est aussi largement influencée par le contexte phonétique proche et surtout dominé par le contexte droit. Ceci explique que les modèles de prédiction de la durée sur le phonème incorporent beaucoup plus d'indices contextuels et phonotactiques que linguistiques dans les prédicteurs (cf. section 1.3.2.1).

L'intensité est un paramètre fortement dépendant de la nature intrinsèque des sons.

Le timbre est le principal corrélat de l'identité des sons notamment des voyelles. L'étude des variations prosodiques du timbre suppose que l'on possède des références ou prototypes. La coarticulation engendre naturellement des réductions. Ainsi, dans les séquence [gag] ou [iai], le spectre instantané prélevé au centre du [a] correspond généralement à la cible prototypique du [ɛ] sans que le phonème /a/ soit mal perçu. Ceci est dû au contexte avant. L'atteinte de la cible prototypique du [a] correspond plutôt à une version "hyperarticulée" et peut être interprêtée comme l'un des corrélats importants de l'accentuation.

Notons cependant que la dichotomie *contrôlé/non contrôlé* a bien des limites que nous allons expliciter par un exemple bien connu : la notion de fréquence intrinsèque des voyelles mentionnée ci-dessus semble être un universel (Whalen & Levitt 1995). Mais le fait que les voyelles hautes [iu] aient une fréquence intrinsèque supérieure d'environ 10 Hz aux voyelles basses [a], va cependant à l'encontre de ce que la théorie acoustique prédirait. En effet, un abaissement de la pression transglottique, dû à une diminution de l'aire de constriction principale du conduit vocal, devrait logiquement faire baisser la fréquence fondamentale (comme c'est le cas pour les occlusives voisées ou les "glides"). L'hypothèse qui semble retenue est l'interaction langue-larynx (Honda, Hirai & Dang 1994) : la langue élevée pour les voyelles hautes exerce une tension musculaire importante qui va élever le cartilage thyroïdien et provoquer ainsi un changement de la longueur des cordes vocales. L'effet combiné des couplages articulatoires et acoustiques de l'articulation vocalique sur le fonctionnement laryngé est donc très complexe. Notons simplement que cette différence de 10 Hz n'est observée, dans la majorité des références utilisées par Whalen & Levitt, que sur les syllabes, mots isolés ou logatomes dans une phrase porteuse. Nous n'avons quant à nous jamais observé de telles différences, et peu de systèmes de stylisation mélodique effectuent une quelconque correction liée à la nature des voyelles de l'énoncé. Nous pensons que la fréquence intrinsèque des voyelles est compensée volontairement à des fins communicationnelles de manière à laisser une courbe de fondamental peu entachée par le contenu segmental. Le fait que ces contraintes communicationnelles soient absentes et que la tâche de lecture soit plus centrée sur l'identité phonétique des segments dans les items cités par Whalen & Levitt, peut inciter le locuteur à utiliser ce contrôle non plus à des fins de neutralisation mais à des fins de mise en valeur phonétique, voire de segmentation.

La question qui se pose donc dans une étude des fonctions linguistiques de la prosodie est la suivante : faut-il éliminer des paramètres prosodiques les variations non contrôlées afin de ne pas les attribuer à tort à une valeur linguistique ? La réponse à cette question n'est pas claire comme nous venons de le voir et comme le note Vaissière (1988) sur des applications en reconnaissance automatique de la parole : « Present systems only include partial compensation of the phonetically conditionned variations, *i.e.* of the ones which are easiest to integrate. It is not clear whether or not such partial normalisation is better than no compensation at all (there is no known comparative studies). »

Certains auteurs comme Silverman (1987) avancent que les variations micromélodiques locales peuvent être d'une ampleur plus grande que les variations macromélodiques. D'autres ('t Hart, Collier & Cohen 1990), dans leurs études sur la perception, les négligent car selon eux il n'a pas été prouvé qu'elles influencent clairement la perception globale de la parole.

Dans la stylisation de la fréquence fondamentale que nous allons utiliser, nous n'appliquerons pas de correction locale des phénomènes micromélodiques. Nous essaierons plutôt de faire varier suffisamment le contenu phonémique des phrases de nos corpus afin d'éviter de capter certaines régularités liées à la micromélodie[5].

1.2 Représentations phonologiques et phonétiques de l'intonation

Les travaux théoriques portant sur l'intonation opposent généralement deux grandes familles :

- La première, de tradition phonologique, qui propose une description généralement ascendante des contours mélodiques à partir d'une représentation tonale discrète. Ces tons, éventuellement regroupés en séquences ("tunes"), sont alors supposés être des manisfestations de structures phonologiques plus profondes, construites sur les principes générativistes. Il reste ensuite à relier ces unités phonologiques au contenu linguistique et au sens global de l'énoncé.

[5]Ces régularités seront introduites dans notre modèle lors de la dernière phase de son apprentissage automatique.

– La deuxième approche suppose la primauté du sens qui donne les clés de lecture de l'apparente variabilité de la substance prosodique. Il convient ainsi de distinguer deux niveaux de lecture : le *quoi* et le *comment* ; c'est-à-dire identifier ce que l'intonation est capable de véhiculer et comment tous ces contenus, qui structurent fortement l'énoncé, négocient entre eux leur *temps ou fréquence de parole* sur le canal langagier. Cette hypothèse de partage de canal d'information inclut donc les notions de distribution de l'information, de redondance et de degrés de liberté, propriétés si appréciables des organismes vivants.

Récemment Rossi (1997) a commenté ces deux approches en distinguant : la tradition européenne, axée sur le contenu, et la tradition américaine, supposant une relative indépendance de la phonologie au contenu linguistique et au sens. Il est indiscutable que notre approche "européenne" du problème a été largement inspirée par la sémiotique Saussurienne et par l'Ecole de Prague. Notons que ce projet de théorie linguistique de l'intonation est de nouveau alimenté par les données récentes sur l'ontogénèse. Morgan et Demuth dans l'introduction du livre "Signal to syntax" dédié à l'acquisition du langage notent que : « [...] it is the signal that conveys words, phrases, sentences, indeed grammars to each new generation. » (Morgan & Demuth 1996, p.20)

Cette section évoque les travaux de chercheurs appartenant aux deux traditions et met en avant la nécessité de bâtir des systèmes de représentation fondés sur un traitement morphologique adapté aux fonctions linguistiques véhiculées par l'intonation (Bailly, Aubergé & Morlec 1997).

1.2.1 Emergence des représentations

1.2.1.1 Le point de vue phonologique : analyse ascendante

L'analyse ascendante vise à extraire de la cinématique des trajectoires un ensemble de paramètres prosodiques supposés véhiculer l'intonation, une suite d'événements saillants et définis *a priori* : de nombreuses méthodes de stylisation automatique de la ligne mélodique délivrant des suites de mouvements élémentaires ou de point-cibles ont été ainsi proposées.

Modèles configurationels

Crystal (1969) définit l'intonation d'une phrase en terme d'unités tonales. La plupart des théories phonologiques de l'intonation incluent ce type d'unité primaire appelée aussi "tonal group" par Halliday (1970), "word group" par O'connor & Arnold (1973) ou encore "intonation phrase" par Pierrehumbert (1980).

Crystal décompose chaque unité tonale en quatre composantes "prehead", "head", "nucleus" et "tail", propose quatre mouvements élémentaires 'rise', 'fall', 'rise-fall','fall-rise' et définit quatre combinaisons permises et phonologiquement distinctes de ces tons primaires par unité. O'connor & Arnold utilisent sept catégories de tons primaires. Halliday, quant à lui, classe les mouvements de fréquence fondamentale en cinq catégories de tons primaires auxquelles il ajoute deux tons composés obtenus à partir des cinq précédents.

Le modèle de description proposé par O'connor & Arnold a d'ailleurs été utilisé pratiquement tel quel, dans une des rares applications de ce modèle, à un système de décodage intonatif en identifiant chaque composante à un état d'un Modèle de Markov caché par Jensen et al (1994).

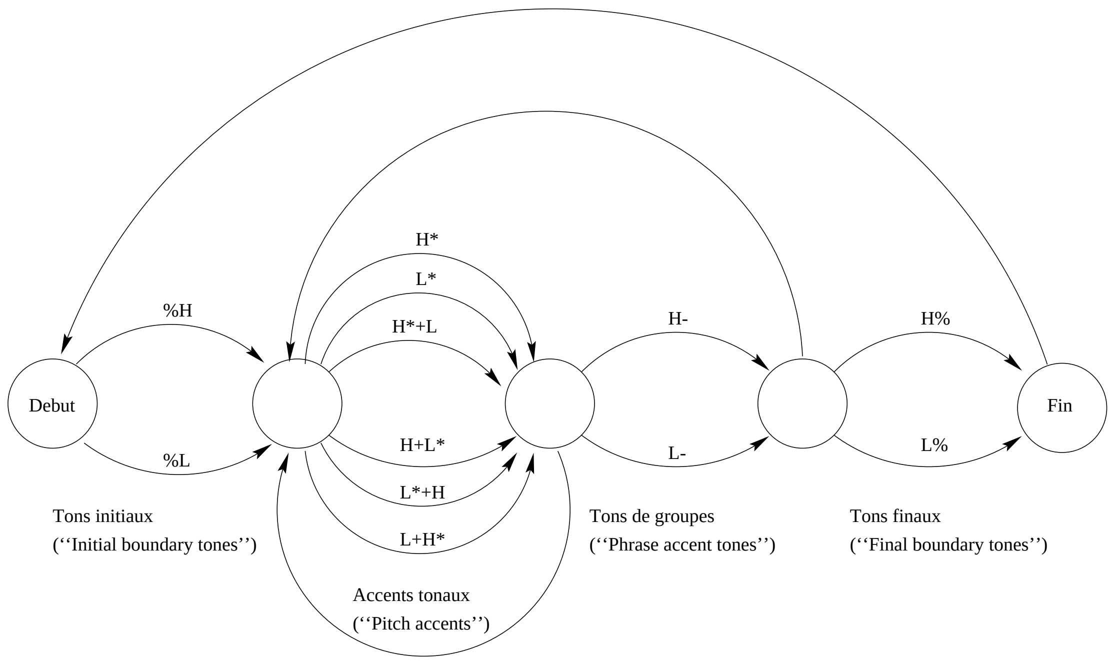


FIG. 1.4: Grammaire d'états finis spécifiant les séquences de tons de base acceptables (d'après Pierrehumbert 1980).

Modèles de séquences de tons

Pierrehumbert (1980) propose un modèle phonologique fondé sur une description de la courbe de F0 en une séquence de point-cibles ("pitch target") reliés par des fonctions d'interpolations. Ces point-cibles correspondent à la réalisation phonétique de segments tonaux distincts : des tons hauts (H) et des tons bas (L). On distingue classiquement les tons liés à l'accent tonique (et synchronisés à celui-ci par la marque '*'), les tons concluant une suite d'accents toniques (L- et H-) qui ont la mission de réinitialiser le mécanisme de descente graduelle des amplitudes accentuelles ("downstep"), et de préparer les tons de frontière finale (L% et H%). La séquence de segments tonaux possible est spécifiée par une grammaire d'états finis (voir figure 1.4). Ces séquences de tons sont associés à des unités phonologiques comme le mot prosodique ("prosodic word"), la phrase intermédiaire ("intermediate phrase"), et, la phrase intonative ("intonation phrase") caractérisée par une déclinaison et un ton final (Pierrehumbert 1979). Dans une perspective hiérarchique (Selkirk 1984), la phrase intonative est constituée de phrases intermédiaires, elles-mêmes constituées de mots prosodiques.

Le nombre de niveaux de la structure intonative est toutefois un sujet controversé. Wightman et al. (1992), sur la base de résultats de perception catégorielle, postulent trois niveaux au dessus du mot prosodique (contre deux pour Pierrehumbert). Ladd & Campbell (1991) montrent que la distribution des durées syllabiques serait modélisée plus précisément en utilisant quatre niveaux de structure prosodique au dessus du mot prosodique.

Silverman et al. (1992) ont proposé un outil de transcription de la prosodie *Tones and Break Indices* (TOBI) fondé non seulement sur des unités tonales mais également sur des indices de rupture correspondant au degré de connexion entre mots adjacents. Mais ce système de transcription ne propose plus, à proprement parler, une description phonétique et l'étiquetage fait appel aux jugements et aux interprétations linguistiques d'un expert.

Le modèle proposé par Hirst & Di Cristo, dont les contours sont esquissés dans de nombreuses publications et thèses (Hirst 1991, Sabio 1996, Campione, Flachaire, Hirst & Véronis 1997), vise à proposer une transcription formelle et inversible de la structure mélodique. Il s'appuie pour ceci sur un modèle de trajectoires, MOMEL (Hirst, Nicolas & Espesser 1991), permettant d'extraire automatiquement un certain nombre de point-cibles connectés par des fonctions splines quadratiques. Chaque point-cible est codé par un symbole soit indiquant un ton absolu : T ("top"), B ("bottom") ou M ("mid"), soit un ton relatif : U ("up"), D ("down"), S ("same"), H ("higher"), L ("lower"). Ce premier étage d'abstraction est ensuite complété par un niveau d'abstraction phonologique comprenant les niveaux classiques : unités tonales (TU) et unités intonatives (IU) dont les rapports et la fonction dans le discours sont indiqués respectivement par un choix approprié de tons relatifs et absolus.

L'analyse de l'intonation en terme de séquence de tons a été appliquée à plusieurs langues comme, par exemple, le hollandais (Gussenhoven 1991), le japonais (Pierrehumbert & Beckman 1988), le suédois (Bruce 1977) ou le français (d'Alessandro & Mertens 1995, Hirst & Di Cristo A paraître).

Ce rapide aperçu montre que l'analyse ascendante a essentiellement pour but de réaliser une description phonétique discrète des événements prosodiques. Il reste ensuite à mettre en correspondance cette description, centrée sur la proéminence, avec des constructions phonologiques servant d'interface entre la description phonétique et des représentations morpho-syntaxiques voire sémantiques ou pragmatiques.

1.2.1.2 Analyse descendante et représentations cognitives

Par une analyse descendante, on cherche à extraire des formes prosodiques qui ont une représentativité statistique suffisante pour véhiculer des contrastes linguistiques ou para-linguistiques. Contrairement à la méthode précédente où les unités intonatives sont contraintes dans leur forme et leur séquencement qui assurent la connexité des structures phonologiques sous-jacentes, les mouvements prosodiques, corrélés à la présence/l'absence de certains types de contrastes linguistiques ou para-linguistiques étudiés, émergent de l'analyse statistique de larges corpus. L'introduction de connaissances a priori dans le processus de formation des modèles est donc limitée à la conception soigneuse de larges corpus. Les items doivent être conçus de manière à favoriser l'*émergence a posteriori* des corrélats prosodiques correspondant aux contrastes implicitement testés dans le corpus. Cette méthodologie d'étude contrastive par paires minimales (cf. section 2.1.2) est bien décrite par Contini & Profili (1987).

Ainsi, la méthodologie d'extraction de contours mélodiques proposée par Aubergé (1992) nécessite la construction de phrases dans lesquelles la position et la longueur des unités linguistiques sont systématiquement distribuées. Ces analyses contrastives permettent de caractériser les réalisations par un nombre illimité de paramètres et d'établir un lien direct entre paramètres prosodiques et représentations cognitives. Cette approche contraste donc avec les méthodes d'analyse visant à associer formes et unités sans se préoccuper *a priori* de leur fonction. Ici, **la forme fait fonction**.

Les modèles de génération qui découlent de cette approche, sont des modèles statistiques allant de simples modèles de superposition (Aubergé 1993, Abe & Sato 1992), à des modèles dynamiques non-linéaires (cf. section 1.3.1.4).

1.2.2 Vers une morphologie intonative

Nous envisageons dans notre étude, la production, la perception et la compréhension du langage comme des opérations portant sur des représentations mentales. Les configurations prosodiques produisant des informations, tant sur le contenu phonologique, lexical et syntaxique du discours que sur l'engagement du locuteur dans le processus de communication, pourraient faire partie de ces représentations mentales et servir de formes de référence auxquelles sont comparées les réalisations sonores actuelles.

Nous pensons donc que l'auditeur a des attentes sur ce qu'est une intonation "normale", comme il existe une syntaxe "normale" et qu'il peut ainsi interpréter comme une mise en relief quelque chose qui émerge du fond, qui dévie d'une forme attendue. Dans cette lignée, Sorin (1989) évoque l'hypothèse d'un catalogue de formes prosodiques pré-stockées en mémoire qui seraient des indications relatives permettant l'utilisation de ces informations avec des locuteurs divers et dans des contextes variés.

Cette notion d'attente est à rapprocher des propositions faites par le mathématicien Jean Petitot (1986) élève de René Thom qui introduit le **paradigme morphologique** dans le schéma général du passage du monde physique au monde mental, du domaine du concret, du continu et du variable à celui des représentations symboliques et des concepts, atomes de connaissance discrets et invariables. Si Smolensky (1992) avait déjà introduit le **paradigme sous-symbolique**, réceptacle des outils de reconnaissance des formes et de l'apprentissage probabiliste, comme complément du paradigme symbolique de l'Intelligence Artificielle (« le rêve booléen » Smolensky 1992, p.83), Jean Petitot plaide pour l'introduction d'un “filtre” morphologique entre les paramètres primaires que le traitement des signaux nous permet d'extraire du monde sensible, et les systèmes d'apprentissage. En effet, les représentations mentales n'ont pas été construites au hasard et sont profondément ancrées sur le concrêt par une sorte **d'histoire naturelle** que notre passage sur le continuum émotions-attitudes illustre. Ce filtre morphologique permet par ailleurs d'expliquer les mécanismes de prédiction, d'attente que nous envisagions plus haut et sur lesquels nous revenons maintenant.

La présupposition de l'existence d'un lexique prosodique permet ainsi d'avancer cette hypothèse : dans les actes de parole, les interlocuteurs devraient être capables de prévoir l'évolution des variations des paramètres acoustiques en se référant aux représentations mentales des configurations acoustiques stockées en mémoire (voir section 2.2.4.2). Ces mécanismes d'attentes ont été mis en évidence dans la perception tonique et rythmique de la musique (Boltz 1992, Schmuckler 1989). Par analogie, l'auditeur aurait deux types d'attentes par rapport aux configurations prosodiques : à court terme, il exploite les régularités phonétiques comme, par exemple, l'allongement progressif des unités de base à l'intérieur d'un groupe rythmique (Pasdeloup 1992, Barbosa & Bailly 1994) ; à long terme, il exploite les représentations prosodiques qu'il a progressivement acquises. Cette hypothèse est renforcée par certaines expériences qui montrent que l'auditeur est capable de prédire précocement la durée d'un énoncé (Grosjean 1983), sa modalité (Boë & Contini 1975), voire certaines attitudes (Aubergé, Grépillat & Rilliard 1997) ou formes prosodiques à fonction illocutoire (Clément & Gérard 1996).

Cette analyse holistique des phénomènes prosodiques favorise l'usage de représentations phonétiques globales. Dès 1966, Delattre propose un inventaire de 10 contours de base obtenus par l'intermédiaire d'oppositions de sens fondées sur la seule intonation. Fónagy (1984) introduit la notion de « clichés mélodiques » pour définir les contours mélodiques typiques des attitudes du français en affirmant que :

« La belle thèse selon laquelle tout énoncé serait une création du moment [...] ne s'applique qu'à l'univers de la poésie moderne (postromantique). En dehors de ce domaine limité nos énoncés sont forgés de grosses pièces préfabriquées, s'il ne sont pas préfabriqués de toutes pièces. »

1.3 Les modèles de génération de la prosodie

Dans cette partie, nous présentons, en regard de la section précédente, différents modèles de génération de la fréquence fondamentale et de la durée.

1.3.1 Les modèles de génération de la fréquence fondamentale

1.3.1.1 Les modèles par point-cibles

Les points-cibles

Ces modèles considèrent que l'intonation d'une phrase est essentiellement véhiculée par la distribution *temps/fréquence* des extrema de la courbe mélodique. Ces point-cibles sont ainsi reliés entre eux par des fonctions de transition dont les formes sont considérées comme non informatives en général. Les cibles sont définies à l'intérieur d'une enveloppe qui détermine la dynamique globale.

L'exemple classique de ce type de modèle est celui de Pierrehumbert (1981) qui est constitué de trois éléments fondamentaux :

Le registre (pitch range) : Pour déterminer la dynamique, une ligne de base et une ligne haute sont calculées indépendamment. Elles forment l'enveloppe du contour mélodique. Le contour sera ensuite exprimé comme une fraction de la distance entre ces deux lignes.

Les point-cibles : les valeurs cibles peuvent appartenir à deux catégories, *haut* (H : *high*) et *bas* (L : *low*).

Les fonctions de transition : deux H consécutifs sont interpolés par une fonction de transition qui abaisse la fréquence entre les deux cibles. Par contre, la transition entre deux cibles L sera toujours monotone. Notons que le choix de plusieurs auteurs s'est porté sur des fonctions quadratiques (Pierrehumbert 1981, Hirst et al. 1991) malgré le fait qu'on trouve aussi l'emploi de droites ('t Hart et al. 1990, Beaugendre 1994).

Si la gestion du positionnement *temps/fréquence* de ces point-cibles est bien adaptée à la synthèse par règles, toute la performance intonative dépendra de la compétence du "marionnettiste phonologique" qui devra notamment coordonner les réalisations et négocier les conflits inévitables entre structures intonatives et rythmiques.

La gestion des gabarits

Le reproche localiste fréquemment adressé à ce type d'approche fait souvent oublier que l'évolution des point-cibles est contrainte par de multiples heuristiques globales telles que registres, lignes de déclinaison hautes et basses, grilles tonales, mécanismes de descente graduelle ("down-stepping")…. La gestion dynamique de ces instruments de contrôle peut parfois donner l'illusion de l'existence de deux canaux prosodiques indépendants permettant de coder des niveaux différents de structuration de l'énoncé. Le modèle proposé par Gårding (1991) assigne à des *pivots* le soin de contrôler le positionnement de grilles tonales dans lesquelles les point-cibles vont s'inscrire. Les pivots sont globalement synchronisés avec le début d'actes de parole ("speech acts") et gèrent ce qu'elle appelle, l'intonation phrastique, par contraste avec la structuration accentuelle et tonale assurée par les point-cibles. Cette approche constitue la première entorse à la trop grande linéarité et au localisme des descriptions en termes de point-cibles.

1.3.1.2 Les modèles superpositionnels

Les filtres

Le premier modèle superpositionnel a été proposé par Öhman (1967). Il conçoit la courbe mélodique comme une somme de deux composantes obtenues par la réponse de filtres linéaires à des commandes échelons évoquant l'initialisation et la fin d'un phénomène physique. Ces composantes sont directement liées à des niveaux linguistiques ; respectivement le niveau du mot ("word intonation") et de la phrase ("sentence accent"). Ce modèle initial sera repris et raffiné par Fujisaki & Sudo (1971) qui donnent une interprétation plus physiologique des composantes. Ils modifient la commande de phrase (impulsion *vs* échelon) et mettent en évidence une organisation plus complexe de ce contrôle. Le contour de fréquence fondamentale est ainsi défini comme la réponse du double mécanisme d'étirement, direct et indirect, des cordes vocales respectivement lié à la pression subglottique (P_s) et la tension des cordes vocales (T_v). Les commandes visent donc à rendre compte de l'activité musculaire laryngienne mise en jeu lors de la production de la mélodie. Les composantes de phrases correspondent à la réponse d'un système linéaire du second ordre à des impulsions. Les composantes d'accents sont obtenues par la réponse d'un autre système linéaire du second ordre à des fonctions échelon. Les deux types de commande génèrent des contours sur une échelle logarithmique et sont ajoutés pour former le contour mélodique final. Il sera appliqué au français par Bailly (1983).

Bien qu'il reste à démontrer que les activités observées sur les lignes de commande correspondent bien à des activités musculaires spécifiques, l'intérêt de ce modèle est double :

- Tout d'abord, ce n'est plus un modèle local : une commande est audible jusqu'à ce qu'une autre commande de même type vienne la masquer.

– Ensuite, la négociation entre niveaux de représentations est simple. On remplace ici l'opération complexe de modulation des gabarits par une simple addition.

Les lignes

Le modèle de Thorsen (1980, 1983) conserve l'idée de modèle superpositionnel mais rompt avec l'ancrage physiologique des modèles précédents : les variables contrôlées ne sont plus des commandes mais des amplitudes de début et de fin de contours. Ces contours sont de simples droites. Cependant, le nombre de niveaux supposés est porté à quatre :

– Une composante de phrase qui produit l'intonation de phrase.

– Sur ce contour, est superposé la composante de groupe accentuel fournissant un patron de groupe accentuel.

– A la résultante de ces deux composantes sont ajoutés des contours de mots.

– Enfin, intervient la composante microprosodique.

Ce modèle préfigure le modèle de Véronique Aubergé qui assignera à ces contours une fonction clairement linguistique alors que Thorsen, redevenue Grønnum énonce : « I have never tried to attach specific meanings to specific intonation contours and I deem it a futile attempt. I do not think that intonation carries any meaning in and off itself. » (Grønnum 1992, p.78).

1.3.1.3 Les modèles par concaténation

Concaténer des sons (diphones, demi-syllabes, mots ou plus généralement polysons) pour générer des signaux synthétiques puis les déformer prosodiquement est une méthode de synthèse qui permet de générer la parole synthétique la plus intelligible et la mieux acceptée. Il n'est donc pas étonnant que cette méthode ait été appliquée très tôt à la génération prosodique et plus particulièrement à celle de la fréquence fondamentale.

Concaténation de mouvements : l'approche IPO

La technique IPO ('t Hart & Collier 1973, 't Hart et al. 1990) est une des rares approches véritablement orientée vers l'auditeur. Appliquée au français par Beaugendre (1994), elle consiste à simplifier la courbe mélodique globale en une suite de segments de droites (selon une échelle logarithmique). Cette simplification est réalisée sur la base de tolérances perceptives.

Le processus nécessite deux étapes de simplification :

La stylisation : la première étape de simplification consiste à obtenir des contours réduits à un nombre minimum de segments linéaires, de telle sorte qu'ils soient "indiscernables" des contours naturels.

La standardisation : La seconde étape est une classification des mouvements obtenus en un nombre restreint de classes standards.

Par cette méthodologie ascendante, les auteurs veulent obtenir une classification purement phonétique qu'ils mettront ensuite en correspondance avec un modèle phonologique. Une grammaire fixe les contraintes structurelles sur l'enchaînement des mouvements mélodiques.

Mais ce qui sépare cette approche d'un modèle point-cibles, c'est la détermination d'un inventaire de mouvements mélodiques **standards** ou **prototypiques** qui seront alors combinés et associés à une fonction particulière au sein de la structure accentuelle et intonative. Certains mouvements représenteront par exemple un corrélat acoustique de l'accent ou encore du continuatif mineur ou majeur.

Concaténation de signaux : le système CHATR

Les modèles de génération par contours stockés nécessitent en aval l'utilisation d'une technique de synthèse pour modifier, à partir de ces contours, un signal acoustique obtenu en parallèle par concaténation de diphones ou de triphones. Ces unités stockées n'ont généralement qu'un représentant dans chaque système de synthèse.

L'alternative que propose le système de synthèse CHATR (Campbell 1997) consiste à utiliser des unités segmentales plus variées extraites d'une base de données contenant plusieurs réalisations de chaque son (typiquement plusieurs centaines). Cette variabilité permet une modélisation plus fine des différents styles d'élocution et améliore la naturalité en s'affranchissant, si la base est suffisamment complète, d'un quelconque traitement du signal.

Campbell avance donc que pour reproduire les caractéristiques de la parole naturelle, il faut stocker des unités qui portent à la fois des attributs de nature segmentale et prosodique. La difficulté principale de cette approche réside dans la sélection de l'unité optimale. Une génération adéquate se réduit alors à la qualité de l'étiquetage automatique et à la détermination d'une hiérarchie entre les attributs permettant d'extraire de la base l'unité qui conduit au minimum de distorsion avec ses voisines.

Du marquage prosodique aux rendez-vous

Notons une fois de plus que les modèles phonétiques cités dans cette section ne résolvent pas de manière satisfaisante le rapport du signifiant au signifié, de la substance au code. Tout au plus, ils introduisent de manière diffuse la notion de lexique prosodique.

Ce n'est cependant pas par hasard qu'une longue tradition française de recherches a postulé un rapport beaucoup plus étroit entre forme et sens. Comme nous l'avons déjà remarqué, la typologie accentuelle finale du français tend peut-être à biaiser notre analyse des rapports qu'entretiennent intonation, rythme, accentuation et morpho-syntaxe. Cependant, les recherches sur la prosodie en France de l'école Grenobloise (Vaissière 1971, Emerard 1977, Bailly 1983, Aubergé 1991) et de l'école Aixoise (Di Cristo 1985, Martin 1980a, Martin 1980b, Rossi 1985, Rossi 1993) ont postulé très tôt un rapport étroit entre forme et sens. C'est ainsi qu'apparaît dans les deux écoles la notion de **marquage prosodique** ou de **marques prosodiques**, où formes prosodiques, contours prosodiques et fonctions structuratrices de l'énoncé par l'intonation sont étroitement liés. Ainsi, Emerard (1977), Di Cristo (1985), Martin (1980a), Bailly (1983), pour ne citer que quelques auteurs, proposent des systèmes de marques, où divers contours prosodiques signalent des jonctures énonciatives spécifiques, souvent mise en rapport avec la notion de dépendance, de présupposition syntaxique des unités que les marques séparent. Dans une proposition récente, Rossi (1993) va même jusqu'à proposer une ségrégation des différentes fonctions de l'accentuation (cf. section 1.1.1.2) par le paramètre prosodique le plus affecté.

Toutefois, la notion de marquage et de contours prosodiques associés manque de souplesse : elle suppose une synchronisation trop rigide entre syntaxe et intonation. Or cette synchronisation, même lorsqu'elle est limitée à la fonction démarcative, peut être relativement lâche, notamment lorsqu'il s'agit d'indiquer une cohésion syntaxique entre unités (Fónagy 1980, Campbell 1993). On trouvera donc chez Aubergé et dans notre travail la notion plus flexible de rendez-vous entre intonation et structuration énonciative, qui suggère la possibilité de rendez-vous manqués et de niveaux linguistiques attracteurs de la forme intonative.

Morphologie intonative

Chez Aubergé (1991), le présupposé de base à la construction du lexique de contours est celui de **rendez-vous** structurels entre l'intonation et la syntaxe. A travers l'étude exhaustive des contours mélodiques, issue d'un corpus établi en fonction de ces présupposés et classés selon certains critères linguistiques, l'auteur se propose de factoriser un certain nombre d'éléments phonétiques et de les remplacer par un prototype. Les hypothèses principales de son travail sont, d'une part, la coïncidence d'unités intonatives avec chacun des niveaux syntaxiques (la phrase, la proposition, le groupe et éventuellement le sous-groupe) et, d'autre part, l'indépendance des contours entre ces différents niveaux.

L'analyse consiste donc, en partant du niveau de la phrase jusqu'au sous-groupe, à extraire du corpus de référence des classes de contours contenant la substance intonative caractéristique de chaque unité de ce niveau. Chaque classe sera caractérisée par un contour moyen.

La génération de l'intonation procède de manière inverse. A partir du lexique de contours moyens et après comparaison de la liste des attributs sur le texte d'entrée, sont calculés successivement : le patron de phrase, le patron de proposition (aligné sur le patron de phrase), le patron de groupe (aligné sur le patron de proposition) et, dans le cas où le groupe dépasse quatre syllabes, le patron de sous-groupe.

Notons que l'on retrouve chez Abe & Sato (1992) cette association entre forme et sens dans un modèle superpositionnel. Un modèle global calcule la fréquence moyenne de chaque groupe prosodique ("minor phrase") selon des critères syntaxiques, paralinguistiques et phonotactiques. Puis un modèle local assigne à chaque syllabe une correction et une pente.

Nous reviendrons de façon plus détaillée sur les principes méthodologiques de l'approche Aubergé (section 1.4) car ils ont joué un rôle primordial dans la réalisation de notre modèle.

Discussion

L'approche par contours stockés présente un certain nombre d'inconvénients :

– D'abord une certaine rigidité dans la mesure où elle peut difficilement tenir compte de la dynamique subtile des mouvements intonatifs qui sont sujets à des distorsions liées au débit d'élocution, au nombre de constituants de l'unité ou à la réorganisation en fonction des unités adjacentes. L'unique solution pour tenir compte de ces facteurs est d'enrichir la paramétrisation des unités stockées et donc d'augmenter le nombre de contours du lexique et la base de données qui a permis de les extraire. Aubergé, par exemple, introduit systématiquement l'attribut de longueur syllabique pour chaque contour de son lexique. Une solution consisterait à utiliser des outils capables de généraliser à des unités plus longues l'évolution des contours extraits sur des unités de petites tailles.
– L'extraction des contours associés à un ensemble d'attributs passe par des moyennes ou des fonctions d'interpolation (Morlec 1994, Morlec, Bailly & Aubergé 1995) de l'ensemble des éléments de la base portant les mêmes attributs. Si l'on ne veut pas se limiter à décrire ces contours par des fonctions simples, nous sommes amenés à présumer de leur forme.

L'utilisation des systèmes dynamiques de prédiction de fonctions non linéaires permet de s'affranchir de ces inconvénients.

1.3.1.4 Modèles dynamiques

L'emploi de modèles dynamiques dans la génération des paramètres prosodiques est récent. La fréquence fondamentale est généralement le paramètre le plus concerné dans ce type de système à apprentissage supervisé. Ces modèles permettent d'associer plus directement description linguistique et substance prosodique. C'est surtout dans la préparation des paramètres de caractérisation que réside la plus grande difficulté de cette approche. Si ces paramètres sont bien choisis, on peut alors arriver à générer des contours mélodiques de très bonne qualité sans pour autant savoir comment se fait le passage entre les niveaux de représentation supérieurs et la substance elle-même.

Les exemples d'application de ces modèles, que nous proposons ici, ont été organisés en fonction de la taille de l'unité dont la fréquence fondamentale est générée : mot, syntagme ou phrase complète.

Ljolje & Fallside (1986) ont travaillé sur l'apprentissage automatique de réseaux Markoviens pour la prédiction des mouvements de F0 portés par des mots monosyllabiques contenant des structures phonétiques variées. Les auteurs montrent que les contours synthétiques obtenus par cette méthode sont perceptivement indifférentiables des contours naturels, et voient dans les réseaux Markoviens un très grand potentiel pour capturer les caractéristiques essentielles de la structure prosodique. Pourtant ce type d'approche, utilisé massivement en reconnaissance de la parole, ne semble pas avoir fait son chemin en génération automatique de la prosodie. Cette technique de modélisation est plus apte à mémoriser les invariances que les indices discriminants entre les formes à traiter. Les réseaux connexionnistes répondent pleinement à cette double recherche - interpolation et séparabilité - sans qu'il soit nécessaire d'avoir un modèle de génération explicite.

Scordilis & Gowdy (1989) ont utilisé une combinaison de deux réseaux neuromimétiques pour prédire la fréquence fondamentale de mots isolés en anglais :

- Le premier réseau est macrophonémique. Il délivre une valeur de F0 pour chaque phonème du mot courant. Il s'agit d'un perceptron à une couche cachée qui possède en entrée une fenêtre de description des propriétés acoustiques de 12 phonèmes et prédit la fréquence fondamentale du onzième (10 contextes gauches et un contexte droit).
- Le second est microphonémique. Il applique une correction à la valeur prédite par le premier toutes les 10 ms en fonction des effets de coarticulation entre phonèmes entraînant des variations micromélodiques. Il s'agit d'un réseau de neurones récurrent (RNR) de type Jordan (cf. section 3.2) qui possède en entrée la description de 3 phonèmes (1 contexte gauche et 1 droit) et prédit 6 valeurs de sortie décrivant les microvariations de la fréquence fondamentale appliquée au phonème central à partir du niveau moyen prédit par le premier réseau.

Les erreurs de prédictions de l'ordre de 4 Hz permettent aux auteurs de conclure au succès de leur approche en insistant sur l'importance d'une conception soignée du corpus d'apprentissage.

Sagisaka (1990) propose un modèle de génération qui prédit 3 valeurs de F0 pour chaque patron *montée/descente* du japonais associé à un syntagme mineur. Lorsqu'une suite de patron de ce type est globalement descendante, le nouveau patron est associé à un syntagme majeur.

Le réseau utilisé est un perceptron à une couche cachée dont les entrées sont définies par 5 paramètres :

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

La sortie prédit, quant à elle, la forme générale de F0 du syntagme mineur en délivrant sa position initiale, maximale et finale.

Les prédictions obtenues par Sagisaka restent proches des contours moyens d'apprentissage mais sa stylisation par trois valeurs pour un syntagme est très simpliste.

Nous avons vu, jusqu'à maintenant, des modèles dynamiques se limitant à la génération de contours mélodiques ne dépassant pas le syntagme. Nous proposons maintenant de détailler un modèle connexionniste qui génère la fréquence fondamentale de phrases complètes.

Le Système de Traber (1992) génère la fréquence fondamentale de phrases complètes en langue allemande. Son modèle est un réseau récurrent de type Elman (cf. section 3.2) à deux couches cachées (figure 1.5). Durant l'apprentissage supervisé, l'auteur présente en entrée du réseau une fenêtre de la chaîne phonologique contenant 13 phonèmes (6 contextes gauches et 6 contextes droits). Le réseau dispose donc d'une double mémoire - métaphore spatiale (fenêtre sur la tâche) et rétroaction - qui lui permettra d'interpoler des contours mélodiques de phrase lors de la phase de prédiction.

L'entrée comporte les informations suivantes :

- quatre niveaux accentuels par rapport au syntagme ;
- frontières de syntagme et de mot ;
- marque *final / non final* indiquant le type de syntagme ;
- position *avant / après* par rapport à l'accent de syntagme.

Elle fournit également des informations d'ordre microprosodique :

- voyelle *longue / courte* ;
- voyelle *haute / basse*, par rapport à la fréquence fondamentale intrinsèque ;

– voisement du contexte consonantique gauche et droit.

La sortie correspond au contour stylisé du symbole central de la fenêtre d'entrée par 8 valeurs de F0.

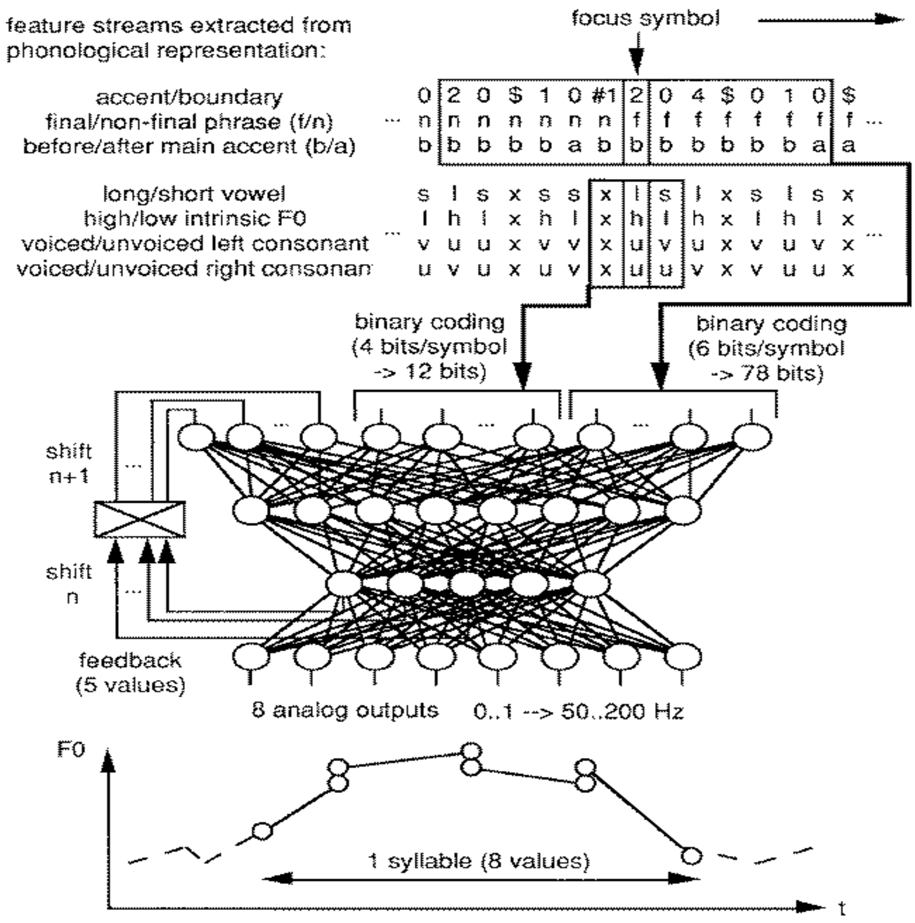


FIG. 1.5: Réseau neuromimétique utilisé par Traber (1992) pour la génération de la courbe de fréquence fondamentale.

Traber réalise une comparaison entre les résultats découlant de son modèle connexionniste et ceux qu'il a obtenus à l'aide d'une approche par concaténation de contours stockés se rapprochant de celle d'Aubergé (1991). L'auteur conclut que les deux méthodes de génération fournissent des contours de fréquence fondamentale de haute qualité mais qu'une approche par réseau nécessite beaucoup moins d'efforts humains pour un résultat équivalent.

Discussion : Dans chacun des exemples précédents, nous avons vu que, pour donner aux perceptrons une mémoire du passé et pour leur permettre d'anticiper les événements à venir, leur paramétrisation d'entrée est classiquement représentée par une fenêtre glissante contenant la donnée en cours de traitement et couvrant les données voisines. Par exemple, l'entrée du perceptron macrophonémique de Scordilis & Gowdy comporte des informations sur 10 phonèmes en contexte gauche. Cette surcharge de paramètres sur le passé peut être facilement éliminée moyennant l'utilisation de réseaux récurrents. C'est vers cette solution que nous nous sommes orientés dès le départ pour prédire la cinématique des mouvements mélodiques.

Les exemples que nous venons de montrer illustrent les travaux précurseurs de certains auteurs dans l'utilisation de modèles connexionnistes pour la génération de la fréquence fondamentale. D'autres chercheurs ont profité de la brèche ouverte dans ce domaine pour tenter avec succès de générer d'autres paramètres prosodiques. Par exemple, Riedi (1995) a conçu un modèle à base de réseaux de neurones pour la génération de la durée segmentale. Des auteurs tels que Campbell (1992b) ou Barbosa (1992) ont fait de même dans les modèles hybrides que nous détaillons dans la section suivante. Hwang et al. (1995) proposent un modèle combinant plusieurs réseaux de neurones chargés de prédire d'une part les contours de F0 pour chaque syllabe d'une phrase mais également deux autres paramètres : durée des syllabes et des pauses, ainsi que l'intensité. De Tournemire (1997) développe, pour le système de synthèse du CNET, deux réseaux générant F0 et durée syllabique.

Comme nous pouvons le remarquer, l'approche connexionniste connaît un succès grandissant dans le domaine de la génération automatique de la prosodie. Cependant, même si plusieurs paramètres sont générés dans un modèle, cela n'est jamais réalisé dans un seul et même réseau. Seul Todd (1989), à notre connaissance, a développé un réseau unique pour générer simultanément la durée et la mélodie dans le cadre d'une application musicale.

Nous proposerons, plus loin, un modèle évolutif de génération multiparamétrique de la prosodie dont chaque réseau neuromimétique prédit à la fois la fréquence fondamentale et le macrorythme mais, avant d'en arriver là, présentons, après les modèles de génération de la fréquence fondamentale, les différentes approches de génération de la durée segmentale.

1.3.2 Les modèles de génération de la durée segmentale

Cette section est d'abord destinée à donner un aperçu de quelques modèles de la durée segmentale. Nous séparerons les modèles qui utilisent directement le phonème comme unité de base, de ceux qui privilégient la structuration rythmique de l'énoncé. En fin de cette section nous examinerons plus précisément le modèle de Campbell (1992a) qui a servi de point de départ aux travaux de recherche de Plinio Barbosa (1994) à l'ICP. Nous expliquerons enfin les raisons principales qui nous ont amenés à apporter quelques modifications à ce modèle en vue de son intégration dans le modèle de génération multiparamétrique qui sera détaillé au chapitre III.

1.3.2.1 Modèles fondés sur le phonème

Les modèles qui prédisent directement la durée phonémique peuvent être classés en trois catégories : additif, multiplicatif, additif & multiplicatif. Dans chacun des cas, on part d'une durée phonémique de base (souvent la durée intrinsèque) que l'on va déformer en l'ajoutant/retranchant ou en la multipliant/divisant par une quantité définie par des facteurs qui influent sur la durée segmentale tels que le contexte phonétique ou l'environnement syntaxique.

Le modèle de Klatt

L'un des premiers modèles de génération de la durée segmentale, et qui fait souvent figure de référence, est celui de Klatt (1987). Ce modèle, dans sa version la plus récente, est à la fois additif et multiplicatif. La durée phonémique est obtenue par l'équation suivante :

$$Dur = MinDur + \frac{(InhDur - MinDur) \times PRNCT}{100} \tag{1.1}$$

où $InhDur$ correspond à la durée intrinsèque du phonème, $MinDur$ est environ égale à $InhDur$ pour les réalisations accentuées et réduite de moitié pour les réalisations non accentuées. Le coefficient $PRNCT$ est le pourcentage de rétrécissement obtenu en multipliant cumulativement des facteurs liés au contexte phonétique et à l'environnement syntaxique.

Les modèles d'O'Shaughnessy

O'Shaughnessy a développé deux types de modèles :

Le premier est multiplicatif (1981) : En partant d'une durée de base, l'auteur applique un pourcentage de modification qui dépend, pour la voyelle, de sa hauteur, du trait de nasalité, du voisement et du mode d'articulation de la consonne qui le suit. Pour la consonne, ce pourcentage dépend du trait de voisement, du mode d'articulation et de ces deux mêmes facteurs liés aux éventuelles consonnes adjacentes.

Le second est additif (1984) : L'auteur part toujours d'un ensemble de durées de base et ajoute ou retranche des durées liées aux facteurs déjà mentionnés et également à des facteurs associés à la position dans la phrase et à la présence de frontières mineures et majeures.

O'Shaughnessy n'intègre pas dans son modèle des facteurs tels que le débit, le *stress* et l'environnement syntaxique.

Le Modèle du CNET

Il s'agit d'un modèle multiplicatif (Bartkova & Sorin 1987) pour lequel les durées intrinsèques du phonème (obtenues à partir d'un corpus de logatomes CVC et VCV) sont multipliées par un ensemble de coefficients dont les valeurs dépendent de la position dans le mot, du contexte phonétique et d'informations grammaticales et syntaxico-sémantiques. Les durées des voyelles et des consonnes sont obtenues par deux équations différentes :

$$Durée\,voyelle = DI.V_i.m_c \tag{1.2}$$

$$Durée\,consonne = DI.C_{ij} \tag{1.3}$$

où DI, durée intrinsèque du phonème concerné est multiplié par des coefficients V_i (pour les voyelles) ou C_{ij} (pour les consonnes) qui dépendent de leur contexte. Pour les voyelles, ce produit est encore multiplié par des coefficients co-intrinsèques m_c.

Le modèle de Van Santen

Nous terminons cette section par le modèle de van Santen (1992) qui propose une méthodologie pour analyser les interactions entre facteurs influençant de façon pertinente la durée phonémique. L'auteur va soulever un certain nombre de problèmes liés à la génération directe de la durée phonémique.

En partant de sept facteurs (*stress*, emphase, phonèmes précédents et suivants, position dans le syntagme, position au sein du mot et nature du phonème), et en construisant une grille de référence associée à ces facteurs, van Santen montre, en premier lieu, que de nombreuses cases vont rester vides même si l'on utilise une très grosse base de données. Il explique ensuite que les facteurs ne sont pas complètement décorrélés. Il propose donc une analyse destinée à déterminer l'ordre des facteurs qui doit être considéré dans la caractérisation de la durée phonémique.

Afin d'établir des ensembles homogènes à l'intérieur desquels les effets d'interactions entre facteurs peuvent être ordonnés, l'auteur distingue trois catégories : les voyelles, les consonnes, les consonnes dans les "clusters". Pour chacune de ces catégories, il estime l'ensemble des paramètres du modèle de prédiction (généralisation du modèle de Klatt) : 32 pour les voyelles, 196 pour les consonnes. Ces paramètres sont calculés à partir d'une segmentation manuelle d'un corpus de 2162 phrases courtes fournissant 41588 réalisations phonémiques.

Discussion

Comme on le voit, la génération rigoureuse de la durée par des modèles prédisant directement la durée phonémique nécessite un travail humain et calculatoire colossal. Nous pensons que ces modèles peuvent être simplifiés en procédant par une génération en deux étapes :

- Une première étape de génération des niveaux supérieurs de la structuration rythmique - prédiction de la durée syllabique ou unité apparentée - qui tiendra compte de facteurs tels que l'environnement syntaxique ou la position de l'unité dans le syntagme.
- Une seconde étape de distribution de la durée de l'unité rythmique choisie parmi ses constituants grâce à des facteurs tels que la durée intrinsèque ou le contexte phonétique.

Outre le fait que ce type d'approche hybride permet de séparer en deux catégories les facteurs influençant la durée des réalisations phonémiques (nous verrons plus loin que leur distinction n'est pas si claire), elle favorise l'intégration de la composante perceptive au niveau macrorythmique.

C'est de ce type de modèles hybrides dont nous allons parler dans la section suivante.

1.3.2.2 Modèles fondés sur des unités apparentées à la syllabe

Modèle Campbell

Campbell (1992a) propose un modèle combiné de génération fonctionnant en deux étapes :

- La première consiste à prédire la durée de chaque syllabe du message à partir d'informations phonotactiques et phonologiques constituant les entrées d'un perceptron multicouche.
- Le durée de chaque phonème de la syllabe est ensuite calculée à partir d'un algorithme de répartition.

Cet algorithme suppose que la durée de la syllabe est égale à la somme de la durée de chacun de ses constituants et que chacun d'entre eux est soumis à une même déformation appelée $z-score$. La durée de chacune des réalisations phonétiques est déterminée selon les paramètres de distribution statistique spécifique à chaque phonème.

La première étape consiste à calculer les moyennes et les écart-types des durées des réalisations phonémiques du locuteur de référence à partir d'un corpus de logatomes. Le $z-score$ associé à chaque syllabe est obtenu en résolvant l'équation suivante :

$$Durée(syllabe) = \sum_{i=1}^{n} \exp(\mu_i + z\sigma_i) \tag{1.4}$$

ou z est le $z-score$ associé à la syllabe et la paire (μ_i, σ_i) correspond à la moyenne et l'écart-type des durées log-transformées des réalisations du phonème i[6]. La durée du phonème i dans cette syllabe sera alors égale à $\exp(\mu_i + z\sigma_i)$.

Modèle Barbosa-Baillly

Le GIPC comme unité de programmation rythmique : Ce modèle (Barbosa 1994) est similaire au modèle hybride de Campbell à la différence près que l'unité macrorythmique n'est plus la syllabe mais le Groupe Inter P-Center (GIPC) défini comme l'ensemble des unités phonémiques entre deux P-centers (cf. page 10) adjacents. Dans le cadre de son travail, le P-center coïncide avec l'établissement vocalique.

L'auteur cherche à démontrer que le GIPC constitue l'unité de programmation rythmique du français :

- Il met d'abord en évidence que l'hypothèse d'élasticité est mieux respectée au sein d'un groupe délimité entre les débuts consécutifs de deux voyelles (1994, p. 78).
- Il constate ensuite que les corrélations des durées entre 5 débits différents sont plus élevées si le GIPC est pris comme unité de référence (1994, p. 84).
- Enfin, il calcule les $z-score$ de chaque GIPC et de chaque syllabe de son corpus, et vérifie que les patrons rythmiques obtenus à partir du GIPC évoluent de façon monotone et croissante au sein d'un groupe accentuel (Pasdeloup 1992), et que le marquage des frontières prosodiques est clairement signalé. Ces phénomènes ne sont pas vérifiés quand la syllabe est l'unité concernée (1994, p. 92).

Modèle complet de génération automatique de la durée : Le modèle fonctionne en deux étapes (cf. figure 1.6) :

- Dans la première étape, un réseau de neurones récurrent délivre la durée du GIPC (notée IIPC pour *Intervalle Inter P-center*) à partir de la description phonologique de la phrase et d'une mesure du débit : le réseau fourni un nombre de PH qui correspond au rapport entre le GIPC courant d'une phrase et la période d'horloge (PH) de cette même phrase. La PH est égale à la moyenne des IIPC non-associés à une marque prosodique. Le nombre de PH permet donc de réaliser une normalisation entre les différents débits.

[6] $durée_{log-transformée} = \ln(durée_{exprimée\ en\ millisecondes})$

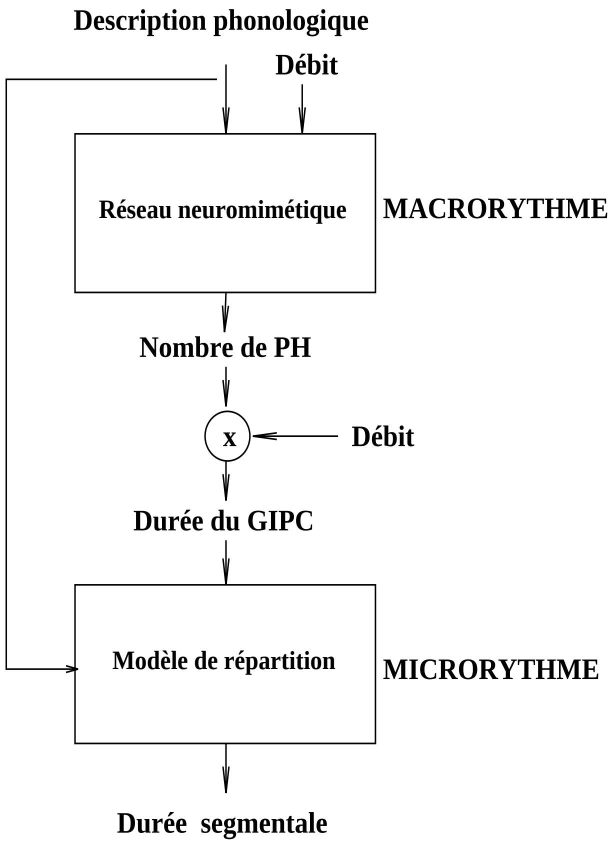


FIG. 1.6: Schéma du modèle de génération automatique de la durée phonémique Barbosa-Bailly.

– Dans une deuxième étape, le modèle de répartition distribue la durée prédite du GIPC entre chacun des ses composants. Ces composants sont les réalisations phonémiques et éventuellement les pauses silencieuses de l'énoncé (voir Barbosa & Bailly (1994) pour le détail du modèle de génération dynamique de la pause).

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

TAB. 1.1: Moyennes (et écart-types) des durées, exprimées en millisecondes, des réalisations des phonèmes français du locuteur EV (d'après Barbosa 1994, p.77).

Modèle Morlec-Baillly

Notre modèle de prédiction de la durée segmentale est très voisin du modèle Barbosa-Bailly : l'algorithme de répartition reste inchangé. Notons toutefois que nous n'avons pas enregistré de corpus de logatomes afin de calculer les moyennes et les écart-types des durées des réalisations phonémiques de notre locuteur de référence. Afin de ne pas biaiser notre étude rythmique, nous avons simplement exploité les caractéristiques des réalisations phonémiques du locuteur EV étudié par Barbosa (cf. tableau 1.1) en vérifiant que celles de notre locuteur, évaluées sur les corpus de phrases détaillés au chapitre II, restaient voisines.[7]

Nous avons repris, dans une première phase, les travaux de Barbosa sur le choix de l'unité de programmation rythmique (syllabe ou GIPC ?). Notre analyse nous a conduit à opter, comme lui, pour le GIPC :

- La figure 1.7 montre les moyennes (traits épais) des contours rythmiques (traits fins) exprimées en *Ratio syllabe* (courbes du haut) et en *Ratio GIPC*[8] (courbes du bas) pour des phrases déclaratives de structures syntaxiques [GNsujet + verbe] (à gauche) et [GNsujet + verbe + GNobjet] (à droite) issues des corpus détaillés au chapitre II. Nous pouvons remarquer que pour ces deux structures, l'unité GIPC, contrairement à la syllabe, fournit des patrons rythmiques monotones et croissants et que les marques prosodiques et leur hiérarchie sont clairement révélées : les maxima locaux coïncident toujours avec une marque prosodique ; l'allongement est plus prononcé avant une frontière *sujet/verbe* qu'avant une frontière *verbe/objet* ; après avoir atteint un maximum, il y a un phénomène de remise à zéro ("reset") du patron rythmique.
- Nous avons analysé les erreurs de répartition des durées phonémiques (cf. figures 1.8) en utilisant l'unité syllabique et GIPC. Ces deux figures représentent l'histogramme de répartition des erreurs $|Durée_{phonème} - Durée_{prédite}|$ sur 3722 phonèmes de notre corpus de phrases déclaratives. La moyenne des erreurs est de 14.05 ms pour la syllabe, et de 11.23 ms pour le GIPC. L'écart-type vaut 13.54 ms pour la syllabe, et 9.64 ms pour le GIPC. Ces erreurs peuvent provenir de deux facteurs : les imprécisions de la segmentation manuelle du corpus, et l'utilisation des caractéristiques des réalisations phonémiques d'un autre locuteur. Mais ces erreurs se répercutent quelle que soit l'unité rythmique choisie.

La principale différence de notre modèle par rapport au modèle initial, provient du module de génération de la composante macrorythmique qui ne prédit plus directement le IIPC (durée du GIPC). Cette modification a été motivée par des résultats de recherche

[7]Barbosa a montré que les auditeurs étaient plus sensibles à une prédiction fine de la structure macrorythmique qu'à celle des durées phonémiques (1994, p. 146).

[8]Ces Ratios correspondent à une unité macrorythmique normalisée que nous définirons un peu plus loin.

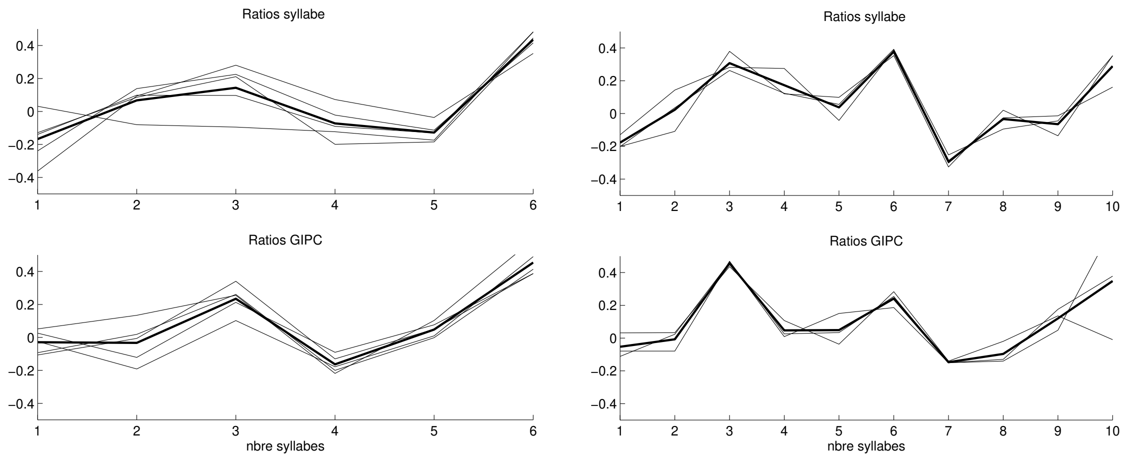


FIG. 1.7: Comparaison de quelques contours macrorythmiques en utilisant l'unité syllabique (en haut) et l'unité GIPC (en bas) : [GNsujet(3 syllabes) + verbe(3 syllabes)] à gauche, et [GNsujet(3 syllabes) + verbe(3 syllabes) + GNobjet(4 syllabes)] à droite. Le contour moyen est superposé en gras.

récents montrant que l'indépendance du macrorythme vis-à-vis du segment n'était pas vérifiée. Fant & Kruckenberg (1996) note une augmentation de la durée des syllabes non accentuées avec le nombre de phonèmes qu'elles contiennent. Van Santen (1997) infirme le modèle initial de Campbell en montrant sur de larges corpus d'apprentissage que la durée de la syllabe dépend systématiquement de la nature de ses constituants.

Nous avons cherché, dans un premier temps, à vérifier les propos avancés par van Santen en réalisant une analyse statistique sur la base du corpus utilisé par Aubergé (1991). Ce corpus comporte des phrases déclaratives, interrogatives et impératives de 1 à 18 syllabes CV essentiellement. Pour l'ensemble des 164 phrases produites par le locuteur EV, nous avons d'abord calculé l'évolution de leur durée totale en fonction du nombre de syllabes qu'elles comportent. Nous avons ensuite réalisé le même calcul, mais en remplaçant la durée réelle de chaque phonème par sa durée moyenne.

Les résultats sont reportés en figure 1.9. Nous observons, sur cette figure, que l'évolution des moyennes des durées réelles (en pointillés) est quasi linéaire ($R = 0.9847$). Cela montre que le locuteur n'a pas tendance à accroître son débit d'élocution lorsqu'il produit des phrases longues. Ce résultat va dans le sens de l'hypothèse d'une horloge interne (Lame 1990, Mates 1992) agissant comme un repère de périodicité lors de l'énonciation, et utilisée par Barbosa dans son modèle. La valeur de la PH vaut ici 165 ms.

Mais, en regardant de plus près cette courbe, on constate deux "accidents" : la moyenne des durées des phrases pour 14 et 17 syllabes et inférieure à celle des phrases possédant une syllabe de moins. De plus, nous retrouvons les mêmes accidents sur l'évolution des durées moyennes des phrases obtenues à partir des durées phonémiques moyennes du locuteur

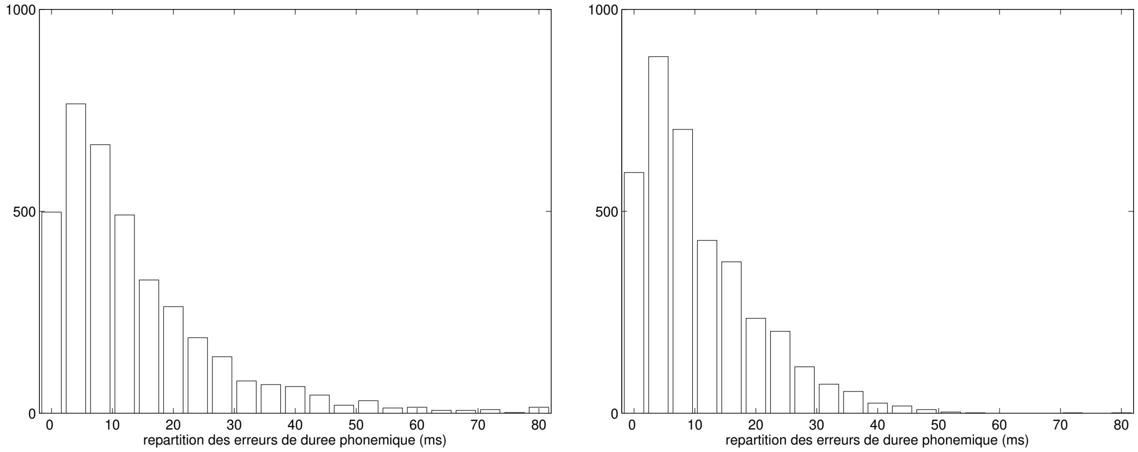


FIG. 1.8: Répartition des erreurs de prédiction des durées phonémiques (exprimées en millisecondes) en utilisant comme unité la syllabe (à gauche) et le GIPC (à droite).

EV (trait pleins). Nous pouvons donc conclure que ce phénomène ne provient pas d'une accélération du débit particulière à ces deux longueurs, mais bien de la nature des phonèmes présents dans les phrases de 15 et 17 syllabes. Nous corroborons, par cette analyse globale sur des durées totales d'énoncés, les conclusions de van Santen et nous proposons maintenant une alternative à la référence *horloge interne*.

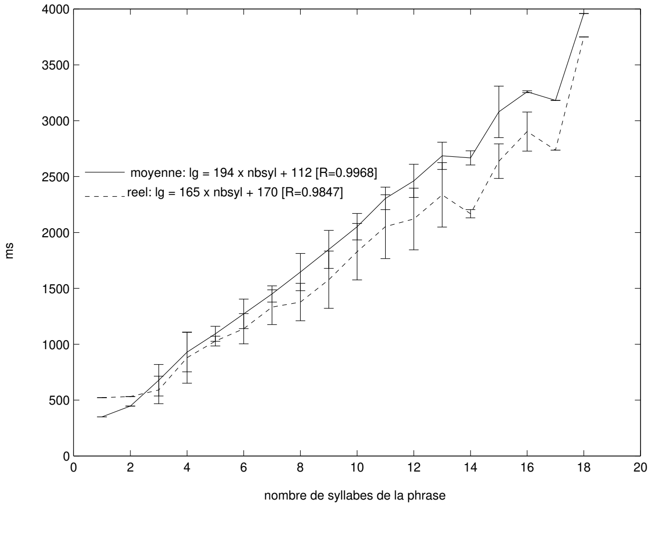


FIG. 1.9: Régression linéaire sur la durée des phrases de 1 à 18 syllabes du corpus d'Aubergé. La courbe en pointillés représente l'évolution de la moyenne des durées réelles des phrases. Pour obtenir la courbe en traits pleins, nous avons remplacé la durée de chaque phonème d'une phrase par les durées phonémiques moyennes du locuteur EV (cf. tableau 1.1).

Pour tenir compte à la fois des caractéristiques intrinsèques des phonèmes et de leur nombre dans le GIPC, nous avons remplacé la PH initiale par la somme des unités moyennes du GIPC. Chaque GIPC présent dans le corpus va donc être caractérisé par un coefficient d'élasticité k_{GIPC}, que l'on notera également k ou *Ratio GIPC*, et qui s'exprime selon la formule suivante :

$$k_{GIPC} = k = Ratio\,GIPC = \frac{d_{GIPC} - \sum_{p_i \epsilon GIPC} dmoy_{p_i}}{\sum_{p_i \epsilon GIPC} dmoy_{p_i}} \tag{1.5}$$

Ce coefficient k_{GIPC} sera prédit par notre modèle de génération biparamétrique de la fréquence fondamentale et du macrorythme (voir le détail au chapitre III). Il s'agit ensuite de retrouver la durée phonémique de chaque constituant du GIPC. Les différentes étapes de ce calcul sont les suivantes :

- On calcule la durée du GIPC courant en multipliant la durée moyenne de ce GIPC par *(k+1)* .
- On détermine le $z - score$ associé au GIPC courant par l'algorithme de répartition (Campbell 1992b) obéissant à l'équation suivante :

$$Durée(GIPC) = \sum_{i=1}^{n} \exp(\mu_i + z\sigma_i) = (k+1) \sum_{i=1}^{n} \exp(\mu_i) \tag{1.6}$$

  ou la paire (μ_i,σ_i) correspond à la moyenne et l'écart-type des durées *log-transformées* des réalisations du phonème i.
- La durée de chaque réalisation phonémique est alors obtenue par l'équation suivante :

$$durée\,réalisation = \exp(\mu_{phonème} + z\sigma_{phonème}) \tag{1.7}$$

1.4 Principes de modélisation adoptés

Notre analyse de la prosodie est guidée par l'approche morphologique que nous avons détaillée en section 1.2.2 : chaque niveau de description linguistique sur l'axe syntagmatique génère des formes prosodiques qui lui sont propres, et qui maintiennent entre elles des oppositions propres à ce niveau (Delattre 1969). Ainsi, une réalisation prosodique observée est le résultat de la superposition de ces formes. Nous nous limitons, pour le moment, à l'étude de phrases isolées et distinguons les niveaux de représentation suivants :

Le niveau de phrase (niveau porteur) dont les contours prosodiques seront essentiellement destinés à véhiculer la modalité de la phrase ou l'attitude du locuteur vis-à-vis de son propos ou de son interlocuteur.

Des sous-niveaux portés (proposition, groupe, sous-groupe...) permettant de représenter la fonction syntaxique de la prosodie de la phrase.

Un niveau mot marquant les phénomènes de mise en relief des mots importants de l'énoncé.

Un niveau morphologique mettant en évidence la structure interne d'un mot complexe (comme par exemple « cytopathogénicité »).

Nota Bene : *Nous emploierons fréquemment le terme* **modulation** *dans ce rapport pour faire référence à la contribution contours portés du niveau du groupe, lesquels, combinés avec le contour prosodique du niveau de phrase, engendrent la réalisation prosodique observée. Ce terme, qui peut prêter à confusion de prime abord, est utilisé pour établir une analogie entre notre description de la prosodie et un système de modulation. L'émetteur (le locuteur) émet un signal modulé composé d'une* **porteuse** *(contours de phrase) modulée par un signal* **modulant** *(contours portés de groupe). Le* **démodulateur** *(fonction du modèle) identifie d'abord les caractéristiques de la porteuse (les contours prosodiques du niveau de phrase) pour ensuite retrouver le modulant (contours du niveau groupe).*

Notre hypothèse initiale est celle de l'indépendance entre les différents niveaux linguistiques. Nous verrons néanmoins en fin de ce mémoire qu'elle sera reformulée afin de rendre compte de certains effacements - locaux ou globaux - des contours associés aux sous-niveaux pour laisser le niveau porteur délivrer son contenu informationnel.

Par une approche descendante, du **signifiant linguistique** vers le **signifié prosodique**, nous avons conçu et analysé des corpus dédiés à chacun de ces niveaux (chapitre II). Ces corpus sont construits de façon à favoriser, par un apprentissage automatique et hiérarchique (du niveau porteur le plus large au niveau porté le plus petit), l'émergence des corrélats prosodiques associés aux attributs linguistiques des représentants de chaque niveau. Nous nous limiterons dans le cadre de cette thèse aux niveaux de la *phrase* et du *groupe*.

L'apprentissage automatique de ces corpus doit permettre de développer un modèle de génération multiparamétrique. Nous avons donc besoin d'une caractérisation cohérente de la prosodie pour révéler les stratégies de coopération entre les différents paramètres étudiés. Notre analyse est à rapprocher des propos de Mehler et al. (1996) qui proposent que l'acquisition précoce de la structure prosodique d'une langue émerge d'une "Time-Intensity Grid REpresentation" (TIGRE) et dans laquelle la voyelle joue un rôle prépondérant. Notre unité prosodique est le *Groupe Inter P-Center* (GIPC) défini comme l'intervalle entre deux *P-center* localisés au voisinage de deux établissements vocaliques consécutifs. Le mouvement mélodique associé à ce GIPC sera porté par son noyau vocalique, et la durée normalisée de ce GIPC (cf. formule 1.5 du coefficient d'allongement k_{GIPC}) constituera l'unité macrorythmique de notre modèle.

Ajoutons que chaque phase de l'apprentissage hiérarchique de notre modèle sera sanctionnée par un test de perception pour évaluer sa capacité à capturer les informations nécessaires et suffisantes pour communiquer les fonctions linguistiques données *a priori* dans chaque corpus (chapitre III).

Chapitre 2

Corpus d'apprentissage : conception, validation et analyse acoustique

2.1 Méthodologie de conception

2.1.1 Corpus dédié à chaque niveau linguistique

Comme nous l'avons décrit à la fin du chapitre précédent, notre méthodologie issue de la phonétique expérimentale (Contini & Profili 1987) repose sur l'analyse de corpus dont la structure doit révéler la nature des **rendez-vous structurels** entre les niveaux de description linguistique (et éventuellement para-linguistique) et la prosodie. L'intérêt principal de notre méthode superpositionnelle réside dans la possibilité de traiter successivement les différents niveaux avec des corpus adaptés.

Dans le cadre de notre étude, nous nous intéresserons, en premier lieu, aux prototypes prosodiques correspondant aux unités phrastiques, et destinés à véhiculer l'attitude du locuteur vis-à-vis de son discours. Cette étude sera réalisée à l'aide d'un premier corpus de phrases relativement courtes et pour lesquelles nous aurons contrôlé au maximum la non réalisation des contours portés.

Une fois que ces prototypes auront émergé de l'apprentissage de notre modèle, nous aborderons alors la modélisation des contours prosodiques d'un niveau inférieur : *le groupe*. Nous appelons groupe (Aubergé 1991, p. 96) tout syntagme maximal, c'est-à-dire tout mot ou groupe de mots qui est un constituant direct de la phrase (ou de la proposition si le niveau propositionnel est représenté dans la phrase). Par exemple, un syntagme nominal (GN) complexe aussi bien qu'un adverbe autonome (GA) sont intégrés au niveau groupe. Le groupe verbal (GV) est défini comme le syntagme constitué uniquement par le verbe et ses composés (verbe conjugué, verbe outil, verbe + infinitif, auxiliaire + participe, épithète, verbe précédé d'un pronom inférieur à la syllabe). Le terme de groupe, employé ici, se place donc plutôt sur une dimension syntagmatique et peut être rapproché de la notion de groupe de sens définie par Emerard (1977).

L'étude du niveau *groupe* passera par la conception d'un second corpus de phrases contenant des unités syntagmatiques plus longues que celles présentes dans le premier corpus, afin de mettre en évidence la modulation introduite par ce niveau sur les contours de phrase.

Comme nous allons le voir par la suite, ce type d'approche nécessite une conception très fine de chaque corpus et un contrôle optimal des différents paramètres qui entrent en jeu dans sa réalisation. Elle passe donc forcément par des enregistrements *in vitro* en laboratoire où le contenu énonciatif et la situation d'élocution sont imposés au locuteur.

2.1.2 Représentativité statistique : contraintes de paires minimales

La méthodologie de conception de ces corpus, héritée des travaux de Delattre (1969) et systématisée par Aubergé (1991), est guidée par une combinaison d'oppositions minimales par paire entre les attributs de chaque niveau linguistique :

- L'opposition au niveau de la phrase sera, dans le cadre de notre recherche, celle de l'attitude du locuteur.

- Au niveau *groupe*, on veillera à faire varier systématiquement la nature du groupe (Groupe Nominal (GN), Groupe Verbal (GV)...), sa fonction (pour un GN : GN sujet ou GN objet), sa position syntagmatique (déplacement d'un adverbe en position initiale, interne et finale ; GN pré-verbal, post-verbal), et sa longueur en terme de syllabes.

- A l'intérieur d'un groupe, nous distinguerons les attributs de nature (déterminant, nom, adjectif, pronom personnel...) et de position relative (adjectif pré-nominal, post-nominal...).

Ce type de méthode permet d'équilibrer la représentativité statistique des différents attributs.

2.2 Corpus dédié à l'apprentissage des contours de phrase

2.2.1 Le corpus "attitude" : conception, enregistrement et évaluation

2.2.1.1 Conception

Nous nous intéressons ici à la manière par laquelle les locuteurs véhiculent par la parole diverses attitudes. Notre but n'est pas de faire un inventaire descriptif des attitudes en français. Nous cherchons simplement à démontrer que, sur un nombre restreint d'attitudes toutes réalisables sur un corpus de phrases non typées, un locuteur produit des formes mélodiques bien identifiées et prototypiques analysables dans le cadre du modèle que nous développons.

Le corpus "attitudes" comporte 322 phrases isolées dont la longueur varie entre 1 et 8 syllabes. Ces phrases seront produites selon six attitudes intonatives référencées dans des ouvrages d'enseignement du français tels que Calbris & Montredon (1981) et Callamand (1973) soit un total de 1932 réalisations. Le contenu du corpus est détaillé dans l'annexe A.1. La longueur réduite des phrases permet de limiter l'influence des contours portés du niveau groupe sur les contours prosodiques[1].

Le choix des attitudes

Parmi les six attitudes retenues, deux se ramènent plutôt à la catégorie des modalités pour laquelle le locuteur a pour tâche de neutraliser son implication vis-à-vis de son propos : la *déclaration* et la *question simple*. Ces deux modalités seront respectivement notées DC et QS.

Les quatre autres sont des intonations de discours faisant apparaître des informations sur la situation dans laquelle se trouvent les interlocuteurs, sur les dispositions du locuteur vis-à-vis de l'objet dont il parle : l'*exclamation de surprise* (EX), le *doute-incrédulité* (DI), l'*ironie de soupçon* (SC) et l'*évidence* (EV).

Comme nous le verrons dans la suite de ce mémoire, la déclaration et la question simple ne semblent pas poser de problème quant à leur définition. Par contre les autres consignes attitudinales nécessiteront quelques précisions. Quoi qu'il en soit, n'ayant pas encore d'information sur l'interprétation des locuteurs-auditeurs sur ces attitudes, nous avons dû, au départ, nous contenter de relever les définitions fournies dans les ouvrages cités plus haut, à savoir :

[1]Elle permet en outre de limiter l'explosion combinatoire liée à la technique des paires minimales.

- L'exclamation de surprise : le locuteur exprime sa stupéfaction.

- Le doute-incrédulité : il exprime le désaccord partiel et à la limite une négation de ce qui a été préalablement affirmé.

- L'ironie de soupçon : on reprend l'affirmation de l'interlocuteur en lui faisant comprendre qu'on la met en doute, qu'on la contredit, ceci par le biais d'une intonation dont la paraphrase serait « Tiens, tiens ! Vous m'étonnez fort ».

- L'évidence : le locuteur est sûr de lui-même.

Pour évoquer d'une manière générale ces 6 catégories, nous utiliserons le terme d'*attitudes* car comme nous l'avons vu en page 13, la distinction entre modalités et attitudes n'est pas claire. Nous englobons donc les modalités dans les attitudes et nous établirons une distinction seulement si leur comportement en production ou en perception sont différents.

Constitution des phrases du corpus

Nombre de syllabes	1	2	3	4	5	6	7	8
Nombre de phrases	5	6	28	58	93	122	5	5

TAB. 2.1: Nombre de représentants du corpus "attitudes" par nombre de syllabes des phrases.

Les 322 phrases du corpus se répartissent en 8 catégories de longueur dont le nombre de représentants est indiqué dans le tableau 2.1. Vous pouvez remarquer que ce nombre est croissant jusqu'à 6 syllabes en raison de l'augmentation des oppositions minimales avec le nombre de syllabes de la phrase. Les phrases de 7 et 8 syllabes n'ont que 5 représentants chacune. Elles seront simplement destinées à tester notre modèle en généralisation à partir des apprentissages réalisés sur le reste du corpus. Pour ces phrases nous n'avons retenu que trois niveaux linguistiques : la *phrase*, le *groupe* et le *sous-groupe*. Le niveau propositionnel, étudié par Aubergé (1991), n'a pas été représenté essentiellement en raison de la faible longueur des phrases du corpus.

Pour chaque longueur de phrase, on trouve, dans l'ordre, des :
- Monomots qui à cause du critère de longueur contiennent, pour certains, des préfixes et des suffixes.
- GN isolés.
- GV seuls introduits par un pronom inférieur à la syllabe (« j' »).
- GV constitué d'un auxiliaire et d'un participe passé avec un adverbe en position interne et finale (GV + Adv).
- Auxiliaire suivi d'un adjectif éventuellement précédé d'un adverbe.
- GNsujet suivi d'un verbe (GNsujet + GV).

– GNsujet suivi d'un verbe avec un adverbe en position initiale, interne et finale (GNsujet + (Adv) + GV).
– GNsujet suivi d'un verbe, tournures négatives (GNsujet + GV + neg).
– GV introduit par un pronom inférieur à la syllabe et suivi par un GNobjet (GV + GNobjet).
– GNsujet précédant un verbe suivi d'un GNobjet (GNsujet+GV+GNobjet).

Voici maintenant, pour les GN et les GV, la liste des catégories lexicales des items qu'ils contiennent :

– Le GN peut être constitué d'un pronom personnel (« tu »), d'un nom propre (« Roger ») ou d'un (article défini / article indéfini / adjectif possessif / pronom pluriel / pronom indéfini / numéral) suivi d'un nom avec parfois un adjectif pré-nominal et post-nominal.
– Le GV est constitué d'un verbe seul, d'un auxiliaire suivi d'un participe passé ou d'un verbe suivi d'un infinitif.

Pour toutes ces catégories, et toutes les unités syntagmatiques contenues dans le corpus "attitudes", le nombre de syllabes a été systématiquement distribué conformément à notre méthodologie de constitution de corpus. Notons enfin que le choix du lexique a été guidé par des contraintes phonétiques privilégiant l'alternance Consonne - Voyelle afin de faciliter l'étiquetage du signal et l'étude rythmique du corpus.

2.2.1.2 Protocole d'enregistrement

Nous avions sélectionné au départ deux locuteurs. L'importance du travail de mise en forme des corpus utilisés nous a contraint à n'étudier qu'un seul locuteur masculin : le jeune thésard Yann Morlec enregistré au tout début de sa thèse. Imprégné des clichés mélodiques proposés par Fónagy (1984) ou Calbris (1981), il avait une bonne connaissance de la façon par laquelle il pouvait véhiculer, par la prosodie uniquement, les 6 attitudes. Il sera vérifié *a posteriori* que les prototypes produits correspondent bien à des universaux (Moroni 1997).

L'enregistrement des 1932 phrases s'est étalé sur 33 sessions de 60 phrases. Les phrases ont été tirées aléatoirement en respectant deux contraintes : (1) ne jamais présenter deux fois de suite la même phrase, (2) ne jamais présenter deux fois de suite la même attitude. Nous avons également veillé à alterner, dans la mesure du possible, la longueur des phrases.

Durant les sessions, chaque phrase était accompagnée de l'attitude à réaliser et d'un texte introducteur destiné à mettre le locuteur en situation. Une fois que le locuteur avait produit son message, il pouvait se réentendre afin de contrôler sa performance, et recommencer éventuellement sa dernière réalisation. Les phrases ont été enregistrées à 10 kHz. L'ensemble du corpus représente environ une heure de parole.

2.2.2 Validation perceptive

Le but de ce premier test psycholinguistique (Grépillat 1996) est triple. Il est destiné en premier lieu à contrôler la capacité du locuteur à transmettre les six attitudes représentées dans le corpus. Nous cherchons également à évaluer si ces six attitudes font partie du code conventionnel du français oral. Enfin, nous voulons obtenir une première mesure de la pertinence des définitions proposées durant le test.

2.2.2.1 Matériel

Pour ce test de perception, nous avons réenregistré à 16 kHz notre locuteur sur des phrases de structures syntaxiques similaires à celles présentes dans le corpus "attitudes". Nous avons choisi des phrases de 1 à 5 syllabes. Pour éviter les effets de focalisation, ou au contraire de réinterprétation par les sujets, nous avons tâché de contrôler le contenu sémantique de chaque phrase afin de le "neutraliser" au mieux.

Le principe syntagmatique de construction de chaque énoncé est le déplacement de la frontière entre le Groupe Nominal (GN) et le Groupe Verbal (GV). Le GN est constitué d'un déterminant inférieur à la syllabe (« l' ») suivi d'un nom, le GV est réduit au verbe, éventuellement introduit par un pronom inférieur à la syllabe (« j' »).

Un énoncé sera constitué d'un GN sujet - dont la longueur augmentera de 0 à 5 syllabes - et d'un verbe dont la longueur diminuera de 5 à 0 syllabes. Nous passerons donc d'une phrase réduite à un verbe précédé du pronom « j' » à une phrase elliptique ne contenant qu'un groupe nominal.

Nous avons essayé de contraindre les formes phonétiques pour produire des énoncés de durées équivalentes variants seulement selon des attributs intonatifs. Cependant, des phrases homophones auraient certainement induit des phénomènes intonatifs particuliers de désambiguïsation. C'est pourquoi nous avons réduit les contraintes phonétiques aux seuls préfixes, choisis phonétiquement dans ([lo], [la] ,[ja]) dans tous les énoncés. Le corpus de phrases utilisé durant ce test, et nommé "identification & gating", est décrit en annexe A.3.

2.2.2.2 Protocole

Douze auditeurs ont participé à cette expérience : étudiants en lettres âgés de 20 à 27 ans, tous de langue maternelle française. Ces auditeurs sont répartis en deux groupes : six effectuent une phase d'entraînement et six commencent le test directement. Cette méthode nous permettra d'observer les éventuelles différences de performance entre les sujets entraînés et les sujets non-entraînés.

S'agissant d'une expérience de perception d'attitudes associées à une morphologie prosodique, et non d'association sémantique, il nous est apparu nécessaire de donner à l'écran la phrase orthographiée.

Le test est précédé d'un message acoustique prononcé par le locuteur de façon à habituer l'auditeur à cette voix. Chaque auditeur, après l'écoute d'un stimulus, doit effectuer un choix parmi l'une des six attitudes (choix fermé et forcé). Il garde sous les yeux une feuille, reproduite ci-dessous, sur laquelle sont notées les définitions des attitudes.

Les auditeurs qui effectuent la phase d'entraînement peuvent reporter, durant celle-ci, leur propres mots-clefs pour définir les stimuli qu'ils ont perçus. Cet entraînement est constitué d'une phrase n'appartenant pas au corpus de test (« j'attends l'autobus ») et prononcée suivant les 6 attitudes prosodiques.

Déclaration `Affirmation neutre (Ex. « Il mange. »)`

Mots-clefs :

Question simple `Question neutre (Ex. « Il mange ? »)`

Mots-clefs :

Exclamation de surprise `Phrase prononcée après un effet de surprise (Ex. « Quoi, il mange ! »)`

Mots-clefs :

Doute-Incrédulité `Exprime un désaccord et à la limite une négation de ce qui a été préalablement affirmé (Ex. « Ah bon, il mange ? »)`

Mots-clefs :

Ironie de soupçon `On reprend l'affirmation de l'interlocuteur en lui faisant comprendre qu'on la met en doute, qu'on la contredit (Ex. « Tiens, tiens ! Vous m'étonnez fort »)`

Mots-clefs :

Evidence `Le locuteur est sûr de lui-même (Ex. « Il mange, c'est évident »)`

Mots-clefs :

Les 19 phrases du corpus sont présentées deux fois (afin de connaître les facteurs d'apprentissage durant le test) mais toutes les phrases sont distribuées dans un ordre aléatoire avec toutefois deux contraintes : deux phrases identiques ne peuvent pas être présentées l'une à la suite de l'autre, deux attitudes identiques ne peuvent pas non plus se succéder.

En résumé, le test est composé de 19 phrases produites selon six attitudes et présentées deux fois soit un total de : $19\,phrases * 6\,attitudes * 2\,répétitions = 228\,stimuli$. Ce test dure environ 45 mn.

2.2.2.3 résultats

(a)

	DC	QS	EX	DI	SC	EV
DC	99.6	0	0	0.4	0	0
QS	0.9	93.9	0.4	3.5	0.4	0
EX	0	0	88.6	8.3	0	3.1
DI	0	0	3.1	82.0	14.5	0.4
SC	0	0	0.4	25.4	72.4	0.4
EV	0	0	1.3	0.4	0	98.2

(b)

	DC	QS	EX	DI	SC	EV
DC	96.9	0	0.4	0	0	2.6
QS	0	90.4	1.8	7.9	0	0
EX	0	0.9	70.6	24.6	3.9	0
DI	1.3	3.5	4.8	69.3	20.6	0.4
SC	1.8	1.3	3.9	39.5	51.3	2.2
EV	2.2	0.9	7.0	1.3	1.8	86.8

TAB. 2.2: Matrices de confusion du test d'identification. (a) avec apprentissage et (b) sans apprentissage.

Résultats généraux

Les résultats principaux que nous pouvons tirer de cette expérience sont les suivants :

- Les attitudes sont bien reconnues dans les deux conditions expérimentales : les sujets non entraînés obtiennent un score global de 78 % et les sujets entraînés un score de 90 % (à comparer avec les 17 % correspondant à un choix aléatoire). Les deux catégories les mieux identifiées sont les DC et QS qui correspondent aux deux modalités de notre corpus.
- La longueur de la phrase a peu d'influence sur le taux d'identification (cf. tableau 2.3) excepté pour les phrases de 1 syllabe.
- La structure syntaxique des phrases n'influence pas non plus les résultats.

Nombre de syllabes	1	2	3	4	5
% identification	67	82	83	84	85

TAB. 2.3: Pourcentage d'identification correct pour chaque longueur de phrase toutes attitudes confondues.

Facteur d'entraînement

Il est apparu de façon très nette (cf. matrices de confusion 2.2) que les sujets ayant suivi la phase d'entraînement ont obtenu des taux de bonnes réponses supérieurs aux sujets non entraînés, et ceci toutes attitudes confondues. Remarquons toutefois que l'amélioration des résultats d'identification est bien plus remarquable pour les attitudes (EX, DI, SC, EV) que pour les modalités (DC, QS) qui ont déjà un score élevé sans apprentissage.

Analyse des reports

On s'aperçoit qu'il existe des reports réguliers selon les attitudes quel que soit le protocole de test (avec *vs* sans entraînement du sujet). Par exemple, les confusions entre DI et SC sont très fréquentes avec toutefois une préférence pour les DI. DI est également reconnu comme une QS ou une EX. Moins significativement, EV est confondue avec EX. Une explication partielle de ces différents reports peut être fournie en dépouillant les feuilles de test des sujets entraînés.

Dépouillement des feuilles de test

La modalité *déclaration* (DC) est définie majoritairement comme étant une affirmation neutre. Pour la *question* (QS), les réponses sont également convergentes : les auditeurs reconnaissent bien une interrogation "neutre" *i.e.* sans codage d'attitude particulier.

En observant les autres réponses, on s'aperçoit que les définitions sont bien moins systématiques :

- L'*évidence* (EV) n'est décrite comme telle que par une seule personne. Les autres sujets la définissent comme étant une « exclamation / affirmation » ou le plus souvent comme une « réponse à une question ».
- De même l'*exclamation de surprise* (EX) n'est définie selon les termes que nous avions employés que par un sujet. Les autres réponses, quant à elles, sont variées : elles vont de « l'affirmation / énervement » à « l'interrogation / surprise » en passant par « l'étonnement après avoir reçu un ordre ou conseil ».
- Le *doute-incrédulité* (DI) est massivement défini comme une « interrogation à laquelle le locuteur ajoute de la surprise ».
- Enfin dans toutes les définitions concernant l'*ironie de soupçon* (SC), on note la présence de négation « désaccord, pied de nez, moquerie » et d'énervement.

2.2.2.4 Discussion

Les six catégories représentées dans le corpus sont bien identifiées différemment mais la définition des attitudes semble varier d'un individu à l'autre.

D'après les résultats, certaines attitudes apparaissaient très proches les unes des autres. C'est le cas par exemple de DI et SC ou bien encore de QS et DI bien que leurs mélodies soient différentes.

Doit-on alors seulement remettre en cause la description des attitudes *i.e.* la définition du code linguistique sous-jacent. Est-ce un problème de distance morphologique ou, au contraire, de distance linguistique (ambiguïté des deux codes) ?

Afin de répondre à cette question, Moroni (1997) a mené une enquête socio-linguistique sur une cinquantaine de sujets. Une tâche d'identification, puis deux tâches de production, ont été proposées à ces sujets :

- Après avoir écouté des exemples de production du corpus "identification & gating" (phrases de 2 et 5 syllabes), l'enquêtrice demandait aux participants de décrire la situation dans laquelle se trouvait le locuteur et, si possible, de donner une définition caractérisant l'attitude qu'il venait de produire.
- Les participants devaient ensuite imaginer, caractériser et produire d'autres attitudes.
- Après une tâche de mémorisation de numéros de téléphone destinée à évacuer dans leur esprit les phrases qu'ils venaient d'entendre, ils devaient, à partir des définitions, produire sur la phrase bisyllabique « L'eau bout » leur propre interprétation des 6 attitudes du corpus.

Les premiers résultats de cette enquête montrent qu'une grande source d'ambiguïté sur l'identification des attitudes provient effectivement des définitions que nous avons proposées *i.e.* d'un problème de distance linguistique. On constate, de manière générale, que la déclaration et la question sont les seules valeurs pour lesquelles les réponses concordent entre les différents sujets. Mais pour la question et les autres attitudes, les auditeurs remettent l'énoncé en contexte dialogique ou l'attribuent à une émotion. Les résultats de l'enquête sont cohérents avec le test d'intelligibilité : SC, qui obtient les plus faibles taux durant le test, reçoit des descriptions très hétéroclites. DI qui joue le rôle d'attracteur reçoit des réponses plus ciblées.

Au niveau de la tâche de production, la déclaration, et dans une proportion moindre la question, sont réalisées avec des contours mélodiques cohérents avec ceux de notre locuteur de référence. Ce n'est pas le cas pour les attitudes qui présentent une grande diversité de réalisations et pour lesquelles on note, chez certains sujets, l'utilisation de détails phonétiques (friction, coup de glotte…) ou d'items prélexicaux comme des interjections (« oh ! », « quoi ! »…), des adverbes ou même un énoncé introducteur. L'apport de ces items d'amorçage peut provenir en partie de la longueur réduite de la phrase proposée (2 syllabes) et on peut se demander s'ils viennent renforcer lexicalement l'attitude prosodique ou bien s'ils l'amenuisent. L'étude de ces phénomènes doit être réalisée dans la continuité des travaux de Alt (1997) sur les liens entre les particules illocutoires de l'allemand et la prosodie.

2.2.3 Analyse acoustique

2.2.3.1 Stylisation de la F0 et extraction des patrons rythmiques

Chaque GIPC de la phrase est caractérisé par 4 paramètres :

1. Les réalisations mélodiques sont caractérisées par trois valeurs de F0 par GIPC. Ces valeurs sont prélevées sur le noyau vocalique (Boë & Contini 1976) : on considère qu'en français, le GIPC reçoit un mouvement mélodique porté par la voyelle et que nous stylisons par une fonction polynomiale du second ordre (Emerard & Benoît 1987). Nous n'avons appliqué aucune correction pour neutraliser les effets de la microprosodie. Ce choix a été guidé par le calcul du nombre d'occurrences des différents phonèmes présents dans le corpus (figure 2.1). Il nous montre que le contenu phonétique des phrases est globalement bien équilibré et donc que l'analyse statistique sur l'ensemble des phrases du corpus devrait "gommer" les variations locales de F0 liées au contexte consonantique. On note que les semi voyelles sont sous-représentées. Ce phénomène résulte de l'alternance Consonne - Voyelle quasi systématique. Les plosives et les fricatives sont, quant à elles, en nombres à peu près équivalents.

   Les trois valeurs de F0 sont exprimées sur une échelle logarithmique en quart de tons. Le passage des Hertz aux quarts de tons nécessite, en premier lieu, l'évaluation de la fréquence laryngienne moyenne du locuteur de référence. Nous avons prélevé cette valeur en calculant la F0 moyenne des syllabes non accentuées sur toutes les déclarations du corpus "attitudes". Le registre obtenu pour notre locuteur est $F0_{moyHz} = 126Hz$. Les valeurs de F0 en quart de tons sont calculées à l'aide de la formule suivante :

$$F0_{1/4tons} = 24log_2(F0_{Hz}/F0_{moyHz}) \tag{2.1}$$

2. Les patrons rythmiques sont extraits à l'aide de l'algorithme détaillé dans le premier chapitre de ce mémoire. Rappelons tout de même l'expression du facteur d'allongement k_{GIPC} associé à chaque GIPC du corpus :

$$k_{GIPC} = \frac{d_{GIPC} - \sum_{p_i \epsilon GIPC} dmoy_{p_i}}{\sum_{p_i \epsilon GIPC} dmoy_{p_i}} \tag{2.2}$$

   Dans la suite de notre étude, le dernier GIPC de phrase sera réduit à la voyelle finale de celle-ci.

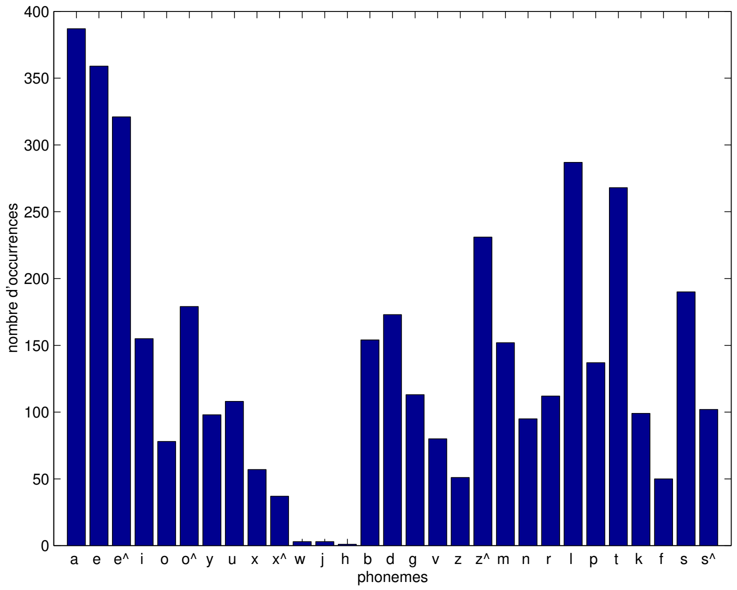


Fig. 2.1: Occurrences des phonèmes du corpus de référence. Dans l'ordre : voyelles, semi voyelles et consonnes. Les phonèmes sont notés en code CPA.

Avant d'entamer la section suivante sur l'analyse du corpus découlant de cette stylisation mélodique et de cette normalisation du GIPC, faisons un bref aparté sur la représentation des patrons mélodiques et rythmiques que nous emploierons dans la suite de ce mémoire.

Nous utiliserons une échelle logarithmique dont l'unité est le *1/4 de ton* pour représenter les contours de F0. La valeur 0 correspondra donc au registre moyen de notre locuteur soit 126 Hz. Chaque GIPC sera caractérisé par 3 valeurs reliées entre elles par des segments de droite.

Pour représenter le facteur d'allongement k_{GIPC}, noté Ratio GIPC sur les figures, les valeurs discrètes seront connectées entre elles de façon à révéler plus clairement les caractéristiques rythmiques des phrases.

2.2.3.2 Mise en évidence des mouvements porteurs

Les figures 2.2 permettent d'illustrer le principe superpositionnel que nous avons adopté :

- La première représente les contours stylisés de F0 d'une vingtaine de phrases déclaratives de 5 syllabes tirées au hasard dans le corpus "attitude".
- Dans la seconde, nous n'avons conservé que les contours de monomots c'est-à-dire les unités qui n'engendrent pas de contours portés (niveau morphologique excepté). Cette figure permet de mettre en évidence l'allure du contour de phrase déclarative

marquée par une ligne de déclinaison.

– Dans la troisième, nous avons gardé une partie de ces monomots (gris clair) auxquels nous avons ajouté des contours de phrase [GNsujet+verbe] contenant un sujet de 2 syllabes (pointillés), trois syllabes (gris foncé), 4 syllabes (traits pleins). Nous notons alors, sur ces phrases, l'apparition de la modulation de groupe qui vient se superposer au contour de phrase.

Une représentation logarithmique de la F0 nous permet de faire l'hypothèse que ces contours mélodiques de phrases isolées peuvent être le résultat de l'addition d'un contour porteur de phrase - bien identifiable sur les monomots - et de contours portés de groupe signalant ici la démarcation entre le GNsujet et le verbe.

Nous voulons montrer, dans la suite de ce chapitre, que l'étude des phrases monomots de notre corpus permet de faire émerger des contours prototypiques de l'attitude du locuteur. Nous étudierons également par une analyse acoustique d'un second corpus (le corpus "modulation") les variations subies par ces contours. Enfin, nous mettrons en évidence que l'approche superpositionnelle peut également s'appliquer aux patrons rythmiques.

2.2.3.3 Caractérisation des mouvements

L'analyse acoustique qui suit est principalement destinée à définir la notion de mouvement prosodique. La caractérisation de ces mouvements est un passage nécessaire en vue de définir une paramétrisation adéquate de notre modèle de génération de la prosodie.

Cette section s'appuie sur les figures 2.4 et 2.5 représentant pour les six attitudes l'évolution des mouvements mélodiques (colonne de gauche) et rythmiques (colonne de droite) des monomots de 1 à 6 syllabes. Sur ces figures sont reportés les valeurs moyennes (en pointillés) et les écart-types (traits pleins). Ce type de représentation facilite la lecture des données mais présente néanmoins certains risques. En effet, l'écart-type est simplement le reflet d'une déviation globale par rapport à la valeur moyenne *i.e.* un ensemble de contours mélodiques évoluant parallèlement au contour moyen peut générer le même écart-type que des mouvements en opposition autour de ce contour. Nous resterons donc très prudents quant à l'interprétation quantitative des courbes.

Pour les phrases de 6 syllabes, trois paramètres principaux ont été extraits (figure 2.6) : le registre moyen (moyenne par attitude des valeurs de F0), le débit moyen (moyenne par attitude de tous les ratios GIPC excepté le dernier de chaque phrase) et la durée moyenne du dernier GIPC correspondant à la dernière voyelle de la phrase.

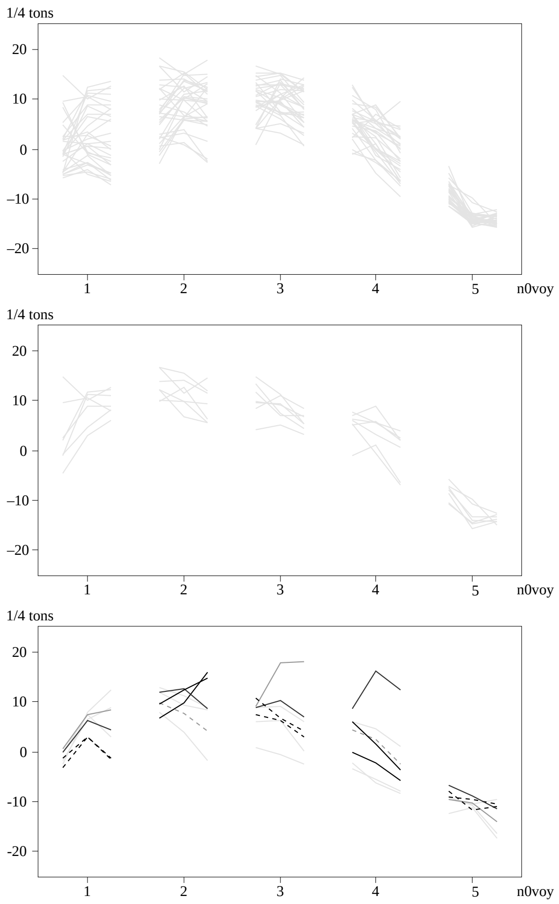


Fig. 2.2: Superposition de contours de F0 (1/4 de tons) pour des *déclarations* de 5 syllabes : tirage aléatoire (en haut), monomots (au milieu) et mise en évidence des contours porteurs/portés (en bas).

Caractérisation intra-attitudes

Caractéristiques mélodiques :

DC : Les monomots déclaratifs se traduisent mélodiquement par un mouvement continu : partant d'une valeur initiale de 5 $1/4tons$, d'abord croissant sur le premier GIPC, il décroît progressivement jusqu'à une valeur finale de -15 $1/4tons$. La constance de ces deux extrêmes avait déjà été remarquée par Aubergé (1991). Elle avait donc modélisé les contours de déclaration par une droite reliant ces deux points d'ancrage mais, ici, on voit une évolution plus complexe.

QS : Les questions sont caractérisées par un mouvement globalement montant : d'une valeur initiale sur le premier GIPC proche du registre du locuteur, le contour croît faiblement jusqu'au dernier GIPC qui reçoit un mouvement montant rapide atteignant 30 $1/4tons$.

FIG. 2.3: Deux versions de courbes mélodiques produites sur la phrase « Bob's going out with Anna. » (reproduit d'après Beckman 1997). D'après l'auteur, la courbe mélodique du milieu traduit l'incertitude ("uncertainty about the pragmatic appropriateness of the information conveyed"), et celle du bas une stupéfaction incrédule ("incredulous amazement").

EX : Le contour mélodique des exclamatives est beaucoup plus modulé. Le mouvement mélodique sur les deux derniers GIPC *descendant-(montant-descendant)* est caractéristique de cette attitude. Il est précédé par des variations importantes liées principalement à la frontière morphologique interne pour les monomots polymorphémiques.

DI : Le contour mélodique des DI est sans doute le plus particulier. Il débute par un ton bas : d'abord à -10 $1/4tons$, il évolue progressivement vers le registre du locuteur quand la phrase s'allonge. La deuxième partie du contour, sur les deux derniers GIPC, est caractérisée par une mise en relief de l'avant-dernier GIPC suivie d'un mouvement final montant sur le dernier.

SC : La mélodie de cette attitude est caractérisée par un ton haut (15 $1/4tons$), presque plat jusqu'à l'avant-dernier GIPC puis, brusquement, sur le dernier GIPC, par une chute de fondamental (-15 $1/4tons$).

EV : Pour l'évidence le mouvement évolue de façon régulière : d'une valeur initiale de 15 $1/4tons$, il décroît progressivement jusqu'à l'anté-pénultième GIPC (-10 $1/4tons$) pour ensuite remonter sur l'avant-dernier GIPC et décroître de nouveau sur le dernier.

Deux remarques découlent de cette description :

- Tout d'abord, et c'est particulièrement le cas pour les DI, les contours de phrases véhiculant l'attitude du locuteur peuvent être sujets à des variations de pente très brusques. Les modèles qui opposent les variations lentes des contours ou des gabarits de phrase à celles plus rapides liées à la composante accentuelle, auront donc du mal à s'accommoder de ces contours de phrase accidentés. Ainsi, dans le modèle de Fujisaki et al. (1971), on devra générer ce contour spécifique à DI par une action coordonnée de deux commandes de phrase (début et fin de contour) et d'une commande d'accent (ictus mélodique sur la pénultième).

  Nous verrons plus loin que la proéminence mélodique sur la pénultième ne dépend pas de la structure syntaxique de la phrase. Ce type de contour existe également en anglo-américain. Nous faisons donc une brève parenthèse sur la description des DI selon l'approche tonale. Nous reproduisons en figure 2.3 une version de notre Doute-Incrédulité en anglo-américain. En suivant le formalisme TOBI, Beckman attribue pour la phrase « Bob's going out with Anna. » la séquence de tons suivante : L*+H sur la première syllabe de **A**nna, puis deux tons L- et H% qui se partagent la deuxième et dernière syllabe de A**nna**. Sur le spectrogramme, la dernière voyelle est donc littéralement coupée en deux. Deux questions se posent alors : (a) deux tons peuvent-ils donc s'appliquer à la même syllabe ? (b) « Anna » a effectivement une syllabe forte en initial, mais que serait-il advenu si cela n'avait pas été le cas ? Nous montrons, dans notre analyse, que nous aurions sans nul doute un accent mélodique sans que celui-ci puisse être prédit par une quelconque unité accentuelle.

- Cette description met en évidence des différences claires entre les six contours mélodiques. Il est cependant intéressant de noter, pour chacun d'eux, la précision du mouvement mélodique final sur la dernière voyelle (écart-type faible).

Caractéristiques rythmiques : Contrairement aux contours mélodiques, les patrons rythmiques ont une évolution globale similaire toutes attitudes confondues. On retrouve le schéma classique caractérisé par une évolution monotone : faiblement croissant au début les ratios GIPC augmentent essentiellement sur les deux derniers GIPC.

A l'inverse de la précision des valeurs finales de F0, nous constatons cette fois une grande variabilité du Ratio GIPC associé à la dernière voyelle. Barbosa (1994, p. 100) avait déjà remarqué ce phénomène suggérant, comme l'avaient préalablement confirmé Nishinuma & Santi (1992), que l'oreille est moins sensible aux allongements de fin de phrase. Ces écart-types importants peuvent également provenir de la difficulté, durant l'étiquetage du corpus, d'indiquer précisément quand se termine la dernière voyelle.

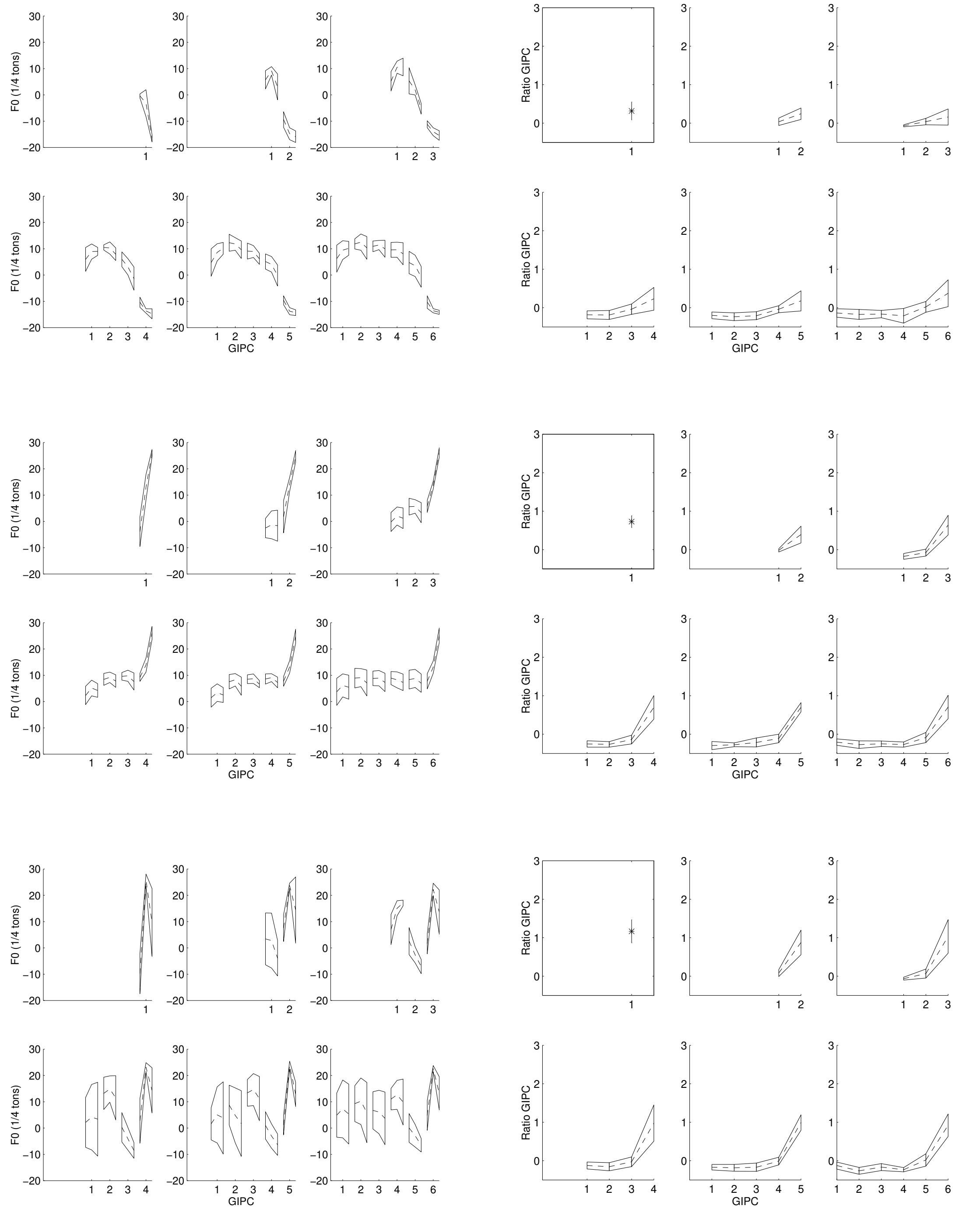


FIG. 2.4: Moyennes et écart-types. F0 (à gauche) et Ratio GIPC (à droite) de monomots : *déclaration*, *question* et *exclamation* de 1 à 6 syllabes.

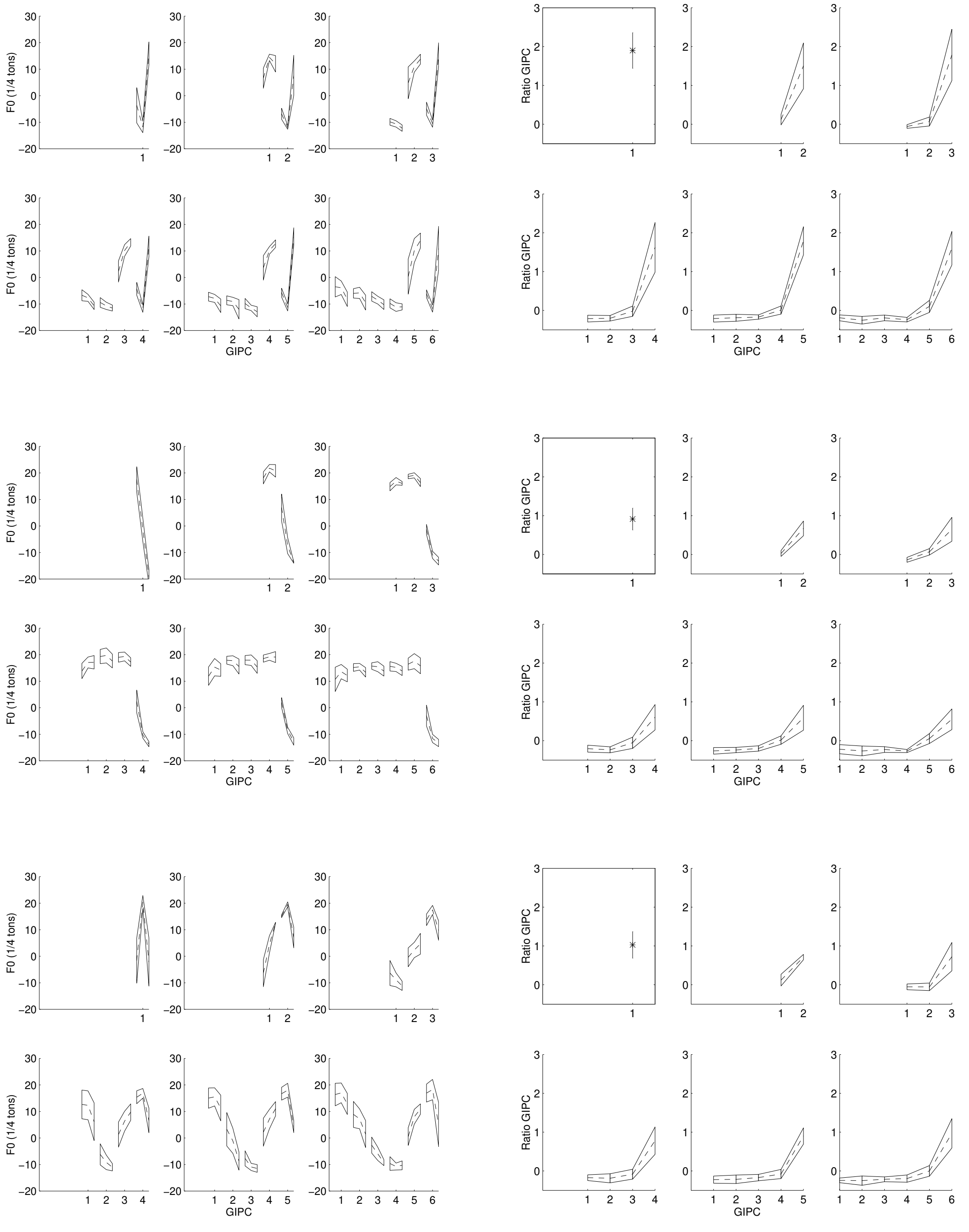


FIG. 2.5: Moyennes et écart-types. F0 (à gauche) et Ratio GIPC (à droite) de monomots : *doute-incrédulité*, *ironie de soupçon* et *évidence* de 1 à 6 syllabes.

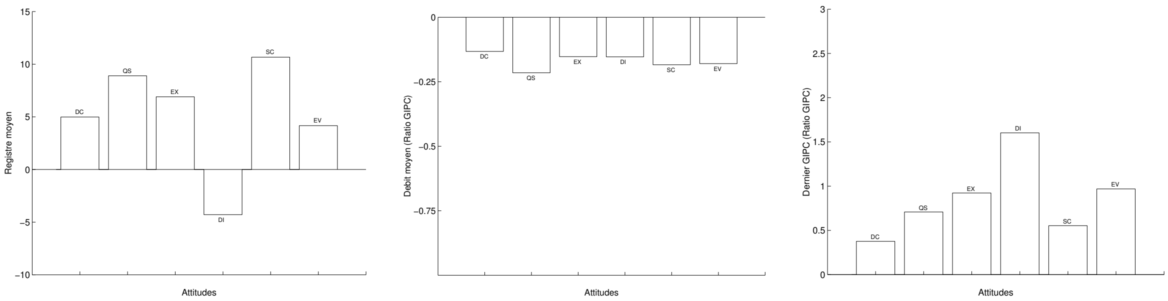


FIG. 2.6: De gauche à droite : registre moyen, débit moyen, durée du dernier GIPC pour des monomots de 6 syllabes.

Caractérisation inter-attitudes

Caractéristiques mélodiques : Nous cherchons ici à comparer les registres moyens du locuteur lorsqu'il produit les différentes attitudes (figure de gauche 2.6). Ce paramètre global indique que pour toutes les attitudes, excepté DI, le registre moyen pour des monomots de 6 syllabes est bien au dessus de la fréquence laryngienne moyenne du locuteur (126 Hz). DI et SC sortent clairement du lot : DI par son registre très bas et SC par son registre élevé.

Caractéristiques rythmiques : Toutes les attitudes ont un débit moyen supérieur à la valeur de référence $d_{GIPC} = \sum_{p_i \epsilon GIPC} dmoy_{p_i}$ ce qui indique un débit rapide sur des phrases monomots.

L'allongement de la dernière voyelle est un facteur plus discriminant. On note les valeurs les plus élevées pour les EX, EV et surtout DI, attitudes pour lesquelles les mouvements mélodiques terminaux sont les plus complexes. Il faut du temps pour les réaliser et les percevoir[2].

Les effets de la modulation

La figure 2.7 représente les contours mélodiques et rythmiques moyens pour DC, EX, QS, DI et toutes les phrases du corpus “attitudes” de 1 à 6 syllabes. La figure 2.8 correspond au registre moyen, débit moyen et allongement terminal pour chaque attitude sur l'ensemble des phrases de 6 syllabes.

Au niveau des contours mélodiques, nous pouvons constater que QS et DI ne subissent

[2]Ce type de mouvement est pris en compte dans l'approche tonale en intégrant, à la description de l'intonation, des tons dynamiques par opposition au tons statiques ou plats (d'Alessandro & Mertens 1995).

pas de modulation liée aux variations de la structure syntaxique des phrases. Ce sont principalement les contours rythmiques qui vont remplir cette fonction structuratrice. Ce phénomène que nous révèle l'analyse multiparamétrique devra être pris en considération par le modèle de génération que nous présentons au chapitre suivant.

Nous constatons que, par rapport aux monomots, il n'y a pas de variations sensibles du registre moyen, Notez simplement une diminution du registre des DC essentiellement lié à l'abaissement de la valeur initiale des contours mélodiques qui, maintenant, se situe autour du registre moyen du locuteur.

Complétons cette analyse en remarquant la modification de comportement des contours rythmiques de QS et DI qui ont évolué, d'un mouvement monotone croissant, vers un mouvement non monotone (cf. phrases de 6 syllabes) pour lequel le rythme commence par accélérer pour, ensuite, ralentir fortement. Ce comportement transparaît dans les valeurs très faibles des ratios GIPC moyens (donc un débit rapide) pour ces deux attitudes, et un allongement très important de la dernière voyelle.

2.2.3.4 Expansion du mouvement

Le but de cette section n'est pas uniquement descriptif. Nous cherchons principalement à définir les informations nécessaires et suffisantes que nous fournirons à notre modèle en vue de générer les différents contours mélodiques et rythmiques que nous avons rencontrés précédemment, et ceci quel que soit le nombre de syllabes de la phrase considérée.

La description que nous venons de réaliser sur les monomots nous permet d'avancer un certain nombre d'hypothèses sur la réalisation de l'expansion du mouvement prosodique lorsque le nombre de syllabes de la phrase augmente :

- Nous remarquons, tout d'abord, que l'hypothèse de l'existence d'un contour global correspondant à l'attitude du locuteur semble être vérifiée. Cette hypothèse sera validée dans l'expérience de dévoilement progressif que nous décrivons dans la section suivante.
- Pour toutes les attitudes, l'allongement final est toujours réalisé et son amplitude dépend de l'attitude. Quand le nombre de syllabes de la phrase augmente, le contour rythmique reste globalement monotone et croissant.
- Le mouvement mélodique final est caractéristique de l'attitude. Quasi invariant, il est porté dans certains cas par la dernière syllabe uniquement (DC, QS), dans d'autres, il se répartit sur les deux dernières syllabes (DI). Ce phénomène pourrait expliquer que le score global d'identification sur des phrases ne comportant qu'une seule syllabe, et donc ne permettant pas toujours la réalisation complète du mouvement, est inférieur aux autres (cf. tableau 2.3). Ce contour final est ensuite complété pour couvrir l'ensemble de l'unité phrastique.

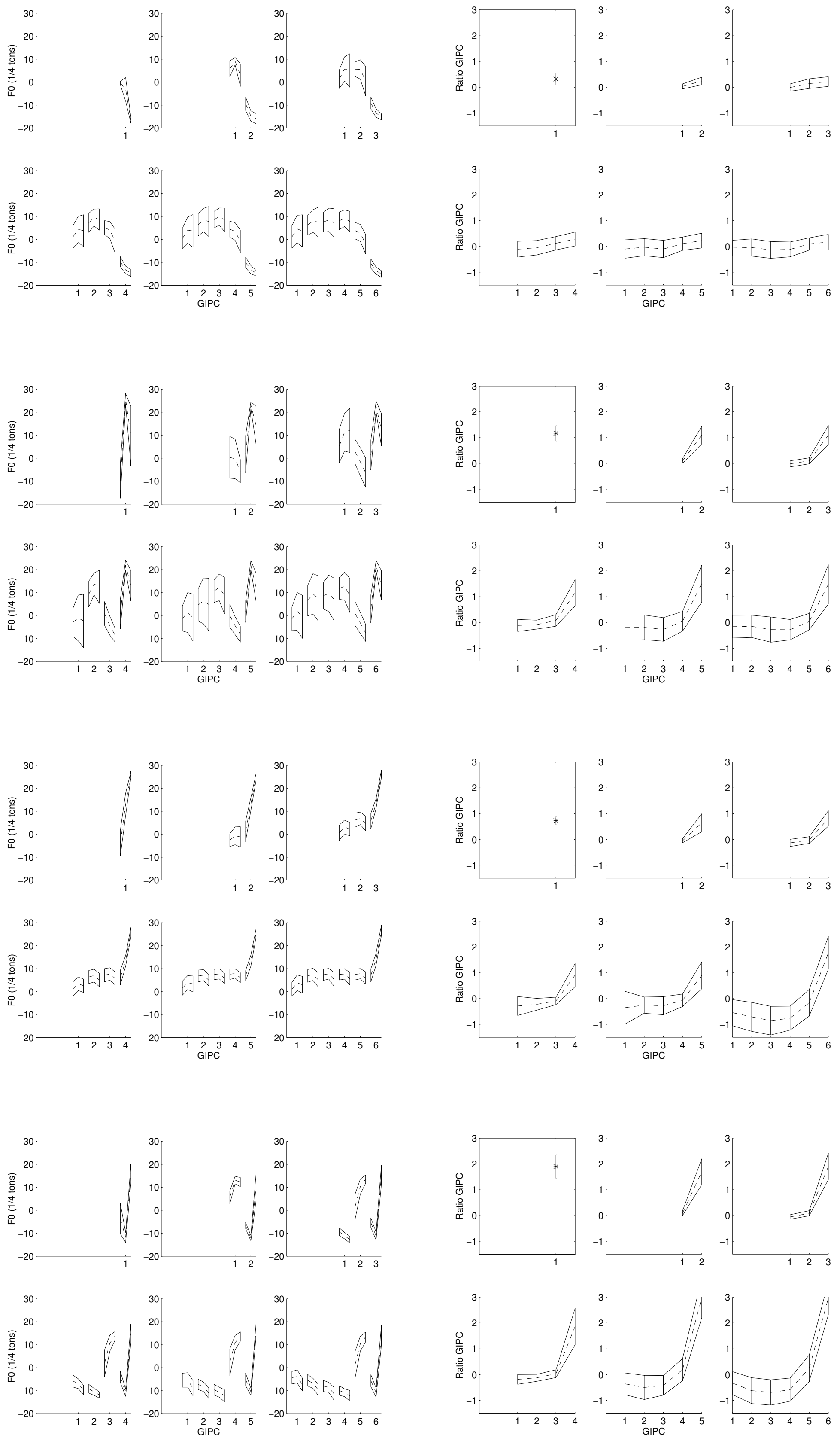


FIG. 2.7: Moyennes et écart-types. F0 (à gauche) et Ratio GIPC (à droite) de tout le corpus "attitudes" : *déclaration*, *exclamation*, *question* et *doute-incrédulité* de 1 à 6 syllabes.

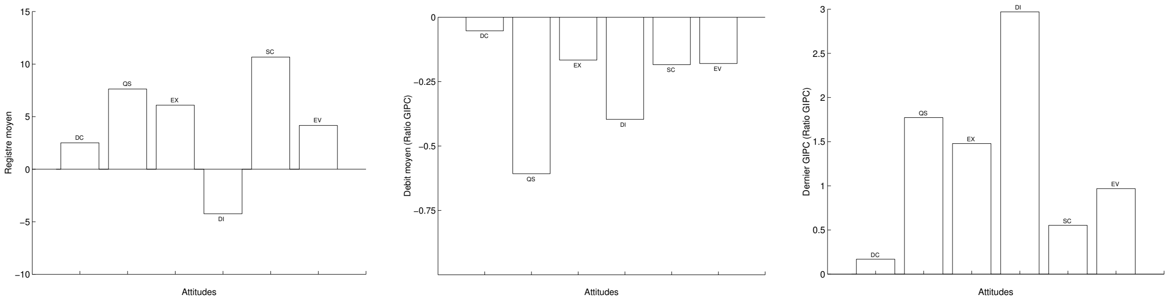


FIG. 2.8: De gauche à droite : registre moyen, débit moyen, durée du dernier GIPC pour toutes les phrases de 6 syllabes.

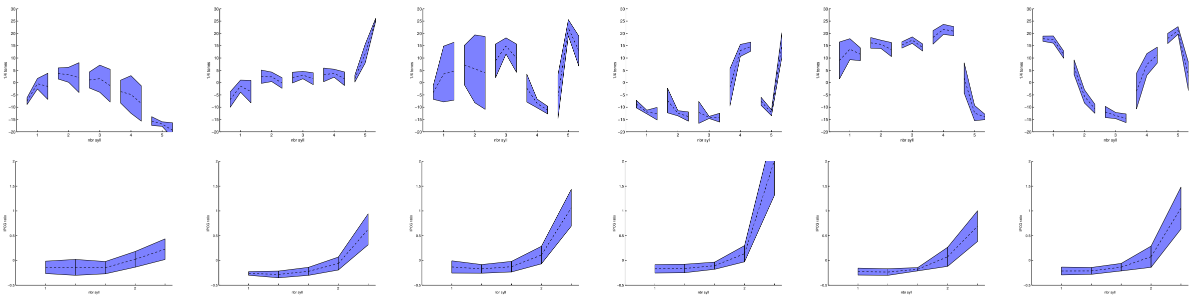


FIG. 2.9: Moyennes et écart-types de F0 (en haut) et de Ratios GIPC (en bas) : *déclaration*, *question*, *exclamation*, *doute-incrédulité*, *ironie de soupçon* et *évidence* du corpus "identification & gating".

– Les contours mélodiques prototypiques de phrase paraissent donc se diviser en deux parties : une phase - que nous nommerons *préparation* - où le mouvement est réparti sur un certain nombre de syllabes. Cette phase prépare la réalisation d'un mouvement mélodique de fin de phrase ancré sur la structure syllabique et couvrant la dernière ou bien les deux dernières syllabes. Nous l'appellerons *capture*.

Le modèle que nous allons développer devra tenir compte de ces deux phases du mouvement.

2.2.4 Contours comme monnaie d'échange

2.2.4.1 Le paradigme du dévoilement progressif ("gating")

2.2.4.2 Introduction

Nous venons de montrer dans la section précédente que l'analyse statistique des phrases du corpus "attitudes" permet d'extraire des patrons intonatifs et rythmiques caractéristiques de l'attitude du locuteur. L'enjeu essentiel de notre travail est de démonter que l'information linguistique est répartie sur ces contours prototypiques de phrase, en ayant conscience que certaines parties de ces contours porteront plus d'information que d'autres. En d'autres termes, nous cherchons à évaluer dans quelle mesure ces contours, dans leur globalité, constituent l'une des monnaies d'échange d'informations entre locuteur et auditeur.

En suivant les propos de Grosjean (1983), nous pensons que l'auditeur utilise les informations passées et courantes, non seulement pour traiter la phrase jusqu'au point atteint par le locuteur, mais également pour prédire les informations à venir. Ce phénomène peut se nommer : anticipation, prévision, attente... Afin de tester cette capacité de l'auditeur et éventuellement de trouver à partir de quel moment (point d'unicité) il sera capable d'identifier l'attitude du locuteur, nous avons appliqué le paradigme du **dévoilement progressif** ("gating") aux contours prosodiques de phrase.

Avant de développer l'expérience de Grépillat (1996) menée à l'ICP, rappelons les principes de ce paradigme et résumons brièvement quelques travaux de recherche effectués dans ce domaine.

Le paradigme du "gating" consiste à dévoiler progressivement aux auditeurs un stimulus de parole depuis son commencement jusqu'à sa présentation complète. Ce stimulus peut être une syllabe, un mot, un syntagme ou même une phrase complète suivant le type de test que l'on veut réaliser. Le découpage du stimulus peut être temporel - on présente à chaque fois une portion du signal plus longue en ajoutant un intervalle de temps régulier - syllabique - on coupe le signal à la fin de chaque syllabe - ou encore par mot. Nous appellerons cette portion de signal une *fenêtre* ("gate").

Ce paradigme a été essentiellement employé pour étudier le phénomène d'amorçage lexical (à partir de quel moment est-t-on en mesure d'identifier un mot ?) comme par exemple dans les expériences de Cotton & Grosjean (1984). Dans ce type d'expériences, l'étude fine du point d'unicité favorise généralement un découpage temporel allant jusqu'à des portions de syllabes (Dupoux 1989).

D'autres études, utilisant ce paradigme, ont été menées dans le domaine du discours spontané afin, d'une part, d'en étudier certains phénomènes comme les répétitions où les faux départs appelés "disfluencies" (Lickley, Shillcock & Bard 1991, Lickley & Bard 1993), et, d'autre part, de trouver les indices qui permettent aux auditeurs de différencier parole lue et parole spontanée (Blaauw 1995).

Mais ce sont les travaux de Thorsen sur le danois (1980) qui se rapprochent le plus de l'expérience de dévoilement progressif que nous décrivons dans cette section. Thorsen réalise, dans un premier temps un test d'identification à choix forcé sur des réalisations d'une même chaîne phonétique qui ne diffèrent que par leur fréquence fondamentale. Elles correspondent soit à des phrases complètes - déclaratives ou interrogatives - soit à des phrases coupées à une frontière de syntagme. Ces réalisations contiennent chacune 3 groupes accentuels. L'auteur montre que les réponses sont très corrélées avec la fréquence fondamentale. Les contours les plus descendants sont identifiés comme déclaratifs, les contours les moins descendants comme interrogatifs, et les contours intermédiaires comme non-finaux. Dans une seconde expérience, l'expérience de "gating" proprement dite, les sujets doivent identifier ces trois catégories en écoutant les mêmes réalisations mais dont la partie finale est tronquée de plus en plus précocement (découpage syllabique). Thorsen montre, après analyse des résultats, que le taux d'identification décroît très lentement avec ce découpage et devient subitement très faible quand il ne reste que le premier groupe accentuel. L'auteur conclut que la première partie de la phrase porte déjà de l'information sur la modalité. Le mouvement final n'est pas indispensable à l'identification.

Plus récemment, Van Heuven (1997) a réalisé une expérience de "gating" sur le hollandais. Pour son test, il considère quatre versions de la même phrase : déclaration, question introduite, question inversée et question déclarative. Nous ne citerons ici que l'expérience concernant la comparaison entre déclaration et question déclarative qui ne se différencient que par leurs paramètres prosodiques. L'auteur utilise une seule phrase de référence de 11 syllabes découpée en six fenêtres irrégulières définies en fonction des positions des accents dans la phrase et du mouvement final. Ses résultats montrent que les auditeurs sont capables d'exploiter très tôt dans la phrase des indices prosodiques pour identifier les deux modalités, et ceci bien avant le mouvement final.

Plusieurs expériences de dévoilement progressif ont donc déjà prouvé l'importance de la prosodie dans le déchiffrage du message en montrant que l'information qu'elle porte est redondante sur celui-ci. Mais peu d'entre elles n'a encore traité le problème de l'identification précoce des catégories attitudinales d'une langue. C'est ce que nous nous proposons de faire maintenant.

2.2.4.3 Matériel

Pour cette expérience de dévoilement progressif, seules les 3 phrases de 2 syllabes, et les 6 phrases de 5 syllabes du corpus "identification & gating" (cf. annexe A.3) ont été retenues. Elles représentent les deux longueurs extrêmes du corpus. Les contours mélodiques et rythmiques pour les phrases de 5 syllabes sont représentés en figure 2.9.

Le découpage est syllabique : les (N-1) jusqu'à la dernière syllabe d'une phrase de N syllabes sont masquées par un bruit blanc. La durée de ce bruit a été fixée de façon à obtenir des stimuli d'une durée constante égale à 2 secondes, et donc bien supérieure à la durée des stimuli. Cette méthode permet d'empêcher l'auditeur d'avoir un indice sur la durée totale de la phrase ainsi masquée.

Pour chaque phrase, les six attitudes sont représentées. On obtient alors 2 stimuli par phrase de 2 syllabes, et 5 stimuli par phrase de 5 syllabes soit, au total, 216 stimuli par auditeur.

2.2.4.4 Protocole

Douze auditeurs entre 20 et 42 ans de langue maternelle française, et n'ayant pas réalisé le test d'identification, participent à cette expérience.

Six d'entre eux écoutent d'abord les phrases de 2 syllabes puis celles de 5 syllabes (nous les nommerons sujets "normaux") alors que les six autres écoutent d'abord les phrases de 5 syllabes puis celles de 2 syllabes (sujets "inversés").

Cette fois, contrairement au test d'identification, les auditeurs n'effectuent pas de phase d'entraînement puisque nous voulons connaître les mécanismes spontanés chez l'auditeur-sujet. Or, comme nous l'avons remarqué durant le test d'identification, les phases d'entraînement augmentent le taux de bonnes réponses.

Durant toute l'expérience, l'auditeur a, à l'écran, la trace écrite de la phrase qu'il est en train de traiter. Ceci permet d'éviter les ambiguïtés liées à la sémantique. L'auditeur conserve devant lui une feuille sur laquelle sont inscrites les définitions des attitudes (cf. test d'identification). Le test est précédé d'un message acoustique prononcé par le locuteur de façon à habituer l'auditeur au son de sa voix. Le sujet doit ensuite attribuer à chaque stimulus tronqué (choix forcé) l'une des 6 attitudes et, ensuite, donner son degré de certitude sur l'échelle suivante : *Tout à fait sûr*, *Moyennement sûr* et *Pas du tout sûr*.

2.2.4.5 Résultats généraux

Notre but n'est pas ici de faire l'inventaire de tous les résultats qui ont découlé de ce test (vous les trouverez dans Aubergé et al. 1997) mais simplement de mettre en lumière la capacité des auditeurs à identifier l'attitude du locuteur avant la fin de son énoncé, et d'évaluer l'effet de l'ordre de présentation sur cette identification (comparaison entre sujets "normaux" et "inversés"). Les résultats concernant ces deux points précis sont les suivants :

La figure 2.10 représente les pourcentages d'identification fenêtre par fenêtre pour les phrases de 5 syllabes et les 6 attitudes. Nous pouvons constater clairement la capacité d'anticipation des auditeurs : dès la deuxième fenêtre, le pourcentage est supérieur à un choix aléatoire (tracé en pointillés). Remarquez que les attitudes les mieux reconnues sont DC et QS comme pour le test d'identification. Précisons simplement que le taux d'identification très élevé des DC quelle que soit la taille du masquage de l'énoncé provient, pour une part importante, d'un choix préférentiel vers cette catégorie dans les cas d'indécision. Notons enfin que, pour DI essentiellement, le taux d'identification n'évolue pas linéairement avec le nombre de fenêtres.

Le tableau 2.4 montre que le taux d'identification moyen, sur toutes les attitudes et toutes les fenêtres des phrases de 5 syllabes, n'est pas affecté par l'ordre de présentation des stimuli. *A contrario* les sujets "inversés" identifient bien mieux les stimuli masqués de 2 syllabes que les sujets "normaux".

Nous pouvons penser que les sujets qui ont déjà entendu les contours prosodiques tronqués et complets de 5 syllabes, sont mieux à même de les réduire à 2 syllabes tandis que qu'il est plus difficile pour les auditeurs d'étendre leur apprentissage sur 2 syllabes aux phrases de 5 syllabes. Cette hypothèse doit néanmoins être nuancée car les phrases de 5 syllabes ont beaucoup plus de représentants que celles de 2 syllabes.

Nombre de syllabes de la phrase	2	5
% ordre "normal"	43	55
% ordre "inversé"	61	56

TAB. 2.4: Pourcentage global d'identification en fonction de l'ordre de présentation pour toutes les fenêtres et toutes les attitudes.

2.2.5 Discussion

Dans cette section, nous avons montré qu'à partir de la conception systématique d'un corpus de phrases simples, de 1 à 6 syllabes essentiellement, nous pouvions extraire des contours prosodiques globaux véhiculant l'attitude du locuteur. Ces contours, soumis à des tests d'identification et de dévoilement progressif, ont été dans l'ensemble bien reconnus et bien "anticipés" par les auditeurs. L'analyse du corpus a révélé certaines caractéristiques majeures de l'expansion du mouvement prosodique de phrase liée à l'allongement progressif de la phrase en terme d'unités syllabiques. L'influence des modulations de groupe a déjà été partiellement discutée. La section qui suit est destinée à prolonger cette discussion à l'aide d'un corpus dédié à l'analyse des contours portés de groupe.

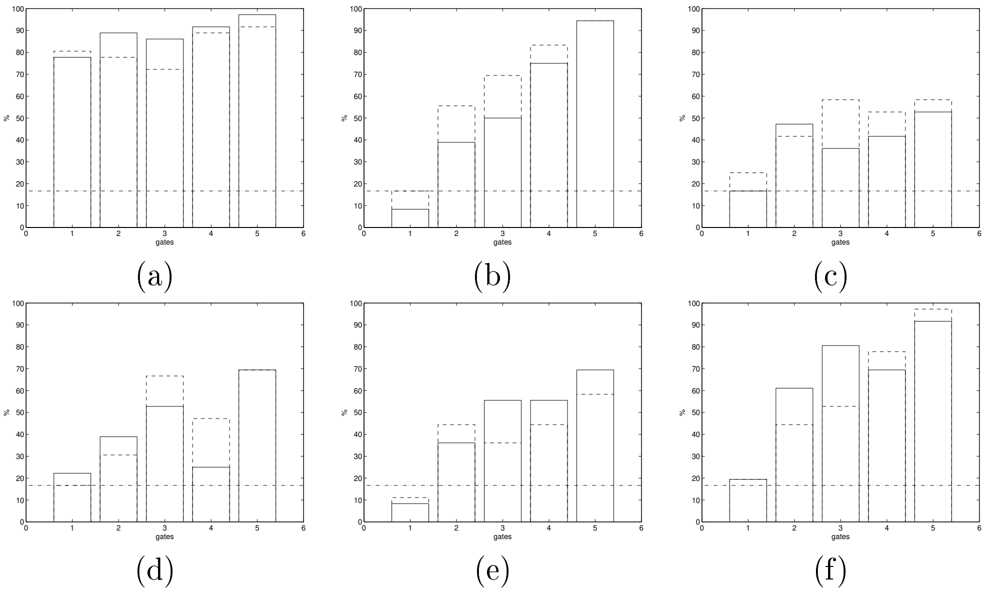


FIG. 2.10: Pourcentages d'identification fenêtre par fenêtre des sujets "normaux" (traits pleins) et "inversés" (traits pointillés) pour les phrases de 5 syllabes. (a) *déclaration*, (b) *question*, (c) *exclamation*, (d) *doute-incrédulité*, (e) *ironie de soupçon* et (f) *évidence*.

2.3 Corpus dédié à l'apprentissage des contours de groupe

2.3.1 Principes de construction : objectifs et limitations

Le corpus "modulation" est destiné à compléter la panoplie des phrases déclaratives (DC) du corpus "attitudes" afin de faire émerger des contours prosodiques de groupe révélateurs de la structure syntaxique de la phrase. Les deux structures principales qui nous intéressent à ce niveau sont [GNsujet+verbe] et [GNsujet+verbe+GNobjet]. L'objectif de corpus est donc essentiellement double :

1. Ajouter aux GNsujets de 1 à 5 syllabes présents dans le corpus "attitudes" des représentants de 6 syllabes dans les structures [GNsujet+verbe] afin de décrire et de modéliser l'expansion du mouvement du GN.
2. Ajouter dans les structures [GNsujet+verbe+GNobjet], pour lesquelles la longueur de la paire *verbe+GNobjet* est pour l'instant limitée à 5 syllabes, des paires de 5, 6 et 7 syllabes. Ceci devrait nous permettre de prédire efficacement les contours signalant les frontières prosodiques entre ces deux groupes.

Ce corpus est constitué de trois groupes distincts (cf. annexe A.2) :

- 12 phrases de 7 et 8 syllabes de structure syntaxique [GNsujet+GV]. Le GNsujet, variant entre 1 et 6 syllabes, est constitué d'un article et d'un nom. Le GV, variant entre 1 et 6 syllabes, est principalement réduit au verbe.

– 10 phrases de structure [GNsujet+GV] de 10 syllabes essentiellement. Le GNsujet, contenant 6 syllabes, est constitué de : [article+nom], [article+nom+adjectif] ou [article+adjectif+nom] avec toutes les combinaisons de longueur pour le nom et l'adjectif.
– 51 phrases de structure [GNsujet+verbe+GNobjet] de 6 à 14 syllabes avec des paires *verbe+GNobjet* de 5 à 7 syllabes principalement.

  Le GNsujet est constitué en majorité de [article+nom] de 4 syllabes, le verbe peut varier de 1 à 5 syllabes, et le GNobjet ([article+nom], [article+nom+adjectif] ,[article+adjectif+nom]) de 1 à 6 syllabes.

Les limitations de ce corpus sont une fois de plus liées à la méthodologie de conception fondée sur les oppositions minimales entre chacun des constituants. Nous nous restreignons donc pour les DC à des phrases de 2 à 12 syllabes (et quelques phrases de 14 syllabes) de structure [GNsujet+verbe] et [GNsujet+verbe+GNobjet].

2.3.2 Analyse acoustique

2.3.2.1 Analyse de la durée totale de la phrase

La première analyse que nous avons menée, sur l'ensemble des déclarations des deux corpus, est le calcul de l'évolution de la durée totale des phrases en fonction du nombre de syllabes qu'elles comportent.

La figure 2.11 montre cette évolution sur toutes les phrases de 1 à 12 syllabes des deux corpus (courbe en pointillés). Cette évolution est quasi linéaire : l'excellente corrélation entre la durée totale et le nombre de syllabes des phrases tend à montrer que, dans ces corpus lus, le locuteur n'accélère pas quand il produit des phrases plus longues. Ce résultat est compatible avec l'hypothèse d'horloge interne agissant comme un repère de périodicité lors de l'énonciation et dont la valeur, indirectement mesurée ici, est de 171 ms.

Rappelons néanmoins que, pour les raisons citées au chapitre I (page 39), notre unité macrorythmique n'est pas le rapport entre la durée du GIPC avec l'horloge interne (comme chez Barbosa), mais le rapport de la durée du GIPC avec la somme des durées moyennes des phonèmes qu'il contient. Ceci a été décidé de manière à répondre à la critique faite par van Santen (1997, p.238) sur l'indépendance, présumée dans le modèle initial, des durées syllabiques vis-à-vis des durées intrinsèques des constituants.

Le remplacement de la référence "horloge" par la référence GIPC moyen conserve les tendances référentielles générales. Dans la figure 2.11, nous avons superposé les durées moyennes des phrases calculées en sommant simplement les durées moyennes des phonèmes qui les constituent (trait plein) : comme ci-dessus, on observe une corrélation très élevée entre ces durées prédites et le nombre de syllabes des phrases[3]. La figure 2.12 permet de comparer divers prédicteurs de la durée de chaque phrase :

(a) le nombre de syllabes de la phrase (référence "horloge"),

(b) le nombre de phonèmes de la phrase,

(c) la somme des durées moyennes des phonèmes qui la constituent (notre hypothèse).

Les corrélations obtenues sont élevées (> 0.95). Elles établissent cependant un ordre croissant de qualité de prédiction ($a < b < c$) qui conforte notre modification. Cet ordre est d'ailleurs conservé si l'on applique les mêmes prédicteurs sur le corpus étudié par Aubergé (1991) (voir figure 2.13).

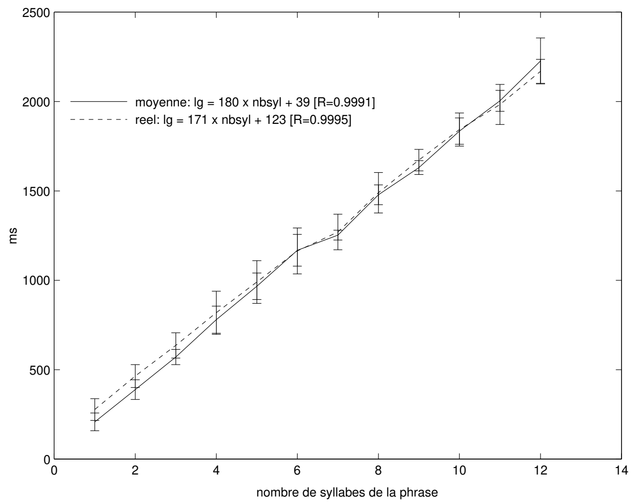


FIG. 2.11: Régression linéaire sur la durée de phrases déclaratives de 1 à 12 syllabes. La courbe en pointillés représente l'évolution de la moyenne des durées réelles des phrases. Pour obtenir la courbe en traits pleins, nous avons remplacé la durée de chaque phonème d'une phrase par sa durée moyenne (cf. tableau 1.1).

2.3.2.2 Mise en évidence de mouvements portés

Pour caractériser les mouvements portés de groupe, nous avons soustrait la contribution des contours prosodiques du niveau de la phrase pour les déclarations : nous avons simplement retranché à chaque réalisation prosodique la moyenne de toutes les phrases contenant le même nombre de syllabes que cette réalisation. Cette méthode n'est évidemment pas d'une grande rigueur scientifique comme l'a déjà souligné Aubergé (1991), mais elle permet néanmoins de dégager les grandes tendances de l'évolution des mouvements de groupe.

[3]Pour les faibles nombres de syllabes ($<$5), la différence entre les deux courbes est due principalement à l'allongement final. Cette différence est minime au-delà : on dispose alors d'une durée suffisante pour moduler le macrorythme.

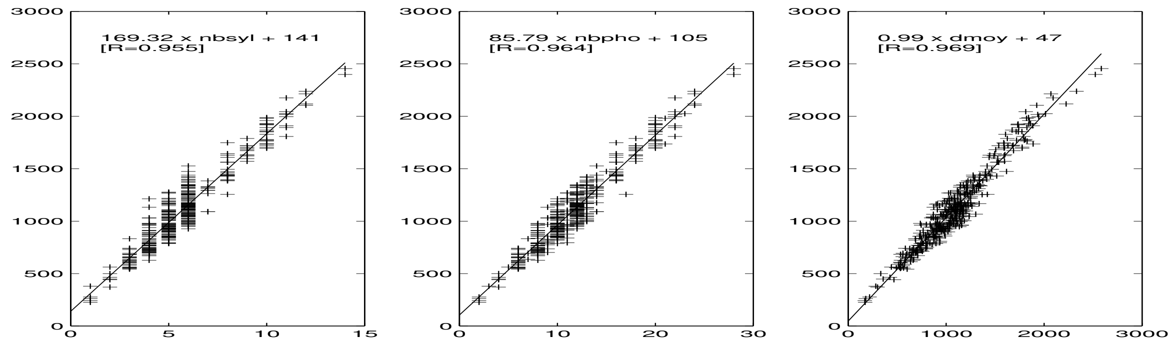


FIG. 2.12: Pour les phrases déclaratives de notre corpus, régression linéaire sur un ensemble de points dont l'ordonnée correspond à la durée réelle de chaque phrase DC, et l'abscisse représente un des trois prédicteurs suivants. De gauche à droite : (a) le nombre de syllabes de la phrase, (b) le nombre de phonèmes de la phrase, (c) la somme des durées moyennes des phonèmes qui la constituent.

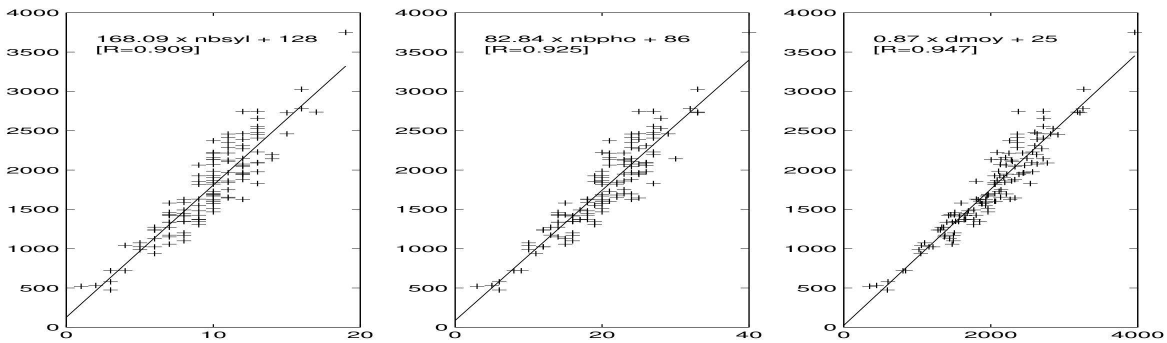


FIG. 2.13: Même représentation que pour la figure 2.12, mais sur les phrases déclaratives du corpus étudié par Aubergé (1991).

2.3.2.3 Caractérisation des mouvements

Les figures 2.14 représentent la superposition des contours prosodiques "résiduels" de groupe pour des phrases [GNsujet (3syllabes) +verbe (3syllabes)] de 6 syllabes (à gauche) et des phrases [GNsujet (3syllabes)+verbe (3syllabes)+GNobjet (4 syllabes)] de 10 syllabes (à droite).

La première figure met en évidence que la jonction *sujet/verbe* est marquée par un mouvement final montant aussi bien par la mélodie que par le rythme. Nous remarquons également que les contours mélodiques du GNsujet, sur le premier GIPC, se divisent en deux groupes qui se rejoignent ensuite. Le premier débute vers -5 1/4*tons* et correspond aux GN comportant un clitique en initial. Le second, centré autour de 5 1/4*tons* est associé

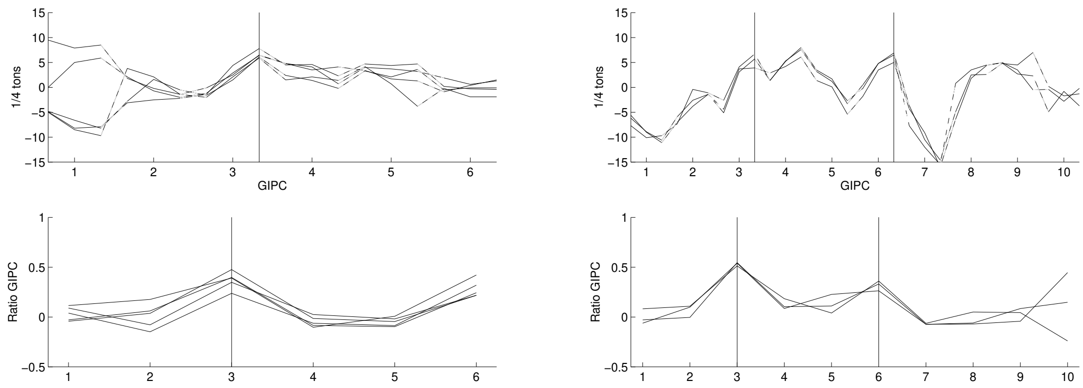


FIG. 2.14: Superposition des contours de groupe pour des phrase [GNsujet+verbe] (à gauche) et [GNsujet+verbe+GNobjet] (à droite) de longueurs de constituants identiques.

aux GN débutant par un non clitique.

La seconde figure illustre certaines caractéristiques de la structuration rythmique que nous avons déjà mentionnées mais qu'il est bon de rappeler ici : la hiérarchie des allongements des Ratios GIPC avec un allongement plus prononcé avant la frontière *sujet/verbe* qu'avant la frontière *verbe/objet* ; la variabilité de la durée de la dernière voyelle de la phrase.

2.3.2.4 Expansion du mouvement

Les figures 2.15 sont destinées à décrire l'expansion du mouvement prosodique liée à l'allongement du GN dans des phrases [sujet+verbe] de 3 à 6 syllabes. Chaque figure représente, pour une longueur de GN donnée, la superposition des moyennes des contours prosodiques de groupe. Nous pouvons remarquer que :

– Les frontières sont toujours marquées mélodiquement et rythmiquement quelle que soit la taille du GN et la longueur de la phrase dans laquelle il se trouve.

– Les variations du rythme sur des GN de 3 à 6 syllabes sont régulières et terminées par un allongement plus important sur les deux derniers GIPC de l'unité (voir les GN de 4 et 5 syllabes).

– L'évolution de la F0 est beaucoup moins systématique mais on peut néanmoins déceler qu'à partir d'une certaine longueur du GN (4 et 5 syllabes), le mouvement mélodique montant sur le dernier GIPC est précédé par un mouvement descendant sur le pénultième.

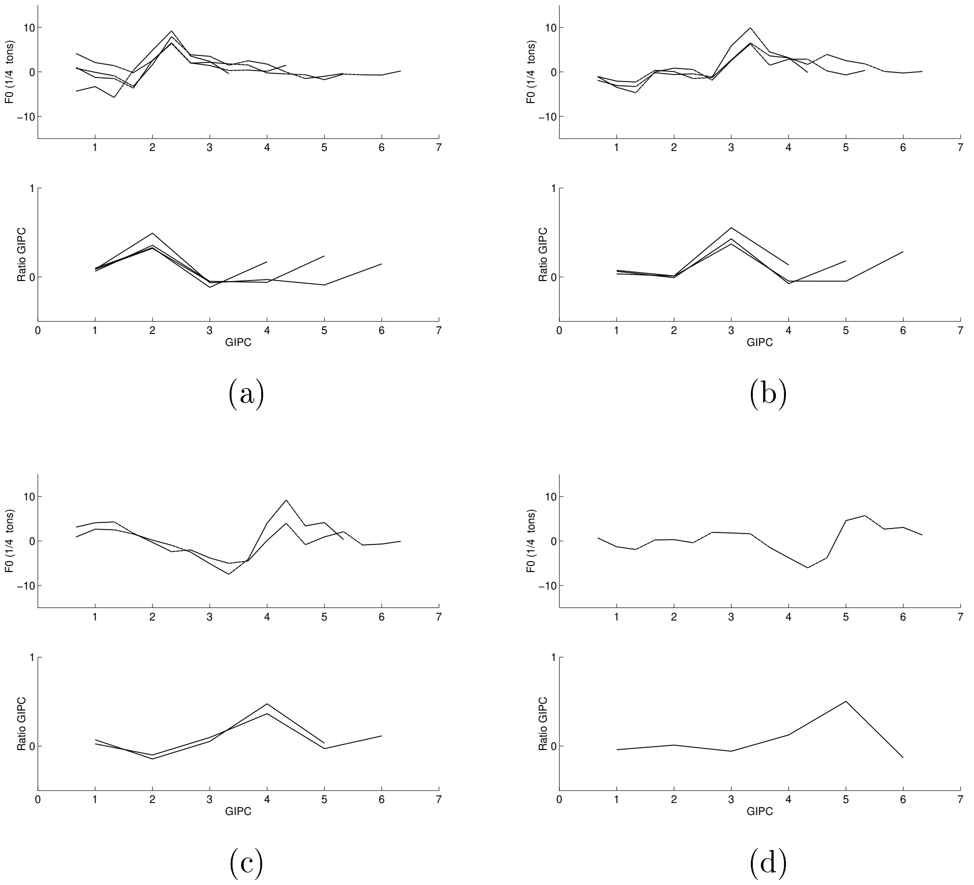


FIG. 2.15: Moyennes de GNsujets de 2 (a), 3 (b), 4 (c) et 5 (d) syllabes pour des phrases [GNsujet + verbe] de 3 à 6 syllabes.

2.4 A retenir

Dans ce chapitre, nous avons détaillé la conception, l'enregistrement, l'évaluation perceptive et l'analyse acoustique de deux corpus de phrases isolées. Cette analyse ascendante a révélé l'existence de prototypes prosodiques véhiculant les attributs des différents niveaux linguistiques considérés dans notre étude (la *phrase* et le *groupe*). L'observation de l'évolution de ces prototypes, en fonction du nombre de syllabes qui les portent, nous a permis d'émettre l'hypothèse que ces mouvements se décomposent en deux phases : *préparation* et aboutissement du mouvement ou *capture*, coïncidant avec des *rendez-vous* : la fin de la phrase pour le niveau *phrase* et les frontières syntaxiques pour le niveau *groupe*.

Le **Modèle d'Expansion du Mouvement** prosodique (MEM) que nous proposons au chapitre suivant doit être capable de modéliser les différentes phases du mouvement. Il faudra donc lui donner des indications sur la position de chaque GIPC de la phrase par rapport au point de rendez-vous qui le suit. Il devra également prédire les rendez-vous manqués entre prosodie et syntaxe qui dépendent, d'une part, de la nature du contour porteur véhiculant l'attitude du locuteur et, d'autre part, de contraintes phonotactiques (nous pouvons également ajouter le débit). La première cause d'effacement de marques pourra être modélisée par une limitation de la modulation en fonction de l'attitude générée ; la seconde, par une information supplémentaire permettant de gérer les effets de coarticulation entre les contours d'un même niveau (information chevauchante). Le chapitre III détaille l'implémentation, l'apprentissage automatique et les performances de ce modèle.

Chapitre 3

Modèle de génération et apprentissage hiérarchique

3.1 Architecture générale du modèle

Le modèle dynamique de génération de la prosodie que nous proposons découle directement de notre approche morphologique et superpositionnelle. Il consiste en une association modulaire de réseaux de neurones récurrents (RNR) chargés de prédire les mouvements prosodiques des niveaux linguistiques auxquels ils sont associés. Le contour prosodique final sera obtenu par la somme pondérée de ces prédictions (figure 3.1) ajoutée aux valeurs prosodiques de référence du locuteur (registre, débit moyen...). Ce modèle unique permet de générer simultanément, et pour chaque GIPC d'une phrase, la fréquence fondamentale et le facteur d'allongement k_{GIPC}. La durée de chaque phonème est ensuite obtenue par le modèle de répartition détaillé en section 1.3.2.2.

3.1.1 Architecture modulaire et apprentissage hiérarchique

L'apprentissage de ce modèle de génération se fait de manière itérative - du niveau porteur le plus large vers le niveau porté le plus petit - par la gestion dynamique des pondérations w. A la manière des stratégies observées d'apprentissage supervisé de l'intonation (Strömqvist, Peters & Ragnarsdóttir 1995), le modèle va d'abord apprendre à générer un contour prosodique global de phrase. Pour ceci, toutes les pondérations w sont initialisées à zéro sauf celle du module de phrase. Lorsqu'il a acquis une performance suffisante[1], il va

[1] La compétence est isomorphe à la performance sous notre hypothèse initiale d'indépendance fonctionnelle entre les niveaux linguistiques.

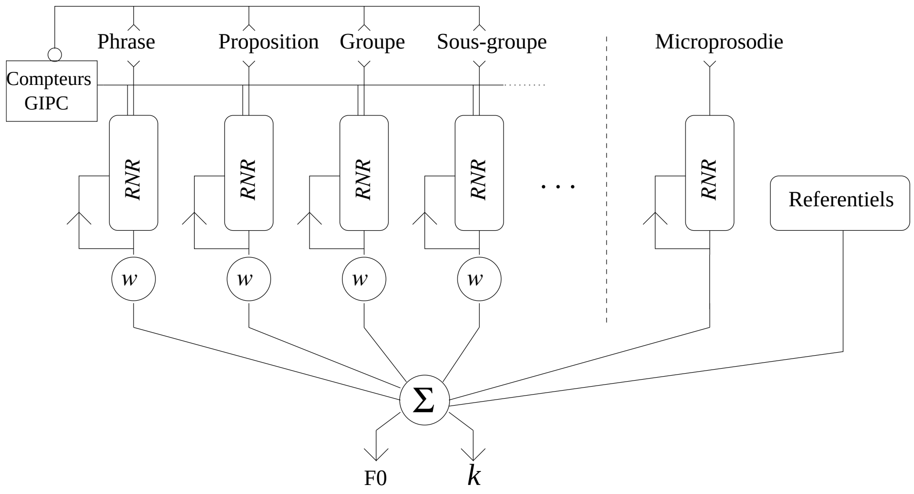


FIG. 3.1: Architecture générale du modèle.

alors moduler ce mouvement prosodique de phrase par l'ajout de la contribution du niveau inférieur le plus proche. Ceci est réalisé concrètement en "gelant" les poids du module de phrase et en positionnant la pondération w correspondant au module de groupe à 1[2]. Cet apprentissage hiérarchique se poursuit jusqu'au dernier module qui ajoutera aux contours prédits les phénomènes microprosodiques spécifiques à chaque phonème.

La modularité de cette architecture permet donc, comme nous l'avons déjà évoqué, d'utiliser un corpus approprié à chaque niveau linguistique étudié. Le modèle d'expansion du mouvement de phrase, générant diverses attitudes, sera entraîné sur un corpus de phrases courtes extraites du corpus "attitudes" et constitué principalement de monomots alors que, dans une seconde phase, l'apprentissage de la typologie des mouvements de groupe pourra être effectué sur les phrases assertives plus longues du corpus "modulation".

L'architecture modulaire du modèle permet en outre d'associer à chaque bloc des fonctions linguistiques distinctes :

- Les contours mélodiques et rythmiques prédits par le module de phrase rempliront les fonctions modale et attitudinale.

- Les modules de groupe et de sous-groupe assureront les fonctions syntaxiques de démarcation et de hiérarchisation.

La génération biparamétrique permettra de révéler les rôles respectifs joués par la mélodie et le rythme dans ces différentes fonctions.

[2]Nous rappelons que le niveau propositionnel n'est pas représenté dans nos corpus.

3.1.2 Somme pondérée de prédictions

Comme nous venons de le voir, les pondérations interviennent déjà au niveau de l'apprentissage supervisé pour mettre progressivement en oeuvre les différents modules du modèle.

Les pondérations sont également primordiales pour la génération :

- La gestion des poids permet de contrôler dynamiquement l'importance relative des niveaux. Nous avons remarqué, au chapitre précédent, que certains contours mélodiques de phrase ne sont pas modulés par le découpage en groupe : par exemple, pour les attitudes correspondant à une reprise mot pour mot de l'assertion de l'interlocuteur (*doute-incrédulité* et *ironie de soupçon*), le locuteur ne semble pas ressentir le besoin d'indiquer par la mélodie la structure de l'énoncé qu'il répète.
- Par contre, la structure syntaxique se projette dans tous les cas sur la composante rythmique. Des pondérations différentes doivent donc être affectées à la fréquence fondamentale et au coefficient d'élasticité pour révéler les stratégies de coopération entre ces deux paramètres.
- Enfin, notre modèle pourra gérer, grâce aux pondérations, les variations de pente de la ligne de déclinaison observées dans des tâches de lecture de textes (Silverman 1987), en modélisant la dynamique des mouvements de niveaux supérieurs, comme par exemple, la pente maximale en début de paragraphe dans le cas d'un changement de thème.

3.2 Le module de phrase : apprentissage des contours porteurs

3.2.1 Architecture et apprentissage du module de phrase

3.2.1.1 Le réseau connexionniste

L'approche de la prosodie que nous proposons est dynamique : nous cherchons à modéliser l'évolution de la fréquence fondamentale et du rythme au cours du temps. Nous allons donc appliquer des méthodes de traitement de formes dynamiques pour la génération automatique de ces mouvements prosodiques. Deux solutions s'offrent à nous pour intégrer la composante temporelle dans notre modèle à base de réseaux connexionnistes (cf. section 1.3.1.4) :

- La métaphore spatiale qui transforme une séquence d'événements temporels en une séquence spatiale. Cela se traduit concrètement, en entrée du réseau, par une fenêtre

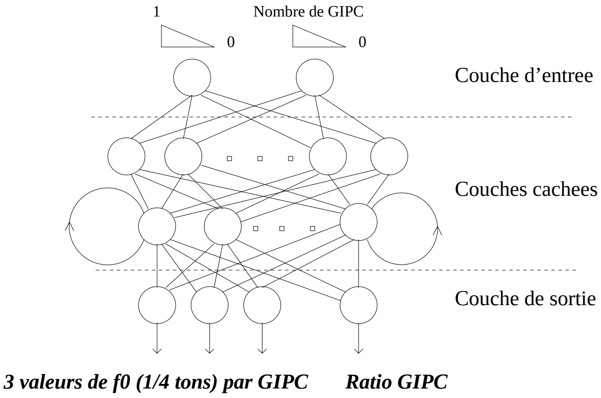


FIG. 3.2: Architecture du module de phrase.

glissante contenant l'échantillon en cours de traitement, et donnant des informations sur le passé et sur l'avenir.

– L'utilisation de réseaux récurrents. Il s'agit de donner au réseau une mémoire sur les informations passées en créant des niveaux de traitement appelés unités de contexte (ou encore unités d'état) qui reçoivent l'état d'un certain nombre de neurones au pas d'apprentissage *(t)*, effectuent un certain traitement, puis délivrent leur activation au pas *(t+1)*. Ces unités de contexte peuvent recevoir l'activation des unité de sortie (modèle de Jordan 1988) ou celle des unités cachées (modèle de Elman 1988).

Ces deux types de réseau ont été employés durant notre recherche. Nous avons d'abord utilisé pour notre premier modèle de génération un réseau de type Jordan (Morlec, Aubergé & Bailly 1995). Mais, pour la conception des modules de ce nouveau modèle, nous nous sommes orientés vers des réseaux de type Elman. Le choix de ces réseaux et leur architecture interne a été uniquement guidé par une comparaison systématique des moyennes des erreurs quadratiques que leurs prédictions engendraient sur un même corpus d'apprentissage. Pour chaque architecture, nous avons également testé la capacité de généralisation du modèle. Cette démarche par *essai-erreur* est la seule envisageable étant donné qu'il n'existe pas de règle, dans un réseau récurrent, pour déterminer précisément le nombre de couches cachées et le nombre de neurones dans chacune d'elles en fonction de la taille de l'échantillon d'entrée.

Le module de phrase (Morlec, Bailly & Aubergé 1997a) est donc un réseau de neurones récurrent unique de type Elman destiné à prédire les contours prosodiques véhiculant les six attitudes enregistrées dans le corpus "attitudes". De l'attitude à réaliser dépend la paramétrisation du réseau : activités initiales et poids des neurones. Ces différents poids sont calculés à partir de l'algorithme de rétropropagation du gradient de l'erreur. La fonction neurone utilisée correspond à la fonction arctangente (*atan*) qui varie entre 0 et 1[3].

[3]Nous n'avons pas utilisé, comme on le fait classiquement, la sigmoïde car elle exhibe une pente beaucoup plus raide et semble donc plus appropriée à des tâches de classification qu'à celles, comme ici, d'interpolation de fonctions non linéraires.

Les entrées du réseau donnent une description de la phrase sous forme séquentielle : une description est présentée au réseau à chaque pas de la séquence. Le temps est donc représenté par la séquence de GIPC de la phrase.

L'architecture du réseau (figure 3.2) comporte :

Deux rampes en entrée L'analyse acoustique de la section 2.2.3.4 nous a permis de faire l'hypothèse que les mouvements prosodiques se décomposaient en deux phases : *préparation* et *capture*. L'entrée du module de phrase est donc constituée au minimum de deux paramètres : une rampe linéaire décroissante relative (composée d'intervalles réguliers entre 10 et 1) donne la position relative du GIPC en court de traitement par rapport à la fin de la phrase (le pourcentage de GIPC restant à traiter). Une rampe linéaire décroissante absolue (du nombre de GIPC à 1) indique pour le GIPC courant le nombre de GIPC manquant pour atteindre la fin de la phrase.

Ces deux entrées permettent de gérer la séquentialité du mouvement et de déclencher la phase de *capture* au bon moment.

Deux couches cachées La première contient 15 neurones, la seconde 11 neurones qui reçoivent au temps *(t)* les activations de la première couche cachée et leurs propres activations calculées au cycle *(t-1)* : c'est la récurrence.

Quatre sorties linéaires Chaque sortie correspond à la somme pondérée des activations de la deuxième couche cachée. Trois d'entre elles fournissent la fréquence fondamentale stylisée du GIPC (exprimée en $1/4tons$), la quatrième délivre le coefficient k_{GIPC} multiplié par 10. Ce facteur multiplicatif permet d'homogénéiser l'excursion des différentes valeurs d'entrée et de sortie entre 0 et 30.

3.2.1.2 Apprentissage supervisé

Le corpus d'apprentissage de ce réseau est constitué, dans une première étape, de l'ensemble des phrases monomots de 1 à 6 syllabes, soit 38 phrases par attitude réparties comme l'indique le tableau 3.1 (rappelons que les monomots permettent de limiter l'incidence des contours portés).

Nombre de syllabes	1	2	3	4	5	6
Nombre de phrases	5	3	6	9	7	8

TAB. 3.1: Nombre de phrases du corpus d'apprentissage extraites du corpus "attitudes" en fonction du nombre de syllabes.

Nous vérifierons ensuite, sur des phrases plus longues, que notre paramétrisation d'entrée permet au réseau de réaliser l'expansion du mouvement prosodique en conservant les caractéristiques principales de chaque attitude.

Pour chaque attitude, le corpus associé est présenté 1000 fois durant l'apprentissage. Cette valeur s'est révélée assez grande pour réaliser des prédictions correctes et suffisamment faible pour éviter un sur-apprentissage qui conduirait à de très mauvaises capacités de généralisation.

3.2.2 Validation objective

3.2.2.1 Prédictions sur le corpus d'apprentissage

Deux types d'erreurs de prédiction ont été calculés :

- Le premier concerne l'erreur de prédiction de F0. Le réseau prédit des valeurs en quart de tons. Ces valeurs sont d'abord exprimées en Hertz d'après :

$$F0_{Hz} = F0_{moyenHz} 2^{\frac{F0_{1/4tons}}{24}} \tag{3.1}$$

  L'erreur relative en Hertz, entre la chaque valeur cible et la valeur prédite correspondante, vaut :

$$\Delta F0 = \frac{|F0_{pred} - F0_{cible}|}{F0_{cible}} \tag{3.2}$$

- L'erreur relative de durée GIPC est également calculée. Rappelons simplement que le réseau prédit le Ratio GIPC, k_{GIPC}, obéissant à la formule 1.5 de la section 1.3.2.2. L'erreur relative de prédiction du réseau sur la durée du GIPC est obtenue d'après l'équation :

$$\Delta d_{GIPC} = \frac{|d_{GIPCpred} - d_{GIPCcible}|}{d_{GIPCcible}} = \frac{|k_{GIPCpred} - k_{GIPCcible}|}{1 + k_{GIPCcible}} \tag{3.3}$$

Longueur	DC	QS	EX	DI	SC	EV
1	5.43	8.65	21.60	10.35	8.57	15.67
2	6.61	9.04	16.73	6.56	6.93	6.26
3	5.97	6.12	10.65	6.01	3.86	6.46
4	5.73	6.03	18.43	4.77	5.21	8.89
5	6.26	5.47	19.27	4.67	4.90	8.01
6	6.78	7.75	17.76	6.22	5.59	7.96
% total	6.25	6.68	17.50	5.63	5.28	8.21

TAB. 3.2: Erreurs relatives de prédiction de la F0 (en Hertz) pour les phrases monomots du corpus d'apprentissage.

Sur le corpus d'apprentissage, les prédictions globales mélodiques et rythmiques générées par le module phrase sont proches des moyennes des échantillons d'apprentissage. Elles

Longueur	DC	QS	EX	DI	SC	EV
1	–	–	–	–	–	–
2	8.85	6.67	4.72	7.93	15.38	10.35
3	6.77	7.04	6.26	8.70	8.00	7.12
4	10.05	10.62	9.39	9.09	15.40	11.18
5	10.41	11.69	9.18	8.39	9.97	10.18
6	12.70	11.18	9.77	8.42	12.66	10.18
% total	10.72	10.60	9.01	8.59	12.21	10.10

TAB. 3.3: Erreurs relatives de prédiction de la durée GIPC (dernier GIPC exclu) pour les phrases monomots du corpus d'apprentissage.

Attitude	DC	QS	EX	DI	SC	EV
% total	17.01	11.95	14.70	15.09	14.00	14.36

TAB. 3.4: Erreurs relatives de prédiction de la durée du dernier GIPC pour les phrases monomots du corpus d'apprentissage.

restent toujours comprises entre $(\mu - \sigma)$ et $(\mu + \sigma)$ quelle que soit l'attitude. Les figures 3.3 et 3.4 illustrent ce résultat.

L'erreur relative globale par valeur prédite de F0 (en Hertz) est d'environ 6 % (voir tableau 3.2) excepté pour les exclamatives 17.5 % pour lesquelles des modulations mélodiques importantes apparaissent. Trois hypothèses peuvent expliquer cette variance particulière à cette attitude :

Variantes : le contour prédit n'est qu'un représentant prototypique de la classe des contours associée à un attribut linguistique, ici une attitude. Autour de ce prototype s'organisent donc des variantes tout aussi acceptables. La variance peut donc être interprétée comme une distance intra-classe importante.

Point d'ancrage : le mouvement de phrase exclamative présente des points d'ancrage dans l'énoncé autres que le début et la fin comme, par exemple, une syllabe particulière d'un item lexical voire un groupe de mots cibles.

Niveau porté : la variance peut plus probablement être expliquée par un niveau porté. En effet, en français, on peut difficilement construire un long monomot sans avoir recours à une construction polymorphémique. Dans le cas de monomots portés par une exclamative, la modulation morphologique serait amplifiée. Cette hypothèse de contrôle d'amplitude de modulation des portés par les porteurs sera examinée en section 3.4.

Les erreurs relatives pour les durées GIPC ont été décomposées en deux tableaux séparés :

– Le premier (tableau 3.3) représente les moyennes des erreurs relatives de tous les GIPC excepté le dernier et classées par attitude et par le nombre de syllabes que

comporte la phrase.
- Le second (tableau 3.4) contient les moyennes des erreurs relatives du dernier GIPC de chaque attitude.

Cette différenciation est nécessaire en raison de la variabilité de l'allongement du dernier GIPC de la phrase qui accroît l'erreur de prédiction à des valeurs qui ne sont pas significatives.

Le premier tableau indique que l'erreur moyenne des durées GIPC prédites pour les six attitudes vaut 10.2 %. Le second, sur la durée de la dernière voyelle de la phrase, donne 14.5 % d'erreur bien que le mouvement rythmique final reste sur une trajectoire médiane - à l'intérieur des données.

3.2.2.2 Expansion du mouvement et extrapolation

La généralisation des réseaux a été étudiée sur les phrases du corpus de test de 7 et 8 syllabes et également sur des phrases plus longues. Les résultats montrent que le réseau réalise effectivement l'expansion du mouvement prosodique. Dans ce mémoire, nous ne présentons que deux cas typiques d'expansion sur des phrases comportant jusqu'à 10 syllabes.

Pour l'attitude de doute-incrédulité (DI à gauche) et la modalité déclarative (DC à droite), la figure 3.5 représente la généralisation des mouvements prosodiques pour des phrases de 7 à 10 syllabes.

Notez, pour le DI, que notre modèle a su reproduire fidèlement les caractéristiques mélodiques essentielles de ce type d'attitude : on retrouve la mise en relief de l'avant-dernière syllabe et le mouvement montant sur la dernière (*capture*). Le réseau a été également capable de modéliser des événements fins comme l'évolution du mouvement de *préparation* qui démarre d'abord à -10 $1/4tons$ sur des phrases courtes pour progressivement évoluer vers 0. Les patrons rythmiques sont fidèlement prédits avec un allongement progressif aboutissant à un GIPC de fin de phrase très long permettant la réalisation d'un mouvement mélodique complexe.

Pour DC, on retrouve une évolution continue avec un allongement rythmique beaucoup plus faible que pour DI. On note également que le mouvement mélodique de *préparation*, entre 1 et 6 syllabes, se place sur un registre de plus en plus élevé. En extrapolation, cette caractéristique est conservée avec une première partie de contour mélodique située vers 10 $1/4tons$. Ce phénomène spécifiquement lié aux monomots DC, pour lesquels le locuteur a systématiquement réalisé un accent en initial de mot, risque d'introduire des divergences importantes lorsque l'on utilisera ces prédictions comme contours porteurs de phrases syntaxiquement plus complexes. Nous avons donc entrepris un nouvel apprentissage des DC en ajoutant au corpus de monomots des phrases de 3 à 6 syllabes [GNsujet+verbe]

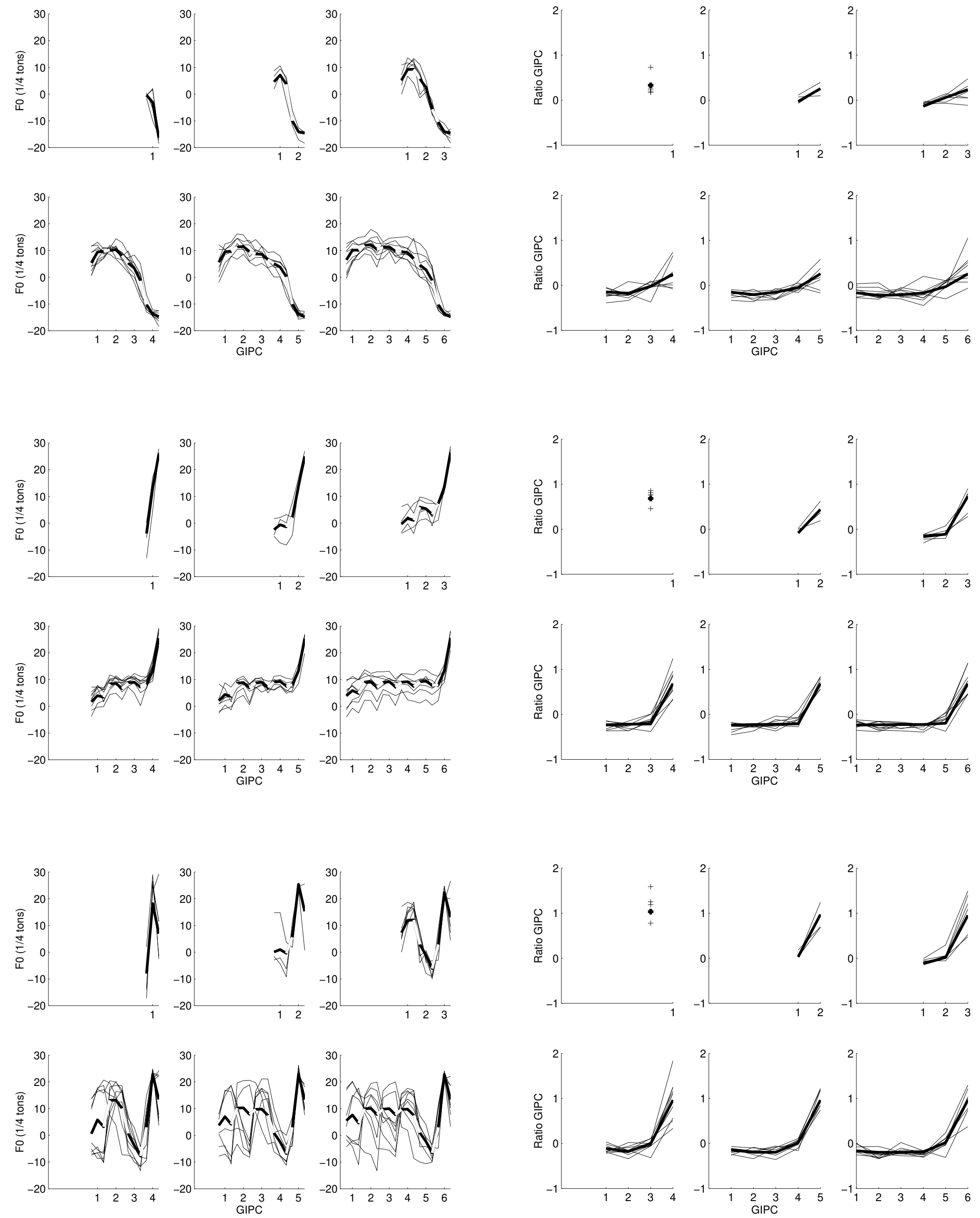


FIG. 3.3: Prédictions (trait épais) sur le corpus d'apprentissage (traits fins) de monomots. F0 (à gauche) et Ratio GIPC (à droite) : *déclaration* (DC), *question* (QS) et *exclamation* (EX) de 1 à 6 syllabes.

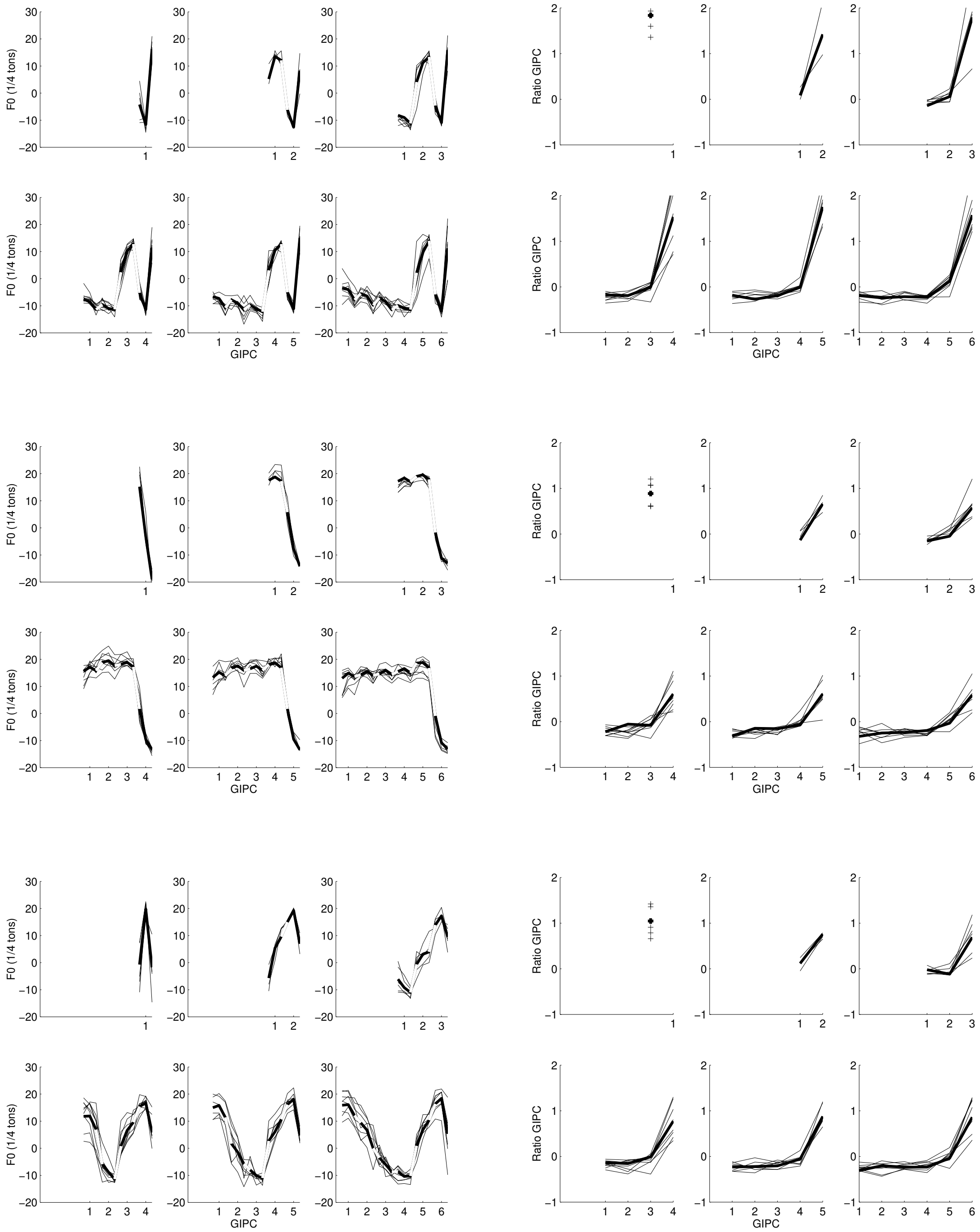


FIG. 3.4: Prédictions (trait épais) sur le corpus d'apprentissage (traits fins) de monomots. F0 (à gauche) et Ratio GIPC (à droite) : *doute-incrédulité* (DI), *ironie de soupçon* (SC) et *évidence* (EV) de 1 à 6 syllabes.

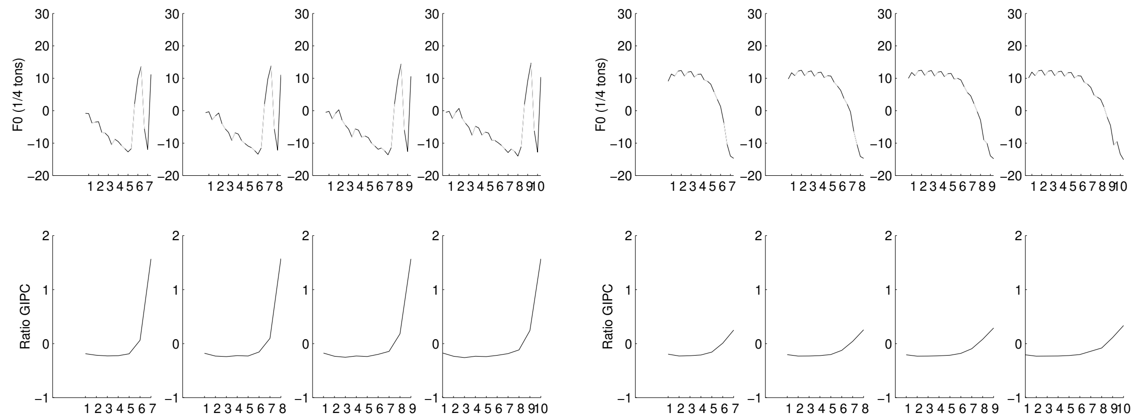


FIG. 3.5: Prédictions de l'expansion du mouvement prosodique pour des *doute-incrédulités* et des *déclarations* de 7 à 10 syllabes.

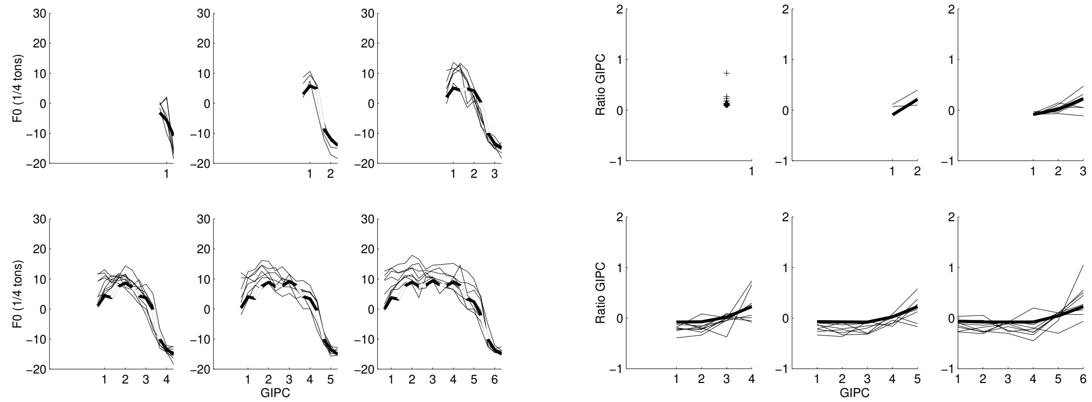


FIG. 3.6: Prédictions de l'expansion du mouvement prosodique pour les *déclarations* de 1 à 6 syllabes : monomots et [GNsujet+verbe]

peu modulées dont le sujet est réduit à [article+nom]. Cette méthode nous permet d'obtenir des contours mélodiques dont la valeur d'attaque est égale au registre du locuteur mais elle engendre également une légère diminution de débit (cf. figure 3.6).

On pourrait croire que cette démarche contredit l'hypothèse selon laquelle les monomots permettent, à eux seuls, d'extraire un prototype de phrase. Mais nous pouvons argumenter qu'il est très rare de rencontrer des monomots de plus de 8 syllabes, et donc qu'une généralisation des contours de monomots au delà de cette longueur a finalement peu de sens.

3.2.3 Validation perceptive : test d'identification

Les résultats chiffrés, présentés ci-dessus sous le titre "validation objective", attestent de la capacité du module de phrase à prédire des valeurs de F0 et de durée GIPC proches des valeurs originales.

Cependant, ils ne sont pas le reflet de la réalité perçue étant donné que certaines déviations importantes peuvent avoir peu d'influence sur l'auditeur tandis que d'autres, plus minimes, ne permettront pas à celui-ci d'identifier l'attitude. C'est pourquoi, les prédictions générées par notre modèle doivent être systématiquement sanctionnées par une évaluation perceptive.

Nous avons donc réalisé un test d'identification sur les différents contours prédits par le module de phrase, afin de montrer, par comparaison avec des stimuli naturels, que ces prédictions ont capté suffisamment d'information prosodique pour permettre aux auditeurs de retrouver l'attitude du locuteur.

3.2.3.1 Matériel

- Le premier test, correspondant à une phase d'entraînement, comporte 24 phrases monomots naturelles de longueurs comprises entre 3 et 6 syllabes.

- Le second est composé de 48 phrases monomots "naturelles"[4] de 1 à 8 syllabes, et des 48 versions synthétiques correspondantes (la prosodie prédite d'une attitude donnée est générée par analyse-synthèse TDPSOLA à partir de la phrase naturelle énoncée selon cette attitude), soit une phrase par attitude et par longueur syllabique. Les phrases employées durant ce test correspondent au premier monomot de chaque longueur du corpus "attitudes".

3.2.3.2 Protocole

Vingt sujets ont participé à cette expérience. Les sujets ont, durant les deux tests, une feuille de définitions à leur disposition. Son contenu est décrit ci-après :

[4]Les contours mélodiques de ces phrases correspondent à la stylisation des contours originaux.

Déclaration « `Phrase déclarative neutre.` »

Mots-clefs :

Question « `Question neutre.` »

Mots-clefs :

Exclamation de surprise « `Etonnement, stupéfaction du locuteur.` »

Mots-clefs :

Doute-Incrédulité « `Il exprime une mise en doute de ce qui a été préalablement affirmé. Le locuteur est sceptique et perplexe.` »

Mots-clefs :

Ironie de soupçon « `On reprend une affirmation en faisant comprendre qu'on la contredit, qu'on se moque de l'interlocuteur, ceci par le biais d'une intonation hautaine et narquoise.` »

Mots-clefs :

Evidence « `Le locuteur insiste sur la véracité de son affirmation. Il rappelle à son interlocuteur l'évidence de son propos.` »

Mots-clefs :

Une première phase d'entraînement a été mise en place afin d'habituer les sujets aux définitions des 6 attitudes. Dans cette tâche à choix forcé, les sujets doivent associer à chaque phrase (présentée une seule fois) une des 6 définitions. Une fois qu'ils ont donné leur réponse, la réponse correcte est affichée à l'écran. Les sujets peuvent à tout moment enrichir les définitions proposées avec leurs propres mots-clefs.

Durant la seconde phase de cette expérience, les sujets écoutent les phrases naturelles et synthétiques dans un ordre aléatoire. Ce tirage est soumis à deux contraintes : la même attitude n'est jamais présentée deux fois de suite ; on ne présente jamais plus de 3 stimuli naturels ou synthétiques consécutivement. Comme pour le premier test, les sujets choisissent une des 6 réponses proposées. Ce test a été réalisé deux fois par chaque auditeur afin de révéler d'éventuels effets d'apprentissage. Ces deux répétitions seront notées *t1* et *t2*.

3.2.3.3 Résultats

Les principaux résultats de cette tâche d'identification sont les suivants :

- Le pourcentage global d'identification durant la phase d'entraînement est de 72.9 % avec de grosses fluctuations selon l'attitude (93 % pour DC, 36 % pour SC)
- Le pourcentage d'identification moyen des stimuli naturels pour *t1* et *t2* est de 72.8 %. La matrice de confusion sur ces stimuli est donnée dans le tableau 3.5.a.

- Le pourcentage d'identification moyen des stimuli synthétiques pour *t1* et *t2* est de 68.6 %. La matrice de confusion sur ces stimuli est proposée dans le tableau 3.5.b.
- Le pourcentage d'identification de *t2* est plus élevé que celui de *t1* : 3,4 % de plus pour les stimuli naturels et 5.6 % de plus pour les synthétiques.

Avant d'entrer dans le détail, il est intéressant de comparer la matrice de confusion 3.5.a. sur les échantillons naturels et celle obtenue après le test préliminaire d'identification pour les sujets entraînés (tableau 2.2 de la section 2.2.2.3). Nous remarquons que les auditeurs du premier test d'identification ont été nettement plus performants que pour celui-ci et surtout sur l'identification des attitudes EX, DI et SC. Une comparaison directe est toutefois risquée :

- Le test décrit au chapitre II a été réalisé par 12 auditeurs dont seulement 6 étaient entraînés (mêmes conditions que ce test). 20 auditeurs ont participé à cette nouvelle expérience. Les résultats observés sont donc statistiquement beaucoup plus fiables.
- Les stimuli sont différents entre les deux expériences. Dans ce test nous mélangeons des stimuli - naturels-stylisés et prédictions - qui sont tous générés par analyse-synthèse.
- Les définitions données aux auditeurs ont évolué. En effet, nous nous sommes interdits cette fois de donner certaines phrases typiques de la réalisation d'une attitude comme « Ah bon » pour le doute-incrédulité ou « Tiens, tiens » pour l'ironie de soupçon. De plus la présence de marques de ponctuation, sur la feuille de définition, a certainement aidé les sujets dans leur tâche d'identification.

(a)

	DC	QS	EX	DI	SC	EV
DC	88.7	0.3	0.0	0.6	2.5	7.8
QS	2.2	81.4	4.8	7.1	3.2	1.3
EX	1.6	1.6	72.9	15.0	4.4	4.5
DI	5.7	1.9	8.0	58.7	17.0	8.7
SC	9.6	3.9	3.5	25.0	48.5	9.7
EV	5.8	0.3	2.9	2.9	1.9	86.3

(b)

	DC	QS	EX	DI	SC	EV
DC	90.2	0.0	0.0	1.9	5.0	2.9
QS	2.9	83.5	3.8	5.7	3.2	0.9
EX	2.2	9.0	54.6	19.5	5.1	9.6
DI	8.0	5.2	8.0	57.0	15.7	6.1
SC	15.3	4.8	3.5	22.8	45.6	8.1
EV	10.2	1.3	1.6	2.9	3.2	80.8

TAB. 3.5: Matrices de confusion des phrases naturelles (a) et synthétiques (b) pour *t1* et *t2*.

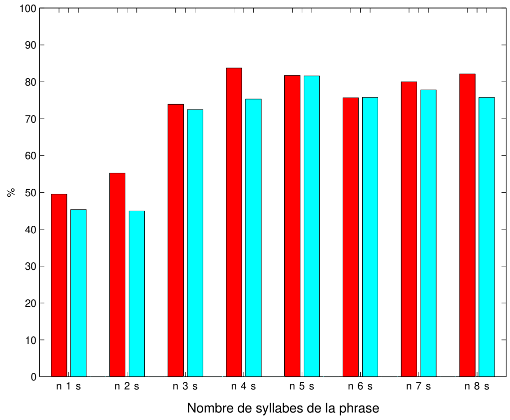


FIG. 3.7: Pourcentage d'identification global de *t1* et *t2* pour les versions naturelles (n) en noir et synthétiques (s) en gris de phrases comportant 1 à 8 syllabes. Les phrases synthétiques de 7 et 8 syllabes correspondent à la généralisation du modèle.

En allant plus loin dans notre analyse des résultats, nous pouvons remarquer quelques phénomènes intéressants :

- Durant la première expérience, l'association d'une attitude intonative donnée avec le code linguistique qui lui correspond est réalisée très rapidement comme nous avons déjà pu le constater pendant le test d'identification décrit en section 2.2.2.

- Le *doute-incrédulité* et *l'ironie de soupçon* sont souvent confondus malgré une différence très claire de leurs caractéristiques prosodiques. Ce résultat est probablement du à la définition de ces attitudes étant donné que les contextes situationnels dans lesquels elles apparaissent sont très similaires.

- Le pourcentage d'identification augmente sur des phrases de 1, 2 et 3 syllabes et reste stable au delà de 3 syllabes (voir figure 3.7).

- Le pourcentage d'identification global montre que les contours naturels sont mieux identifiés que les contours synthétiques mais la différence entre les deux taux reste minime quelle que soit la longueur de la phrase proposée. De plus, si l'on écarte les résultats associés aux *exclamatives* pour lesquelles l'absence de modulation intra-mot est clairement perceptible dans les versions synthétiques, les 5 attitudes restantes sont aussi bien reconnues que les originales. On peut donc se demander si la modulation intra-mot est nécessaire à la perception des exclamatives ou s'il s'agit simplement d'un distracteur sans lequel l'auditeur trouve le stimulus très peu naturel.

3.2.4 Discussion

Nous avons montré dans la partie précédente que notre module de phrase avec sa paramétrisation d'entrée dépouillée était capable de générer les prototypes prosodiques associés aux six attitudes du corpus "attitudes" :

- Par le biais d'un apprentissage supervisé sur des phrases monomots de 1 à 6 syllabes, le réseau récurrent a prédit des contours mélodiques et rythmiques très voisins des échantillons d'apprentissage, a capturé les caractéristiques essentielles de chaque attitude, et a réalisé l'expansion du mouvement prosodique tout en conservant celles-ci.
- Un test d'identification a montré que ces contours synthétiques de phrase permettaient aux auditeurs d'accéder au code linguistique presque aussi efficacement qu'avec des contours naturels. Cela nous a permis de valider perceptivement la première phase de l'apprentissage hiérarchique de notre modèle.

Il s'agit maintenant pour nous de "geler" les poids du réseau connexionniste prédisant la morphologie prosodique des attitudes, et de mettre en action le module de génération des contours portés de groupe.

3.3 Le module de groupe : apprentissage de la modulation

3.3.1 Architecture et apprentissage du module de groupe

3.3.1.1 Le modèle *phrase+groupe*

La figure 3.8 représente l'évolution du modèle de génération biparamétrique maintenant que le module de phrase a été entraîné de manière satisfaisante. Sur ce schéma d'architecture, on voit apparaître un nouveau réseau récurrent qui vient s'ajouter au réseau "figé" de phrase. Cette nouvelle contribution est obtenue concrètement par l'initialisation des 4 pondérations w qui passe de 0 à 1. Elle sera ajoutée à la prédiction du module de phrase.

Le module de groupe (Morlec, Bailly & Aubergé 1997b) est presque identique au module de phrase. L'architecture interne (nombre de couches cachées, nombre de neurones dans chaque couche, fonction neurone *atan*, récurrence) reste inchangée. Par contre les entrées sont cette fois adaptées à la génération des contours mélodiques et rythmiques indiquant la structure syntaxique de la phrase :

Trois rampes Deux rampes décroissantes indiquent la position du GIPC traité. Elles se comportent comme un compte à rebours entre la marque prosodique qui précède ce GIPC et celle qui le suit, ou entre le début de la phrase et la première marque,

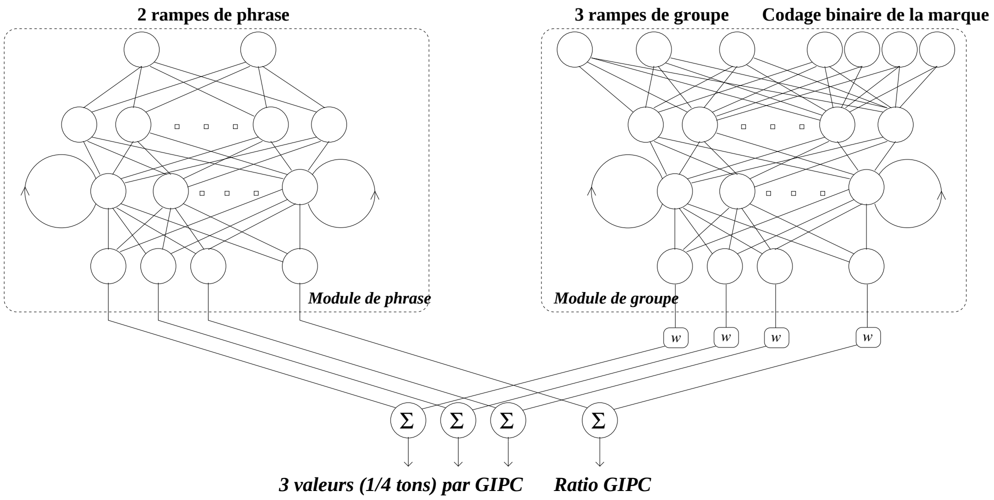


FIG. 3.8: Architecture du modèle *phrase+groupe*.

ou encore entre la dernière marque et la fin de la phrase. La première rampe est absolue et correspond au nombre de GIPC qui séparent le GIPC courant de la marque prosodique la plus proche. La seconde, relative, est calculée comme le rapport entre la rampe absolue et le nombre de GIPC du groupe courant.

Enfin, la troisième rampe décroissante et absolue indique la position du GIPC courant par rapport à la marque qui suit immédiatement la marque la plus proche. Cette rampe est destinée à modéliser les effacements de marque liés à la proximité, en terme de syllabes, de deux marques consécutives. Elle permettra au modèle de générer les effets de coarticulation de contours associés aux unités adjacentes d'un même niveau.

Quatre entrées binaires Ces entrées codent la nature de la marque prosodique qui borne à droite le GIPC courant. Le codage reste fixe tout au long du groupe courant. Chaque marque est codée par une seule entrée initialisée à 1, les autres restant à 0. Par exemple, La marque prosodique entre un sujet et un verbe recevra la séquence [0010], celle associée à une jonction *verbe/objet* [0001]. Ajoutons que la séquence de fin de phrase correspond à [0000].

Pour plus de clarté dans l'évolution de ces rampes et du codage des frontières, le tableau 3.6 définit les différentes entrées pour une phrase [GNsujet+verbe+GNobjet] de 11 syllabes.

3.3.1.2 Apprentissage supervisé

Durant l'apprentissage, les sorties du module de groupe correspondent strictement à la différence entre chaque contour prosodique cible et la prédiction du module de phrase.

Nous avons limité le corpus d'apprentissage à l'ensemble des phrases déclaratives des corpus "attitudes" et "modulation" contenant les structures syntaxiques [GNsujet+verbe],

Nombre de syllabes	1	2	3	4	5	6	7	8	9	10	11
Rampe phrase absolue	11	10	9	8	7	6	5	4	3	2	1
Rampe phrase relative	10	9.1	8.2	7.3	6.4	5.5	4.6	3.7	2.8	1.9	1
Rampe groupe absolue	4	3	2	1	2	1	5	4	3	2	1
Rampe groupe relative	10	7	4	1	10	1	10	7.75	5.5	3.25	1
Rampe coarticulation	6	5	4	3	7	6	5	4	3	2	1
Codage marque	0010	0010	0010	0010	0001	0001	0000	0000	0000	0000	0000

TAB. 3.6: Séquences d'entrée du modèle *phrase+groupe* pour une phrase [GNsujet (4syllabes) + verbe (2 syllabes) + GNobjet (5 syllabes)]. Exemple : *« Les japonais // mangeaient // du Tiramissu ».*

[GNsujet+verbe+GNobjet] et [GV+GNobjet] (Le GV est réduit au verbe précédé du pronom « j' » inférieur à la syllabe). Ce choix a été guidé par l'analyse acoustique des modulations de groupe du chapitre II qui nous a permis de constater que DC était la seule attitude pour laquelle des contours de groupe émergeaient clairement aussi bien mélodiquement que rythmiquement et ceci quelle que soit la position de la fin du groupe accentuel dans la phrase. Nous verrons plus loin que les modulations prosodiques de groupe des déclarations constitueront une *référence* à partir de laquelle nous prédirons les contours portés des autres attitudes.

Ce corpus d'apprentissage comporte 198 phrases de 2 à 12 syllabes et représente 74 phrases [GNsujet+verbe], 49 phrases [GV+GNobjet] et 75 phrases [GNsujet+verbe+ GNobjet]. La répartition des longueurs des différents syntagmes est donnée dans les tableaux 3.7. Seront donc présents dans ce corpus : 149 GNsujet de 1 à 6 syllabes, 123 GV de 1 à 6 syllabes et 124 GNobjet de 1 à 6 syllabes. Cette répartition homogène va nous permettre d'étudier les marques prosodiques liées aux frontières syntagmatiques entre *sujet/verbe* et *verbe/objet*, et de mettre en évidence la hiérarchie entre ces deux frontières lorsqu'elles sont toutes deux présentes dans une même phrase.

3.3.2 Validation objective

3.3.2.1 Prédictions sur le corpus d'apprentissage

Nous avons d'abord mesuré l'apport du module de groupe dans la génération des trois structures représentées dans le corpus d'apprentissage. Le tableau 3.8 donne les erreurs de prédiction sur le corpus d'apprentissage pour le module de phrase seul et celles qui résultent du modèle *phrase+groupe*. On constate une diminution substantielle de cette erreur (environ 3,6 % pour la F0 et pour la durée GIPC).

La deuxième analyse que nous avons menée consiste à observer les prédictions mélodiques et rythmiques de chaque structure et chaque distribution syllabique afin de les

(a)

Nombre de syllabes	1	2	3	4	5	6
GNsujet	5	11	17	19	12	10
verbe	28	19	18	13	3	1

(b)

Nombre de syllabes	1	2	3	4	5	6
GV	22	15	8	3	1	–
GNobjet	4	7	14	16	8	–

(c)

Nombre de syllabes	1	2	3	4	5	6
GNsujet	15	13	38	5	3	–
verbe	31	22	13	5	3	1
GNobjet	12	11	16	17	12	7

Tab. 3.7: Répartition des unités syntagmatiques par nombre de syllabes : (a) [GNsujet+verbe], (b) [GV+GNobjet] et (c) [GNsujet+verbe+GNobjet].

Erreur relative	F0	durée GIPC
Phrase	10.6	16.0
Phrase & Groupe	7.0	12.3

Tab. 3.8: Erreurs relatives sur la F0 et la durée GIPC pour les phrases du corpus d'apprentissage sur les prédictions du module de *phrase* seul et du modèle *phrase+groupe*.

comparer avec l'ensemble de leurs représentants dans le corpus d'apprentissage.

Les trois figures 3.9 font partie de cette analyse. Sur chaque figure sont représentés, en traits fins, les contours d'apprentissage permettant au modèle de réaliser des prédictions complètes (traits épais) qui correspondent à la somme des contours de phrase (pointillés) et des contours de groupe.

Nous remarquons sur ces figures que les prédictions du réseau marquent clairement les frontières syntaxiques et que les contours rythmiques respectent la hiérarchie des allongements en attribuant un allongement plus important pour les frontières *sujet/verbe* que pour les frontières *verbe/objet* (c). On peut également constater que le phénomène de « remise à zéro » est correctement modélisé : le modèle attribue une durée GIPC légèrement inférieure à la durée prédite par le module de phrase juste après la réalisation d'un accent “d'intonation”.

Ainsi, peut-on raisonnablement avancer que notre modèle est capable de générer des contours prosodiques dont le comportement est voisin de celui des contours d'apprentissage. Cependant, une observation plus fine nous amène à constater que le réseau prédit un allongement trop grand sur une frontière *sujet/verbe* pour les phrases [GNsujet+verbe] - (a) : Ratio GIPC=0.4 - et trop faible pour les phrases [GV+GNobjet] - (b) : Ratio GIPC=0.2 - alors que la valeur moyenne de cet allongement semble être autour de 0.3 pour (a) et (b). On voit donc que notre modèle conserve la hiérarchie des accents même quand la phrase n'en possède qu'un seul. Ce comportement pourrait être corrigé à l'avenir en enrichissant la description de l'entrée du module de groupe. Donner une information supplémentaire sur la nature de la marque prosodique suivant celle qui borne à droite le GIPC courant pourrait être une solution. Cette nouvelle entrée, ajoutée à la troisième rampe, permettrait de tenir compte plus efficacement des effets de coarticulation entre deux mouvements prosodiques consécutifs.

3.3.2.2 Expansion du mouvement et extrapolation

Comme pour l'expansion du mouvement de phrase, nous observons ici l'évolution des prédictions du module de groupe avec l'allongement des syntagmes et également l'allongement de la phrase.

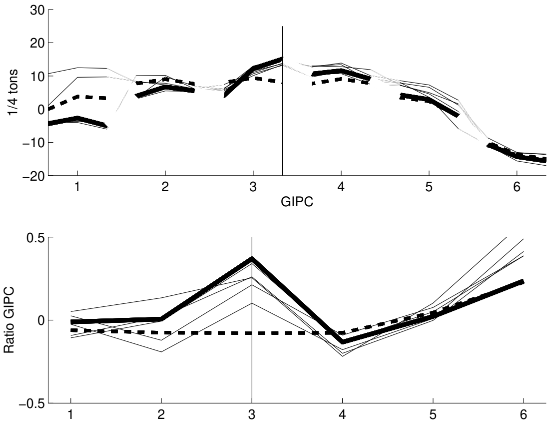
La figure 3.10 représente l'expansion des mouvements prosodiques du GNsujet allant de 2 à 5 syllabes pour des phrases variant entre 3 et 6 syllabes. Nous rappelons que ces contours sont obtenus en retranchant les prédictions globales aux prédictions du module de phrase. Ils correspondent donc exactement aux contours prédits par le module de groupe. Cette figure est à comparer avec la figure 2.15 de la section 2.3.2.4.

Notons d'abord quelques caractéristiques globales. Comme nous l'avons déjà dit précédemment, les frontières syntagmatiques sont toujours clairement marquées prosodiquement quel que soit le nombre de syllabes de l'unité. Nous constatons également que ces contours portés débutent par une valeur quasi nulle et que toute la partie finale, après la "remise à zéro" suivant la dernière marque prosodique est également nulle. Cela signifie simplement que le module de phrase a capturé efficacement les contours prosodiques de la déclaration. Le module de groupe n'a donc pas d'ajustement supplémentaire à réaliser[5].

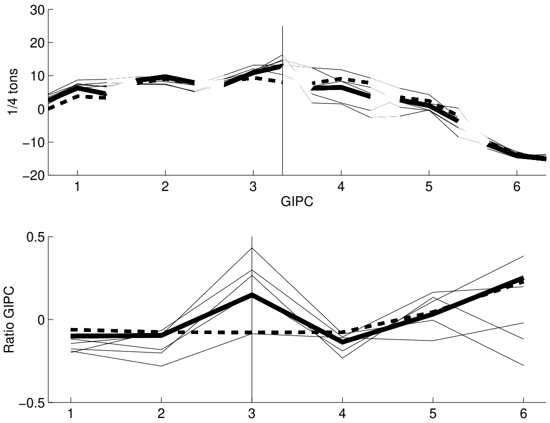
Nous remarquons également un allongement progressif du Ratio GIPC à l'intérieur du groupe rythmique. Cependant, nous n'observons pas un accroissement exponentiel sur la fin du groupe comme nous l'avons remarqué lors de l'analyse acoustique du corpus mais plutôt un allongement linéaire jusqu'à l'avant dernier GIPC du GNsujet suivi par un allongement beaucoup plus prononcé sur le dernier.

Au niveau des contours de F0, nous pouvons remarquer que l'on retrouve, à partir de 4 syllabes (c), un mouvement descendant qui prépare la montée finale du dernier GIPC du groupe. On note également, dès un GNsujet de 3 syllabes un "décrochage" du premier

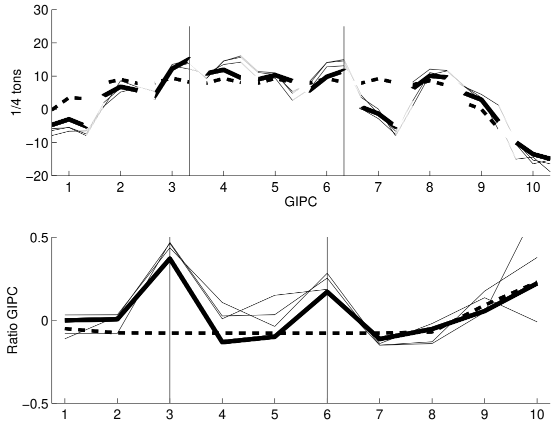
[5]Nous avons confirmé ce résultat en proposant des monomots au module de groupe. Les prédictions obtenues étaient toutes égales à zéro.



(a)



(b)



(c)

FIG. 3.9: Prédictions du modèle *phrase+groupe* pour des *déclarations* des corpus “attitudes” et “modulation” : (a) [GNsujet(3 syllabes) + verbe(3 syllabes)], (b) [GV(3 syllabes) + GNobjet(3 syllabes)] et (c) [GNsujet(3 syllabes) + verbe(3 syllabes) + GNobjet(4 syllabes)].

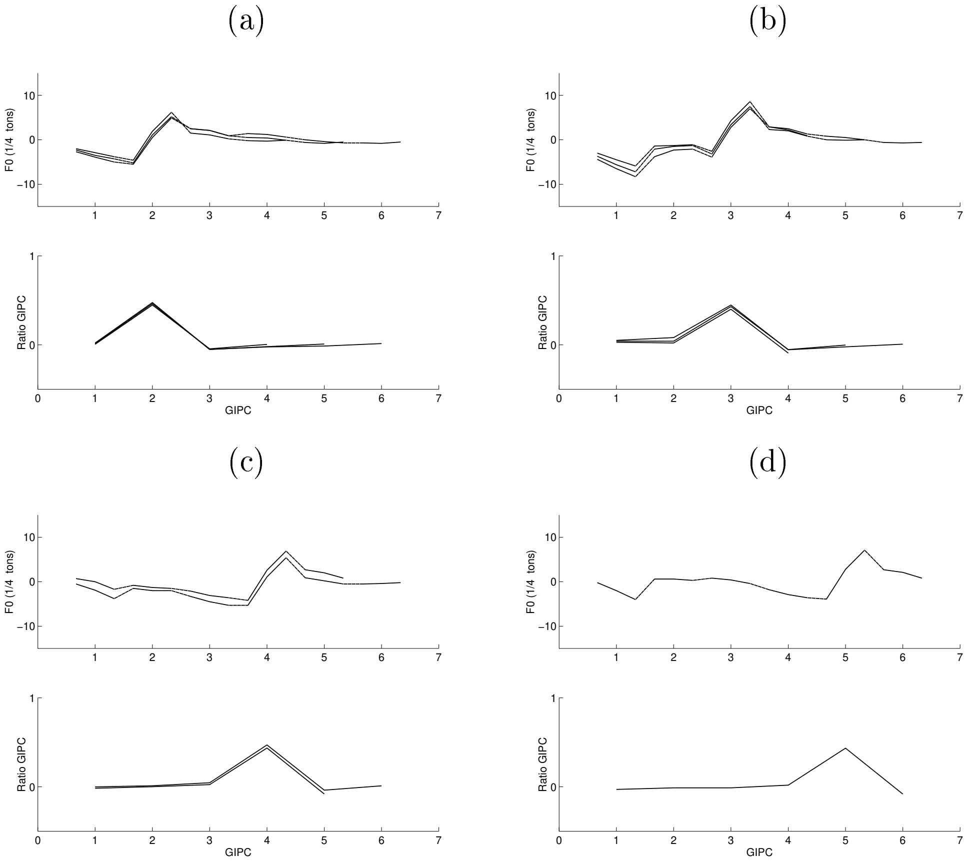


FIG. 3.10: Prédictions des GN sujets de 2 (a), 3 (b), 4 (c) et 5 (d) syllabes pour des phrases [GNsujet+verbe] de 3 à 6 syllabes.

GIPC dû à la prédominance de clitiques en initial de GN dans le corpus "modulation".

Nous terminerons notre analyse avec les phrases [GNsujet+verbe+GNobjet] dont les prédictions de groupe sont données en 3.11. Les 4 figures représentent l'expansion des mouvements prosodiques portés par le verbe. Outre le fait que l'ensemble des remarques faites plus haut s'appliquent également au cas présent, on peut ajouter que l'on observe, comme pour le GNsujet, une chute de fondamental en GIPC initial du GNobjet toujours liée à la présence d'un clitique.

3.3.3 Validation perceptive

Dans la section précédente, nous avons analysé les contours prosodiques de groupe prédits par le modèle *phrase+groupe* à partir des phrases déclaratives du corpus "attitudes" et du corpus "modulation". Nous avons mis en évidence que ces prédictions capturent de

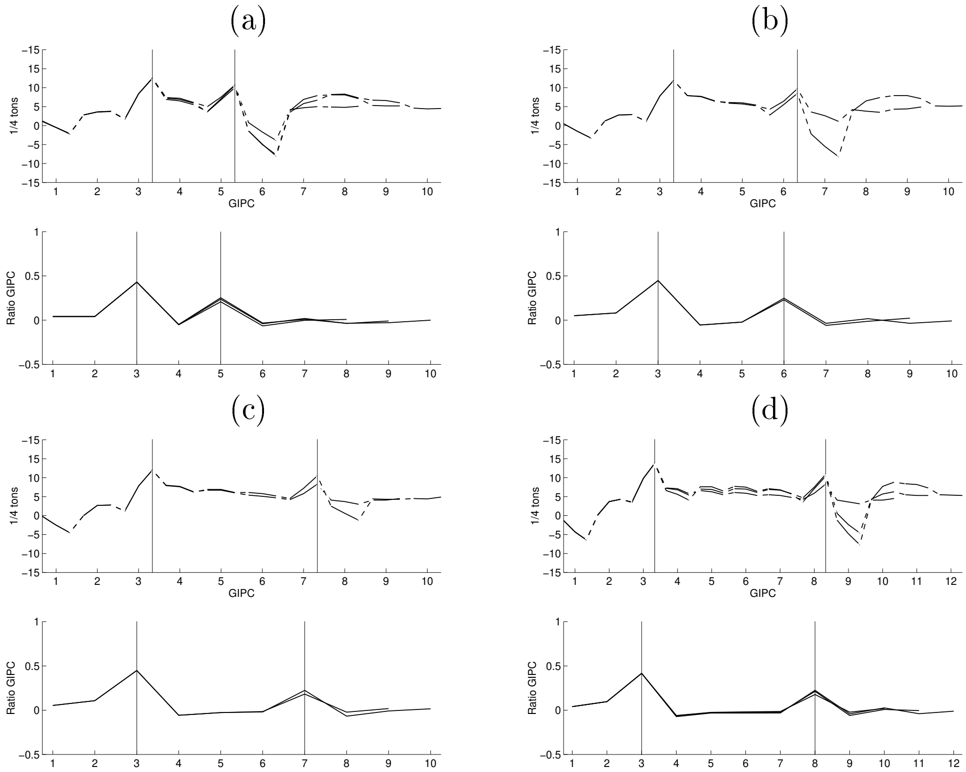


FIG. 3.11: Prédictions de GV de 2 (a), 3 (b), 4 (c) et 5 (d) syllabes pour des phrases [GNsujet + verbe + GNobjet] de 8 à 12 syllabes.

façon satisfaisante les événements prosodiques caractérisant trois structures syntaxiques : [GNsujet+verbe], [GV+GNobjet] et [GNsujet+verbe+GNobjet].

Comme pour le niveau de phrase seul, nous proposons maintenant une validation perceptive de ces prédictions. Toutefois, l'entreprise s'avère bien plus délicate que pour le test précédent. En effet, il nous est impossible, cette fois, de tout tester en même temps à savoir : identification et discrimination des trois structures, localisation des frontières entre les syntagmes. . . Il faut donc faire un choix tant sur l'objectif du test que sur le protocole à mettre en place.

3.3.3.1 Quelle méthode ?

Si l'on se réfère au rapport EAGLES (1994), on s'aperçoit que l'évaluation de la prosodie dans les systèmes de synthèse s'est tour à tour intéressée aux aspects formels et aux aspects fonctionnels de la prosodie.

Tests de jugement

L'évaluation des propriétés formelles de la prosodie, comme par exemple celle des mouvements mélodiques d'une phrase, est réalisée en général en demandant à un groupe d'auditeurs de donner sa préférence pour un stimulus donné, soit en le comparant à un autre stimulus, soit, dans l'absolu, en le quantifiant sur une échelle de "naturalité" ou de "bonne forme". Il s'agit là de tests de jugement.

Tests fonctionnels

L'évaluation de la prosodie utilisant des tests fonctionnels n'en est qu'à son balbutiement. Etant donné que la prosodie est en partie redondante avec l'information segmentale (excepté pour sa fonction expressive), les tests fonctionnels ne peuvent être réalisés que si l'on réduit cette redondance. Trois méthodes au moins permettent d'arriver à ces fins.

Désambiguïsation : La première consiste à utiliser des structures segmentales qui seraient ambiguës sans la prosodie. Le test consiste alors à demander aux auditeurs de résoudre cette ambiguïté. Si la désambiguïsation est satisfaisante, on peut admettre que le système testé possède les fonctions prosodiques appropriées. Ce type d'approche est utilisé dans les tests SUS (Benoît, Grice & Hazan 1996) qui proposent des phrases sémantiquement imprédictibles. Ces phrases sont obtenues à partir d'un lexique de mots contenant le moins de syllabes possible et de 5 structures syntaxiques différentes. Par exemple, pour une phrase [sujet+verbe+objet], la phrase présentée peut être « Le verre vrai ouvre le coin ». La tâche de l'auditeur consiste alors à noter par écrit le contenu de l'énoncé qu'il a entendu. Cette méthode est essentiellement destinée à évaluer et à comparer des systèmes de synthèse complets tant au niveau segmental qu'au niveau prosodique.

Dégradation : La seconde possibilité (Carlson, Granström & Klatt 1979) consiste à dégrader la qualité segmentale de telle sorte que, sans prosodie, l'intelligibilité de la parole synthétique soit très réduite. La performance de la prosodie synthétique est alors mesurée en terme de gain d'intelligibilité.

Délexicalisation : Enfin, la troisième méthode consiste à retirer, dans la phrase, tout accès aux informations linguistiques autres que prosodiques. Il existe plusieurs façons d'arriver à ce résultat. On peut citer l'usage de glottogramme (Lhote, Filleau & Grange 1975), la production de grommelis ou encore la parole réitérée par un ou plusieurs locuteurs sur une syllabe canonique telle que « ma » (Liberman & Streeter 1978) ou « ba » (Oller 1973). L'avantage de cette dernière méthode est de conserver la structure syllabique. Larkey (1983) avance que la méthode de réitération permet de préserver les caractéristiques mélodiques et rythmiques des phrases originales tout en neutralisant les différences intrinsèques des syllabes.

Ce type d'expérience utilisant la réitération a été mené au sein du laboratoire (Rilliard, Aubergé, Bailly & Morlec 1997) afin de montrer que la méthode peut s'avérer efficace pour l'évaluation de la prosodie "pure" des systèmes de synthèse actuels. Cependant, d'autres travaux doivent être mis en oeuvre pour mesurer précisément comment se fait le passage d'une phrase de référence à une phrase réitérée. Nous allons donc éviter pour notre expérience d'utiliser ce type de stimuli "non écologiques", et conserverons le support lexical sur lequel nous allons faire varier la prosodie.

L'objectif essentiel du test que nous allons présenter est principalement de vérifier l'adéquation de la structure prosodique réalisée par le locuteur et générée par le modèle avec les trois structures syntaxiques que nous avons étudiées. En ce sens, il s'agit d'un test fonctionnel.

3.3.3.2 Matériel

Cet objectif est réalisé en proposant un test de préférence contenant des paires de stimuli que l'on peut décomposer en trois catégories :

- Une paire représentée par un énoncé de structure syntaxique donnée sur lequel on applique (1) sa prosodie originale, (2) sa prosodie prédite. Nous la noterons (nat-juste,pred-juste). Cette catégorie nous permet une évaluation directe de la prosodie synthétique. Notons, tout de suite, que le terme « prosodie originale » désigne la stylisation des contours naturels telle que décrite en section 2.2.3.1.
- Une paire représentée par un énoncé de structure syntaxique donnée sur lequel on applique (1) sa prosodie originale, (2) la prosodie originale d'un énoncé comportant une autre structure syntaxique (nat-juste,nat-faux). Elle est destinée à montrer dans quelle mesure la prosodie naturelle peut être considérée comme bien formée en regard d'une structure syntaxique donnée et dans quels cas deux contours prosodiques associés à des phrases différentes peuvent être en adéquation avec une même structure syntaxique.
- Une paire représentée par un énoncé de structure syntaxique donnée sur lequel on applique (1) sa prosodie prédite, (2) la prosodie prédite d'un énoncé comportant une autre structure syntaxique (pred-juste,pred-faux). Elle remplira la même fonction que la précédente, mais avec des contours prédits. La mise en correspondance des choix effectués sur cette paire et sur la paire (nat-juste,nat-faux) nous permettra d'évaluer indirectement la capacité de notre modèle à remplir sa fonction de structuration de l'énoncé, et ceci de la même manière que la prosodie originale.

Le corpus de phrases utilisé dans cette expérience est donné en annexe A.4. Il est composé de 5 phrases de 6 syllabes et 4 phrases de 8 syllabes. Les phrases de 6 syllabes comportent un monomot, un [GNsujet+verbe], un [GV+GNobjet] et deux [GNsujet+verbe+GNobjet]. Les phrases de 8 syllabes comportent les mêmes structures, le [GV+GNobjet] mis à part (cette structure n'existe pas pour les phrases de 8 syllabes dans nos corpus).

Le nombre de syllabes des syntagmes de ces phrases a été déterminé de façon à donner une frontière syntagmatique commune (à 3 syllabes du début) aux différentes phrases de ce test comme l'indique le tableau 3.9.

<table>
<tr><td>Nombre de syllabes</td><td colspan="6">6</td><td colspan="8">8</td></tr>
<tr><td>Numéro</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td></tr>
<tr><td>V_O</td><td colspan="3">V</td><td colspan="3">O</td><td colspan="8">non représenté</td></tr>
<tr><td>S_V</td><td colspan="3">S</td><td colspan="3">V</td><td colspan="3">S</td><td colspan="5">V</td></tr>
<tr><td>monomot</td><td colspan="6">Mono</td><td colspan="8">Mono</td></tr>
<tr><td>S_VO</td><td colspan="3">S</td><td colspan="2">V</td><td>O</td><td colspan="3">S</td><td>V</td><td colspan="4">O</td></tr>
<tr><td>SV_O</td><td colspan="2">S</td><td>V</td><td colspan="3">O</td><td colspan="2">S</td><td>V</td><td colspan="5">O</td></tr>
</table>

TAB. 3.9: Structure syntaxique des phrases de 6 syllabes (à gauche) et de 8 syllabes (à droite). Monomot, GNsujet, GV et GNobjet sont notés respectivement Mono, S, V et O. La frontière commune est indiquée dans les identificateurs par "_".

Les stimuli ont été fabriqués de la façon suivante :

- Les 9 phrases ont été réenregistrées à 16 kHz. Le locuteur avait pour consigne de produire des énoncés sans variation prosodique : rythme constant et mélodie la plus "plate" possible. Cette étape est absolument nécessaire car si nous avions fait les transferts de prosodie sur le signal original, l'application d'une prosodie "fausse", ou synthétique, aurait produit un stimulus très dégradé et donc introduit un biais important dans ce test de préférence.

  L'analyse des phrases neutres ainsi obtenues a révélé des contours de F0 complètement plats et situés à $94.82 \pm 0.54Hz$ (soit environ 30 Hz en dessous du registre moyen du locuteur dans les corpus étudiés).
- Nous avons ensuite appliqué (par analyse-synthèse TDPSOLA) à chaque phrase "neutre" le contour prosodique naturel et prédit qui lui correspond, mais également les contours naturels et prédits associés aux autres phrases de même longueur. Par exemple, chaque phrase de 6 syllabes produit 5 versions avec prosodies naturelles - 1 (nat-juste) et 4 (nat-faux) - et 5 versions avec prosodies prédites - 1 (pred-juste) et 4 (pred-faux). On obtient par cette méthode 50 stimuli de 6 syllabes et 32 de 8 syllabes.
- Les stimuli sont rangés par paire suivant les 3 catégories citées plus haut : 9 paires (nat-juste,pred-juste), 32 paires (nat-juste,nat-faux) et 32 paires (pred-juste,pred-faux).

Ce test est de type ABBA et comporte donc à un total de 146 paires.

3.3.3.3 Protocole

12 sujets ont participé à cette expérience. Rappelons la consigne donnée à l'écran aux auditeurs avant d'en détailler le contenu :

```
Ceci est un test de préférence visant à comparer des voix synthétiques.
   - Une phrase est affichée à l'écran.
   - En appuyant sur "Ecoute", vous entendez 2 intonations différentes de cette
     phrase. Attention, vous ne pouvez l'écouter qu'une fois.
   - Laquelle de ces 2 versions est la plus acceptable pour une lecture simple ?
     C'est-à-dire sans insistance particulière sur une partie de la phrase,
     cliquez au choix sur :
        - La première version. (choix1)
        - La deuxième version. (choix2)
        - Si vous n'arrivez pas à vous décider, alors seulement vous répondez :
            - Les deux sont acceptables. (choix3)
            - Aucune des deux n'est acceptable. (choix4)
Ce test est précédé d'un court test blanc pour vous familiariser avec la tâche.
```

La consigne prépare les auditeurs à réaliser une tâche à choix fermé. Toutefois, nous ne leur imposons pas d'opter forcément pour l'un des deux stimuli de la paire présentée. Nous leur suggérons simplement d'éviter, dans la mesure du possible, de se porter sur les choix 3 et 4. La première phrase de la consigne, prévenant les auditeurs qu'ils ne vont pas entendre de la parole naturelle, nous affranchit d'une préférence trop prononcée pour le *choix4* liée aux distorsions des stimuli générés par PSOLA. Nous précisons également que, dans ces phrases, il n'y a pas de focalisation particulière. En effet, l'application d'une prosodie “fausse” peut conduire à l'interprétation des accents “d'intonation” qu'elle possède comme des accents d'insistance et, dans ce cas, la plupart des versions peuvent être considérées comme acceptables. Nous espérons donc, dans le scénario le plus favorable, que le *choix3* soit utilisé principalement lors de la présentation des paires (nat-juste,pred-juste) et, dans le scénario le plus défavorable, que le *choix4* sorte majoritairement avec ces paires auquel cas le locuteur n'aurait pas réalisé sa tâche de lecture correctement, en entraînant ainsi de mauvaises prédictions.

L'expérience débute par une phase d'entraînement de 6 paires. Le test se poursuit alors en respectant une présentation quasi aléatoire pour laquelle deux paires successives n'ont jamais le même contenu lexical.

3.3.3.4 Résultats

Les résultats obtenus à l'issue de cette expérience peuvent être séparés en deux catégories. La première concerne la comparaison entre les choix effectués après la présentation des paires (nat-juste,nat-faux) et ceux attribués aux paires (pred-juste,pred-faux). La seconde se concentre sur les paires (nat-juste,pred-juste) afin d'évaluer la performance de notre modèle.

Prosodie adéquate *vs* prosodie non-adéquate

Pour cette catégorie, l'ensemble de figures 3.12 établit une synthèse de tous les choix observés. Chaque figure doit s'interpréter de la façon suivante :

- On trouve en ordonnée les stimuli avec la prosodie juste (nat-juste ou pred-juste) et en abscisse les stimuli avec la prosodie fausse (nat-faux ou pred-faux).
- Pour chaque paire (juste,faux), une répartition des réponses est proposée. De bas en haut : Choix *juste* en blanc, choix *faux* en noir, choix *juste et faux acceptables* en gris clair et choix *juste et faux inacceptables* en gris foncé.

Voici un exemple avant d'aller plus loin : pour les paires de phrases naturelles de 6 syllabes (figure (n6) en haut à droite), la paire (Mono-juste , S_V-faux) correspond à une paire constituée d'un monomot avec sa prosodie originale et de ce même monomot sur lequel nous avons appliqué la prosodie d'une phrase [GNsujet+verbe]. La répartition - donnée sur la position $(x, y) = (4, 3)$ de la figure - indique que les choix se sont tous portés sur la version *juste* (case complètement blanche).

Points communs entre paire naturelle et paire prédite : Nous proposons d'abord un résultat permettant de comparer les distributions des choix entre paires (nat-juste,nat-faux) et (pred-juste,pred-faux). Le coefficient de corrélation des réponses entre *naturel* et *prédit* vaut $R_{6syll} = 0.844$ pour les phrases de 6 syllabes et $R_{8syll} = 0.968$ pour les phrases de 8 syllabes. Ceci permet d'observer que ces distributions sont globalement voisines, et le sont plus pour les phrases de 8 syllabes que pour les phrases de 6 syllabes.

On constate que pour les phrases de 8 syllabes, les choix se portent en grande majorité sur la version *juste*, que ce soit pour les paires *naturel* et les paires *prédit*. On remarque également que les choix *juste et faux acceptables* sont beaucoup plus représentés pour les paires de phrases de 6 syllabes (p6),(n6) que pour celles de 8 syllabes (p8),(n8). Ce phénomène peut être lié à la modulation de groupe plus marquée lorsque la phrase est longue, ce qui entraîne des énoncés avec prosodie fausse beaucoup moins acceptés pour les 8 syllabes que pour 6 syllabes.

Quelques observations locales permettent d'illustrer ce comportement :

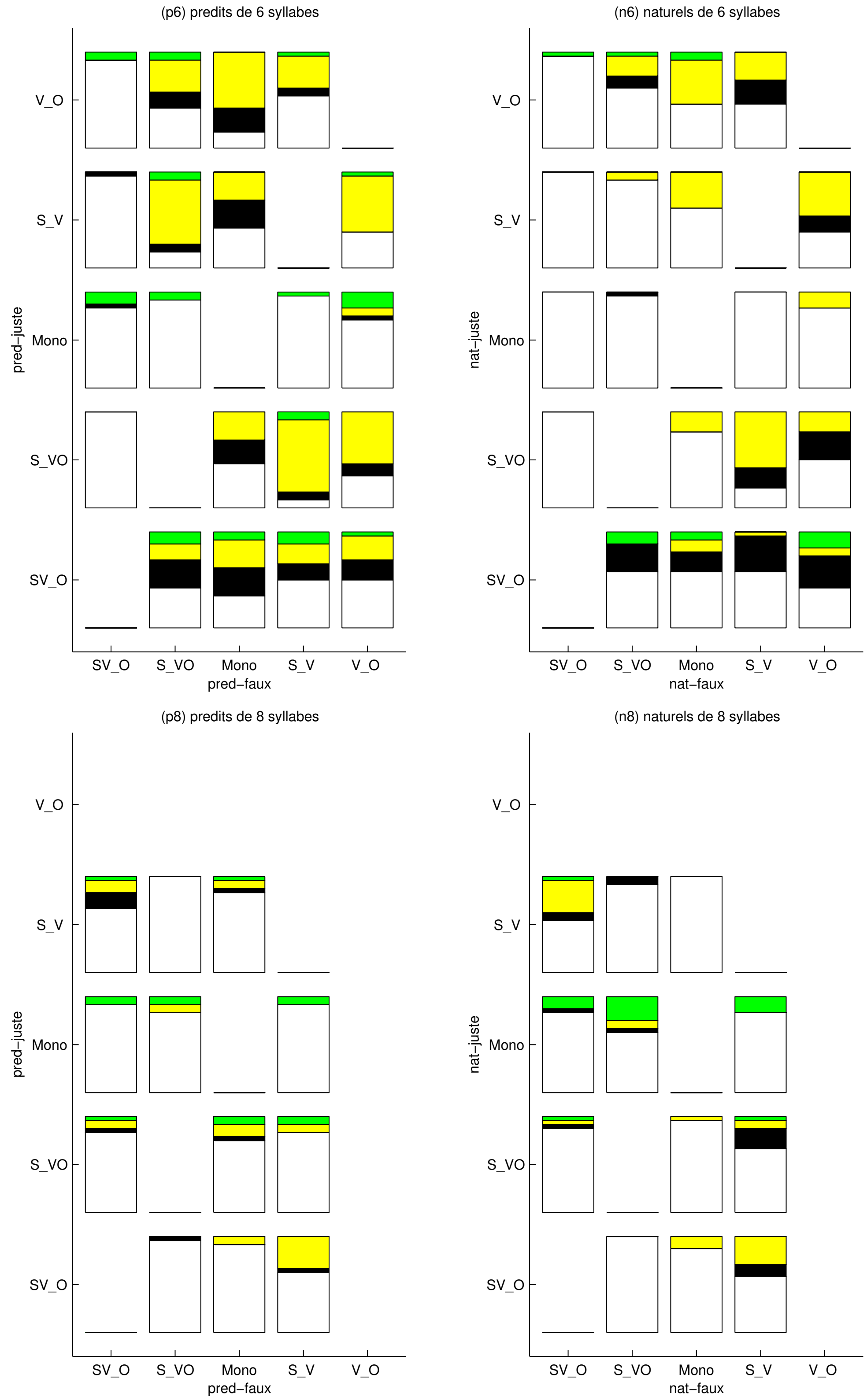


FIG. 3.12: Récapitulatif des choix effectués pour les phrases de 6 syllabes (figures du haut) et de 8 syllabes (figures du bas) pour les paires (pred-juste,pred-faux) (colonne de gauche (p6) et (p8)) et (nat-juste,nat-faux) (colonne de droite (n6) et (n8)).

- Lorsque la version juste est un monomot ($y = 3$), elle est préférée dans tous les cas (p6),(n6),(p8),(n8) à une version fausse qui comporte la modulation de groupe.
- Lorsqu'une version autre qu'un monomot est présentée, elle est largement préférée à une version fausse monomot pour (p8) et (n8) ($x = 3$) et beaucoup moins pour (p6) et (n6).
- Enfin , nous constatons que pour (p6) et (n6) les versions S_VO-juste et S_V-faux (4,2) sont considérées comme équivalentes alors que ce n'est pas de cas pour (p8) et (n8). Cela met en évidence le marquage plus important de la frontière *verbe/objet* quand le nombre de syllabes total des couples [verbe+objet] augmente.

Remarquons, pour terminer, que quel que soit le stimulus juste de 6 syllabes (p6)(n6), il est presque toujours choisi lorsqu'on le met en concurrence avec une structure SV_O (x=1). Ceci est en accord avec le fait que la frontière *sujet/verbe* est toujours marquée et ne peut être interprétée comme une frontière interne du GV, du GNsujet ou du monomot sur lesquels on l'applique (cf. tableau 3.9).

Différence entre paire naturelle et paire prédite : Les principales différences entre les choix portés sur les paires naturelles et les paires prédites se situent au niveau des phrases de 6 syllabes. Les choix *juste et faux acceptables* sont sensiblement plus représentés dans (p6) que dans (n6). L'écart le plus important se situe au niveau de la paire (S_V-juste,S_VO-faux) (2,4) : pour (p6), nous avons une majorité de *juste et faux acceptables* et pour (p8) c'est la réponse *juste* qui prédomine. Cela provient du fait que le modèle ne marque pas en durée la frontière *verbe/objet* lorsque l'ensemble est trisyllabique alors que le locuteur l'a marqué dans l'item S_VO sélectionné. Ainsi les stimuli prédits sont jugés équivalents alors que l'item *naturel* S_VO-faux est bien jugé inacceptable par les auditeurs.

Prosodie naturelle *vs* prosodie synthétique

La figure 3.13 indique la distribution des choix pour les paires (nat-juste,pred-juste) de phrases de 6 syllabes (a) et de 8 syllabes (b). On peut remarquer sur (a) et (b) que dans la majorité des cas, c'est le choix *naturel et prédit acceptables* qui est attribué à ces paires (plus de 50 % des réponses). Toutefois la catégorie des monomots échappe à la règle. Pour le monomot de 6 syllabes, la version naturelle est préférée alors que pour le monomot de 8 syllabes, c'est la prédiction.

- La première observation s'explique par le fait que notre modèle n'est pas, pour l'instant, suffisamment armé pour la génération de monomots, le niveau du *mot* n'étant pas encore implémenté : ainsi, l'accent d'insistance généralement présent sur la première syllabe des monomots a été gommé par l'ajout de structures [GNsujet + verbe] dans le corpus d'apprentissage des contours de phrase (cf. 3.2.2.2). De plus, un mot plurisyllabique reçoit un contour de durée spécifique qui réduit sa durée perçue. Ces propriétés étant absentes de la version prédite, la version naturelle est donc préférée.

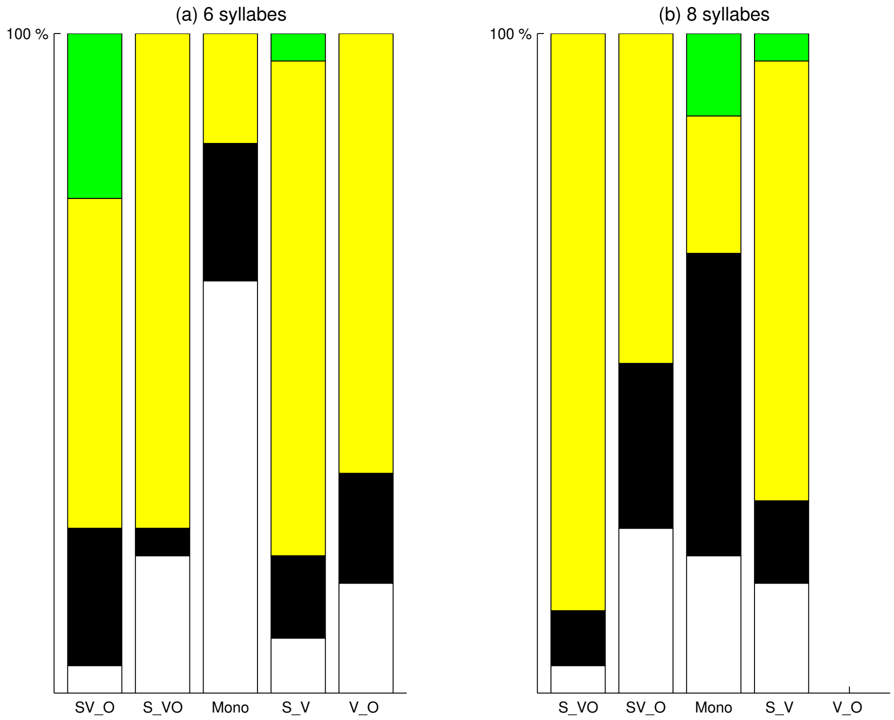


FIG. 3.13: Récapitulatif des choix effectués pour les phrases (a) de 6 syllabes et (b) de 8 syllabes sur les paires (nat-juste,pred-juste). En abscisse est représenté le type de structure syntaxique de la paire. La répartition correspond, de bas en haut, au : choix *naturel* (en blanc), choix *prédit* (en noir), choix *naturel et prédit acceptables* (en gris clair) et choix *naturel et prédit inacceptables* (en gris foncé).

– Pour le monomot de 8 syllabes, le résultat obtenu semble provenir de la consigne qui précise que le locuteur n'insiste pas sur un endroit particulier de la phrase. Mais pour « cytopathogénicité », le locuteur a marqué clairement la structure morphologique du mot. Les sujets se sont donc reportés sur la réalisation moins marquée de la prédiction. L'absence des propriétés lexicales, énoncées ci-dessus, dans la version prédite est encore pénalisante comme l'atteste le taux assez élevé de rejets.

Il reste enfin à noter l'importance inhabituelle du choix *naturel et prédit inacceptables* pour les paires de structure SV_O (« Les gars scient des rondins. »). Ce choix est lié à la qualité médiocre des signaux synthétiques générés par PSOLA sur cette phrase[6].

[6]Le *i* de « scient » est grandement dévoisé dans l'énoncé “plat” de référence ce qui implique des distorsions importantes liées à la technique de synthèse lors de l'allongement de la syllabe.

Discussion

Deux points importants ressortent de ce test :

- L'analyse des corrélations entre les réponses pour les paires (nat-juste,nat-faux) et (pred-juste,pred-faux) révèle que la répartition des choix est très semblable dans les deux cas et surtout pour les phrases de 8 syllabes.
- Les résultats pour les paires (nat-juste,pred-juste) ne révèlent aucune préférence globale pour les stimuli naturels.

Nous pouvons donc penser, au vu de l'analyse acoustique précédente et de ce test de préférence, que les contours prosodiques de groupe que nous prédisons sont bien formés et qu'ils remplissent leur fonction de structuration de l'énoncé de façon satisfaisante.

3.4 De l'indépendance des contours porteurs/portés

Maintenant que nous sommes parvenus à modéliser les contours porteurs de phrase et portés de groupe, nous nous intéressons aux relations éventuelles qu'ils peuvent entretenir. L'hypothèse de base de notre démarche postule l'indépendance entre les différents niveaux linguistiques, et donc l'indépendance des contours véhiculant les attributs de chaque unité de ces niveaux. Cependant, nous avons entrevu, lors de l'analyse acoustique du chapitre II, que les formes globales véhiculant l'attitude du locuteur pouvaient influencer le comportement des sous-niveaux, et qu'elles n'agissaient pas de la même façon sur chaque paramètre prosodique. L'analyse statistique que nous allons développer, sur la base de nouvelles observations, va nous permettre d'approfondir cette information et nous conduira à apporter des modifications essentielles à notre modèle afin de générer les 6 attitudes du corpus, contours de phrase et modulations de groupe compris.

3.4.1 Validation sur le corpus "modulation"

La première étape de cette analyse consiste à évaluer, sur les phrases déclaratives, l'indépendance entre les niveaux de phrase et de groupe. A cet effet, nous avons "dégelé" le réseau de phrase (en laissant à nouveau évoluer ses poids) et réalisé, en partant des poids déjà appris, un apprentissage supervisé du modèle *phrase+groupe*. Le corpus est cette fois constitué des monomots du corpus d'apprentissage du module de phrase, et des phrases [GNsujet+verbe], [GV+GNobjet] et [GNsujet+verbe+GNobjet] du corpus d'apprentissage du module de groupe.

Nous avons pu constater, après une centaine d'itérations de ce corpus, que les contours prédits de phrase et de groupe n'évoluent pratiquement pas durant l'apprentissage[7]. Cela signifie simplement que chaque module a déjà capturé les régularités spécifiques au niveau auquel il est associé et qu'il n'y a pas de phénomène d'empiétement entre niveaux pour réajuster les prédictions du modèle sur des valeurs plus proches des contours cibles.

3.4.2 Contrôle de l'amplitude de la modulation des portés par les porteurs

3.4.2.1 Mise en évidence

Pour mettre en évidence l'influence des contours porteurs sur les portés, nous avons d'abord entrepris quelques observations locales sur les attitudes qui nous paraissaient les plus représentatives de ce phénomène.

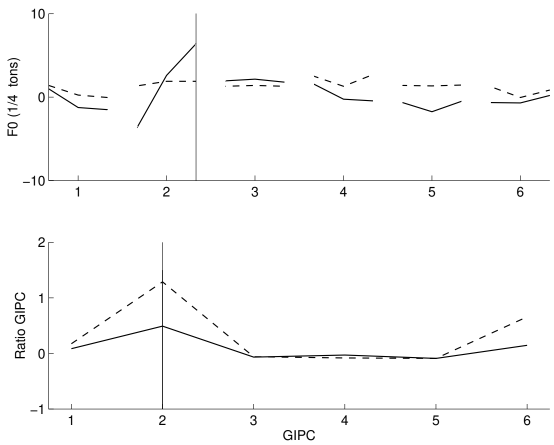
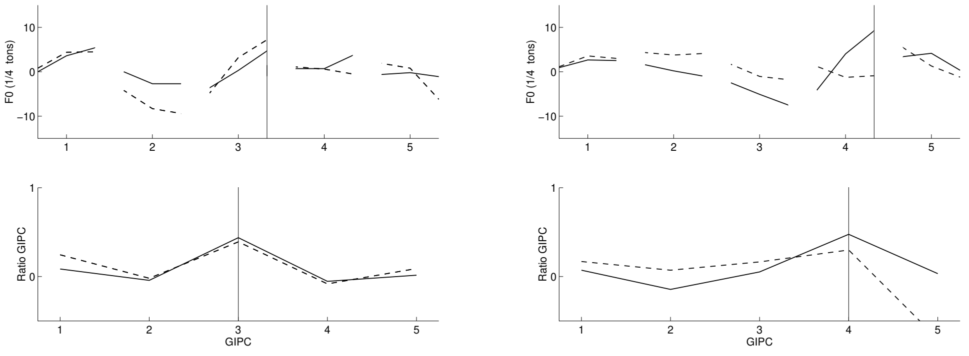


FIG. 3.14: Contours de groupe pour des *déclarations* (traits pleins) et des *questions* (pointillés) : [GNsujet(2 syllabes) + verbe(4 syllabes)].

Nous avons sélectionné les questions pour lesquelles nous avions remarqué l'insensibilité des contours mélodiques vis-à-vis de la structure syntaxique de la phrase (cf. figure 2.7). La figure 3.14 représente les contours moyens de groupe[8] pour des déclarations (trait plein) et des questions (pointillés) de 6 syllabes et de structure [GNsujet (2 syllabes) + verbe (4 syllabes)]. Cet exemple montre clairement que, pour QS, les modulations mélodiques de groupe sont quasi nulles. Par contre, l'allongement rythmique qui marque la frontière syntagmatique est plus important pour QS que pour DC.

[7]Cette observation qualitative reste à être appuyée par une analyse quantitative de ces variations.

[8]Nous rappelons que ces courbes correspondent à la différence entre les contours naturels et les contours moyens de toutes les phrases du corpus de même longueur et de même attitude.



(a) (3+2) (b) (4+1)

FIG. 3.15: Modulations de groupe pour des *déclarations* (traits pleins) et des *exclamations* (pointillés) dans les énoncés [GNsujet+verbe] pour deux distributions syllabiques.

Après cet exemple simple où nous n'observons pas de modulation mélodique de groupe sur l'ensemble de la phrase, nous passons à l'exemple plus complexe des exclamatives. La figure 3.15 compare les contours de groupe des EX aux contours des DC pour des phrases dont la position accentuée est à (a) 2 syllabes de la fin de la phrase et (b) sur l'avant dernière syllabe. Nous notons que, pour (a), la marque prosodique est réalisée mélodiquement et rythmiquement alors que pour (b) seul l'allongement rythmique est présent. Cette particularité démontre bien l'influence du contour de phrase sur la modulation. Nous avons vu au chapitre II que l'attitude du locuteur était principalement transmise par le mouvement de fréquence fondamentale. Pour qu'un contour d'exclamation soit effectivement identifié, le mouvement mélodique sur les deux dernières syllabes *descendant-(montant-descendant)* semble être indispensable. La modulation de groupe qui modifierait l'allure du contour de F0 porté par l'avant dernière syllabe en le réhaussant est donc empêchée (figure (b)).

Nous allons maintenant étendre ces observations sur le contrôle de la modulation des portés par les porteurs par une analyse statistique sur l'ensemble des phrases du corpus et deux attitudes : EX et DI. Mais, d'abord, énonçons clairement les hypothèses de l'analyse qui suit : nous considérons la modulation de groupe issue de la déclaration comme la *référence* de la capacité du locuteur à segmenter son énoncé. Nous allons comparer les modulations de groupe des autres attitudes à cette référence, et chercher si celles-ci peuvent effectivement en être déduites.

Nous avons donc reporté sur un diagramme les contours de groupe mélodiques et rythmiques de chaque attitude en fonction des contours de groupe mélodiques et rythmiques des déclarations[9]. Cette méthode nous permet, pour chaque position du GIPC dans la phrase, d'obtenir un nuage de points que nous avons représenté par une ellipse de disper-

[9]Ces valeurs sont obtenues cette fois en retranchant les prédictions du module de phrase aux contours naturels.

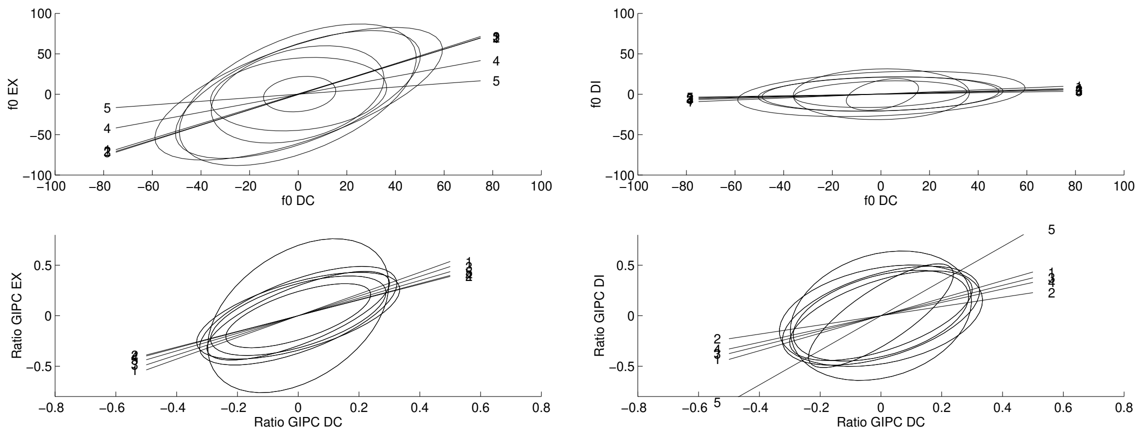


FIG. 3.16: Ellipses de dispersion et régressions linéaires pour chaque GIPC d'*exclamations* et de *doute-incrédulités* par rapport aux *déclarations* (phrases de 6 syllabes).

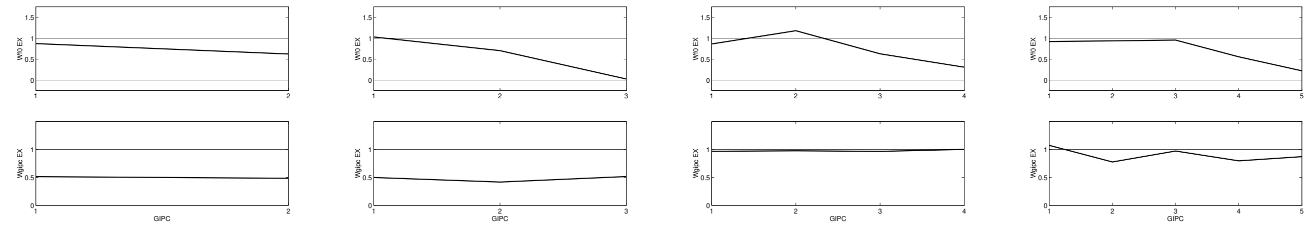
sion afin de faciliter la lecture des figures. Pour chaque ellipse associée à une position de GIPC donnée, nous avons effectué une régression linéaire. Ces ellipses sont centrées sur l'origine car le prototype de phrase est pratiquement égal à la moyenne des contours. Les pentes des droites de régression correspondent alors aux facteurs multiplicatifs qu'il faut appliquer aux contours de groupe des DC pour obtenir les contours de groupe des autres attitudes.

La figure 3.16 représente deux de ces diagrammes pour les exclamatives (à gauche) et les doute-incrédulités (à droite) de 6 syllabes. Chaque diagramme est constitué de deux figures : celle du haut correspond aux contours de groupe mélodiques et celle du bas aux contours rythmiques.

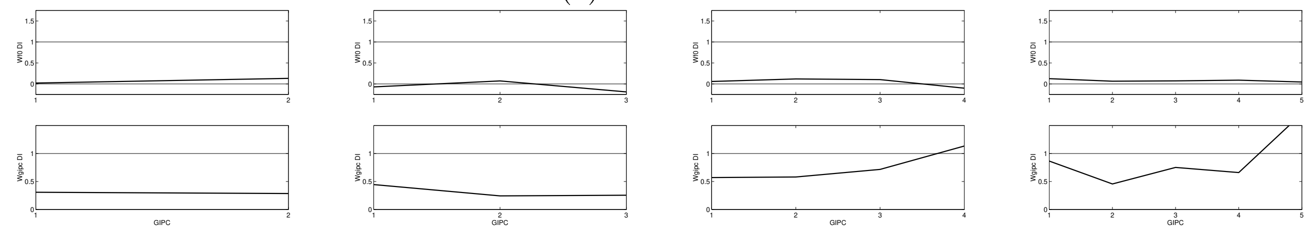
Nous observons en premier lieu que, dans les deux cas, l'ellipse de dispersion mélodique la plus petite correspond à l'ellipse de dispersion rythmique la plus grande. Les deux sont associées au GIPC final de la phrase. Ceci traduit un phénomène que nous avons évoqué au chapitre II : le mouvement mélodique final est très précis alors que l'allongement final est très variable quelle que soit l'attitude. Pour ces deux ellipses, nous n'avons pas calculé les droites de régression étant donné le regroupement des points de la première autour de zéro et la dispersion importante de la seconde.

Pour DI, on remarque que la modulation mélodique est inexistante (droites de régression horizontales) alors qu'elle est présente au niveau rythmique et même très marquée sur l'avant dernier GIPC de la phrase. Pour EX, la modulation mélodique, présente en début de phrase, s'amenuise peu à peu pour être presque nulle sur l'avant dernier GIPC. Au niveau rythmique, la modulation est semblable à celle des DC.

Sur la figure 3.17, nous avons reporté les pentes des différentes droites de régression associées à chaque GIPC (du premier à l'avant-dernier - cf. remarque précédente sur les



(a) Exclamatives



(b) Doute-incrédulité

FIG. 3.17: Pondérations de groupe obtenues par régression linéaire pour des phrases de 3 à 6 syllabes.

ellipses de dispersion) pour des EX (a) et des DI (b) de 3 à 6 syllabes .

Cette figure représente, de façon plus claire que la précédente, les facteurs multiplicatifs (que nous nommerons *pondérations*) w_{F0} et w_{GIPC} qu'il va falloir appliquer au module de référence, entraîné sur les DC, pour générer les modulations de groupe de ces deux attitudes. Pour les EX, les pondérations respectent l'évolution suivante :

- Les pondérations w_{F0} décroissent globalement de 1 vers 0. La valeur minimale est atteinte pour l'avant-dernier GIPC qui correspond au démarrage du mouvement de capture. Cette allure globale pourrait donc être décrite par un modèle d'expansion du mouvement possédant les mêmes phases que le modèle mélodique correspondant.

- Les pondérations w_{GIPC} sont voisines de 0.5 pour les phrases de longueur inférieure à 5 syllabes puis voisines de 1 au-delà. Rappelons cependant que les phrases contenant moins de 5 syllabes ne comportent qu'un nombre faible de réelles modulations de groupe.

Pour les DI, l'évolution est la suivante :

- L'observation des pondérations w_{F0} permet de généraliser nos précédents commentaires sur les DI de 6 syllabes : les pondérations mélodiques sont pratiquement nulles. Cela confirme donc l'absence de modulation de groupe pour cette attitude.

- les pondérations w_{GIPC} sont faibles pour les phrases comportant moins de 5 syllabes. Au delà, elles augmentent surtout sur l'avant dernier GIPC.

3.4.2.2 Une solution : le contrôle des pondérations des portés par les porteurs

La génération automatique des pondérations revient concrètement à attribuer aux attitudes un contrôle sur les w de notre modèle *phrase+groupe* pour chaque position de GIPC.

Ce contrôle peut être réalisé par apprentissage automatique en ajoutant au modèle un nouveau réseau connexionniste générant l'évolution des pondérations de groupe (Morlec, Bailly & Aubergé 1997c) et contrôlé par les deux rampes de phrase (figure 3.18). Ce réseau fournit des activations w_{F0} et w_{GIPC} qui sont multipliées aux prédictions du module de groupe de référence avant que celles-ci ne soient ajoutées au contours de phrase.

L'apprentissage de ce nouveau module comporte pour chaque attitude les étapes suivantes :
- On "gèle" les poids du module de phrase correspondant à cette attitude.
- On "gèle" les poids du module de groupe de référence entraîné sur les DC.
- On réalise un apprentissage supervisé du module de pondérations de groupe en présentant à la sortie du modèle les contours de [GNsujet+verbe], [GV+GNobjet] et [GNsujet+verbe+GNobjet] associés à cette attitude.

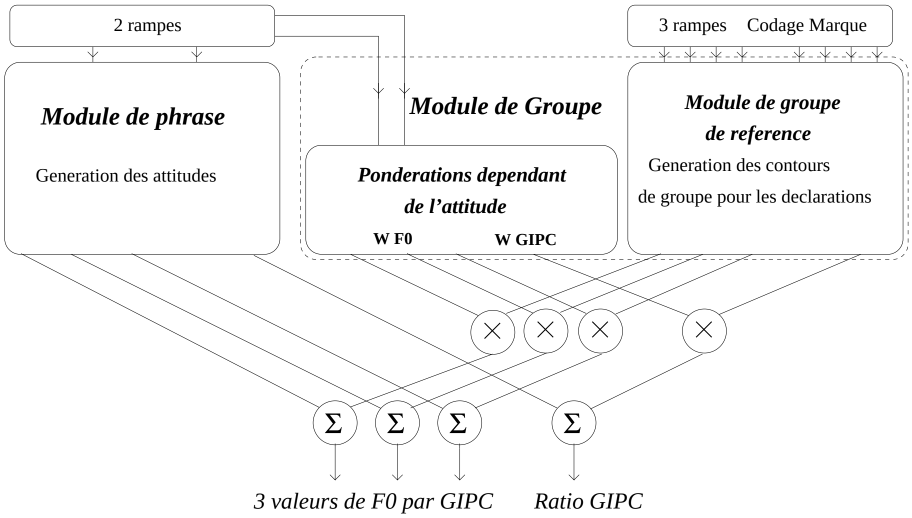


FIG. 3.18: Architecture du modèle complet *phrase+groupe*.

3.5 A retenir

Nous avons proposé une architecture modulaire de modèles connexionnistes calquée sur l'hypothèse de codage prosodique de la structuration de l'énoncé par des modulations spécifiques à chaque niveau linguistique. Deux modules ont été élaborés, entraînés et évalués par des tests spécifiques :

Un module de phrase : Les propriétés dynamiques des réseaux séquentiels nous ont permis d'implémenter un module de phrase qui génère des expansions du mouvement compatibles avec celles observées lors de l'analyse du corpus "attitudes". La génération automatique de ces attitudes a été évaluée par un test d'identification qui montre des structures de matrices de confusion entre stimuli *naturels* et *prédits* quasiment identiques.

Un module de groupe : Ce module est pour l'instant limité aux structures [(GNsujet+) GV (+GNobjet)]. Il a été évalué par un test fonctionnel original opérant par transfert de prosodie. Ce test permet de mettre en lumière l'importance perceptive relative des frontières syntaxiques, et de comparer les performances des prosodies naturelles et prédites.

Enfin, l'hypothèse d'indépendance forte des mouvements porteurs et portés a été reformulée en lui substituant une hypothèse d'indépendance faible : nous avons mis en évidence la nécessité d'un contrôle d'amplitude de la modulation de groupe suivant le mouvement porteur, afin de ne pas masquer le contenu informationnel de ce dernier. L'importance de ce contrôle dépend donc de la position du GIPC dans la phrase et du paramètre prosodique considéré ($F0$ ou k_{GIPC}).

Cette dernière évolution permet d'adapter notre modèle aux seules données du corpus "attitudes" qui, répétons-le, n'est représentatif que d'un locuteur dans une situation artificielle. Il nous reste à évaluer si cette nouvelle hypothèse d'indépendance faible est nécessaire et suffisante, en comparant les performances des contours prosodiques naturels avec ceux issus d'une somme directe de prédictions (indépendance forte), ou d'une somme pondérée dynamiquement (indépendance faible).

Conclusions et perspectives

Nous avons, durant ces trois années de recherche, développé un modèle connexionniste de génération biparamétrique de la prosodie. Cette étude s'est appuyée sur l'approche théorique d'Aubergé qui postule un **encodage direct** des informations linguistiques et para-linguistiques *via* un lexique de formes intonatives prototypiques. L'intégration de la composante rythmique a été facilitée par les travaux de Barbosa. Nous avons montré, après lui, la pertinence de la considération d'un événement psychoacoustique privilégié, le *Perceptual-center* (PC), deux PC consécutifs définissant les frontières du groupe *inter-Perceptual-center* (GIPC) et dont la durée normalisée k_{GIPC} a constitué notre unité de programmation rythmique.

Afin de proposer une approche cohérente, notre **analyse descendante** s'est appuyée sur une **morphologique superpositionnelle** selon laquelle chaque niveau de description linguistique sur l'axe syntagmatique génère des formes globales qui lui sont propres et qui maintiennent entre elles des contrastes propres à ce niveau. Nous avons donc entrepris la génération des **prototypes mélodiques et rythmiques** du niveau de la *phrase* véhiculant sa modalité ou l'attitude du locuteur vis-à-vis de son propos ou de son interlocuteur. Nous avons également étudié les modulations du niveau porté de *groupe* prenant en charge la fonction de structuration de l'énoncé.

Chaque étape d'apprentissage de notre modèle a été sanctionnée par un **test de perception** destiné à évaluer sa capacité à capturer les informations nécessaires et suffisantes pour communiquer les fonctions linguistiques données a priori dans chaque corpus.

Un certain nombre d'incertitudes théoriques a été levé. En outre, un ensemble de propriétés non entrevues dans nos propositions initiales a été mis en évidence et ajouté au modèle :

- L'étude de l'encodage prosodique des attitudes a révélé deux phénomènes :

  Le lexique prosodique : Des mouvements prototypiques de phrase existent, sont différenciés, et semblent bien organisés en ensembles catégoriels que nous pensons être modélisables sous la forme d'un lexique prosodique. Cette hypothèse est confortée par la présence et l'exploitation perceptive de points d'unicité révélés dans notre expérience de dévoilement progressif.

Le Modèle d'Expansion du Mouvement : Ces mouvements prosodiques de phrase se déploient (*préparation*) et sont ancrés (*capture*) autour de repères relativement fixes qui ne dépendent pas des constituants de plus bas niveau.

– L'étude de la modulation de groupe a permis de décrire et de modéliser comment la superposition des mouvements de phrase et de groupe se **négocie** en fonction de l'attitude prosodique du locuteur.

Notre recherche sur les attitudes a suggéré de nombreuses pistes de recherches centrées sur le lien supposé entre les formes que nous avons fait émerger, et leur interprétation conventionnelle. Les travaux de Tuulikki Grépillat (1996), Véronique Moroni (1997) et Susanne Alt (1997) en sont la concrétisation.

Le modèle, dans sa version actuelle, ne réalise pas une génération complète de la prosodie pour la synthèse à partir du texte. Nous avons, jusqu'ici, testé la capacité de ce modèle fonctionnel modulaire à capturer les formes prosodiques véhiculant quelques fonctions attitudinales et énonciatives de l'intonation. L'utilisation de réseaux connexionnistes nous a permis de mener à bien ce projet qui doit, à présent, évoluer vers l'apprentissage automatique complet des sous-niveaux mais aussi dans l'autre sens, c'est à dire vers une génération dépassant le cadre de la phrase. La "boîte à outils" que nous avons développée étant à présent bien maîtrisée, il ne reste plus qu'à lui fournir les corpus adéquats.

Le modèle devra également prendre en compte des phénomènes de réductions prosodiques révélés par Barbosa au niveau rythmique et qui se traduisent par l'effacement progressif de certaines marques prosodiques en fonction du débit d'élocution.

Nous avons entamé une réflexion sur le problème de modélisation et de génération automatique des **variantes** autour des prototypes que notre modèle génère. L'approche neuromimétique nous permettra de structurer spatio-temporellement la variance sur la base de ces prototypes.

Ces variantes peuvent être la conséquence d'une fonction essentielle de l'intonation mise entre parenthèses dans cette étude, quoique bien présente dans nos tests de perception : la mise en valeur ou **focus**. Deux questions essentielles peuvent être posées à propos de la focalisation sur un élément ou sur un contraste entre éléments de l'énoncé :

– N'a-t-elle qu'une incidence acoustique locale sur les mouvements prosodiques ou bien les influence-t-elle dans leur globalité ?
– Si l'effet est morphologiquement local, a-t-il une incidence perceptive locale ou globale ?

Nous avons proposé un paradigme de couplage faible entre intonation de phrase et intonation morpho-syntaxique. Il reste à évaluer l'efficacité du couplage proposé et à démontrer que les larges ellipses de dispersion observées autour de la stratégie proposée sont bien interprétables en termes de degrés de liberté ou de variantes.

Bibliographie

Abe, M. et Sato, H. (1992). Two-stage F0 control model using syllable based F0 units, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 53–56.

Abry, C., Benoît, C., Boë, L. J. et Sock, R. (1985). Un choix d'évènements pour l'organisation temporelle du signal de parole, *Actes des XIVèmes Journées d'Etude sur la Parole* pp. 133–137.

Abry, C., Orliaguet, J.-P. et Sock, R. (1990). Patterns of speech phasing. their robustness in the production of a timed linguistic task: single versus double (abutted) consonants in French, *European Bulletin of Cognitive Psychology* **10**.

Allen, G. D. (1975). Speech rhythm: its relation to performance universals and articulatory timing, *Journal of Phonetics* **3**: 75–86.

Alt, S. (1997). Prosodie attitudinale de l'allemand : indices sur la globalité des réalisations intonatives de quelques particules illocutoires, *D.E.A. sciences du langage*, Université Stendhal, Grenoble, France.

Aubergé, V. (1991). La synthèse de la parole : "des règles aux lexiques", *Thèse de troisième cycle*, Université Pierre Mendès-France, Grenoble, France. Sous la direction de J. Rouault.

Aubergé, V. (1992). Developing a structured lexicon for synthesis of prosody, *in* G. Bailly et C. Benoît (eds), *Talking Machines: Theories, Models and Designs*, Elsevier B.V., pp. 307–321.

Aubergé, V. (1993). Prosody modeling with a dynamic lexicon of intonative forms: Application for text-to-speech synthesis, *Working Papers of Lund University* **41**: 62–66.

Aubergé, V., Grépillat, T. et Rilliard, A. (1997). Can we perceive attitudes before the end of sentences ? The gating paradigm for prosodic contours, *Proceedings of the European Conference on Speech Communication and Technology*, Vol. 2, Rhodes, Greece, pp. 871–874.

Austin, J. L. (1962). *Quand dire c'est faire*, Seuil, Paris.

Bailly, G. (1983). Contribution à la détermination automatique de la prosodie du français parlé à partir d'une analyse syntaxique. Établissement d'un modèle de génération, *Thèse de troisième cycle*, Institut National Polytechnique de Grenoble, Grenoble, France. Sous la direction de M. Rossi et R. Carré.

Bailly, G. (1989). Integration of rhythmic and syntactic constraints in a model of generation of French prosody, *Speech Communication* **8**: 137–146.

Bailly, G., Aubergé, V. et Morlec, Y. (1997). Des représentations cognitives aux représentations phonétiques de l'intonation, *Polyphonie pour Iván Fónagy*, Editions L'Harmattan, pp. 19–28.

Barbosa, P. (1994). Caractérisation et génération automatique de la structuration rythmique du français, *Thèse de troisième cycle*, Institut National Polytechnique de Grenoble, Grenoble, France. Sous la direction de G. Bailly.

Barbosa, P. et Bailly, G. (1992). Generating segmental duration by P-centers, *in* C. Auxiette, C. Drake et C. Gérard (eds), *Fourth Rhythm Workshop: Rhythm Perception and Production*, Ville de Bourges, Bourges, France, pp. 163–168.

Barbosa, P. et Bailly, G. (1994). Characterisation of rhythmic patterns for text-to-speech synthesis, *Speech Communication* **15**: 127–137.

Bartkova, K. et Sorin, C. (1987). A model of segmental duration for speech synthesis in French, *Speech Communication* **6**: 245–260.

Beaugendre, F. (1994). Une étude perceptive de l'intonation du français: Développement d'un modèle et application à la génération automatique de l'intonation pour un système de synthèse à partir du texte, *Thèse de troisième cycle*, Université Paris XI, Paris, France. sous la direction de C. d'Alessandro.

Beckman, M. E. (1997). Speech Models and Speech Synthesis, *in* J. P. H. van Santen, R. W. Sproat, J. P. Olive et J. Hirschberg (eds), *Progress in Speech Synthesis*, Springer Verlag, New York, pp. 185–209.

Beckman, M. E. et Edwards, J. (1992). Intonational categories and the articulatory control of duration, *in* Y. Tohkura, E. Vatikiotis-Bateson et Y. Sagisaka (eds), *Speech, perception, production and linguistic structure*, IOS Press, pp. 359–376.

Benoît, C., Grice, M. et Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences, *Speech Communication* **18**: 381–392.

Berthier, V., Abry, C. et Lalouache, T. (1991). Coordination du geste et de la parole dans la production d'un instrument traditionnel, *Proceedings of the XIIth International Conference on Phonetic Sciences*, Vol. 4, Aix-en-Provence, France, pp. 34–37.

Blaauw, E. (1995). *On the perceptual classification of spontaneous and read speech*, PhD thesis, OTS Dissertation Series. ISBN 90-5434-045-2.

Boë, L. et Contini, M. (1976). Synthèse paramétrique de la phrase interrogative en français, *Actes des VIIèms Journées d'Etude sur la Parole*, Nancy, France, pp. 130–144.

Boë, L.-J. et Contini, M. (1975). Etude de l'intonation de la phrase interrogative en français (question totale). premiers résultats, *Bulletin de l'Institut de Phonétique de Grenoble*, Université de Grenoble III, Grenoble, France.

Bolinger, D. (1951). Intonation: Levels versus configuration, *Word* **VII**: 199–210.

Bolinger, D. (1972). *Intonation - Selected readings*, Penguin Books.

Bolinger, D. (1989). *Intonation and its uses*, Edward Arnold, London.

Boltz, M. (1992). Temporal expectancies and melody recognition, *Fourth International Workshop on Rhythm Perception and Production*, pp. 97–103.

Botte, M.-C., Canévet, G., Demany, L. et Sorin, C. (1989). *Psychoacoustique et perception auditive*, Les Éditions INSERM, Paris.

Bruce, G. (1977). *Swedish word accents in sentence perspective*, Gleerup, Lund.

Calbris, G. et Montredon, J. (1981). *Oh là là. Expression intonative et mimique*, CLE international.

Callamand, M. (1973). *L'intonation expressive*, Collection le français dans le monde, B.E.L.C. Librairies Hachette et Larousse.

Campbell, W. (1992a). *Multi-level Timing in Speech*, PhD thesis, University of Sussex.

Campbell, W. (1992b). Syllable-based segmental duration, *in* G. Bailly et C. Benoît (eds), *Talking Machines: Theories, Models and Designs*, Elsevier B.V., pp. 211–224.

Campbell, W. (1997). Synthesizing spontaneous speech, *in* Y. Sagisaka, N. Campbell et N. Higuchi (eds), *Computing prosody: Computational models for processing spontaneous speech*, Springer Verlag, pp. 165–186.

Campbell, W. et Beckman, M. (1997). Stress, Prominence and Spectral Tilt, *ESCA Workshop on Intonation: Theory, Models and Applications*, Athens, Greece, pp. 67–70.

Campbell, W. N. (1993). Automatic detection of prosodic boundaries in speech, *Speech Communication* **13**: 343–354.

Campione, E., Flachaire, E., Hirst, D. et Véronis, J. (1997). Stylisation and symbolic coding of F0: a quantitative model, *ESCA Workshop on Intonation: Theory, Models and Applications*, Athens, Greece, pp. 71–74.

Carlson, R., Granström, B. et Klatt, D. H. (1979). Some notes on the perception of temporal patterns in speech, *Proceedings of the IX^{th} International Conference on Phonetic Sciences*, Copenhague, Denmark.

Clément, J. et Gérard, C. (1996). Programmation de la production et anticipation de l'identification des formes prosodiques. Etude développementale, *Actes des XXIèmes Journées d'Etude sur la Parole*, Avignon, France, pp. 199–202.

Contini, M. et Profili, O. (1987). Génération automatique des schémas macroprosodiques en italien à partir du texte écrit, *Actes des XVIèmes Journées d'Etude sur la Parole*, Hammamet, Tunisie, pp. 245–248.

Cotton, S. et Grosjean, F. (1984). The gating paradigm : a comparison of successive and individual presentation formats, *Perception and Psychophysics* **35**(1): 41–48.

Crystal, D. (1969). *Prosodic System and Intonation in English*, Cambridge University Press, Cambridge, U.K.

Cutler, A. et Norris, D. (1991). Prosody in situations of communication: salience and segmentation, *Proceedings of the XIIth International Conference on Phonetic Sciences*, Vol. 1, Aix-en-Provence, France, pp. 264–270.

d'Alessandro, C. et Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception, *Computer Speech and Language* **9**: 257–288.

de Tournemire, S. (1997). Identification and automatic generation of prosodic contours for a text-to-speech synthesis system in french, *Proceedings of the European Conference on Speech Communication and Technology*, Vol. 1, Rhodes, Greece, pp. 191–194.

Delattre, P. C. (1966). Les dix intonations de base du français, *French Review* **40**(1): 1–14.

Delattre, P. C. (1969). L'intonation par les oppositions, *Le français dans le monde* **64**: 6–13.

Di Cristo, A. (1978). Méthodes et modèles d'analyse dans les recherches sur l'intonation, *Revue d'Acoustique* .

Di Cristo, A. (1985). De la microprosodie à l'intonosyntaxe, *Thèse de troisième cycle*, Université de Provence, Aix-en-Provence, France. Sous la direction de M. Rossi.

Dubois, J., Giacomo, M., Guespin, L., Marcellesi, C., Marcellesi, J.-B. et Mével, J.-P. (1994). *Dictionnaire de linguistique et des sciences du langage*, Larousse, Paris.

Ducrot, O. et Todorov, T. (1972). *Dictionnaire encyclopédique des sciences du langage*, Editions du Seuil, Paris.

Dupoux, E. (1989). Identification des mots parlés. Détection de phonèmes et unité prélexicale, *Thèse de troisième cycle*, Ecole des hautes études en sciences sociales, Paris, France. Sous la direction de J. Mehler.

EAGLES Spoken Language Systems DRAFT – WORK IN PROGRESS (1994). Eagles document eag-slwg-ir.2 edn.

Elman, J. L. (1988). Finding structure in time, *CRL Technical Report 8801*, University of California, San Diego, Center for Research in Language, La Jolla, CA.

Emerard, F. (1977). Synthèse par diphones et traitement de la prosodie, *Thèse de troisième cycle*, Université de Grenoble III, Grenoble, France.

Emerard, F. et Benoît, C. (1987). De la production à l'extraction, l'état d'un chantier, *Actes des XVIèmes Journées d'Etude sur la Parole*, pp. 224–228.

Fant, G. (1973). *Speech Sounds and Features*, The MIT Press, Cambridge, MA.

Fant, G. et Kruckenberg, A. (1996). On the quantal nature of speech timing, *Proceedings of the International Conference on Speech and Language Processing*, Vol. 3, Philadelphia, USA, pp. 2044–2047.

Fant, G., Kruckenberg, A. et Nord, L. (1991). Durational correlates of stress in Swedish, French and English, *Journal of Phonetics* **19**: 351–361.

Fletcher, H. et Munson, J. (1933). Loudness, its definition, measurement, and calculation, *Journal of the Acoustical Society of America* **5**: 1982.

Fónagy, I. (1980). L'accent en français : accent probabilitaire, *Studia Phonetica* **15**: 123–233.

Fónagy, I. (1987). Semantic diversity in intonation, *Proceedings of the XIth International Conference on Phonetic Sciences*, Vol. 2, Tailinn, Estonia, USSR, pp. 468–471.

Fónagy, I., Bérard, E. et Fónagy, J. (1984). Clichés mélodiques, *Folia Linguistica* **17**: 153–185.

Fougeron, C. et Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains, *Journal of the Acoustical Society of America* **101**(6): 3728–3740.

Fraisse, P. (1974). *La psychologie du rythme*, Presses Universitaires de France, Paris.

Fujisaki, H. et Sudo, H. (1971). A generative model for the prosody of connected speech in Japanese, *Annual Report of Engineering Research Institute* **30**: 75–80.

Garde, P. (1968). *L'accent*, Presses Univeristaires de France.

Gårding, E. (1991). Intonation parameters in production and perception, *Proceedings of the XIIth International Conference on Phonetic Sciences*, Vol. 1, Aix-en-Provence, France, pp. 300–304.

Gee, J. P. et Grosjean, F. (1981). Performance structures: A psycholinguistic and linguistic appraisal, *Cognitive Psychology* **15**: 411–458.

Gérard, C. et Auxiette, C. (1992). The processing of musical prosody by musical and nonmusical children, *Music Perception* **9**: 471–503.

Grépillat, T. (1996). Perçoit-on, par l'intonation, l'attitude du locuteur avant la fin de l'énoncé ?, *T.E.R. de maîtrise d'anglais mention industrie de la langue*, Université Stendhal, Grenoble, France.

Grønnum, N. (1992). *The ground-works of Danish intonation*, Museum Tusculanum Press - Univ. Copenhagen, Copenhagen, Denmark.

Grosjean, F. (1983). How long is the sentence ? Prediction and prosody in the on-line processing of language, *Linguistica* **21**: 501–529.

Gussenhoven, C. (1991). Tone segments in the intonation of Dutch, *in* T. Shannon et J. Snapper (eds), *The Berkley conference on Dutch linguistics*, University Press of America, Lanham, MD, pp. 139–155.

Halliday, M. (1970). *A course in spoken English: Intonation*, Oxford University Press, London.

Hirst, D. (1991). Intonation models: towards a third generation, *Proceedings of the XIIth International Conference on Phonetic Sciences*, Vol. 1, Aix-en-Provence, France, pp. 305–310.

Hirst, D. J. et Di Cristo, A. (A paraître). *Intonation systems: a survey of twenty languages*, Cambridge University Press, Cambridge, UK.

Hirst, D., Nicolas, P. et Espesser, R. (1991). Coding the F0 of a continuous text in French: an experimental approach, *Proceedings of the XIIth International Conference on Phonetic Sciences*, Vol. 5, Aix-en-Provence, France, pp. 234–237.

Hockett, C. F. (1942). A system of descriptive phonology, *Language* **18**: 3–21.

Honda, K., Hirai, H. et Dang, J. (1994). A physiological model of speech production and the implication of tongue-larynx interaction, *Proceedings of the International Conference on Speech and Language Processing*, Vol. 1, Yokohama, Japan, pp. 175–178.

Hwang, S.-H. et Chen, S.-H. (1995). A prosodic model of mandarin speech and its application to pitch level generation for text-to-speech, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 616–619.

Jensen, U., Moore, R. K., Dalsgaard, P. et Lindberg, B. (1994). Modelling intonation contours at the phrase level using continuous density hiddeen Markov models, *Computer Speech and Language* **8**: 247–260.

Jordan, M. I. (1988). Supervised learning and systems with excess degrees of freedom, *COINS Tech. Rep. 88-27*, University of Massachussetts, Computer and Information Sciences, Amherst - MA, USA.

Klatt, D. H. (1987). Review of text-to-speech conversion for English, *Journal of the Acoustical Society of America* **82**(3): 737–793.

Kohler, K. J. (1986). Invariability and variability in speech timing: from utterance to segment in German, *in* J. Perkell et D. H. Klatt (eds), *Invariance and Variability in Speech Processes*, Lawrence Erlbaum Associates, Hillsdale, NJ - USA, pp. 268–298.

Ladd, D. R. et Campbell, N. (1991). Theories of prosodic structure: evidence from syllable duration, *Proceedings of the XIIth International Conference on Phonetic Sciences*, Vol. 1, Aix-en-Provence, France, pp. 290–293.

Lame, G. D. (1990). P-centers and temporal distortion, *Journal of the Acoustical Society of America* **88, Suppl. 1**: S128. 120th Meeting: Acoustical Society of America.

Larkey, L. (1983). Reiterant speech: An acoustic and perceptual validation, *Journal of the Acoustical Society of America* **73**(4): 1337–1345.

Lass, N. J., Hendricks, C. A. et Iturriaga, M. A. (1980). The consistency of listener judgements in speaker height and weight identification, *Journal of Phonetics* **8**(4): 439–448.

Lehiste, I. (1977). Isochrony reconsidered, *Journal of Phonetics* **5**: 253–263.

Leinonen, L., Linnankoski, I., Laakso, M.-L. et Aulanko, R. (1991). Vocal communication between species : man and macaque, *Language and Communication* **11**: 241–262.

Léon, P. (1993). *Précis de phonostylistique : parole et expressivité*, Nathan Université.

Lhote, E., Filleau, M. et Grange, F. (1975). Reconnaissance de patrons intonatifs, *Actes des VIèmes Journées d'Etude sur la Parole*, Toulouse, France, pp. 29–37.

Liberman, M. et Streeter, L. (1978). Use of nonsense-syllable mimicry in the study of prosodic phenomena, *Journal of the Acoustical Society of America* **63**(1): 231–233.

Lickley, R. et Bard, E. (1993). Processing disfluent speech: Recognising disfluency before lexical access, *Proceedings of the European Conference on Speech Communication and Technology*, Vol. 3, Berlin, Allemagne, pp. 935–937.

Lickley, R., Shillcock, R. et Bard, E. (1991). Processing disfluent speech: How and when are disfluences found ?, *Proceedings of the European Conference on Speech Communication and Technology*, Vol. 3, Genève, Suisse, pp. 1499–1502.

Lieberman, P. (1965). On the acoustics basis of the perception of intonation by linguists, *Word* **21**: 40–54.

Linnankoski, I., Laakso, M.-L., Aulanko, R. et Leinonen, L. (1994). Recognition of emotion in macaque vocalizations by children and adults, *Language and Communication* **14**: 183–192.

Ljolje, A. et Fallside, F. (1986). Synthesis of natural sounding pitch contours in isolated utterances using Hidden Markov Models, *TrASSP* **34**: 1074–1080.

Loevenbruck, H. et Perrier, P. (1993). Vocalic reduction: prediction of acoustic and articulatory variabilities with invariant motor commands, *Proceedings of the European Conference on Speech Communication and Technology*, Vol. 1, Berlin, Germany, pp. 85–88.

Löfqvist, A. (1991). Proportional timing in speech motor control, *Journal of Phonetics* **19**: 343–350.

López-Gonzalo, E. et Hernández-Gómez, L. A. (1994). Data-driven joint F0 and duration modeling in text-to-speech conversion for Spanish, *IEEE International Conference on Acoustics, Speech, and Signal Processing* **1**: 589–592.

MacKay, D. G. (1987). Asymmetries in the relationship between speech perception and production, *in* H. Hener et A. Sanders (eds), *Perspectives on perception and action*, Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 301–333.

Marcus, S. M. (1975). Perceptual centres, *Unpublished fellowship dissertation*, King's College, Cambridge, UK.

Martin, P. (1980a). Sur les principes d'une théorie syntaxique de l'intonation, *Studia Phonetica* **17**: 91–101.

Martin, P. (1980b). Une théorie syntaxique de l'accentuation en Français, *Studia Phonetica* **15**: 1–12.

Martinet, A. (1991). *Eléments de linguistique générale*, 3^{e} edn, Armand Collin, Paris.

Mates, J. (1992). Timing and corrective mechanisms in sinchronization of motor acts with a sequence of stimuli, *in* C. Auxiette, C. Drake et C. Gérard (eds), *Fourth Rhythm Workshop: Rhythm Perception and Production*, Ville de Bourges, Bourges, France, pp. 43–48.

Mehler, J., Dupoux, E., Nazzi, T. et Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: the infant's viewpoint, *in* J. L. Morgan et K. Demuth (eds), *Bootstrapping from speech to grammar in early acquisition*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 101–116.

Morgan, J. L. et Demuth, K. (1996). *Signal to syntax: an overview*, Lawrence Erlbaum Associates, Mahwah, NJ, USA.

Morlec, Y. (1994). Apprentissage automatique de la prosodie pour la synthèse à partir du texte, *D.E.A. sciences cognitives*, Institut National Polytechnique de Grenoble, Grenoble, France.

Morlec, Y. (1996). Une approche globale pour l'apprentissage de l'intonation de discours, *IIe Colloque Jeunes Chercheurs en Sciences Cognitives*, Giens, France, pp. 346–349.

Morlec, Y., Aubergé, V. et Bailly, G. (1995). Evaluation of automatic generation of prosody with a superposition model, *Proceedings of the XIIIth International Conference on Phonetic Sciences*, Vol. 4, Stockholm, Sweden, pp. 224–227.

Morlec, Y., Bailly, G. et Aubergé, V. (1995). Synthesis and evaluation of intonation with a superposition model, *Proceedings of the European Conference on Speech Communication and Technology*, Vol. 3, Madrid, Spain, pp. 2043–2046.

Morlec, Y., Bailly, G. et Aubergé, V. (1996a). Generating intonation by superposing gestures, *Proceedings of the International Conference on Speech and Language Processing*, Vol. 1, Philadelphia, USA, pp. 283–286.

Morlec, Y., Bailly, G. et Aubergé, V. (1996b). Un modèle connexionniste modulaire pour l'apprentissage des gestes intonatifs, *Actes des XXIèmes Journées d'Etude sur la Parole*, Avignon, France, pp. 207–210.

Morlec, Y., Bailly, G. et Aubergé, V. (1997a). Apprentissage automatique d'un module de génération multistyle de l'intonation, *1ères JST FRANCIL*, Avignon, France, pp. 407–412.

Morlec, Y., Bailly, G. et Aubergé, V. (1997b). Generating the prosody of attitudes, *ESCA Workshop on Intonation: Theory, Models and Applications*, Athens, Greece, pp. 251–254.

Morlec, Y., Bailly, G. et Aubergé, V. (1997c). Synthesising attitudes with global rhythmic and intonation contours, *Proceedings of the European Conference on Speech Communication and Technology*, Vol. 1, Rhodes, Greece, pp. 219–222.

Moroni, V. (1997). Enquête sur quelques attitudes prosodiques du français : définition et interprétation, *T.E.R. sciences du langage*, Université Stendhal, Grenoble, France.

Moziconacci, S. (1995). Pitch variations and emotions in speech, *Proceedings of the XIIIth International Conference on Phonetic Sciences*, Vol. 1, Stockholm, Sweden, pp. 178–181.

Nishinuma, Y. et Santi, S. (1992). Comment percevons-nous la dernière syllabe d'une phrase ? Relations entre intensité et durée dans la discrimination de segments vocaliques, *Travaux de l'Institut de Phonétique d'Aix* **14**: 149–162.

Nooteboom, S. G. (1991). Some observations on the temporal organisation and rhythm of speech, *Proceedings of the XIIth International Conference on Phonetic Sciences*, Vol. 1, Aix-en-Provence, France, pp. 228–237.

O'Connor, J. et Arnold, G. (1973). *Intonation of Colloquial English*, second edition edn, Longman, London.

Ohala, J. J. (1996). Ethological theory and the expression of emotion in the voice, *Proceedings of the International Conference on Speech and Language Processing*, Vol. 3, Philadelphia, USA, pp. 1812–1815.

Öhman, S. E. G. (1967). Word and sentence intonation: a quantitative model, *Technical Report 2-3*, Speech Transmission Laboratory - Departement of Speech Communication and Music Acoustics - KTH, Stockholm - Sweden.

Oller, D. (1973). The effect of position in utterance on speech segment duration in english, *Journal of the Acoustical Society of America* **54**(5): 1235–1247.

O'Shaughnessy, D. (1981). A study of French vowel and consonant durations, *Journal of Phonetics* **9**: 385–406.

O'Shaughnessy, D. (1984). A multispeaker analysis of durations in read French paragraphs, *Journal of the Acoustical Society of America* **76**(6): 1664–1672.

Pasdeloup, V. (1990). Modèle de règles rythmiques du français appliqué à la synthèse de la parole, *Thèse de troisième cycle*, Université de Provence, Aix-en-Provence, France. Sous la direction de M. Rossi.

Pasdeloup, V. (1992). Durée inter-syllabique dans le groupe accentuel en français, *Actes des XIXèmes Journées d'Etude sur la Parole*, Bruxelles, Belgique, pp. 531–536.

Petitot, J. (1986). *Le « morphological turn » de la phénoménologie. Chapitre I, II et III de morphogénèse du sens II*, Centre d'analyse et de Mathématique Sociales, EHESS-CNRS.

Pierrehumbert, J. (1979). The perception of fundamental frequency declinaison, *Journal of the Acoustical Society of America* **66**: 363–369.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*, PhD thesis, MIT, Cambridge, MA.

Pierrehumbert, J. (1981). Synthetizing intonation, *Journal of the Acoustical Society of America* **70**(4): 985–995.

Pierrehumbert, J. et Beckman, M. (1988). *Japanese tone structure*, The MIT Press, Cambridge, MA.

Pike, K. (1945). *The intonation of American English*, University of Michigan Press, Ann Arbor.

Pompino-Marschall, B. (1989). On the psychoacoustic nature of the P-Center phenomenon, *Journal of Phonetics* **17**: 175–192.

Pompino-Marschall, B. (1992). The P-center and the perception of rhythm in connected speech, *in* C. Auxiette, C. Drake et C. Gérard (eds), *Fourth Rhythm Workshop: Rhythm Perception and Production*, Ville de Bourges, Bourges, France, pp. 157–162.

Quilis, A. (1979). Fonction linguistique de l'intonation, *Travaux de l'institut de phonétique de Strasbourg* **11**: 79–108.

Riedi, M. (1995). A neural-network-based model of segmental duration for speech synthesis, *Proceedings of the European Conference on Speech Communication and Technology*, Vol. 1, Madrid, Spain, pp. 599–602.

Rilliard, A., Aubergé, V., Bailly, G. et Morlec, Y. (1997). Vers une mesure de l'information linguistique véhiculée par la prosodie, *1ères JST FRANCIL*, Avignon, France, pp. 481–487.

Rossi, M. (1980). Le cadre accentuel et le mot en italien et en français, *Studia Phonetica* **17**: 9–22.

Rossi, M. (1985). L'intonation et l'organisation de l'énoncé, *Phonetica* **42**(2-3): 135–153.

Rossi, M. (1993). A model for predicting the prosody of spontaneous speech (PPSS model), *Speech Communication* **13**: 87–107.

Rossi, M. (1997). Intonation: Past, Present and Future, *ESCA Workshop on Intonation: Theory, Models and Applications*, Athens, Greece, pp. 1–10.

Rossi, M., Di Cristo, A., Hirst, D., Martin, P. et Nishinuma, Y. (1981). *L'intonation : de l'acoustique à la sémantique*, Klincksieck, Paris.

Sabio, F. (1996). Description prosodique et syntaxique du discours en français : données et hypothèses, *Thèse de troisième cycle*, Université de Provence, Aix-en-Provence.

Sagisaka, Y. (1990). On the prediction of global F0 shapes for Japanese text-to-speech, *IEEE International Conference on Acoustics, Speech, and Signal Processing* **1**: 325–328.

Scherer, K. R. (1996). Adding the affective dimension: a new look in speech analysis and synthesis, *Proceedings of the International Conference on Speech and Language Processing*, Philadelphia, USA.

Schmuckler, M. (1989). Expectation in music: investigation of melodic and harmonic processes, *Music Perception* **7**: 109–150.

Scordilis, M. et Gowdy, J. (1989). Neural Network based generation of Fundamental Frequency contours, *IEEE International Conference on Acoustics, Speech, and Signal Processing* **1**: 219–222.

Scott, S. K. (1993). *Perceptual Centres in Speech - An Acoustic Analysis*, PhD thesis, University College, London, England.

Searle, J. R. (1969 - réédité en 1985). *Du cerveau au savoir*, Hermann, Paris.

Selkirk, E. (1984). *Phonology and syntax*, The MIT Press, Cambridge, MA.

Silverman, K. (1987). *The structure and processing of fundamental frequency contours*, PhD thesis, Cambridge University, UK.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. et Hirschberg, J. (1992). TOBI: a standard for labeling english prosody, *Proceedings of the International Conference on Speech and Language Processing* **2**: 867–870.

Sluijter, A. M. C. et van Santen, J. P. H. (1996). Spectral balance as an acoustic correlate of linguistic stress, *Journal of the Acoustical Society of America* **100**(4): 2471–2485.

Smolensky, P. (1992). IA connexionniste, IA symbolique et cerveau, *in* D. Andler (ed.), *Introduction aux sciences cognitives*, Editions Gallimard, pp. 77–106.

Sock, R. et Löfqvist, A. (1995). Some timing constraints in the production of bilabial stops, *Journal of Phonetics* **23**: 129–138.

Sorin, C. (1989). Perception de la parole continue, *in* M. Botte, G. Canevet, l. Demany et C. Sorin (eds), *Psychoacoustique et perception auditive*, audition, INSERM/SFA/CNET, pp. 123–139.

Strömqvist, S., Peters, A. et Ragnarsdóttir, H. (1995). Particles and prepositions in Scandinavian child language development : effects of prosodic spotlight ?, *Proceedings of the XIIIth International Conference on Phonetic Sciences*, Vol. 4, Stockholm, Sweden, pp. 38–45.

't Hart, J. et Collier, R. (1973). Intonation by rule: a perceptual quest, *Journal of Phonetics* **1**: 309–327.

't Hart, J., Collier, R. et Cohen, A. (1990). *A perceptual study of intonation*, Cambridge University Press, Cambridge, UK.

Thorsen, N. G. (1980). A study of the perception of sentence intonation – Evidence from Danish, *Journal of the Acoustical Society of America* **67**(3): 1014–1030.

Thorsen, N. G. (1983). Standard Danish sentence intonation – Phonetic data and their representation, *Folia Linguistica* **17**: 187–220.

Todd, P. (1989). A connectionist approach to algorithmic composition, *Music Journal* **13**(4): 27–43.

Traber, C. (1992). F0 generation with a database of natural F0 patterns and with a neural network, *in* G. Bailly et C. Benoît (eds), *Talking Machines: Theories, Models, and Designs*, Elsevier Science Publishers B.V., pp. 287–304.

Vaissière, J. (1971). Contribution à la synthèse par règles du français, *Thèse de troisième cycle*, Grenoble III, Grenoble, France.

Vaissière, J. (1980). La structuration acoustique de la phrase française, *Annali della Scuola Normale Superiore di Pisa X*, Vol. 2, pp. 529–560.

Vaissière, J. (1988). The use of prosodic parameters in automatic speech recognition, *in* H. Niemann, M. Lang et G. Sagerer (eds), *Recent Advances in Speech Understanding and Dialog Systems*, Springer-Verlag, Heidelberg, Germany, pp. 71–99.

van Bezooyen, R. (1984). *Characteristics and recognizability of vocal expressions of emotion*, Foris, Dordrecht.

van Heuven, V. J., Haan, J., Janse, E. et van der Torre, E. J. (1997). Perceptual identification of sentence type and the time-distribution of prosodic interrogativity marker in Dutch, *ESCA Workshop on Intonation: Theory, Models and Applications*, Athens, Greece, pp. 317–320.

van Santen, J. P. H. (1992). Deriving text-to-speech durations from natural speech, *in* G. Bailly et C. Benoît (eds), *Talking Machines: Theories, Models, and Designs*, Elsevier Science Publishers B.V., pp. 275–285.

van Santen, J. P. H. (1997). Segmental duration and speech timing, *in* Y. Sagisaka, N. Campbell et N. Higuchi (eds), *Computing prosody: Computational models for processing spontaneous speech*, Springer Verlag, pp. 225–249.

Vigouroux, N. et Caelen, J. (1985). Segmentation en vue de l'organisation d'une base de données acoustique et phonétique, *Actes des XIVèmes Journées d'Etude sur la Parole*, Paris, France, pp. 152–155.

Wenk, B. J. et Wioland, F. (1982). Is French really syllable-timed ?, *Journal of Phonetics* **10**: 193–216.

Whalen, D. H. et Levitt, A. G. (1995). The universality of intrinsic F0 of vowels, *Journal of Phonetics* **23**: 349–366.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M. et Price, P. (1992). Segmental durations in the vicinity of prosodic boundaries, *Journal of the Acoustical Society of America* **91**(3): 1707–1717.

Wightman, C. W. et Talkin, D. T. (1997). The Aligner: Text-to-Speech Alignment Using Markov Models, *in* J. P. H. van Santen, R. W. Sproat, J. P. Olive et J. Hirschberg (eds), *Progress in Speech Synthesis*, Springer Verlag, New York, pp. 313–323.

Annexe A

Corpus de parole naturelle

A.1 Corpus *Attitudes*

Une syllabe

1 Oui
2 Non
3 Deux
4 Jean
5 Bof

Deux syllabes

1 Bonjour
2 Quentin
3 102
4 Les gars
5 J'ai bu
6 l'eau bout

Trois syllabes

1 Notamment
2 Découper
3 Mégabits
4 Baptiser
5 Natacha
6 602
7 Les beaux gars
8 J'agitais
9 J'ai chanté
10 j'avais bu
11 J'ai tout bu
12 J'ai peu bu

13 j'ai bu peu
14 C'est joli
15 C'est très beau
16 C'était beau
17 Jean comptait
18 tu chantais
19 Roger ment
20 Les pies chantent
21 L'eau bout peu
22 J'ai du pain
23 J'ai deux fous
24 J'aidais Jean
25 J'en donnais
26 J'en ai bu
27 Jean fait ça
28 Tu les sens

Quatre syllabes

1 Suffisamment
2 Déboutonner
3 Consolider
4 Convalescent
5 Roman-Photo
6 Monophasé
7 Electrochoc
8 Quasimodo
9 217
10 Les beaux gamins
11 Les gentils nains
12 J'abandonnais
13 J'ai débuté
14 j'avais chanté
15 J'adorais lire
16 J'ai beaucoup bu
17 j'ai bu beaucoup
18 j'ai bien chanté
19 C'est décadent
20 C'est très joli
21 C'est vraiment beau
22 C'était joli
23 C'était très beau
24 Jean papotait
25 Tu débutais
26 Papa chantait
27 Chacun comptait
28 Des gars chantaient
29 Benjamin lit
30 Ton gamin lit
31 Certains gars chantent
32 Les bons gars chantent
33 Les chats blancs jouent
34 Souvent l'eau bout
35 L'eau souvent bout
36 L'eau bout souvent
37 Tu n'as pas bu
38 Tu ne lis pas
39 Jean chantait peu
40 J'ai tes bâtons
41 J'ai certains bouts
42 J'ai des beaux chats
43 J'ai des chats blancs
44 J'avais du pain
45 J'épatais Jean

46 J'en ai donné
47 J'en avais bu
48 Jean vit Quentin
49 Tu joues ton fou
50 Tu les lui rend
51 Tu leur en fais
52 Jean finit l'eau
53 Tu les donnais
54 Tu les as bu
55 Sophie joue l'as
56 Quentin les joue
57 Ce gars vit ça
58 J'en ai vu deux

Cinq syllabes

1 Momentanément
2 Antidérapant
3 Monopoliser
4 Décapuchonner
5 Electrochimie
6 Safari-photo
7 252
8 Les beaux galopins
9 Les gentils gamins
10 Ces fabuleux nains
11 J'académisais
12 J'ai débarrassé
13 J'avais papoté
14 J'adorais chanter
15 J'ai beaucoup chanté
16 J'ai peu décidé
17 J'ai chanté beaucoup
18 J'avais beaucoup lu
19 J'avais lu beaucoup
20 C'est sécurisant
21 C'est très décadent
22 C'est vraiment joli
23 C'est forcément beau
24 C'était décadent
25 C'était très joli
26 Jean débarrassait
27 Quentin papotait
28 Des gars papotaient
29 Benjamin chantait
30 Des gamins chantaient
31 Quelques-uns chantaient
32 Les bons gars chantaient
33 Les chats blancs courraient
34 Des galopins jouent
35 Certains requins tuent
36 Les beaux gamins jouent
37 Deux jolis gars chantent
38 Deux chats marrons jouent
39 Deux beaux chatons jouent
40 Deux chatons blancs jouent
41 Certains beaux gars chantent
42 Certains chats blancs jouent
43 Souvent Jean tombait
44 Jean souvent tombait
45 Jean tombait souvent
46 Souvent les gars jouent
47 Les gars souvent jouent
48 Les gars jouent souvent

49 Posément Jean lit
50 Jean posément lit
51 Jean lit posément
52 J'en joue posément
53 Jean n'a pas chanté
54 Je n'ai jamais bu
55 Quentin n'a pas bu
56 Tu ne chantais pas
57 Quentin ne lit pas
58 Jean décidait peu
59 J'ai tes pantalons
60 J'ai certains chatons
61 J'ai des longs cheveux
62 J'ai des cheveux longs
63 J'ai deux jolis chats
64 J'ai deux chats marrons
65 J'ai certains beaux chats
66 J'ai certains chats blancs
67 J'aimais Rébécca
68 J'aimais tes cheveux
69 J'aimais certains jeux
70 J'aimais tes beaux chats
71 J'aimais tes chats blancs
72 J'adorais papa
73 J'adorais ces jeux
74 J'abominais tout
75 J'en ai vu passer
76 Jean vit Benjamin
77 Jean fait des gamins
78 Jean connaît Quentin
79 Tu connais ces gars
80 Tu leur en faisais
81 Tu leur en as pris
82 Tu défendais ça
83 Tu les baptisais
84 Tu les as donnés
85 Tu les avais bu
86 Ces gars jouent ton jeu
87 Quentin vit les gars
88 Quentin finit l'eau
89 Quentin les finit
90 Des gars visaient Jean
91 Les gamins jouent l'as
92 Les gamins les jouent
93 J'en ai vu 500

Six syllabes

1 Sentimentaliser
2 Décapitaliser
3 Photo-luminescent
4 Electrothérapie
5 Végétominéral
6 Radiculodentaire
7 Considérablement
8 577
9 Les beaux sénégalais
10 Les gentils boulangers
11 Ces fabuleux gamins
12 J'ai comptabilisé
13 J'avais débarrassé
14 J'espérais débuter
15 J'ai beaucoup potassé
16 J'ai peu débarrassé

17 J'ai potassé beaucoup
18 J'avais beaucoup chanté
19 J'avais chanté beaucoup
20 J'espérais beaucoup lire
21 J'espérais lire beaucoup
22 C'est tarabiscoté
23 C'est très sécurisant
24 C'est vraiment décadent
25 C'est fréquemment sympa
26 C'est suffisamment beau
27 C'était sécurisant
28 C'était très décadent
29 C'était vraiment joli
30 C'était fréquemment bon
31 Jean monopolisait
32 Papa débarrassait
33 Des gars débarrassaient
34 Benjamin papotait
35 Des gamins papotaient
36 Certains gars papotaient
37 Des bons gars papotaient
38 Les chats blancs divaguaient
39 Des galopins chantaient
40 Certains savants composent
41 Deux jolis chats dansaient
42 Deux chats marrons dansaient
43 Deux chatons blancs dansaient
44 Deux beaux chatons dansaient
45 Certains bons gars chantaient
46 Certains chats blancs dansaient
47 Les napolitains jouent
48 Certains galopins jouent
49 Ces bons galopins jouent
50 Les chimpanzés blancs jouent
51 Ces gentils gamins jouent
52 Deux chatons marrons jouent
53 Certains beaux chatons jouent
54 Certains chatons blancs jouent
55 Certains mauvais chats tuent
56 Certains chat marrons jouent
57 Souvent Jean papotait
58 Jean souvent papotait
59 Jean papotait souvent
60 Souvent les gars chantaient
61 Les gars souvent chantaient
62 Les gars chantait souvent
63 Souvent les gamins jouent
64 Les gamins souvent jouent
65 Les gamins jouent souvent
66 Posément Jean comptait
67 Jean posément comptait
68 Jean comptait posément
69 Résolument Jean scie
70 Jean résolument scie
71 Jean scie résolument
72 Tu n'as pas débuté
73 Les gens n'ont pas chanté
74 Les gamins n'ont pas bu
75 Tu n'a jamais chanté
76 Les gens n'ont jamais bu
77 Tu ne débutait pas
78 Tu ne chantais jamais
79 Quentin ne chantait pas
80 Benjamin ne lit pas
81 Jean vagabondait peu
82 J'ai des subordonnés
83 J'ai certains pantalons
84 J'ai des beaux pantalons

85 J'ai des pantalons blancs
86 J'ai des jolis chatons
87 J'ai des chatons marrons
88 J'ai deux fabuleux chats
89 j'ai des chats fabuleux
90 J'aimais tes pantalons
91 J'aimais certains chatons
92 J'aimais tes longs cheveux
93 J'aimais tes cheveux longs
94 J'aimais ces jolis chats
95 J'aimais ces chats marrons
96 J'aimais certains beaux chats
97 J'aimais certains chats blancs
98 J'adorais Rébécca
99 J'adorais tes cheveux
100 J'adorais certains jeux
101 J'adorais tes beaux chats
102 J'adorais tes chats blancs
103 J'abominais ceci
104 J'abominais ces jeux
105 J'académisais tout
106 Jean vit les galopins
107 Jean laissait Benjamin
108 Jean laissait ses gamins
109 Jean débauchait Quentin
110 Jean débauchait les gars
111 Tu leur en as compté
112 Tu leur en avais fait
113 L'eau contaminait tout
114 Tu les débarrassais
115 Les gars scient des rondins
116 Les gars boudaient Quentin
117 Les gars débauchaient Jean
118 Les gamins scient du bois
119 Les gamins visaient Jean
120 Les galopins jouent l'as
121 Les galopins les scient
122 J'en ai vu 502

Sept syllabes

1 Communicabilité
2 Les gentils sénégalais
3 Les galopins papotaient
4 Tu leur en avais donné
5 les gars coupaient des rondins

Huit syllabes

1 Cytopathogénicité
2 Les fabuleux sénégalais
3 Les galopins débarrassaient
4 Tu leur en avais proposé
5 Les gamins coupaient des rondins

A.2 Corpus *Modulation*

Déclarations de 7 et 8 syllabes

7 syllabes

Les gars monopolisaient

Les gamins débarrassaient

Les galopins papotaient

Les napolitains chantaient

La tabacomanie tue

8 syllabes

Ce gars sentimentalisait

Les dandys monopolisaient

Les galopins débarrassaient

Les napolitains papotaient

La Méditerranée baissait

Les gars pouvaient fanfaronner

Ce gars peut monopoliser

GN sujet de 6 syllabes

Des mathématiciens téléphonaient

Les beaux napolitains magazinaient

Les napolitains saouls monologuaient

Les galopins rusés monologuaient

Ces maudits galopins téléphonaient

Les gamins boutonneux téléphonaient

Les savoureux gâteaux ravigotaient

Le lait vitaminé ravigotait

Les mathématiciens savaient dériver

Le lait vitaminé pouvait macérer

Allongement du Verbe+GNobjet

Verbe+GNobjet de 5 syllabes

Les gamins jouent les mécanos

Ses parents font des beaux gâteaux

Roger fait des gâteaux roux

Les gamins font des jolis coups

Les gamins font des pas sautés

Jean mangeait des gâteaux

Les gamins mangeaient du bon pain

Les cousins mangeaient du pain chaud

Le miséreux labourait son champ

Verbe+GNobjet de 6 syllabes

Les bambins jouent les napolitains

Les gamins jouent les bons mécanos

Roger fait du chocolat chaud

Les gamins font des jolis gâteaux

Les gamins font des gâteaux légers

Les laborantins font des coups délirants

Les gamins mangeaient du chocolat

Les gamins mangeaient des bons gâteaux

Les nécessiteux mangeaient du boudin chaud.

Les gamins mangeaient des petits pains

Les lapins mangeaient du pain rassi

Les gamins dévoraient les gâteaux

Les gamins dévoraient le bon pain

Les gens dévoraient des pains chauds

Les gamins cataloguaient les mots

Verbe+GNobjet de 7 syllabes

Les gamins jouent la solidarité

Les latins sont des bons karatékas

Les figurants jouent les napolitains saouls

La maman fait du chocolat gouteux

Les parents font du fameux chocolat

Roger fait des gâteaux savoureux

Les bouffons jouent des coups désopilants

Les japonais mangeaient du Tiramissu

Les gamins buvaient du bon chocolat

Les chatons lapaient du chocolat chaud

Jean mangeait les jolis gâteaux

Les cochons mangeaient des gâteaux rassis

Les gamins coupaient du pain farineux

Les japonais dévoraient le chocolat

Les gamins dévoraient les bons gâteaux

Les gamins dévoraient les gâteaux chauds

Ce joli baudet recevait des mauvais coups

Les lapins dévoraient du pain rassi

Les gamins débarrassaient les copains

Les gamins sollicitaient les bons mots

Les nécessiteux sollicitaient du pain chaud

Les gamins comptabilisaient les sous

Verbe+GNobjet de plus de 7 syllabes

Les colons comptabilisaient les canots

Les colons naturalisaient les miséreux

Ses méfaits pourraient faciliter son boulot

Son candidat pourrait paralyser ses capitaux

Son bagou pourrait faciliter la communauté

A.3 Corpus *Identification & Gating*

Une syllabe

1 L'eau

Deux syllabes

1 J'attends

2 L'eau bout

3 L'auto

Trois syllabes

1 J'agitais

2 L'eau chantait

3 L'auto roule

4 L'attaquant

Quatre syllabes

1 J'abandonnais
2 L'eau clapotait
3 L'auto roulait
4 L'attaquant joue
5 L'américain

Cinq syllabes

1 J'académisais
2 L'eau caracolait
3 L'auto dérapait
4 L'attaquant chantait
5 L'américain joue
6 L'académicien

A.4 Corpus *Evaluation des contours portés*

Six syllabes

1 Sentimentaliser (monomot)
2 J'adorais / tes cheveux (GV+GNobjet)
3 Des gamins / papotaient (GNsujet+verbe)
4 Les gamins / visaient Jean (GNsujet+(verbe+GNobjet))
5 Les gars scient / des rondins ((GNsujet+verbe)+GNobjet)

Huit syllabes

1 Cytopathogénicité (monomot)
2 Les dandys / monopolisaient (GNsujet+verbe)
3 Les gamins / jouent les mécanos (GNsujet+(verbe+GNobjet))
4 Roger fait / du chocolat chaud ((GNsujet+verbe)+GNobjet)

Annexe B

Le parcours du combattant en dix épreuves d'un corpus de laboratoire

Cette annexe résume la démarche que nous avons suivie depuis la conception des corpus jusqu'à la modélisation des attitudes prosodiques, prédictions des contours de *phrase* et de *groupe* inclus.

Elle représente, pour nous, le fil conducteur de nos travaux de recherche et sera je l'espère, pour nos lecteurs, le fil d'Arianne qui leur permettra, par cette description plus synthétique, de naviguer sans encombre dans les détails de méthodologie, d'implémentation et de validation présentés dans ce mémoire.

I/ Enregistrements

Nous avons réalisé les enregistrements d'un locuteur masculin en chambre sourde. Le contenu énonciatif et les situations d'élocution lui sont imposés.

Corpus "attitudes" (2.2) : Sur la base d'un corpus de 322 phrases de 1 à 8 syllabes non typées, le locuteur reçoit 6 consignes attitudinales : *déclaration* (DC), *question simple* (QS), *exclamation de surprise* (EX), *doute-incrédulité* (DI), *ironie de soupçon* (SC) et *évidence* (EV). L'enregistrement des 1932 phrases s'étale sur 33 sessions de 60 phrases. Les phrases sont enregistrées à 10 kHz, fréquence d'échantillonage suffisante pour l'étude des paramètres prosodiques retenus dans ce mémoire.

Corpus "modulation" (2.3) : Le même locuteur est enregistré dans les mêmes conditions sur un corpus de 73 phrases destiné à compléter la panoplie des phrases déclaratives (DC) du corpus "attitudes".

Au total, 2005 énoncés du locuteur de référence, correspondant à 1h10 de parole environ, sont collectés.

II/ Tests de perception

Le corpus "identification & gating" est soumis à deux tests de perception :

- Un **test d'identification** (2.2.2) destiné, en premier lieu, à contrôler la capacité du locuteur à transmettre les six attitudes représentées. Il vise également à évaluer si les six attitudes choisies font partie du code conventionnel du français oral. Enfin, il nous permet d'obtenir une première mesure, complétée par une enquête socio-linguistique, de la pertinence des définitions proposées durant le test.
- Un **test de dévoilement progressif** (2.2.4) mis en place pour démontrer, en révélant les capacités d'anticipation des auditeurs, que l'information attitudinale est répartie sur l'ensemble des contours prosodiques au niveau phrastique.

III/ Traitement du signal et stylisation

Des outils de traitement semi-automatique sont utilisés pour segmenter la séquence des réalisations phonémiques de chaque phrase (aligneur programmé sur HTK) et détecter la fréquence fondamentale par insertion de marqueurs de périodes (un exemple est donné en figure B.1). Un progiciel de visualisation et d'édition de signal (ISIS) est alors utilisé pour corriger manuellement les étiquettes ainsi obtenues.

La stylisation de la F0 et l'extraction des patrons rythmiques sont ensuite effectués. A partir de l'étiquetage réalisé, les GIPC (1.1.1.3) de chaque phrase sont caractérisés par 4 paramètres (2.2.3.1) :

- Les réalisations mélodiques sont caractérisées par trois valeurs de F0 (exprimées en quart de tons) par GIPC .
- Les patrons rythmiques sont extraits à l'aide d'un algorithme fournissant pour chaque GIPC un facteur d'allongement k_{GIPC} correspondant à sa durée normalisée (1.3.2.2).

IV/ Le fichier prosodie

A partir de l'étiquetage et de la stylisation réalisés, on extrait un fichier de référence par phrase appelé **fichier prosodie**. Ce fichier constitue le point de passage obligé de toutes nos analyses et de toutes nos modélisations. Il contient les informations suivantes :

- En entête, la fréquence laryngienne moyenne du locuteur.

- Les trois valeurs de F0 par phonème voisé exprimées en dixième de quart de tons.
- L'énergie.
- Le facteur d'allongement k_{GIPC}.
- Le temps.
- La durée phonémique.
- La nature du phonème.
- La syllabation.
- La catégorie lexicale.
- l'orthographe.
- Les différents niveaux linguistiques précisant la modalité de la phrase, la nature du syntagme et les frontières syntagmatiques - **DG** : dépendance gauche, **DD** : dépendance droite, **IT** : interdépendance et **ID** : indépendance (Bailly 1989) - susceptibles d'être marquées prosodiquement (*rendez-vous*).

Voici un exemple de fichier prosodie pour la phrase déclarative « Des gars débarrassaient. »[1] :

F0_Ref= 126 Hz

F1	F2	F3	ENR	KGIPC	GIPC	TPS	DUREE	PHON	SYLL	LEX	ORTHOGRAPHE	PHRASE	PROP	GROUPE
*	*	*	17	*	*	0	90	_	*	*	*	*	*	*
-59	-97	-54	44	*	*	90	67	d	Syl	Det	DES	PHR	PRNC	GN
-25	-25	-24	69	0.102	eg	157	82	e	*	*	*	*	*	*
-26	-47	22	48	0.102	eg	239	98	g	Syl	Nom	GARS	*	*	*
16	94	116	73	0.173	ad	338	131	a	*	*	*	*	*	*
105	54	96	51	0.173	ad	470	67	d	Syl	Vrb	DEBARRASSAIENT	*	*	GV:DG
83	92	74	70	-0.212	eb	537	65	e	*	*	*	*	*	*
74	56	87	53	-0.212	eb	602	65	b	Syl	*	*	*	*	*
79	74	63	74	-0.198	ar	667	83	a	*	*	*	*	*	*
53	47	41	71	-0.198	ar	751	51	r	Syl	*	*	*	*	*
35	13	-34	71	0.0312	as	802	81	a	*	*	*	*	*	*
*	*	*	43	0.0312	as	883	154	s	Syl	*	*	*	*	*
-120	-146	-144	62	0.294	e ^	1037	108	e^	*	*	*	*	*	*
*	*	*	17	*	*	1146	229	_	Syl	*	*	:DC	:FF	:FF

V/ Analyse acoustique

A partir des fichiers prosodie, nous réalisons une analyse acoustique des contours prosodiques du niveau *phrase* (2.2.3) et du niveau *groupe* (2.3.2). Cette analyse met en évidence l'expansion des mouvements prosodiques de ces deux niveaux et nous amène à proposer une paramétrisation minimale des entrées de notre modèle sous la forme de deux rampes linéaires décroissantes permettant de générer les deux phases du mouvement : *préparation* et *capture*.

[1]Les niveaux en dessous du groupe n'ont pas été représentés.

VI/ Apprentissage du réseau de neurones récurrent (RNR) du niveau phrase

L'apprentissage hiérarchique de notre modèle débute par le module de phrase (3.2.1.1). Pour chaque attitude, le corpus d'apprentissage est constitué des phrases monomots du corpus "attitudes". Le RNR est un modèle commun aux six attitudes. Seuls changent, en fonction de l'attitude générée, ses activations initiales et les pondérations obtenues après apprentissage. Ce RNR est simulé par le logiciel de réseaux de neurones XRBP développé à l'ICP. Il reçoit comme *patterns* d'apprentissage un fichier semblable à celui qui est décrit un peu plus loin.

VII/ Evaluation objective et perceptive des prédictions du RNR de phrase

Une fois que le module de phrase a été entraîné sur les six attitudes, une première analyse est réalisée en comparant les contours prosodiques prédits avec les contours d'apprentissage. Nous testons aussi la capacité du modèle à réaliser l'expansion prosodique de phrase (3.2.2). Une validation perceptive est également entreprise (3.2.3) pour montrer que ces contours synthétiques de phrase permettent aux auditeurs d'accéder au code linguistique sous-jacent aussi efficacement qu'avec des contours naturels. Pour tous les tests de perception, les stimuli synthétiques sont obtenus par analyse-synthèse PSOLA en appliquant, à partir d'un *fichier prosodie de prédiction*, la prosodie générée par le modèle sur le signal d'origine.

VIII/ Apprentissage du modèle *phrase+groupe*

A partir du moment où les analyses acoustiques et perceptives des prédictions du module de phrase indiquent des résultats satisfaisants, l'apprentissage hiérarchique peut se poursuivre. Le module de phrase est "gelé" (ses poids sont fixés aux valeurs obtenues pour les "déclarations") et l'apprentissage du module de groupe relatif aux *déclarations* (module de *référence*) peut commencer. La première étape consiste à regrouper l'ensemble des phrases de structures syntaxiques [GNsujet+verbe], [GNsujet+verbe+GNobjet] et [GV+GNobjet] des corpus "attitudes" et "modulation" dans un fichier de *patterns* d'apprentissage. Un *pattern* correspond à une phrase et contient les informations suivantes (3.3.1) :

En entrée Deux rampes de phrase, trois rampes de groupe et le codage sur 4 bits de la marque prosodique qui borne à droite le GIPC courant. La marque prosodique entre un sujet et un verbe (GV:DG) reçoit la séquence 0010, celle associée à une jonction *verbe/objet* (GN:DD) 0001. Ajoutons que la séquence de fin de phrase correspond à 0000.

En sortie Pour chaque GIPC de la phrase : 3 valeurs de F0 exprimées en $1/4 tons$ et le coefficient d'allongement k_{GIPC} multiplié par 10.

Par exemple, le *pattern* pour la phrase « Les gars boudaient Quentin » de structure [sujet(2) + verbe(2) + objet(2)] correspond aux séquences d'entrée et de sortie suivantes :

Sequence ENTREE

units	phrase_abs	phrase_relative	groupe_abs	groupe_rel	groupe_suivant	bin1	bin2	bin3	bin4
	6.0	10.0	2.0	10.0	4.0	0.0	0.0	1.0	0.0
	5.0	8.2	1.0	1.0	3.0	0.0	0.0	1.0	0.0
	4.0	6.4	2.0	10.0	4.0	0.0	0.0	0.0	1.0
	3.0	4.6	1.0	1.0	3.0	0.0	0.0	0.0	1.0
	2.0	2.8	2.0	10.0	2.0	0.0	0.0	0.0	0.0
	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0

Sequence SORTIE

units	f01	f02	f03	kgipc
	-3.9	-2.1	-4.1	0.873
	2.6	8.0	7.7	1.714
	2.3	3.2	1.9	-2.188
	5.5	6.8	5.2	0.839
	1.3	-0.3	-5.7	-0.311
	-12.3	-15.9	-17.2	0.924

IX/ Evaluation objective et perceptive des prédictions du modèle *phrase+groupe*

Comme pour le module de groupe, nous réalisons une comparaison entre les prédictions des contours prosodiques de groupe et le corpus d'apprentissage (3.3.2). L'expansion des mouvements de groupe est également étudiée.

Cette phase est accompagnée d'un test de perception (3.3.3) qui montre que la prosodie synthétique est bien formée et remplit sa fonction de structuration de l'énoncé dans la même proportion que la prosodie naturelle.

X/ Les contraintes de l'attitude sur la modulation de groupe

Maintenant que le module de groupe de *référence* est correctement entraîné sur les déclaratives, nous figeons ses poids pour débuter la dernière phase de l'apprentissage du modèle *phrase+groupe* (3.4.2.2). Les poids du module de phrase restent figés eux aussi.

Pour cette dernière étape, nous introduisons un réseau de neurones supplémentaire chargé de gérer dynamiquement pour chaque attitude la contribution des contours de groupe sur la réalisation prosodique finale. Ce réseau est commandé par les deux rampes de phrase et son apprentissage est réalisé à partir de corpus associés aux 5 attitudes restantes et contenant les trois structures testées : [GNsujet+verbe], [GNsujet+verbe+GNobjet] et [GV+GNobjet].

L'influence des contours porteurs de phrase sur les contours portés de groupe est alors modélisée et l'apprentissage des sous-niveaux peut débuter mais, pour cette thèse, l'histoire s'arrête là...

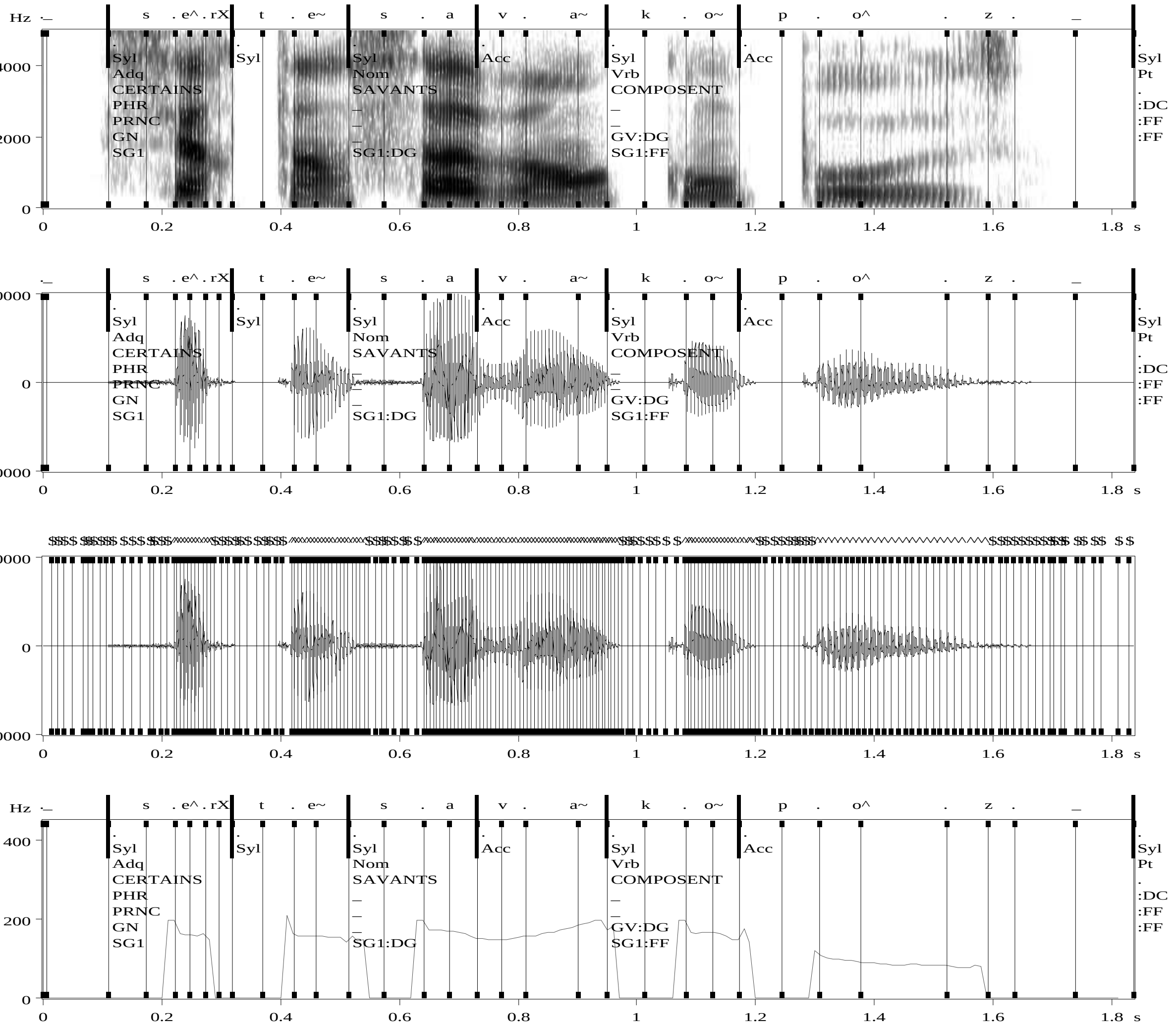


FIG. B.1: Exemple d'étiquetage pour la phrase « Certains savants composent. ». De haut en bas : le spectrogramme, le signal étiqueté, les étiquettes de voisement, le contour de F0. L'étiquetage des phonèmes est complété par des informations sur le lexique et les niveaux linguistiques : l'attitude, la nature des unités syntagmatiques et les relations entretenues par les unités adjacentes.

Table des figures

Liste des tableaux