ACADÉMIE D'AIX-MARSEILLE

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

THÈSE

Présentée à l'Université d'Avignon et des Pays de Vaucluse
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : Informatique

Traitement de la prosodie en reconnaissance automatique de la parole

par

Philippe Langlais

Soutenue le 11 octobre 1995 devant le Jury composé de :

MM	Gérard Chollet	Examinateur
	René Collier	Examinateur
	Albert Di Cristo	Rapporteur
	Marc El-Bèze	Président
	Joseph-Jean Mariani	Rapporteur
	Henri Méloni	Directeur de Thèse

À mes parents
À Patricia

ABSTRACT

In this study, we investigate the treatment of prosody in automatic speech recognition systems. The major steps of a classical prosodic approach (parametrization, microprosodic and perceptive corrections, application of suprasegmental rules) are discussed, thus introducing the choices made for each of them.

In the first part of this dissertation, the segmental variations are analyzed in detail. A selection of the phenomena which have been most investigated in the past is proposed. Each is then studied on isolated words corpora for two specific purposes : firstly, to evaluate if the automatic parameters extraction techniques used here allow the use of microvariations to improve an acoustico-phonetic decoding process ; secondly, to measure the pertinence of automatic microprosodic corrections. The study shows that only few phenomena can be significantly observed by means of these techniques. The reliable phenomena have been successfully integrated in a lexical access system.

The second part presents the major difficulties of prosodic analysis carried out by experts and attempts to explain why statistical methods are more widely applied. An automatic correlative system has been elaborated ; firstly, it upholds an assistance to prosodic analysis (providing vizualization and query tools), and secondly, it gives a predictive function of the linguistical structure of the message to decode. Two applications of this system are proposed ; a first one for the recognition of decimal numbers (our system is able to locate the word “virgule”in an unknown number, only by means of prosodic information) and a second one for the recognition of read isolated sentences. The results obtained fully validate the approach we proposed.

Pour l'intérêt qu'ils ont spontanément témoigné à mes activités de recherche, je tiens à exprimer mes remerciements et ma sincère reconnaissance aux membres du jury :

Henri Méloni, professeur à l'Université d'Avignon et des Pays de Vaucluse, qui a dirigé mes recherches et m'a encouragé de ses précieux commentaires, conseils et critiques ;

Albert Di Cristo, professeur à l'Institut de Phonétique d'Aix-en-Provence, qui m'a fait l'honneur d'examiner ce travail à la lumière de ses compétences en prosodie ;

Joseph-Jean Mariani, directeur du Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, qui malgré ses très nombreuses responsabilités a pris le temps d'étudier minutieusement ce mémoire ;

Gérard Chollet, directeur de recherche CNRS à l'École Nationale Supérieure des Télécommunications, qui m'a chaleureusement accueilli au sein de l'équipe parole de l'IDIAP et qui n'a eu de cesse de me voir mener à bien ce travail ;

Marc El-Bèze, professeur à l'Université d'Avignon et des Pays de Vaucluse, qui a toujours fait preuve de gentillesse et de disponibilité ;

René Collier, professeur à l'Institute for Perception Research, spécialiste en prosodie, qui a eu l'amabilité de faire également partie de ce jury.

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1

Position du problème

Ce premier chapitre a pour objectif d'introduire le travail présenté dans ce mémoire, en prenant soin d'expliquer les choix méthodologiques retenus à la lumière de la littérature concernée.

1.1 Définition de la prosodie

Durant plusieurs années consacrées à l'étude de la prosodie, j'ai souvent dû répondre à la même question lorsque quelqu'un était suffisamment curieux pour me demander le sujet de ma thèse :

> "La prosodie... C'est quoi ?"

Une première réponse simple mais peu courtoise aurait été de renvoyer mon interlocuteur à un dictionnaire, ce qui ne lui aurait pas rendu grand service tant les définitions que l'on peut y trouver sont confuses :

> PROSODIE [*ρzdi*] n.f. — 1573 "bonne prononciation" 1562 ; gr. *prosôdia* "accent, quantité, dans la prononciation". Caractères quantitatifs (durée) et mélodiques des sons en tant qu'ils interviennent dans la poésie (→ **métrique, versification, mètre, pied**) ; règle concernant ces caractères. "En apprenant la prosodie d'une langue, on entre plus intimement dans l'esprit de la nation qui la parle" (Staël). Règles concernant les rapports de quantité, d'intensité, entre les temps de la mesure et les syllabes des paroles, dans la musique vocale. Ling. Étude de l'accent et de la durée des phonèmes.

J'ai alors très vite adopté un comportement consistant à ne répondre que de manière succincte (par une formule ressemblant à : "la prosodie ? Et bien c'est en quelque sorte la musique de la parole") proposant ensuite aux personnes non satisfaites de cette réponse des explications plus fournies...
Le terme de prosodie est souvent utilisé pour regrouper différents phénomènes liés de la communication orale (comme l'accent, le rythme, l'intonation, les tons). La conception

des éléments prosodiques a évolué avec le nombre des études, de telle façon que le simple rôle de véhicule des expressions et des sentiments qu'on leur accordait avant les années 70 a fait place à un statut linguistique incontestable (au moins pour un sous-ensemble de ces éléments). Il convient de décomposer les faits prosodiques en trois classes bien distinctes[1] [38] :

- le niveau spontané qui permet la manifestation des réactions instinctives (douleur, joie, émotion, état psycho-physiologique, *etc.*),
- le niveau des formes prosodiques expressives employées de façon intentionnelle,
- et le niveau référentiel où les unités prosodiques assument des fonctions morphologiques, syntaxiques et informatives spécifiques à chaque langue.

On sent intuitivement, que la prosodie ne peut être définie de manière rigoureuse que par une spécification des éléments qui la composent ainsi qu'un ensemble de règles permettant de décrire leurs interactions possibles avec les structures des différents niveaux linguistiques étudiés. On évalue aisément l'ampleur de la tâche aussi existe-t-il deux conceptions traditionnelles de la prosodie :

- On la définit généralement comme l'ensemble des phénomènes suprasegmentaux de la parole c'est-à-dire relevant de domaines plus larges que l'unité linguistique distinctive minimale — le phonème — parmi lesquels la syllabe, le mot, les clauses, les phrases, les paragraphes, *etc.* Lehiste ajoutait à juste titre que les traits suprasegmentaux sont établis par comparaison d'items en séquence, alors que les traits segmentaux sont identifiables par la seule étude du segment lui-même [83].
- Une autre conception répandue est de considérer la prosodie comme l'étude des corrélats phonétiques de fréquence fondamentale, de durée et d'intensité.

La nature du travail qui est présenté ici peut fort bien s'accommoder de ces "définitions". Il faut cependant reconnaître qu'un recours à une présentation plus formalisée de la prosodie nous a été salutaire à plus d'un titre ; aussi allons-nous rappeler brièvement quelques points développés par Di Cristo [39] qui ont fortement influencé notre conception de la prosodie.
Si l'on considère le rythme comme relevant de l'étude distributionnelle des accents, et que l'on fait abstraction des langues dites tonales (pour la majorité asiatiques et africaines), alors l'accentuation et l'intonation sont les deux principaux éléments (ou structures) prosodiques.
La figure 1.1 précise les typologies et les fonctions associées des différents accents en référence à diverses terminologies couramment employées. Les fonctions de l'accent dépendent de l'appartenance de la langue considérée à l'un des deux groupes suivants :

[1] La première catégorie qui ne relève pas du domaine de la langue, ne fera l'objet d'aucune investigation dans la suite de cet exposé.

les langues à accent fixe où la place de l'accent est prévisible (comme le français, le tchèque, *etc.*) et les langues à accent libre (comme l'anglais, l'italien, l'allemand, *etc.*) où la place de l'accent — *a priori* imprévisible — est déterminée à partir de critères morphologiques et/ou sémantiques. Pour la langue française — seule considérée dans notre travail — l'accent non-emphatique assume une fonction démarcative (l'accent tombe sur la dernière syllabe du syntagme ou l'avant-dernière dans le cas d'un [∂] terminal), et génératrice d'intonèmes (un intonème étant un fait intonatif à valeur fonctionnelle). L'accent emphatique qui se distingue par son caractère facultatif remplit également une fonction démarcative (en affectant généralement la première syllabe des mots) et peut accessoirement avoir une fonction contrastive. Ce type d'accent ne faisant pas l'objet d'investigations poussées dans la suite de cet exposé, nous emploierons dès maintenant le terme d'accent pour désigner l'accent non-emphatique. La substance acoustique et perceptive de l'accent est à caractère pluriparamétrique (la hiérarchie des divers paramètres étant dépendante de la langue considérée [103]). Ainsi comme le précise Rossi [134] avec des notations légèrement différentes pour une conception similaire, l'*accent lexical* ou interne qui a pour domaines la syllabe et l'unité accentuelle est caractérisé principalement en français par l'indice de durée [47] ; l'*ictus mélodique* (ou accent secondaire pour Di Cristo) a pour domaine la syllabe, il est caractérisé par une proéminence tonale et assume des fonctions essentiellement rythmiques ; et l'*accent de focalisation* (ou accent intellectuel) dont les domaines sont la syllabe et le mot, il se caractérise par des proéminences du fondamental et de l'énergie.

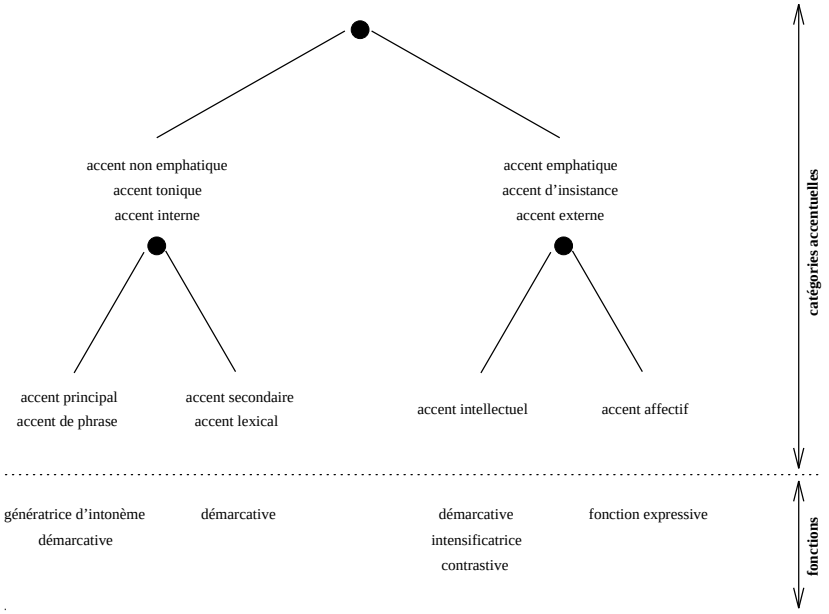


Figure 1.1: Catégories et fonctions accentuelles du français selon Di Cristo. Les dénominations catégorielles attachées à un même nœud sont équivalentes et reprennent les principales terminologies.

Le rôle de l'intonation est d'expliquer le comportement d'un locuteur en fonction d'une situation et des personnes à qui s'adresse le message. Les fonctions de l'intonation recensées par Di Cristo [39] sont au nombre de douze, témoignant — si besoin était — du rôle important de la prosodie et principalement de l'intonation dans la communication linguistique.

Fonction intégrative : l'intonation permet d'accorder à une suite non grammaticalement correcte (ex: Max travaille d'une façon[2]) le statut d'énoncé intono-syntaxiquement bien formé.

Fonction modale : de nombreuses études [74, 111] et [126, pp. 149–177] ont mis en relief l'importance de la configuration intonative pour la détermination de la modalité de la phrase (affirmative, impérative, interrogative, *etc.*). La figure 1.3 présente la courbe du fondamental de trois réalisations de la même phrase avec des modalités différentes.

Fonction informative : la participation de l'intonation à la structuration du message (sémantique) a été abondamment étudiée, notamment pour l'identification du thème (fonction *thématique*) et le rhème (fonction rhématique) [120, 93, 126, 133], mais également dans le cas d'une emphase non expressive ou à valeur fonctionnelle comme dans l'exemple : "*C'est* celle-ci *que je veux*" (fonction *contrastive*). La figure 1.4 propose un exemple de cette fonction.

Fonction identificatrice : l'intonation participe également à la structuration de l'énoncé en délimitant les frontières de certains constituants syntaxiques [129, 120, 93, 31, 16, 21] (voir figure 1.5).

Fonction hiérarchisante : certains auteurs s'accordent à dire que la structure intonative a pour rôle de proposer une hiérarchie des constituants syntaxiques [126, pp. 223,233], [39].

Fonction de désambiguïsation : cette fonction illustrée par l'exemple classique : *la belle ferme le voile*[3] est d'autant plus intéressante qu'elle est souvent la seule à pouvoir trancher sur la structure profonde de la phrase [68, 113, 123, 167]. Les figures 1.6 et 1.7 proposent des exemples de réalisations de phrases ambiguës.

Fonction attitudinelle : l'intonation véhicule des informations sur l'attitude (doute, surprise, *etc.*) du locuteur envers le message énoncé.

Fonction quantificatrice : qui définit une notion d'intensité des différentes modalités expressives.

Fonction d'emphatisation : l'intonation permet finalement au locuteur de mettre en relief une portion d'un énoncé (voir une illustration en figure 1.2).

[2]Exemple proposé par Gross [60].
[3]Exemple proposé par Malmberg [88].

Comme pour l'accentuation, Rossi [134] rappelle l'aspect pluriparamétrique de l'intonation et spécifie que le domaine de l'intonation est la phrase et ses constituants. La table 1.1 résume les intonèmes (morphèmes de l'intonation) qu'il distingue (et que l'on retrouve habituellement dans la littérature) ainsi que leur caractérisation paramétrique.

intonèmes	Désignation	Fonction	*f0*	durée	loudness
CT	continuatif majeur	hiérarchisation pragmatique et syntaxique	+	+	+
ct	continuatif mineur	démarcatif	+	+	
CA	vocatif non terminal	marqueur du topique			
PAR	parenthétique	marqueur d'apposition	stable		
CC	conclusif majeur	marqueur terminal et rhématique	−	+	−
cc	conclusif mineur	marque la fin d'une parenthèse droite ; marqueur de disjonction entre les syntagmes internes d'une parenthèse			

Table 1.1: Intonèmes répertoriés par Rossi avec leurs caractérisations paramétriques.

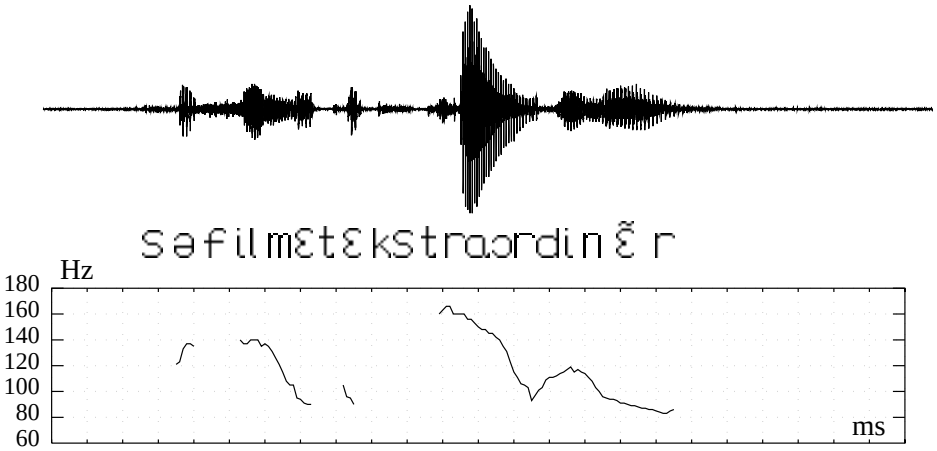


Figure 1.2: Illustration de la fonction d'emphase assurée par la prosodie pour une réalisation de la phrase : "Ce film est ex**tra**ordinaire" : l'emphase est ici marquée par des valeurs maximales de l'intensité et de la fréquence fondamentale sur la voyelle [*a*] du mot *extraordinaire*.

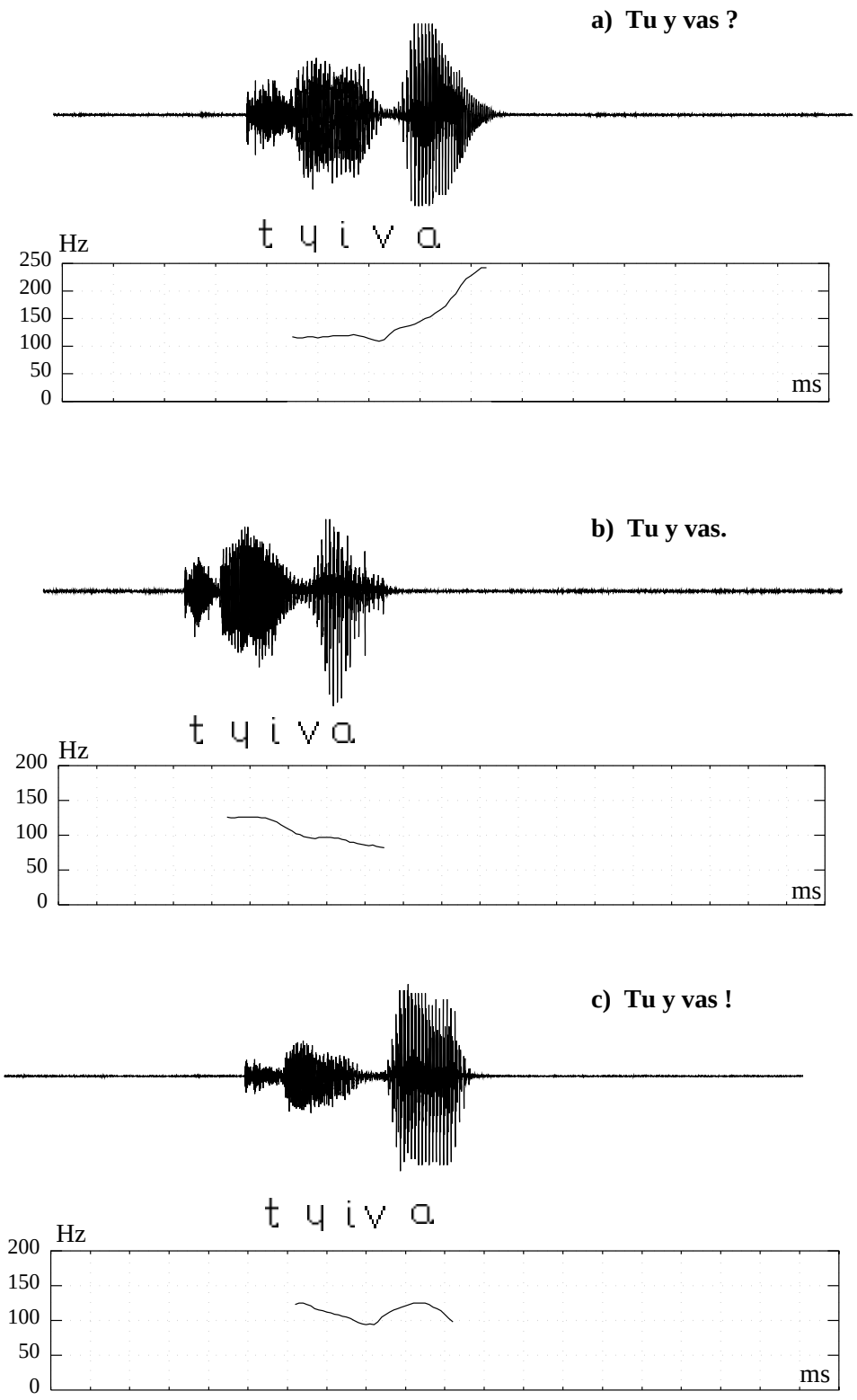


Figure 1.3: Illustration de la fonction modale (non expressive) assurée par la prosodie pour trois réalisations de la phrase “Tu y vas”.

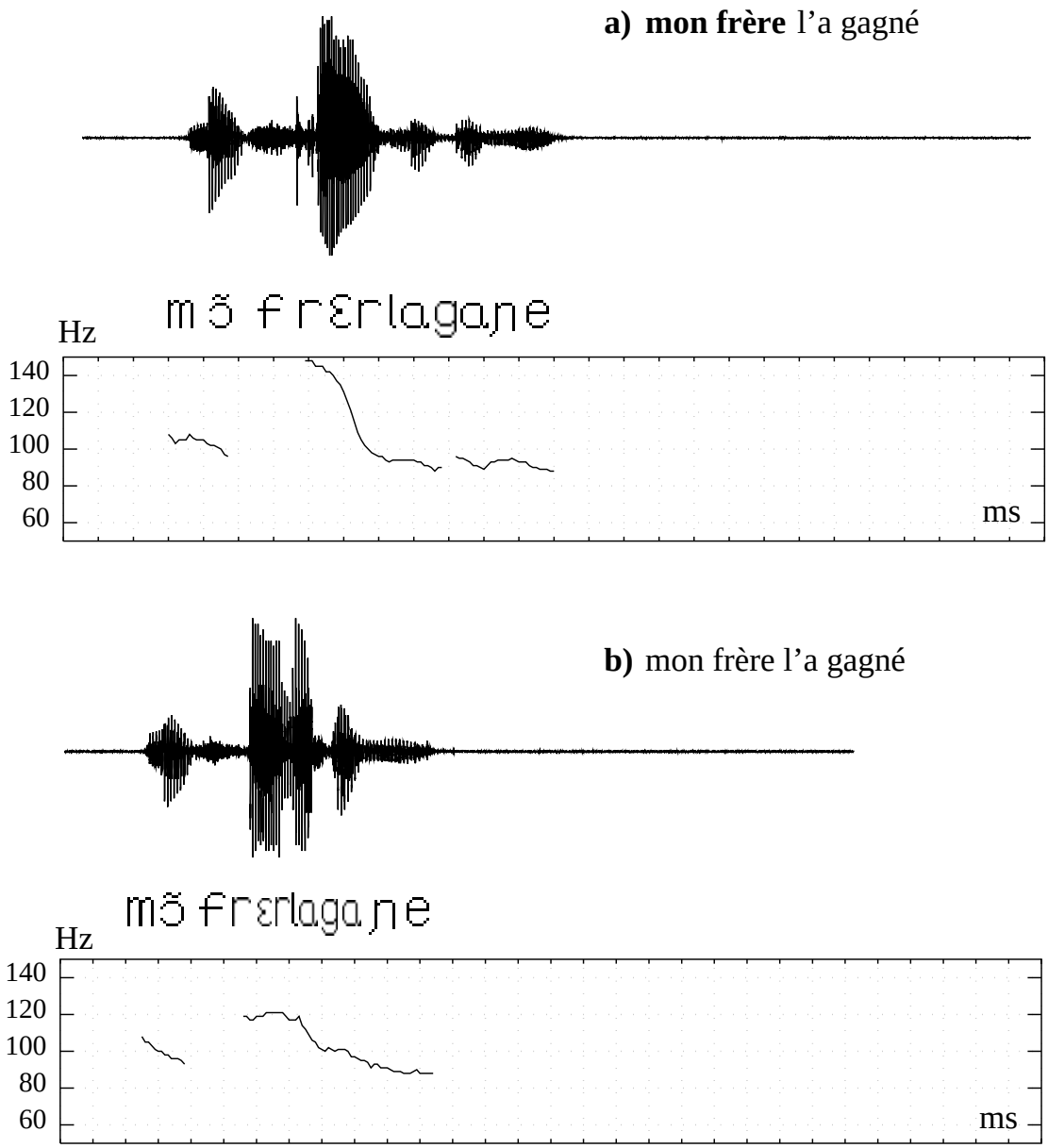


Figure 1.4: Illustration de la fonction informative (sémantique) assurée par la prosodie. Dans la réalisation de la phrase a), le locuteur insiste sur le fait que c'est son frère qui a gagné.

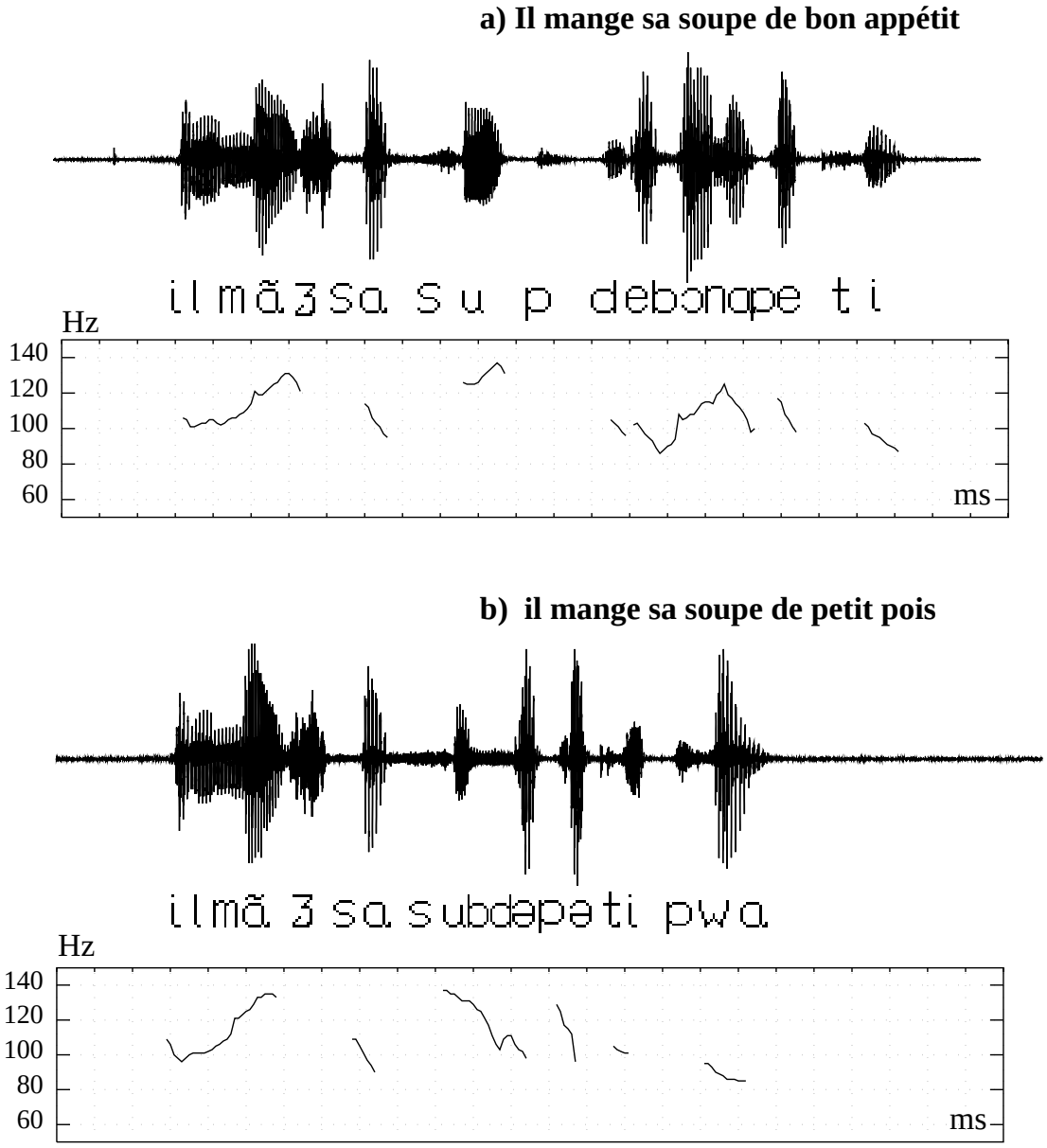


Figure 1.5: Illustration de la fonction identificatrice (syntaxe) assurée par la prosodie. On observe dans le cas a) un allongement de la voyelle [u] du mot *soupe* qui est suivi d'une pause.

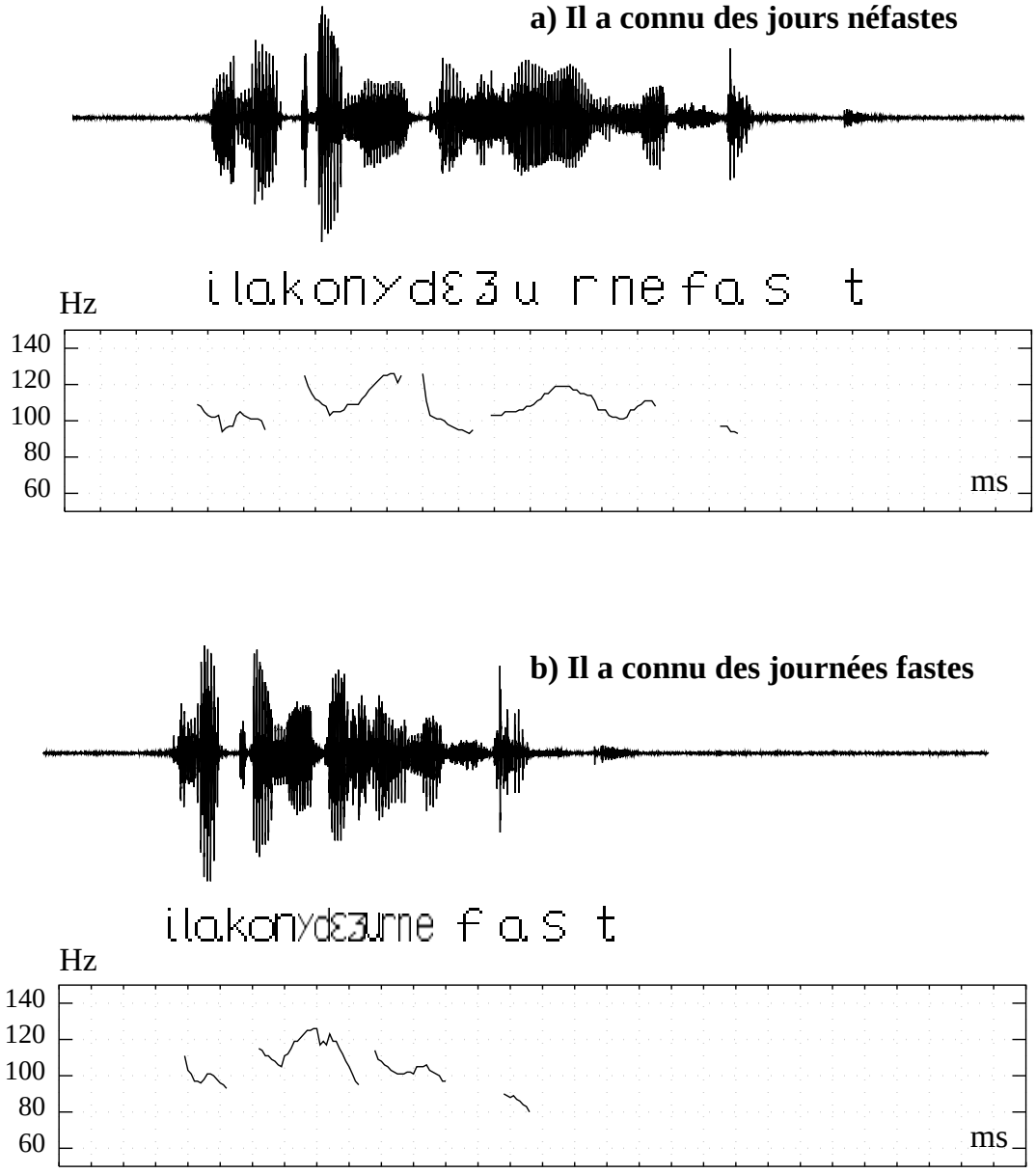


Figure 1.6: Illustration de la fonction de désambiguïsation assurée par la prosodie. Dans la réalisation a), on peut observer un allongement de la voyelle [u] du mot *jour* ainsi qu'une inversion de pente de la fréquence fondamentale après ce mot.

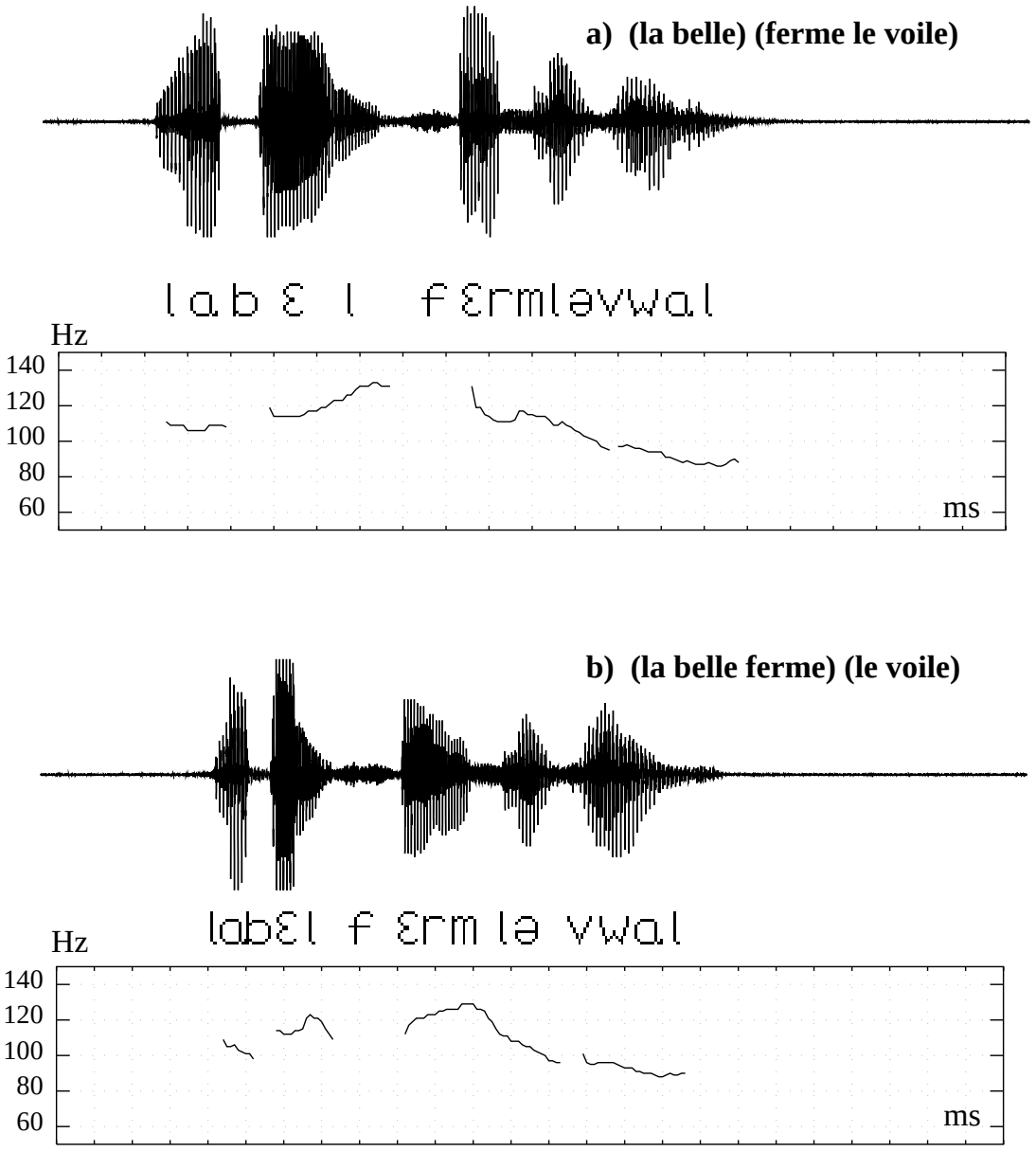


Figure 1.7: Illustration de la fonction de désambiguïsation assurée par la prosodie : l'emplacement du pic de fréquence fondamentale permet à lui seul dans ces deux réalisations de lever l'ambiguïté.

1.2 La prosodie un problème simple ?

L'inventaire réalisé par Di Cristo [37] des études prosodiques entreprises depuis le début du siècle jusqu'aux années 70, témoigne de l'intérêt que l'on portait alors à la prosodie. Si aucun recensement aussi méticuleux n'a depuis été dressé, le nombre des travaux sur la prosodie des langues naturelles n'a cessé d'augmenter et les manifestations dédiées à ce thème sont de plus en plus nombreuses (workshop de Barcelone, école d'été de Lund, workshop de New York, école de Londres ...). Cette profusion d'études n'empêche cependant pas les spécialistes d'arriver à la constatation formulée par Hirst [64, p. 43] :

> "Malgré une quantité impressionnante de recherches récentes sur la prosodie des langues naturelles, on est obligé d'admettre qu'aujourd'hui encore nous ne savons pas grand chose sur elle."

Vaissière [158] très récemment confirme cet état de fait avec une formulation plus directe :

> "Le problème est qu'il n'existe pas de modèle unique pour une langue donnée, et plus une langue est étudiée, et plus les modèles abondent et se contredisent, du moins en partie."

Comme le souligne G. Caelen [21] la prosodie est encore à l'heure actuelle sujet à controverses. Pour les adeptes de l'"école générativiste" par exemple, il est des positions divergentes quant à l'organisation linguistique qui, de la syntaxe ou de la sémantique, conditionne la place de l'accent de phrase. Pour les fervents de l'approche empiriste, le problème ne se pose pas en ces termes puisque la prosodie peut-être considérée comme un canal qui se superpose au message verbal et dont les structures ne sont pas sous l'égide d'une organisation linguistique *a priori*. Les composantes prosodiques et linguistiques interagissent alors épisodiquement en des points souvent nommés *rendez-vous* et une étude corrélative permet alors de préciser et de quantifier la nature de ces relations.

Qui s'intéresse à la prosodie ?

Si l'on considère que la prosodie intervient à tous les niveaux de la communication, il est normal de concevoir que la communauté scientifique qui s'y intéresse est large et constituée de chercheurs de domaines différents aux motivations très variées. Lors d'une conférence invitée du séminaire prosodie de la Baume-lès-Aix 1992, Monaghan [106] remarque très justement le problème concret de cette pluridisciplinarité :

> "Prosody in general, and synthetic prosody in particular, currently interests workers in engineering, computer science, phonetics, linguistics, psychology, philosophy and artificial intelligence to name just some of the fields where research in this area can be found : it is impossible for any one person to be aware of all this work."

Une première distinction peut être faite pour classer la multitude des travaux traitant de la prosodie, selon qu'il s'agit d'études fondamentales ou bien technologiques. Cette dichotomie ne nous semble cependant pas totalement pertinente car elle est génératrice d'un état non fécond où les chercheurs "fondamentalistes" tendent à considérer les travaux des chercheurs "applicatifs" comme une activité de second ordre ; à leur tour ces derniers peuvent s'appuyer sur les lacunes des modèles produits par les premiers pour dénigrer leur études. Fort heureusement, plusieurs auteurs se sont "compromis" dans des recherches appartenant aux deux classes réduisant la polémique à un état d'obsolescence [136, 126, 100, 101]. Considérant que ces deux approches sont complémentaires, nous adopterons cette première distinction en introduisant cependant une classe supplémentaire, ce qui nous permet de définir trois approches en recherche prosodique :

- les études psycho-linguistiques qui tentent de comprendre les mécanismes de planification/perception de la prosodie dans la parole,
- les études phonético-linguistiques qui modélisent et/ou quantifient les relations entre l'organisation prosodique et les structurations linguistiques de la langue,
- les études technologiques pour lesquelles la prosodie est vue comme une structure de performance.

La saga des terminologies

La pluridisciplinarité des chercheurs en prosodie ne va pas sans poser quelques problèmes de vocabulaire, de même que la diversité et la spécificité des langues étudiées ne joue pas non plus en faveur d'une terminologie claire et consensuelle. Il n'est ainsi pas facile d'établir une signification précise de certains termes (stress, accent, focus, *etc.*) qui à l'intérieur d'une même langue donnent déjà lieu à des utilisations hétérogènes. Pourtant nous savons qu'il existe de nombreux invariants universels, qui peuvent globalement s'expliquer par la fonction de communication qui n'est bien sûr pas dépendante d'une langue particulière pas plus que ne le sont les appareils de production et de perception de la parole. Vaissière [156] a dressé une liste détaillée de ces invariants parmi lesquels on peut distinguer :

- le rôle des pauses qui, si leurs distributions peut varier d'une langue à l'autre, sont des indices forts de groupement ; ainsi les pauses respiratoires tombent généralement à des jonctions grammaticales (cf. la notion de groupe de souffle adoptée par Lieberman [86]), et présentent une certaine hiérarchie (les pauses sont plus longues en finale de phrase qu'à l'intérieur)[32],
- la tendance globale de la courbe de fréquence fondamentale à décroître dans le temps, qui s'explique bien par des raisons physiologiques mais aussi phonologiques[4] [17, 122, 144, 55],

[4] Voir une discussion de Rossi [137] sur l'inadéquation du modèle de la ligne de déclinaison pour rendre compte du relief accentuel en français.

- l'allongement des syllabes (qu'il soit final ou progressif) dans les clauses, que l'on peut considérer d'un point de vue physiologique (relâche des articulateurs), ou fonctionnel (marqueur d'une fin d'unité *a fortiori* lorsque celle-ci n'est pas déjà marquée d'une pause par exemple).

Ainsi, le problème de la terminologie ne relève pas d'une revendication purement esthétique, mais correspond à un véritable problème lorsqu'il s'agit de tirer des enseignements des études portant sur la prosodie d'autres langues. On trouve par exemple dans des études récentes des auteurs qui prennent soin de préciser leur terminologie [43]. G. Caelen [21] rappelle à ce sujet une expérience qu'elle a menée à l'occasion du workshop sur la prosodie de Barcelone en demandant aux participants de remplir une grille précisant les domaines (syllabes, mots...) des principaux termes employés (stress, focus ...). Nous ne saurions qu'encourager une extension de ce type d'expérience qui pourrait dans le plus heureux des cas s'achever par la mise à jour d'un "petit lexique à l'usage de l'apprenti prosodicien" aux vertus pédagogiques indéniables et salvatrices !

De l'art de la juste mesure...

Une première difficulté "matérielle" à laquelle est rapidement confronté un chercheur en prosodie est l'extraction des paramètres prosodiques. Cette opération sera d'autant plus critique si l'approche adoptée n'autorise pas une phase d'intervention manuelle (validation/correction) comme c'est généralement le cas dans les études "technologiques" *a contrario* des autres catégories d'études. Dans le cadre d'un traitement automatique, deux stratégies s'ensuivent, selon que l'on considère comme fiable l'extraction des paramètres (et toute erreur est alors comptabilisée comme une erreur du système dans sa globalité) ou au contraire que l'on admet les limites de ces algorithmes d'extraction et que l'on s'attache à répertorier les configurations erratiques en leur associant des probabilités d'occurrence qui seront prises en compte au moment de la résolution du problème. Cette dernière solution a par exemple été retenue dans [69] où les erreurs de détection du fondamental dues aux harmoniques sont prises en compte. La section 2.1 traite des problèmes particuliers liés à l'extraction des paramètres de fréquence fondamentale, de durée et d'intensité, et décrit les algorithmes que nous avons employés pour nos travaux.
Une fois l'extraction des paramètres prosodiques réalisée, une autre difficulté se pose pour le chercheur quant à la décision d'une représentation de ces paramètres. Comment par exemple traiter un paramètre comme la fréquence fondamentale qui par nature est continu, alors que les méthodes ou règles s'appliquent le plus souvent à des entités discrètes ? La réponse à cette question est loin d'être simple[5] ; nous allons essayer d'expliquer quelques raisons à cet état de fait. En tout premier lieu, il convient de distinguer plusieurs causes de variabilité des paramètres prosodiques. Très schématiquement, on peut établir une distinction entre les variations qui sont régies par les instructions linguistiques (les variables contrôlées) et celles qui sont inhérentes au mécanisme de production (les variables non contrôlées) [163]. Lors d'une étude sur les fonctions linguistiques de la prosodie, il serait

[5] Voir par exemple [61] pour une discussion sur ce thème.

donc souhaitable de pouvoir "soustraire" des paramètres prosodiques les variations non contrôlées, afin de ne pas leur accorder à tort une valeur linguistique. Nous verrons au cours du chapitre 4 que de telles corrections — bien que séduisantes — sont loin d'être évidentes à mettre en place surtout dans le cadre d'un traitement automatique. Ainsi Vaissière [157] écrit-elle à ce sujet :

> "Present systems only include partial compensation of the phonetically-conditioned variations, *i.e.* of the ones which are the 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)."

Nous n'aborderons pas ici des problèmes plus délicats, pour lesquels nos connaissances sont pour le moins déficientes, comme les caractéristiques individuelles (sexe, âge, émotions) ou socio-culturelles qui contribuent elles aussi aux variations des paramètres prosodiques, et qu'il conviendrait à ce titre de neutraliser (*i.e.* de contrôler).
En second lieu, il convient de considérer les problèmes non triviaux de conversion perceptive. Nous savons en effet qu'il existe de grande différences entre le niveau de la production et celui de la perception ; ainsi l'appareil auditif ne perçoit-il pas les données objectives en l'état. Des travaux de Rossi sur les glissandos mélodiques, nous savons par exemple qu'en plus d'être dépendants des glissements d'intensité [131], la perception d'un glissando dépend étroitement de sa durée [128] : en dessous d'une certaine durée de l'ordre de 50 ms, aucune variation tonale n'est perçue, de même que les glissandos perceptibles ne sont pas perçus dans leur totalité mais en un point (devenu fameux) qui se situe aux deux tiers de la pente. Munson [107] très tôt démontrait que l'intensité subjective était dépendante de la durée ; Lehiste a également montré que la durée perçue d'une syllabe est plus grande quand cette dernière est porteuse d'une variation mélodique que lorsqu'elle est affectée d'un ton statique [83]. Ainsi de nombreuses études attestent le caractère pluriparamétrique de la prosodie et montrent de toute évidence la nécessité d'interpréter les données brutes des paramètres à la lumière d'un modèle de perception.
Toutefois le grand nombre d'études sur les mécanismes de la perception des paramètres prosodiques ne cache pas les problèmes que nous soulevons maintenant (sans d'ailleurs aucune intention polémique sur l'utilité de tels travaux) et qui nous ont amenés à ne traiter dans ce mémoire que les valeurs objectives (*i.e.* telles qu'elles nous sont données par nos algorithmes d'extraction) des paramètres de fréquence fondamentale, de durée et d'intensité.

- Nous pourrions nous poser la question de savoir si les efforts des études psychoacoustiques sont utilisables directement pour proposer des méthodes d'interprétation perceptives des valeurs objectives des paramètres prosodiques. De nombreuses études ont été en effet le fruit de travaux sur des stimuli artificiels sur lesquels les résultats sont généralement plus fins que ceux obtenus avec de la parole naturelle. On peut également formuler quelques reproches à l'encontre des protocoles expérimentaux qui placent le sujet dans des conditions peu naturelles d'écoute

(répétition à loisir des item, attention soutenue sur un point particulier du message ...). On pourrait également se poser la question de savoir si la compétence des auditeurs ne serait pas un facteur perturbant lors de ces expériences. Plusieurs études tendent à confirmer le fait qu'un auditeur entraîné perçoit les choses plus finement qu'un auditeur naïf, même si des études récentes [48] montrent que la différence pour une tâche donnée n'est pas hautement significative.

- Les résultats des expériences psycho-acoustiques s'ils ne sont pas contradictoires, ne sont pas toujours en parfait accord. Ainsi par exemple Lieberman et McDowall [95] ont mesuré des réponses satisfaisantes lorsqu'ils demandaient à des sujets (entraînés ou pas) de prendre une décision binaire sur la nature proéminente d'une syllabe alors que les résultats s'avéraient mauvais lorsqu'on leur demandait de fournir une appréciation continue. Collier [35] quant à lui, reporte à l'occasion d'expériences sur les PBS (Perceptual Boundary Strength) que les auditeurs sont tout à fait capables de noter ces dernières sur une échelle de valeurs de 10 points.

- Les impressions des auditeurs peuvent être faussées par des indices non pris en compte dans l'expérience ; ainsi au cours d'études sur de la parole naturelle, l'influence du contexte (position syllabique, distribution des mots dans les syntagmes ...) peut conditionner les mesures des seuils différentiels. S'il est vrai que les études présentent fréquemment des tests sur de la parole dégradée, il n'en reste pas moins vrai, d'après les expériences de Blesser [14] qu'un auditeur est capable de percevoir la structure syntaxique des phrases privées de leur contenu phonétique (rotated speech) ce qu'il faudrait donc en tout état de cause considérer dans ces expériences.

- Les méthodes de stylisation des paramètres prosodiques sont très souvent monoparamétriques ce qui — aux vues des remarques précédentes — n'est pas souhaitable. Elles sont généralement compliquées et fastidieuses ; ainsi la méthode bien connue conçue par les chercheurs de l'institut sur la perception d'Eindhoven [145, 146] est-elle pour le moins — si ce n'est soumise à un certain subjectivisme des auteurs lors de la décision d'une équivalence perceptive — contraignante dans sa réalisation (nous ne discutons bien sûr pas de son efficacité qui n'est pas contestable). Une autre fameuse méthode proposée par l'institut de phonétique d'Aix [40, 126] qui consiste en l'élimination des effets microprosodiques puis en la transformation des valeurs objectives par un ensemble complexe d'opérateurs fréquentiels et temporels, si elle présente l'avantage d'être rigoureuse quant aux interactions des différents paramètres n'en est pas moins difficile à mettre en œuvre. Ainsi ces méthodes semblent peu se prêter à un traitement automatique, ce qui pour nous est une condition nécessaire.

- Nous rejetons de même toute méthode partielle de correction perceptive qui selon nous est aussi critiquable qu'une analyse objective des données. De plus si l'argument souvent avancé selon lequel les variations non perçues sont inutiles (voir [39] pages 29 et 30), il n'en reste pas moins qu'elles ont été produites et qu'une étude corrélative

entre des indices prosodiques acoustiques et un niveau d'organisation linguistique donné (ou plusieurs) devrait être à même de vérifier qu'elles ne sont d'aucun intérêt.

- De manière générale, même si des études ont montré qu'il était possible de segmenter un paramètre comme la ligne mélodique en unités discrètes [56] nous sommes obligés de reconnaître que cela ne peut se faire sans une part d'arbitraire. Pour ne citer que cet exemple, le découpage proposé par Delattre [46] en quatre niveaux (grave, médium, infra-aigu, aigu), qui perdure dans les études prosodiques actuelles est réalisable de plusieurs façons selon que l'on considère le registre d'un locuteur sur l'ensemble de ses réalisations ou sur l'entité (phrase ou autre) observée (ce que G. Caelen [21] qualifie respectivement de codage "texte" et de codage "phrase"), ce choix n'étant pas tout à fait innocent.

Ces quelques points sur lesquels on pourrait encore longtemps polémiquer, s'ils n'apportent aucune réponse, permettent simplement de rappeler qu'une des difficultés de la recherche prosodique est d'adopter une représentation adéquate des paramètres prosodiques. Les choix faits à ce moment pourront être motivés par des raisons théoriques et/ou pragmatiques qui seront de toutes façons critiquables.

Mais comment parlons-nous donc ?

Les difficultés de l'analyse prosodique ne sont pas les mêmes selon le type de parole analysée. Il est habituel de distinguer basiquement les études sur la parole lue de celles sur la parole spontanée. Si cette dichotomie reflète intuitivement bien la différence entre ces deux styles, il ne reste cependant pas évident de placer entre ces deux limites des styles de parole comme la lecture avec consignes [21], les monologues à thème [13], les conférences, les interviews, les conversations [87], les requêtes non contraintes à but informatif [108], la "parole contrôlée" [114], *etc.* S'il est certain que tout acte de parole non lue ne relève pas du domaine de la spontanéité, nous pouvons cependant nous contenter d'énumérer quelques caractéristiques générales qui séparent la parole lue de la parole non lue et qui relèvent d'un statut prosodique :

- De manière très générale, dans la parole spontanée, les structures syntaxiques sont relativement simples, de nombreuses phrases ne sont pas syntaxiquement correctes mais possèdent un statut intono-syntaxique qui leur confère un sens. Ce type de communication favorise des phénomènes d'emphases qui sont ordinairement absents dans une situation de lecture. Mariani [91, p. 209] rappelle également que le débit moyen diffère sensiblement entre les deux situations de parole : 4 mots/s pour de la parole lue contre 2,5 mots/s pour de la parole spontanée.

- Blaauw [13] note à la suite d'expériences perceptives, qu'il est tout à fait possible pour un auditeur de distinguer des échantillons de parole lue et spontanée

mieux qu'aléatoirement (taux de 77%) [6]. Elle étudie alors les distributions et les réalisations des marques prosodiques qui contribuent à cette faculté distinctive et permet de dégager certains points. Dans la parole spontanée, les tons montants sont prédominants aux frontières majeures à l'inverse de la parole lue ; les frontières des clauses mineures sont également marquées dans les deux types d'élocution ; l'auteur relève de plus que les indices temporels (pauses et allongements) assument principalement un rôle démarcatif dans des réalisations lues alors qu'il est fréquent dans de la parole spontanée de les rencontrer en des points ne correspondant pas à des limites de constituants structurels (syntaxiques ou prosodiques) mais tendant à se localiser sur les mots les plus informatifs. Cette remarque confirme les travaux de Cooper sur les unités de planifications [36] qui pourraient être différentes lors d'une lecture et d'une prise de parole "naturelle".

- Des études sur de grands corpus de parole spontanée montrent que 10% des phrases contiennent des ruptures (dénommées *repairs* dans la littérature de langue anglaise) [141]. Nakatani et Hirschberg dans une étude très intéressante [108] proposent un algorithme[7] de prédiction de ces *repairs* à partir du modèle *RIM* (Repair Interval Model) en étudiant les configurations acoustico-prosodiques sur les trois intervalles contigus définis par ce modèle : le *reparandum interval* (ri1) qui correspond au segment à corriger, le *disfluency interval* (di) qui s'étend du début de la rupture jusqu'au moment où la réparation est effectuée, cet intervalle abrite aussi bien des silences que des pauses remplies (*euh*) ou des onomatopées caractéristiques d'une prise de conscience d'une erreur (*oops, hum, etc.*) et le *repair interval* (ri2) qui correspond à la correction apportée. Dans cet article, sont présentés un ensemble d'indices pouvant participer à la détection de ces accidents : des indices de fragmentation comme des coups de glotte (dans 30% des cas en finale de ri1), des gestes co-articulatoires sur des fragments se terminant par des voyelles, ou encore des distributions particulières de longueur des fragments (un fragment étant inférieur ou égal à une syllabe dans la majorité des cas), des indices déjà signalés par d'autres auteurs comme la tendance à accentuer la proéminence intonative sur le ri2 [85], *etc.*

Pour autant passionnantes que soient les études qui traitent de la parole non lue, et bien qu'elles soient d'un intérêt certain pour la robustesse des systèmes de reconnaissance [165, 82], il n'en reste pas moins qu'elles sortent très nettement du cadre de ce mémoire qui se restreint à l'étude de corpus de phrases lues.

[6] Tests réalisés sur 21 auditeurs devant se prononcer sur le caractère spontané de 109 paires de phrases (chaque paire étant constituée d'une phrase extraite d'un monologue à thème puis de la même phrase lue par le même locuteur) ne se distinguant pas par des indices particuliers (tels que les pauses remplies, faux départs, *etc.*).

[7] Par des techniques de Classification And Regression Tree (CART) [18].

1.3 Présentation du travail

Notre travail se réclame d'une approche à caractère applicatif. Nous argumentons ce choix de la "performance" comme moyen d'investigation dans les lignes qui suivent :

- Ce travail s'est déroulé au Laboratoire Informatique d'Avignon (LIUAPV) et à l'Institut Dalle Molle d'Intelligence Artificielle et Perceptive (IDIAP) qui favorisent l'aspect applicatif des études qui y sont menées [97, 112, 59, 143, 11, 30, 29].

- S'il nous semblait impensable de pouvoir — durant ces quelques années consacrées à cette thèse[8] — maîtriser suffisamment une science[9] aussi complexe que la prosodie afin d'en proposer une modélisation, il nous paraissait cependant réalisable d'étudier son potentiel applicatif pour l'amélioration des systèmes automatiques existants au laboratoire informatique d'Avignon et à l'IDIAP.

- L'élaboration d'une théorie ne saurait être posée comme *a priori* mais doit selon nous suivre une étape d'observations dont l'expert s'attachera à définir les structurations à la lumière d'hypothèses fondamentales. Léon ira encore plus loin en écrivant [84] :

  > "Seul est sûr le domaine de la "performance" et la procédure du type inductif, qui va des faits à la théorie, évitant les simplifications *a priori*."

- Nous ne prônons cependant pas une approche "atomiste" et sommes conscients que l'analyse statistique ne saurait être suffisante pour la définition d'unités prosodiques. Nous savons effectivement que les données seules n'ont pas de valeur, et qu'elles ne prennent un sens qu'au regard d'une théorie [90, 94, 40] ; aussi concevons-nous ce travail comme une étape préliminaire à une phase de construction d'un "modèle". De plus, nous gardons comme principe sous-jacent à ce travail des hypothèses fondamentales que nous avons présentées précédemment sur les fonctions intonatives et accentuelles.

- Enfin et sans vouloir ré-ouvrir le débat sur le bien-fondé des approches empiristes et fondamentalistes, nous formulerons cependant quelques remarques qui ont renforcé notre approche expérimentale. Il existe un grand nombre de modèles et de théories[10] de l'intonation ou plus généralement de la prosodie [126, pp. 17–39][94, 147, 58, 67, 134]. Il n'est pas aisé d'opter pour un de ces modèles car ils sont par nature difficilement comparables ne modélisant pas tous les mêmes unités (*f0* cible, contours stylisés...). Ils nécessitent bien souvent une intervention manuelle que nous avons écartée dans notre travail et/ou ils font intervenir des niveaux linguistiques qui sont encore difficiles à intégrer dans un système de reconnaissance (sémantique,

[8] J'entends déjà au moment où j'écris le mot *quelques* les railleries de quelques personnes qui ne me reprocheront pas pour une fois mon caractère Marseillais...

[9] Voir la Thèse de Doctorat de G. Caelen pour une discussion sur le statut scientifique de la prosodie[21].

[10] Voir l'étude de Marchal [90] pour un rappel des termes *modèle* et *théorie* et pour une revue de plusieurs théories et modèles.

pragmatique). Nous conclurons par cette citation de Vaissière [158] qui propose une vision très tranchée sur les problèmes de caractérisation des courbes de la fréquence fondamentale :

> "Le cantonnement à des explications relevant de la linguistique ou à la phonologie particulière d'une langue peut aboutir à un obscurcissement des problèmes."

Le présent document se divise en quatre parties distinctes. Dans le premier chapitre nous décrivons notre approche expérimentale en précisant d'une part nos méthodes d'extraction des paramètres de fréquence fondamentale, de durée et d'intensité et, d'autre part, en décrivant notre algorithme d'étiquetage prosodique automatique dont nous nous servirons par la suite.
Comme toute analyse nécessite des données, nous précisons dans le chapitre suivant la nature et le contenu des divers corpus qui sont utilisés dans ce travail.
Le quatrième chapitre, propose une étude des phénomènes microprosodiques dont on parle dans de nombreuses études [152, 81, 126, 40, 42, 157, 160, 50] ; soit qu'on les considère comme des variations "parasites" des divers paramètres prosodiques sans aucune valeur fonctionnelle [43], soit au contraire qu'on leur attribue des potentialités discriminantes. Il nous a donc semblé intéressant de faire le point sur ces positions avec une approche de traitement automatique orientée sur le filtrage lexical de grands vocabulaires.
Dans le cinquième chapitre, nous nous attachons à décrire les travaux que nous avons entrepris dans le cadre élargi de la "macroprosodie". Plus particulièrement, nous présentons les structures logicielles mises en place pour l'analyse de la corrélation entre des indices prosodiques et diverses organisations linguistiques. Nous montrons comment cette architecture permet une utilisation simple des informations prosodiques dans des systèmes de reconnaissance et/ou de synthèse de la parole. Nous illustrerons nos propos par l'étude descriptive d'un corpus de phrases lues et montrerons les améliorations possibles de deux tâches de reconnaissance (l'une "restreinte" de reconnaissance de nombres, l'autre plus générale de reconnaissance de la parole continue).
Nous terminerons — comme l'exige la tradition — par tirer des conclusions de cette première rencontre avec la prosodie et dresserons une liste de problèmes que nous aimerions traiter ultérieurement.

Chapitre 2

Les paramètres prosodiques

Nous allons dans ce chapitre présenter et justifier les choix algorithmiques retenus pour extraire du signal les trois paramètres de fréquence fondamentale, d'intensité et de durée. Nous montrerons que chaque paramètre peut être appréhendé par différentes techniques dont aucune n'est complètement satisfaisante. Nous détaillerons également dans ce chapitre notre système d'étiquetage prosodique ascendant en spécifiant pour chaque paramètre les indices mesurés et terminerons par quelques exemples.

2.1 Extraction automatique des paramètres acoustiques

2.1.1 La fréquence fondamentale

La mélodie de la voix se traduit sur le plan physique par l'évolution de la fréquence laryngienne — caractéristique des sons voisés — en fonction du temps. La plage de variation moyenne de cette fréquence varie d'un locuteur à l'autre en fonction principalement de son âge et de son sexe (de 100 à 160 Hz pour un homme adulte et de 150 Hz à 300 Hz pour une femme adulte), et peut enregistrer d'importantes variations chez un même locuteur. D'un point de vue acoustique, on a coutume de nommer *fréquence fondamentale* ou *f0* l'estimation des variations laryngiennes à partir du signal de parole. Bien qu'il soit possible d'extraire directement la fréquence laryngienne de l'observation de données physiologiques (ou encore avec l'usage d'appareillages) on a le plus souvent recours à des algorithmes de détection de *f0* pour étudier la mélodie à partir du signal de parole. Il en existe plusieurs centaines[1] [62, 27] et les plus performants d'entre eux sont cependant incapables de fournir des valeurs toujours correctes de *f0* dans toutes les circonstances (sons, bruits, locuteurs, *etc.*). Les principaux problèmes rencontrés sont :

les sauts d'octaves : L'analyseur fournit une valeur de *f0* qui ne correspond pas au premier harmonique. Cela peut arriver pour un spectre dont le deuxième har-

[1] voir par exemple Mariani [91, pp. 61–62] pour un bref descriptif des principales méthodes.

monique correspond au premier formant ou dans le cas d'une insuffisance passagère de l'amplitude du fondamental.

les non-détections : Il existe une fréquence "théorique" que l'algorithme n'a pas détectée. Ceci arrive très souvent dans des portions peu énergétiques et/ou bruitées du signal de parole.

la finesse du détecteur : Les valeurs proposées sont éloignées faiblement des valeurs théoriques.

la décision de voisement : Cette décision, bien que difficile à prendre dans certaines situations (faible énergie, parole bruitée, ...) serait cependant fort utile — au-delà du bon fonctionnement du détecteur — à des fins de segmentation du continuum sonore (*ex.* distinction [ρ] / [b]).

On recense deux grandes catégories d'algorithmes de détection :

- ceux qui opèrent dans le domaine temporel comme la technique d'*amdf* que nous décrirons et emploierons par la suite,
- et ceux qui travaillent dans le domaine spectral : les valeurs de *f0* sont calculées à partir des maxima des spectres d'amplitude.

Il existe des études comparatives de quelques algorithmes courants [27] qui ne doivent pas faire oublier que les résultats obtenus dépendent souvent des pré-traitements (filtrage, pré-accentuation, ...) qu'ils effectuent sur le signal avant détection, et que les conclusions quant à la performance de chacun ne doivent être faites qu'après l'analyse méticuleuse de leur comportement dans tous les contextes possibles. Il n'y a donc pas d'algorithme parfait, et bien qu'une solution satisfaisante consisterait à mettre en parallèle une large panoplie de détecteurs dont les performances dans chaque situation seraient automatiquement apprises à partir d'exemples, il nous a fallu faire un choix. Après avoir considéré plusieurs méthodes, nous avons opté pour un algorithme d'*amdf* (*Average Magnitude Difference Fonction*) d'une écriture simple et qui fournit de bons résultats que se soit dans les domaines temporel [151, 9, 6, 79] ou spectral [125].

Le filtre passe-bas

Les valeurs de *f0* sont recherchées dans une bande "basse fréquence" de largeur généralement inférieure à 400 Hz. Un premier traitement consiste alors à éliminer du signal une grande partie des composantes "hautes fréquences" (cf fig 2.1), nous utilisons pour cela un filtre récursif d'ordre 1 d'équation :

$$\begin{cases} \hat{S}_t = S_t C + \hat{S}_{t-1}(1-C) \\ \hat{S}_0 = 0 \end{cases}$$

où $\hat{S}_t$ est la valeur du signal filtré au temps t et C le coefficient (inférieur à 1) donné par la relation liant la fréquence de coupure f_c et la fréquence d'échantillonnage f_e :

$$C = \frac{2\pi f_c}{f_e}$$

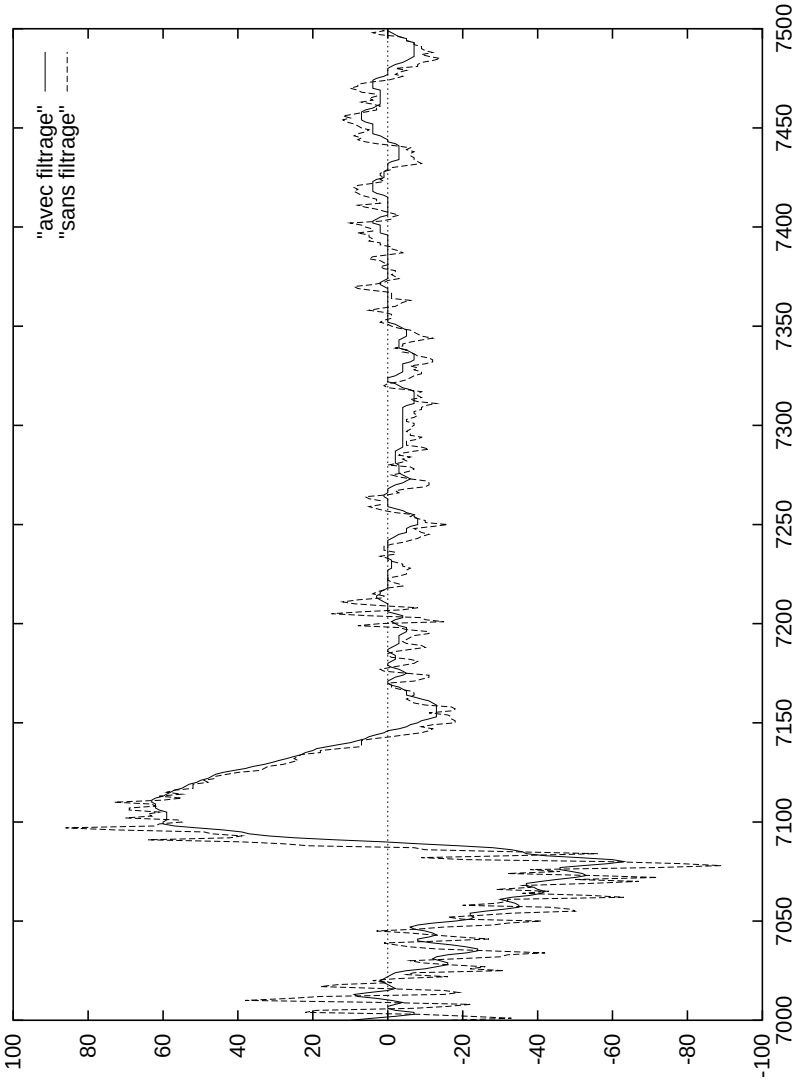


Figure 2.1: Exemple de signal avant et après filtrage.

L'*amdf*

Le principe est simple : si un signal est périodique, alors la différence du signal avec le même signal décalé temporellement d'une période est théoriquement nulle. Dans le cas d'un signal de parole, on peut mesurer sur une portion assez courte la pseudo-périodicité en recherchant le décalage qui minimise la différence du signal avec le signal décalé (différence exprimée par la fonction d'*amdf* $f(dec)$). On réalise pour cela un fenêtrage du signal de parole avec un décalage temporel adéquat (dans notre cas 10 ms). L'utilisation d'une fenêtre de Hamming de longueur t_h (calculée pour contenir au moins deux fois la période fondamentale maximale recherchée), permet une atténuation du signal aux bornes de chaque fenêtre, éliminant ainsi les effets "pervers" du fenêtrage.

$$Ham(i) = 0.54 - 046 * \cos \tfrac{2\pi i}{t_h}, \quad \text{avec} \quad i \in [0, t_h[$$

Le calcul — pour chaque fenêtre de signal — des valeurs d'*amdf* pour tous les décalages du signal envisagés, permet d'obtenir une courbe d'*amdf* (voir la figure 2.3) dont les minima correspondent normalement aux multiples de la fréquence fondamentale. On remarquera sur la figure 2.2 que bien que la précision du détecteur soit inversement proportionnelle à la fréquence (ceci étant dû à la non linéarité de la fonction $1/x$), les plages "usuelles" de *f0* sont correctement balayées par la méthode d'*amdf*.

$$f(dec) = \frac{\sum_{i:dec}^{t_h} |S_i - S_{i-dec}|}{t_h - dec}$$

La décision voisée/non voisée

Ces calculs de la fonction $f(dec)$ sont inutiles sur des portions non voisées du signal de parole. Une méthode habituelle empruntée par de nombreux détecteurs consiste à réaliser une décision voisée/non voisée à partir des valeurs de l'intensité du signal et de la densité de passage par zéro (*dpz*). Les fonctions d'*amdf* ne sont alors calculées que sur les fenêtres dont l'intensité est supérieure à un seuil (généralement fixé en fonction de la dynamique du signal) et où la *dpz* ne dépasse pas une valeur donnée. Si ce procédé obtient de bon résultats pour de la parole peu bruitée, ses performances sont médiocres dans le cas de signaux provenant du réseau téléphonique par exemple où la décision devient par trop dépendante des seuils. Nous lui préférons une méthode plus fiable qui nécessite cependant le calcul des fonctions d'*amdf* sur toutes les fenêtres du signal à analyser. La décision voisée/non voisée n'est plus binaire comme précédemment mais indique le degré de voisement de la fenêtre analysée. L'indice de voisement d'une fenêtre de signal est donné par la profondeur du plus grand pic dans la courbe d'*amdf* préalablement lissée et normalisée en énergie (voir la figure 2.3) [78]. Cette mesure rend compte de manière efficace des situations suivantes :

- Dans le cas d'un signal périodique, il y a répétition du schéma suivant : la fonction d'*amdf* atteint un maximum local lorsque le signal et le signal décalé sont en opposition de phase, puis décroît continûment vers un minimum local correspondant à un décalage entre les deux signaux d'un multiple de la période. La fonction croît ensuite avec la même régularité pour atteindre le maximum local suivant.

- Pour un signal non périodique, la fonction d'*amdf* présente une dynamique faible dénuée de toute régularité.

Nous montrerons dans le chapitre suivant des utilisations de cette courbe de voisement pour l'amélioration d'un module d'accès lexical. Notons que la décision du trait voisé/non voisé reste toujours un problème ouvert [142].

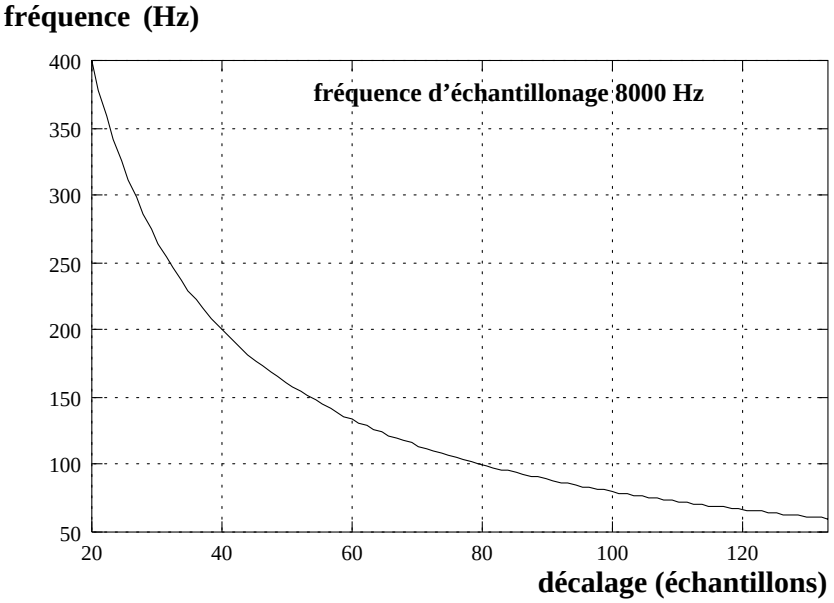


Figure 2.2: Fréquences associées aux décalages (exprimés en échantillons) envisagés par l'algorithme d'*amdf* pour une plage de recherche du fondamental entre 60Hz (133 échantillons) et 400Hz (20 échantillons) pour une fréquence d'échantillonnage de 8000Hz. On observe un étalement des valeurs dans les basses fréquences alors que les faibles décalages (20 à 30 échantillons) engendrent des sauts fréquentiels responsables d'une certaine imprécision dans les fréquences hautes de la plage de recherche du fondamental.

Obtention de la courbe de *f0*

Pour chaque fenêtre de signal périodique, nous retenons au plus les cinq plus faibles minima, bien que généralement, il semble que les deux ou trois premiers coefficients suffisent à déterminer la courbe de fréquence fondamentale. La courbe de fréquence fondamentale est alors déterminée par le calcul d'un chemin dans ces coefficients qui minimise les choix erratiques.
Soient $C_{r,t}$ la valeur du coefficient d'*amdf* de rang r au temps t (avec $r \in [0,5[$ dans notre algorithme), A le nombre de trames d'anticipation et Mo la moyenne sur toutes les portions voisées du signal du premier coefficient d'*amdf*. La recherche du chemin de *f0* sur chaque zone, se déroule en deux étapes :

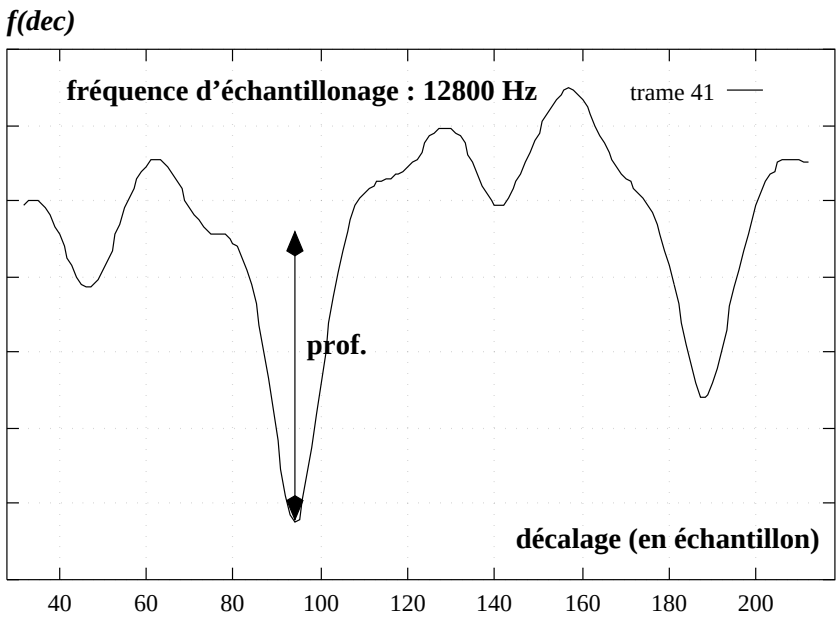


Figure 2.3: Exemple de courbe d'*amdf* obtenue pour un son voisé : pour chaque valeur du décalage (*dec* en abscisse) correspond la valeur de la fonction d'*amdf* ($f(dec)$). La flèche étiquetée *prof* indique la mesure du voisement par la profondeur du plus grand pic d'*amdf*.

- Détection d'un "point d'accrochage" par le calcul du couple (r_i, t_i) qui minimise l'écart à la moyenne Mo sur toute la zone voisée considérée.

- Extension à droite et à gauche à partir de C_{r_i,t_i} par "anticipation/correction" détaillée dans le cas d'une extension à droite :

  - Un premier chemin est calculé sur A trames par minimisation d'une trame à l'autre de l'écart avec la valeur précédente (noté $C_{r_i,t_i} \longmapsto C_{r',t_i+A}$).
  - On sélectionne le coefficient C_{r'',t_i+A} qui minimise la quantité $|C_{r_i,t_i} - C_{r'',t_i+A}|$.
  - Si C_{r'',t_i+A} et C_{r',t_i+A} sont proches, alors la valeur de *f0* retenue pour la trame $t_i + 1$ est donnée par le deuxième coefficient appartenant au chemin $C_{r_i,t_i} \longmapsto C_{r',t_i+A}$
  - Sinon la valeur au temps $t_i + 1$ est calculée par interpolation linéaire entre les coefficients C_{r_i,t_i} et C_{r'',t_i+A}.
  - Continuer en prenant la dernière valeur calculée comme nouveau point d'accrochage.

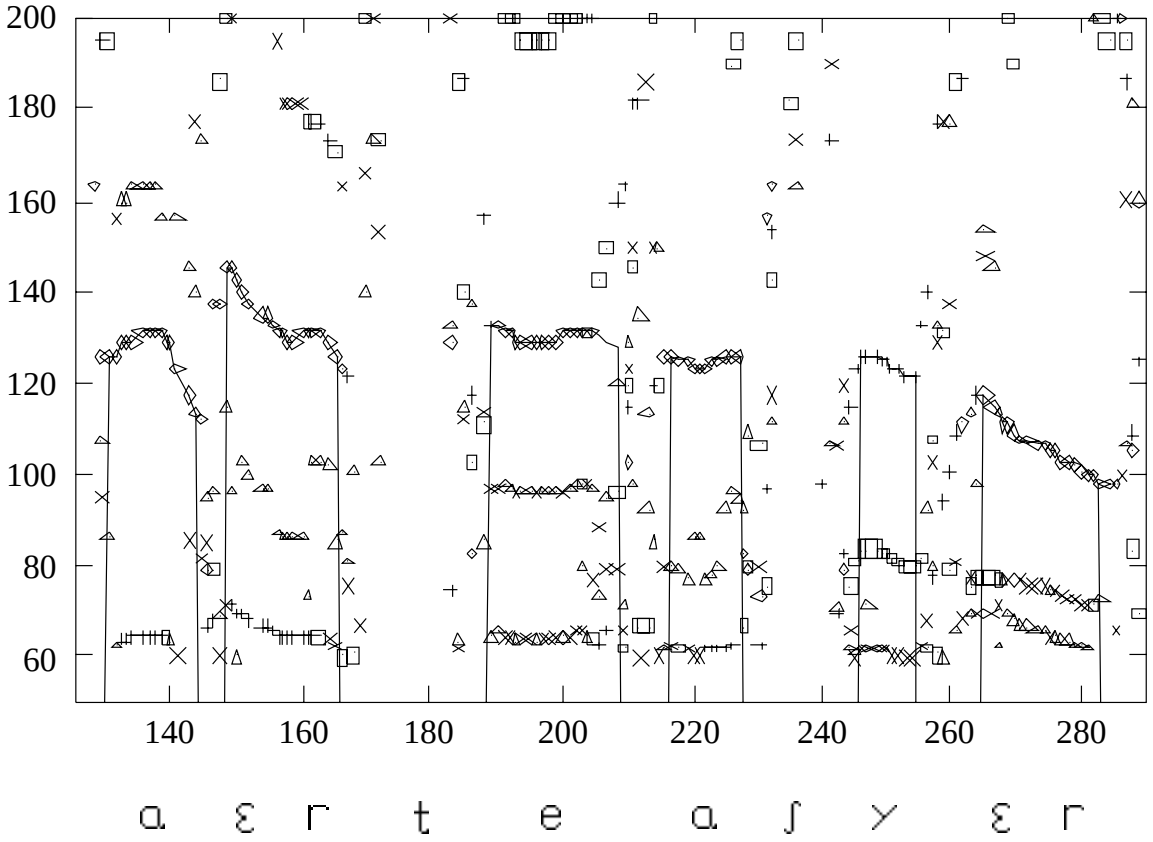


Figure 2.4: Courbe de *f0* et coefficients d'*amdf* calculés pour une portion de signal de la base PolyVar correspondant à l'épellation du mot "Arthur".

2.1.2 Mesure de la durée

Donner une mesure de la durée est un problème délicat qui nécessite en tout premier lieu de décider d'une unité de référence. Il semblerait aux vues de la littérature consacrée que l'unité syllabique recueille l'intérêt du plus grand nombre d'études (voir par exemple [33]) bien que certains auteurs présentent des unités alternatives comme le groupe interperceptual-center [10] ou les pseudo-syllabes [110] dont les limites sont repérables à partir de corrélats purement acoustiques. Deux raisons nous ont cependant conduit à retenir le noyau vocalique comme unité de référence : la première très pragmatique tient à la difficulté — pour le français au moins — d'opérer, dans une phase ascendante, une segmentation syllabique automatique fiable [110], la deuxième plus théorique s'appuie sur le fait que les noyaux vocaliques sont les centres privilégiés des indices prosodiques [40, p. 537].
Proposer l'alignement d'une chaîne phonétique donnée avec le signal de parole associé reste un problème ouvert. Si plusieurs techniques classiques permettent d'obtenir des résultats satisfaisants [6, 148], il faut cependant garder à l'esprit que cette même tâche demandée à plusieurs experts phonéticiens donne lieu à des variations assez sensibles

dans des zones de forte co-articulation [50], ceci s'expliquant par la nature continue du signal de parole. Dans notre travail nous avons donc considéré deux méthodes de calcul de la durée des phonèmes : un module d'accès lexical (SPEX) [11] s'appuyant sur un étage ascendant de décodage acoustico-phonétique [59] et une technique d'alignement par modèles phonétiques markoviens à probabilité d'émission continue [168]. Le mode de fonctionnement du système SPEX est décrit au chapitre 4 traitant du filtrage lexical aussi décrivons-nous maintenant la deuxième méthode.
Parmi les techniques largement employées en reconnaissance de la parole, les modèles de Markov cachés (HMM) connaissent depuis une dizaine d'années un grand succès [70, 7, 51]. Les sources de Markov offrent un formalisme probabiliste qui, appliqué à la reconnaissance de la parole, permet de modéliser des unités variées (phonèmes, diphones, triphones, polysons, syllabes, mots, *etc.*) en s'affranchissant — au moins partiellement — des problèmes rencontrés lors de la mise en œuvre de systèmes à base de connaissances. Notre propos n'est pas de décrire la théorie markovienne mais simplement de montrer comment elle peut être utilisée dans notre travail. Aussi renvoyons-nous le lecteur aux études qui décrivent les sources de Markov, leurs algorithmes d'apprentissage et de décodage [124, 168]. Nous nous contentons dans cette section de les considérer fonctionnellement comme un outil capable de réaliser un alignement d'une suite de symboles donnée (les phonèmes) avec une suite de vecteurs équidistants (un toutes les 10 ms) représentant le signal de parole.
Après avoir fait l'étude de plusieurs paramétrisations du continuum acoustique, nous avons choisi de présenter en entrée à nos modèles des vecteurs de 39 composantes régulièrement espacés toutes les 10 ms :

- 12 coefficients *Mfcc* (Mel-Frequency Ceptral Coefficient) — technique de paramétrisation largement répandue en raison de résultats semble t-il accrus comparés à ceux obtenus par d'autres méthodes — et un coefficient d'énergie.
- la dérivée première des 13 premiers coefficients qui correspond intuitivement à une prise en compte d'un "mouvement" dans la parole,
- et la dérivée seconde des mêmes 13 coefficients répondant à une notion intuitive de dynamique du mouvement.

Nous utilisons les algorithmes fournis par la boîte à outils *HTK* [168] qui propose de modéliser les probabilités d'émission de chaque état d'un modèle par des distributions de gaussiennes (souvent référencées *mixtures* dans la littérature).
Tous nos modèles (38 au total dont 3 modèles de silence, 15 pour les voyelles, 17 pour les consonnes et 3 modèles pour les semi-voyelles) possèdent la même topologie à trois états "émetteurs" (*i.e.* sans compter l'état initial et final), exception faite des modèles de silence pour lesquels plusieurs topologies ont été étudiées. Plus précisément chaque état émetteur est décomposé en trois parties ou *stream* (ce sont celles décrites précédemment), chacune d'elles étant modélisée par deux gaussiennes avec une matrice de variance-covariance complète (voir la figure 2.5).
L'apprentissage de ces modèles a nécessité trois étapes brièvement décrites :

L'initialisation des modèles : Une première estimation des moyennes et variances de chaque modèle est effectuée à partir d'une base segmentée phonétiquement. Chaque suite de vecteurs acoustiques correspondant à un phonème donné est extraite de la base et les vecteurs sont répartis uniformément sur chaque état du modèle correspondant où ses paramètres (moyenne et variance) sont alors calculés. Plus d'une centaine d'exemples de chaque phonème (10 fois répétés) ont permis une première estimation des paramètres de chaque état sans affecter leur probabilité de transition.

Ré-estimation : Sur la même base segmentée phonétiquement, une ré-estimation des moyennes et variances des 39 composantes pour chaque état est effectuée par l'algorithme de Baum-Welch répété 20 fois. Une première estimation des probabilités de transition est alors fournie. Une phase de test sur les données d'apprentissage permet à ce niveau de s'assurer du fonctionnement correct de nos modèles.

Ré-estimation automatique : Cette dernière étape utilise la totalité de la base d'apprentissage pour effectuer la ré-estimation de chaque modèle en parallèle. Il n'est plus utile de disposer à ce niveau d'un alignement phonétique du signal mais simplement de la suite de phonèmes lui correspondant. L'algorithme de Baum-Welch réalise alors les même ré-estimations que dans la phase précédente en considérant maintenant non plus chaque modèle isolément mais un modèle composite obtenu par concaténation des modèles des phonèmes composant le message. Cette phase coûteuse répétée plusieurs fois permet l'estimation définitive des probabilités de transition ou d'occupation des états, ainsi que des moyennes et variances associées à chaque *stream* de chaque état.

Un test de reconnaissance sur une base de 500 phrases représentant plus de 20 000 phonèmes nous permet de valider nos modèles. La reconnaissance est réalisée par un algorithme de Viterbi dont les entrées sont :

- un réseau de reconnaissance constitué de tous les modèles d'allophones en parallèle dans une boucle,
- le signal de parole paramétrisé.

Le calcul du meilleur chemin est effectué en prenant en compte la probabilité de transition d'un modèle vers un autre (*modèle bigramme*) appris sur la base d'apprentissage. Les résultats obtenus (par l'application des formules qui suivent) sont consignés dans la table 2.1 et la matrice de confusion obtenue est donnée en annexe. Comparés à d'autres systèmes du même genre [76],[49], ces résultats nous semblent tout à fait convaincants.

$$\begin{aligned} Correcte &= 100\% \times \frac{total - omission - substitution}{total} \\ Taux &= 100\% \times \frac{total - omision - substitution - insertion}{total} \end{aligned}$$

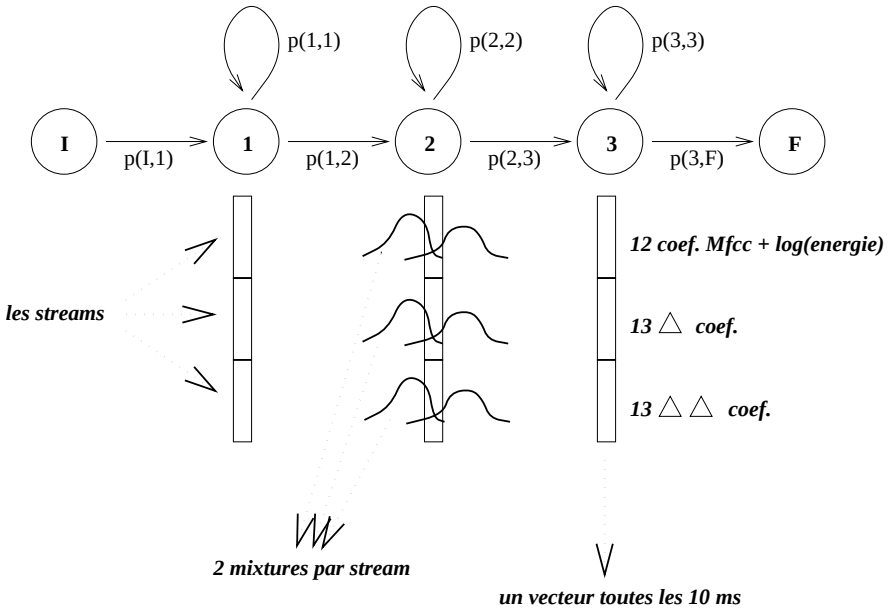


Figure 2.5: Topologie adoptée pour nos modèles de phonèmes à trois états émetteurs chacun modélisant les paramètres acoustiques divisés en trois composantes (coefficients *Mfcc*, dérivée première et dérivée seconde) chacune d'elle étant modélisée par deux gaussiennes.

Résultats	Pourcentage
Correcte	73.03%
Taux	62.52%

Prédiction	*Nombre*	*Pourcentage*
Correcte	15,486	73.03
Substitution	3,464	16.33
Omission	2,254	10.63
Insertion	[2,230]	[10.51]
Nombre d'échantillons	21,204	100

Table 2.1: Résultats obtenus par nos modèles de phonèmes indépendants du contexte.

2.1.3 Le paramètre d'intensité

Nous utiliserons dans ce travail une mesure classique de l'intensité (le carré de l'amplitude du mouvement vibratoire pris sur une fenêtre de taille fixée) pour chaque trame $j \in [0, fin[$:

$$Er0(j) \;=\; 10\ log_{10}\ (\sum_{i=0}^{t_h-1}(Ham(i)\ S^2_{(j\ t_h)+i}))$$

Nous en avons maintenant terminé avec la description de nos algorithmes d'extraction des paramètres de fréquence fondamentale, de durée et d'énergie. Nous allons donc procéder dans la section suivante à la description de notre système d'étiquetage prosodique.

2.2 Obtention automatique d'un étiquetage prosodique

Nous avons déjà mentionné (cf. 1.2) qu'une des difficultés de l'analyse prosodique est d'associer aux paramètres de fréquence fondamentale, de durée et d'énergie une représentation facilitant le traitement des informations qu'ils véhiculent. Nous avons également rappelé que la méthode retenue devait idéalement rendre compte des mécanismes de perception de ces paramètres avec toutes les difficultés que cela entraîne. La rapide revue qui suit présente les indices prosodiques retenus par différents auteurs dans des études récentes.
G. Caelen [21] a consacré ses travaux de thèse de doctorat d'état à l'étude des corrélations entre un faisceau d'indices prosodiques et différents modèles linguistiques (syntaxiques, sémantiques et pragmatique) en fonction de diverses consignes de lecture. Les indices qu'elle décrit sont au nombre de 33 (14 indices de fréquence fondamentale, 13 indices de durée et 6 indices d'énergie). En fait, les indices suivants sont initialement retenus — et détectés de manière semi-automatique — pour chaque paramètre puis réduits par deux codages à quatre niveaux (texte et phrase) multipliant ainsi le nombre d'indice par deux[2] :

7 indices de *f0* : gradient de *f0*, *f0* moyen et *f0* maximum sur le mot lexical et la syllabe finale du mot lexical ; gradient de *f0* sur une unité proche du contour (voir par exemple [39, p. 43] pour une définition du contour dans le groupe intonatif).

7 indices de durée : durée totale du mot, durée moyenne syllabique sur le mot, durée de la syllabe maximale sur le mot, gradient de durée, durée de la dernière syllabe, moyenne phonétique de la dernière syllabe et durée totale de la dernière syllabe augmentée de la durée de la pause subséquente (DSP).

3 indices d'énergie : gradient d'énergie dans le mot lexical, maximum et moyenne sur le mot.

[2] À l'exception d'un indice de durée (DSP) où seul le codage texte a été considéré.

Dans son système de génération automatique d'étiquettes prosodiques, Bailly et *al.* [6] utilisent une technique de décomposition temporelle pour réaliser l'alignement de la chaîne phonétique avec le signal. Ils distinguent alors 4 points particuliers qui leur serviront à la définition de 6 indices prosodiques par phonème :

OT : (Onset Time) défini dans le temps par l'abscisse du point d'intersection entre les fonctions d'interpolation (FI) de deux phonèmes adjacents[3].

OS : (Onset of Stationary part) défini par l'abscisse du point le plus à gauche de la FI qui est supérieur à 0.8.

ES : (End of Stationary part) défini par l'abscisse du point le plus à droite de la FI qui soit supérieur à 0.8.

NC : (Nucleus Center) défini par l'abscisse du centre de gravité de la partie stable de la FI.

Les indices alors retenus sont constitués de deux indices de durée – la durée du phonème (déterminée par deux OT adjacents) et la durée du noyau (ES-OS) – trois indices de fréquence fondamentale – valeurs de *f0* aux points SO, NC et ES – puis un indice d'énergie mesuré en NC.
Delais dans son étude des régularités rythmiques [45] utilise des indices de fréquence fondamentale et de durée qui lui permettent — sauf dans certains cas limites — de distinguer les accents régulateurs rythmiques (dont les indices localisés à l'initiale ou à l'antépénultième syllabe [115] sont : un mouvement montant de *f0* de faible ampleur et un éventuel allongement non significatif de la durée syllabique) des accents démarcatifs de constituants (dont les indices localisés sur la dernière syllabe sont une montée ou une descente assez ample de la courbe du fondamental et un allongement significatif de la syllabe). L'auteur précise cependant que la mesure de la durée de la syllabe est en fait assimilée à celle du noyau vocalique.
Bonin et Pierrel [16] présentent des indices de fréquence fondamentale et de durée pertinents pour la détection de frontières syntagmatiques. Ces indices booléens peu nombreux (présence d'une pause après un noyau vocalique, valeur de *f0* supérieure à un seuil (fonction adaptative de seuillage homographique), durée vocalique une fois et demi à deux fois supérieure à la moyenne, durée comprise entre deux et trois fois la moyenne et enfin durée vocalique plus de trois fois supérieure à la moyenne) leur permettent de définir par combinaison une grille de 16 classes de configurations d'indices dont ils évaluent ensuite la corrélation avec les frontières syntagmatiques.
Nasri et *al.* [110] proposent un ensemble de règles prosodiques utilisable en reconnaissance de la parole. Ils mesurent pour cela les paramètres prosodiques sur les noyaux vocaliques après avoir précisé la gageure que constituerait une segmentation syllabique dans une traitement automatique ascendant (le noyau vocalique est défini par le centre des voyelles

[3] Nous renvoyons le lecteur au nombreux travaux traitant de la technique de décomposition temporelle [3, 28, 12]

émergeant suffisamment de la courbe d'intensité étendue de part et d'autre). Les indices pertinents retenus sur chaque noyau vocalique (n) sont au nombre de neuf : le niveau d'énergie et de fréquence fondamentale (codage à quatre niveaux), valeur moyenne de *f0*, $Fo(n)$, la durée entre deux noyaux vocaliques ,$D(n)$, l'accélération/ralentissement du débit ($\delta D(n) = D(n) - D(n-1)$), $\delta D(n)$ codée sur 16 niveaux locaux à la phrase, la différence entre la ligne de déclinaison de la durée et $D(n)$ ($DLC(n)$ codée sur 8 niveaux) et enfin la différence entre la ligne de déclinaison de *f0* et $Fo(n)$.
Pour Vaissière [158], fournir une caractérisation de la courbe du fondamental de phrases lues en anglais nécessite trois étapes : la première qui consiste à segmenter sur la base du passage de *f0* par la ligne de déclinaison des unités qu'elle appelle HP (en référence aux fameux hat-patterns [146, 154]), une phase de marquage du degré de frontière entre les HPs sur la base d'indices classiques (insertion d'une pause, cassure de la ligne de déclinaison par nivellement ou ré-haussement généralement appelé resetting dans la littérature anglo-saxone), élévation ou diminution de la valeur d'un pic de *f0* d'un HP pour se démarquer du HP voisin, *etc.*), la troisième exprime le degré de jointure d'un même HP (estimé à partir de la rupture de *f0* sur le plateau).
Dans leur système de génération de corpus phonétiques et prosodiques, Pérennou et *al.* [119] proposent des règles de définition de marques abstraites à partir de la combinaison des indices booléens d'allongement, de présence de pause silencieuse et de registre du fondamental infra-aigu.
Dans la réalisation et l'exploitation de bases de données prosodiques à finalité de synthèse à partir du texte, Emerard et *al.* retiennent des indices constitués par les valeurs brutes de durée des phonèmes et des valeurs initiale, médiane et finale de voyelle [54, 53].
Niemann et *al.* [111] présentent quatre indices leur permettant de distinguer la nature déclarative, interrogative ou continuative des requêtes de demande d'horaire de trains :

- pente de la droite de régression rg de *f0* sur l'ensemble de la phrase,
- différence entre la valeur finale de *f0* et la valeur finale de rg,
- pente de la droite de régression rgl de *f0* sur la dernière portion voisée du signal,
- différence entre les valeurs finales de *f0* et de rgl sur cette portion.

Ils reportent un taux moyen de plus de 86% de reconnaissance sur une base de test de 90 phrases (30 interrogatives, 30 déclaratives et 30 continuatives) prononcées par 4 locuteurs. Ces indices sont alors utilisés dans leur système de compréhension et de dialogue EVAR. Ils précisent ensuite de manière sommaire qu'ils utilisent également la prosodie pour classer les frontières des phrases prosodiques à partir d'appréciations perceptives (très forte, forte, faible et non marquée) avec 22 indices dont ils fournissent une liste pour le moins incomplète : pente et forme de *f0*, débit, durée des syllabes. Ils reportent un taux de reconnaissance de 67% à l'aide d'un classificateur gaussien.
Nous avons passé sous silence un certain nombre d'études qui ne sont pas moins dignes d'intérêt [98, 24, 23], mais nous avons présenté ici une panoplie suffisamment large

d'approches décrivant l'extraction automatique d'indices prosodiques. Nous ne pourrions cependant pas terminer cette succincte revue sans parler de la très récente et intéressante étude de Wightman et Ostendorf [166] sur l'étiquetage prosodique automatique. Dans cet article, les auteurs présentent des résultats expérimentaux proches de ceux obtenus manuellement sur deux corpus lus par des professionnels. Un ensemble de 12 indices est fourni automatiquement pour chaque syllabe :

- forme de la *f0* sur la syllabe suivante (montante, descendante, montante-descendante et descendante-montante),
- présence ou pas d'une voyelle accentuée,
- syllabe finale d'un mot,
- durée de l'éventuelle pause suivant la syllabe analysée,
- allongement final,
- énergie médiane de la syllabe,
- rapport de la *f0* maximale sur la phrase avec la *f0* moyenne,
- rapport de la *f0* maximale sur la syllabe avec la *f0* moyenne de la syllabe suivante,
- rapport de la *f0* maximale sur la syllabe avec la *f0* maximale de la syllabe précédente,
- rapport des valeurs maximale et moyenne de *f0* dans la syllabe,
- rapport de la valeur minimale à la valeur moyenne de la *f0* dans la syllabe.

Ils présentent également les indices (localisés cette fois-ci sur les mots) qu'ils retiennent pour l'étiquetage des *break index* [123] que l'on peut définir comme l'indice de perception de la disjonction entre deux mots successifs (sur une échelle de 0 à 6) :

- présence ou pas d'une pause respiratoire audible (les auteurs reportent que les résultats obtenus après correction manuelle ne devraient pas différer énormément de ceux obtenus par une détection automatique),
- durée de la pause qui suit le mot éventuellement ajoutée à la pause respiratoire,
- durée normalisée sur la moyenne de la rhyme de la dernière syllabe d,
- différence de d avec le début de la syllabe,
- différence entre les durées normalisées des trois syllabes précédant la syllabe finale de mot avec les durées des trois syllabes suivant la syllabe finale,
- une valeur booléenne indiquant la présence ou l'absence d'une syllabe accentuée dans le mot considéré,

- la probabilité d'une limite tonale sur la dernière syllabe du mot obtenue à partir des indices intonatifs précédemment énoncés.

À la lumière de ces différents choix, nous pouvons maintenant décrire notre approche paramétrique. L'étiquetage prosodique proposé a pour entrée un signal de parole et pour sortie un treillis d'étiquettes valuées marquant les noyaux vocaliques détectés. Cet étiquetage prosodique ne nécessite aucune intervention manuelle, pas plus qu'il n'a besoin de seuil spécifique à un locuteur ou à un style d'élocution donné. Nous précisons dès maintenant que nous considérons seulement un sous-ensemble — somme toute classique — de la panoplie d'indices prosodiques disponibles. L'obtention d'un treillis prosodique ne constitue pas une fin, mais au contraire un moyen pour étudier la corrélation des indices prosodiques avec des niveaux de structuration linguistique comme la syntaxe. À ce titre, l'ajout d'un indice supplémentaire, pour autant qu'il soit automatiquement mesurable, ne pose pas de problème particulier, et ne nécessitera pas de programmation spécifique lors des traitements ultérieurs développés au chapitre 5.

2.2.1 Indices de fréquence fondamentale

Concernant le paramètre de fréquence fondamentale, un ensemble de 22 étiquettes a été retenu pour la caractérisation automatique des noyaux vocaliques de l'énoncé analysé.

MAX_FO : le maximum sur la courbe du fondamental restreinte aux seuls intervalles délimités par les noyaux vocaliques,

MIN_FO : le minimum de *f0* sur l'ensemble des valeurs prises par le paramètre sur les noyaux vocaliques,

EFO1 : émergence de la valeur moyenne d'un noyau vocalique par rapport aux valeurs moyennes des noyaux adjacents ($n-1$ et $n+1$),

EFO1' : émergence de la valeur au 2/3 d'un noyau vocalique par rapport aux valeurs prises au 2/3 des noyaux adjacents,

EFO2 : émergence de la valeur moyenne d'un noyau vocalique n par rapport aux valeurs moyennes des noyaux $n-2$ et $n+2$,

EFO2' : émergence de la valeur prise au 2/3 d'un noyau vocalique n par rapport aux valeurs prises au 2/3 des noyaux $n-2$ et $n+2$,

NIVAx **:** où $x \in [1,4]$, est le niveau *absolu* associé à chaque noyau vocalique en découpant horizontalement la dynamique du paramètre de fréquence fondamentale (mesurée sur la phrase) en quatre bandes de même largeur. Le choix du découpage en quatre niveaux s'appuie autant sur la pratique usuelle que sur des essais multiples sans changement significatif quant au nombre de niveaux,

NIVRx **:** où $x \in [1,4]$, est le niveau *relatif* associé à chaque noyau vocalique. Nous avons adopté cette méthode de codage pour pallier les inconvénients du précédent où deux noyaux vocaliques de valeurs très proches peuvent se voir attribuer deux niveaux différents selon leur position par rapport à la limite artificielle qui sépare les niveaux. Deux noyaux vocaliques ont alors des niveaux de *f0* différents uniquement si un écart d'au moins le quart de la dynamique du paramètre les sépare.

ENIVA1 : étiquette un noyau vocalique dont le niveau absolu est supérieur aux niveaux absolus des noyaux adjacents,

ENIVR1 : caractérise un noyau vocalique dont le niveau relatif est supérieur au niveaux relatifs des noyaux adjacents,

ENIVA2 : indique qu'un noyau vocalique n à un niveau absolu supérieur aux niveaux absolus des noyaux $n-2$ et $n+2$,

ENIVR2 : étiquette un noyau vocalique n dont le niveau relatif est supérieur au niveaux relatifs des noyaux $n-2$ et $n+2$,

+ : indique que la *f0* est à tendance montante sur le noyau vocalique considéré ; nous avons retenu comme mesure de la pente, le coefficient directeur a de la droite de régression calculée sur une zone étendue de quelques trames à droite et à gauche du noyau vocalique ($[d,f]$) :

$$\begin{cases} a & = & \frac{\frac{1}{f-d+1}\left(\sum_d^f x.f0(x)\right) - \overline{x}.\overline{f0(x)}}{\frac{1}{d-f+1}\left(\sum_d^f x^2\right) - \overline{x}^2} \\ \overline{x} & = & \frac{1}{f-d+1}\sum_d^f x \\ \overline{f0(x)} & = & \frac{1}{f-d+1}\sum_d^f f0(x) \end{cases}$$

– : indique la tendance décroissante du paramètre de fréquence fondamentale sur le noyau vocalique examiné,

= : note une configuration plane de *f0* d'un noyau vocalique (en pratique $a \in [-0.1, 0.1]$),

INFO_FO : cette étiquette contient des valeurs qui peuvent être utilisées à des fins diverses parmi lesquelles il y a la *f0* à l'initiale, au milieu et en finale de noyau vocalique ainsi que la moyenne du paramètre sur le noyau délimité ainsi que son gradient de *f0* ($\Delta f0$).

2.2.2 Indices de durée

Neuf étiquettes de durée sont retenues pour caractériser ce paramètre :

MAX_MS : étiquette la durée vocalique maximale,

MIN_MS : étiquette la durée vocalique minimale,

ED1 : indique qu'une voyelle est plus longue que ses voisines (droite et gauche),

ED2 : indique qu'une voyelle n est plus longue que ses voisines (droite $n+2$ et gauche $n-2$),

AL1 : indique qu'une voyelle dépasse de 20ms la moyenne des durées des voyelles dans la phrase,

AL2 : indique qu'une voyelle dépasse de 40ms la moyenne des durées des voyelles dans la phrase,

PAUSE : indique qu'une pause silencieuse d'au moins 200 ms suit le noyau vocalique étiqueté,

INFO_MS : étiquette informative contenant la durée de la voyelle considérée,

STAB : caractérise un noyau vocalique qui présente une stabilité spectrale très marquée ; cette étiquette accompagne souvent les noyaux allongés. La courbe de stabilité *Stab* est obtenue par lissage de la somme normalisée de la valeur absolue de la dérivée première de l'énergie $Er0$ avec une fonction d'instabilité qui mesure la similitude de l'enveloppe énergétique des 64 premiers canaux (des spectres LPC calculés sur 128 points) de deux trames distantes de 20 ms ($C_{i,j-1}$ et $C_{i,j+1}$, i $\in$ [1,64]) :

$$\begin{cases} Stab_j = & H\left(|Er0_{j-1} - Er0_{j+1}|\right) + H\left(\sum_{i=1}^{64} \delta_{ij}\right) \\ avec & \begin{cases} \delta_{ij} = 1 \quad si\,(C_{i,j-1} - C_{i+1,j-1}).(C_{i,j+1} - C_{i+1,j+1}) < 0 \\ \delta_{ij} = 0 \quad sinon \end{cases} \\ et & H\text{, fonction homothétique pour la mise à niveau lors de la sommation} \end{cases}$$

Ajoutons simplement que la moyenne prise en compte pour l'attribution des étiquettes **AL1** et **AL2** est obtenue simplement à partir des durées des voyelles telles qu'elles nous sont fournies par nos modèles phonétiques : aucune normalisation par rapport à des valeurs intrinsèques n'est ici effectuée. Rappelons que dans le cadre d'une phase ascendante de reconnaissance, la nature précise du segment vocalique n'est pas toujours une information fiable rendant pour le moins hasardeuse toute tentative de normalisation [16]. Le chapitre 4 apportera des justifications supplémentaires de ce choix.

2.2.3 Indices d'énergie

Huit étiquettes d'énergie s'ajoutent à celles déjà présentées pour constituer un total de 39 étiquettes caractérisant les trois paramètres :

MAX_ERO : indique le noyau vocalique qui possède la valeur d'énergie la plus forte,

MIN_ERO : indice le noyau vocalique qui possède la valeur d'énergie la plus faible,

EERO1 : émergence de la valeur moyenne de l'énergie d'une voyelle par rapport à ses voisines directes,

EERO1' : émergence de la valeur de l'énergie prise au 2/3 de la voyelle par rapport à ses voisines directes,

EERO2 : émergence de la valeur moyenne en énergie d'une voyelle n par rapport à ses voisines $n-2$ et $n+2$,

EERO2' : émergence de la valeur de l'énergie prise au 2/3 de la voyelle n par rapport à ses voisines $n-2$ et $n+2$,

EEN : émergence du niveau d'intensité d'un noyau vocalique par rapport aux noyaux précédent et suivant,

INFO_ERO : étiquette informative contenant la valeur du paramètre d'énergie à l'initiale, au milieu, au 2/3 et en finale de noyau vocalique ainsi que la moyenne et le gradient ($\delta F0$) sur le noyau.

Nous pourrions dès maintenant nous interroger sur le choix de ces paramètres, sur l'absence de certains ou sur la redondance d'autres, ou bien encore remarquer l'arbitraire d'une partie d'entre-eux ; mais nous avons déjà précisé que la caractérisation d'un paramètre continu était une opération délicate et critiquable, aussi nous contenterons-nous de rappeler que notre système reste ouvert à l'ajout d'autres étiquettes pour autant qu'elles soient automatiquement calculables. Nous laissons donc le soin à notre étude corrélative de souligner les redondances et/ou inadéquations éventuelles de certains de ces choix. Les figures 2.6 et 2.7 illustrent des exemples de treillis prosodiques automatiquement obtenus.

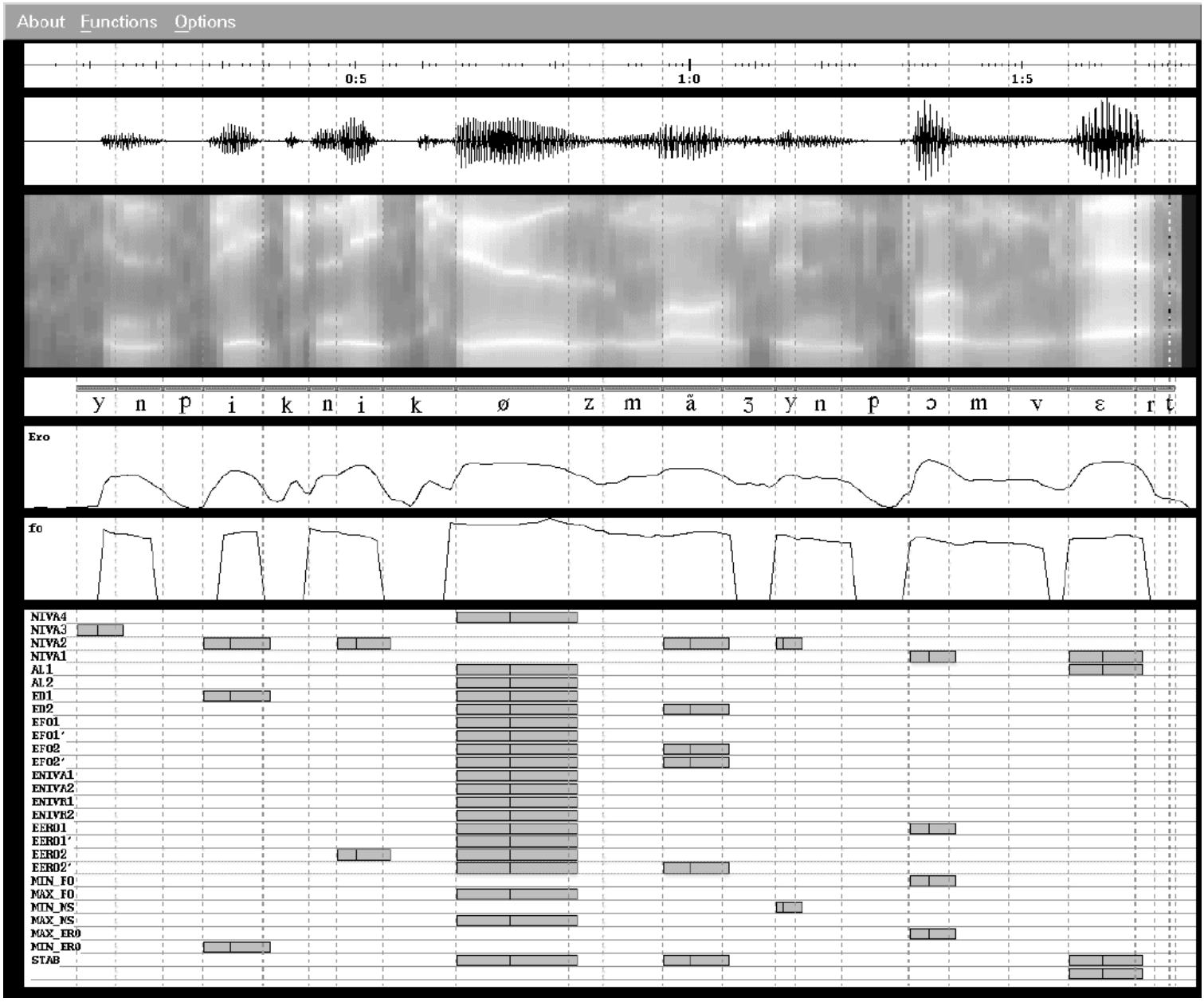


Figure 2.6: Exemple de treillis prosodique obtenu pour une réalisation de la phrase *une pique-niqueuse mange une pomme verte* via une ligne téléphonique.

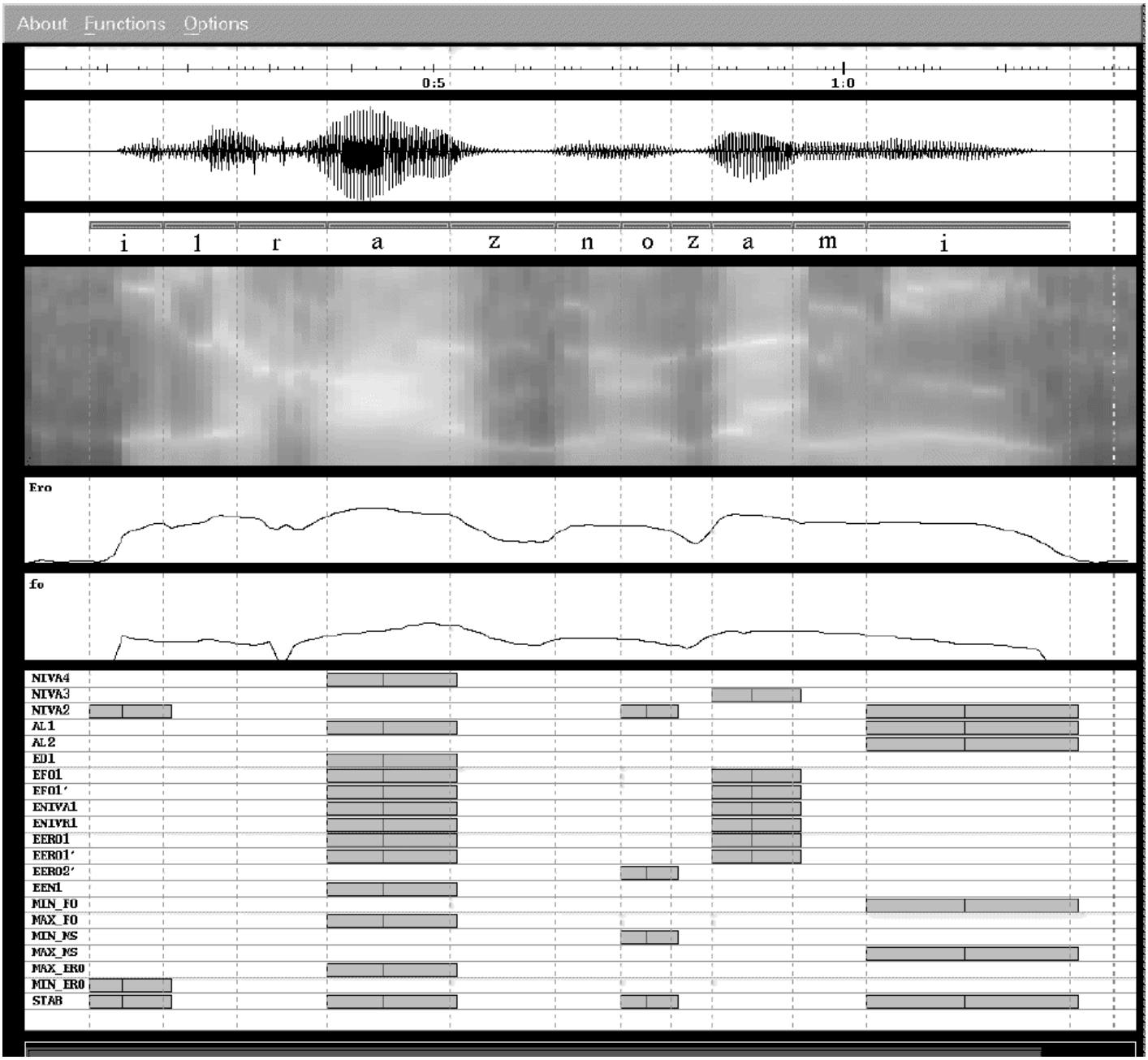


Figure 2.7: Exemple de treillis prosodique obtenu pour une réalisation de la phrase *il rase nos amis* via le canal téléphonique.

Chapitre 3

Les bases de données vocales

Ce court chapitre n'a d'autre prétention que d'introduire les différentes bases de données de parole utilisées dans ce mémoire.

3.1 Les bases de parole continue

3.1.1 PolyVar

Cette base de données est actuellement en cours d'acquisition à l'Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP) et a pour objectif de recueillir 100 appels téléphoniques (échantillonnés à 8000 Hz) pour 50 locuteurs volontaires. Bien qu'elle soit pensée pour répondre aux besoins spécifiques des recherches en reconnaissance et identification du locuteur (et plus particulièrement pour l'étude des variations intra-locuteur) nous aurons recours à cette base à plusieurs reprises pour nos expériences prosodiques. Chaque appelant reçoit quotidiennement (par courrier électronique ou sur support papier) une feuille qui contient les items qu'il sera chargé de prononcer (un exemple d'une de ces feuilles est donnée en annexe page 220). Chaque feuille d'appel contient :

- dix phrases sélectionnées pour leur richesse en unités segmentales variées (phonèmes, diphones, triphones et polysons) susceptibles d'être modélisées dans des systèmes de reconnaissance de parole,
- une vingtaine de mots isolés applicatifs,
- un nombre décimal,
- et une certaine quantité d'épellations, de séquences de chiffres, de montants, ...

Chaque appel (environ 5 minutes de parole) est ensuite écouté puis annoté orthographiquement via une interface graphique conviviale développée spécialement pour cette tâche (voir la figure 3.1).

Deux corpus extraits de la base PolyVar seront étudiés au cours du chapitre traitant de la prosodie en tant que canal linguistique : un corpus de nombres décimaux et un corpus de phrases isolées ; ce sont eux que nous décrivons maintenant.

3.1.2 PolyNombre

Au cours du chapitre 5 nous serons amenés à tester la validité d'un système d'apprentissage que nous présenterons. Nous aurons pour cela recours à une première série de tests sur des nombres. Deux corpus ont été réunis à cette fin : un corpus d'apprentissage et un corpus de test. Seule la disponibilité des données orthographiquement annotées a dirigé notre choix quant au contenu précis de ces deux corpus ; la répartition des nombres dans les deux corpus résulte d'un procédé complètement aléatoire. Notons simplement que ces deux corpus — bien que possédant des locuteurs et des nombres communs — sont d'intersection bien évidemment vide.

Le corpus d'apprentissage

Ce premier corpus réunit 500 nombres (décimaux pour la plupart) prononcés depuis de multiples téléphones[1] par 47 locuteurs différents dont un tiers de locutrices. La partie entière de ces nombres est inférieure au million et la partie décimale — le cas échéant — est inférieure au millier. La table 3.1 reporte les décomptes des différentes réalisations de ces locuteurs. Parmi ces locuteurs, quarante sont de langue maternelle française (dont douze suisses romands) les autres étant essentiellement de langue maternelle allemande ou encore suisse allemande. Seules les réalisations pour lesquelles aucun défaut de prononciation[2] n'a été signalé lors de l'étape d'annotation ont été retenues. La figure 3.2 présente les distributions des item du corpus PolyNombre en fonction de leur nombre de voyelles ainsi que la distribution des différentes structures "grammaticales". Au total 125 nombres différents ont été prononcés ; six d'entre eux étant des nombres entiers.

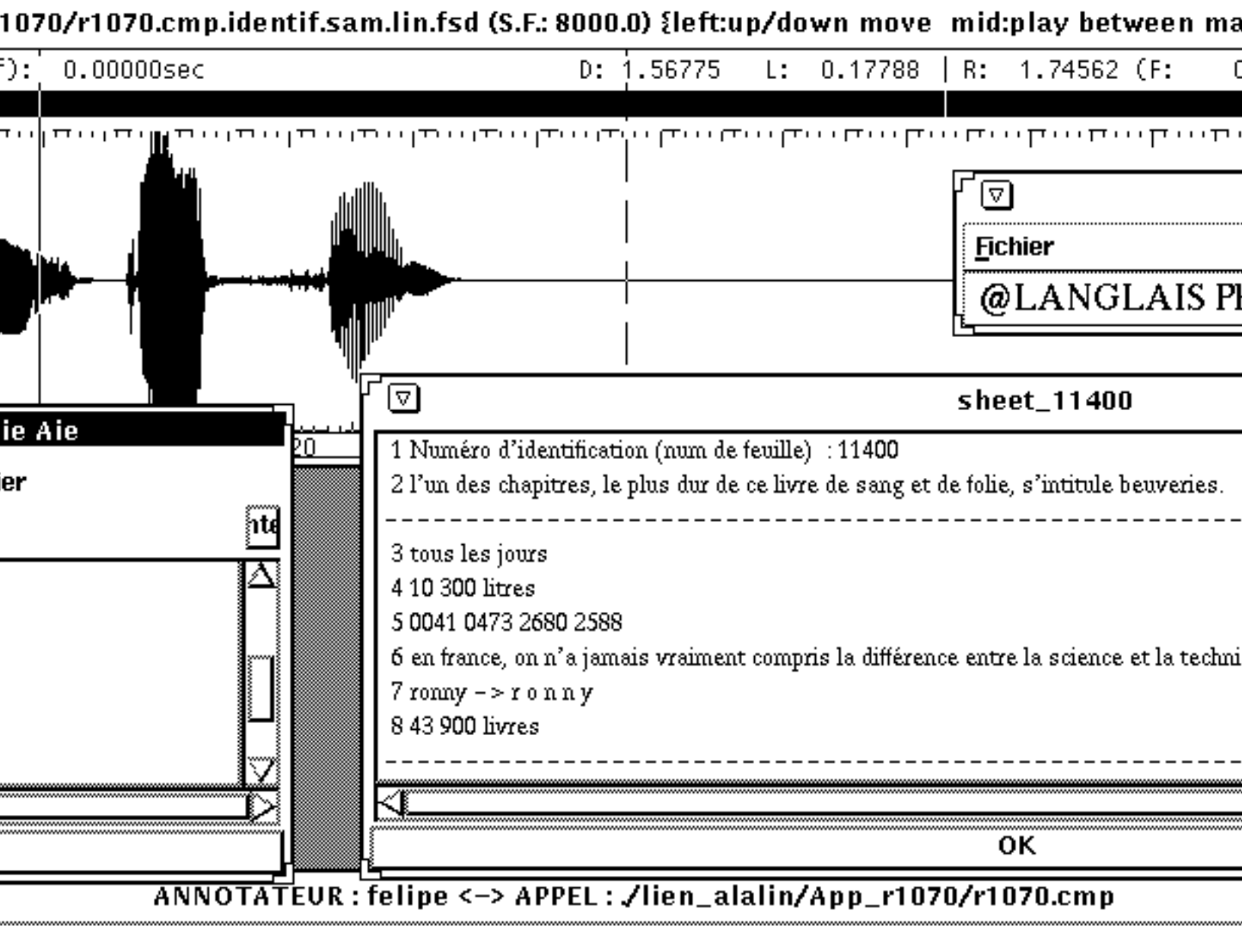
[1] Certains items — notamment ceux du locuteur MG — proviennent d'une cabine téléphonique.

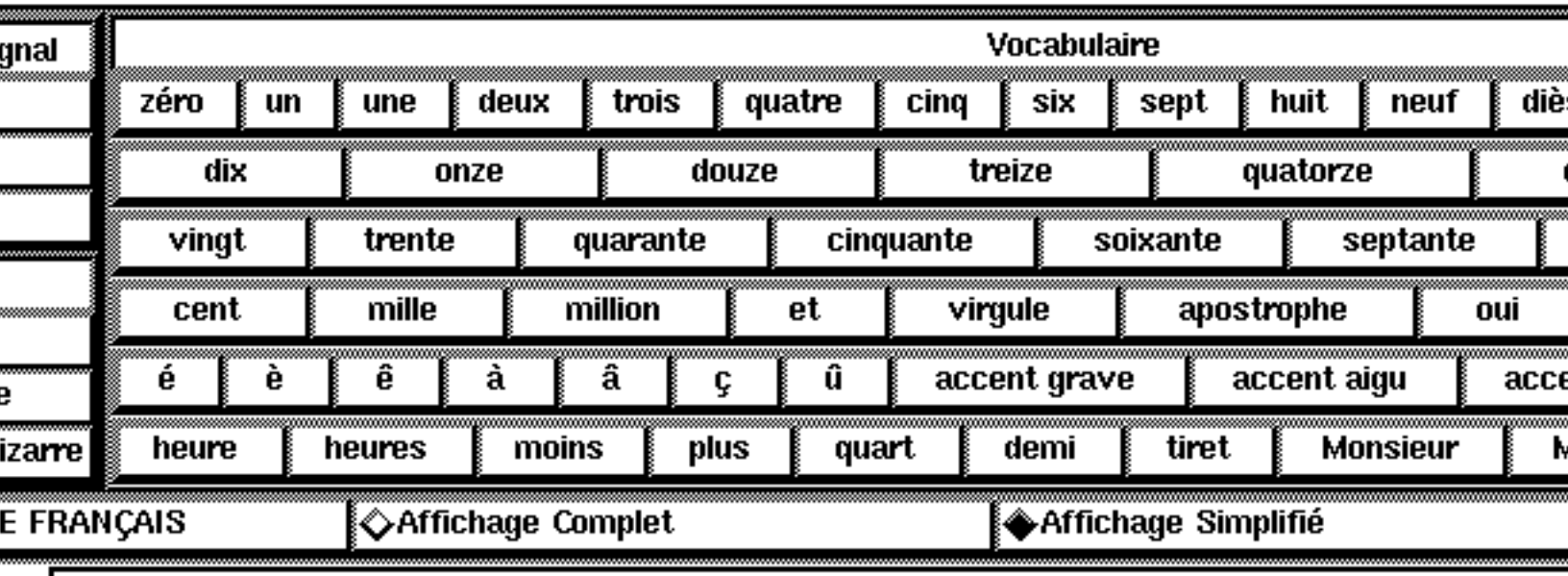
[2] Plus précisément le corpus retenu ne comprend pas d'hésitation, de bégaiement ou encore d'erreur/reprise ; il contient cependant des prononciations parfois peu fluides (dues à des locuteurs étrangers pratiquant peu la langue française) pour autant qu'elles soient intelligibles.

Locuteur	occ.	locuteur	occ.	locuteur	occ.	locuteur	occ.
AC	2	AS	6	BC	8	BV	1
CM	49	CA	2	CN	8	CC	1
FM	1	GM	23	GML	3	GL	1
LS	3	MJ	6	NI	1	OY	1
PI	3	SN	1	SB	1	ZA	1
20 locutrices $\Longrightarrow$ 122 réalisations							

Locuteur	occ.	locuteur	occ.	locuteur	occ.	locuteur	occ.
AH	50	AG	1	BR	15	BE	1
BS	5	BP	1	BO	23	BS	23
CG	51	CK	1	CJ	19	CCA	5
DN	1	DC	24	EJ	1	GD	1
KT	5	LP	50	LJ	3	MC	8
MG	20	ME	51	MM	1	OO	5
RP	1	VK	11	VP	1		
27 locuteurs $\Longrightarrow$ 378 réalisations							

Table 3.1: Décompte des réalisations des différents locuteurs féminins puis masculins de la base PolyNombre.


1070/r1070.cmp.identif.sam.lin.fsd (S.F.: 8000.0) {left:up/down move mid:play between ma
f): 0.00000sec D: 1.56775 L: 0.17788 | R: 1.74562 (F:
Fichier
@LANGLAIS PI
sheet_11400
1 Numéro d'identification (num de feuille) : 11400
2 l'un des chapitres, le plus dur de ce livre de sang et de folie, s'intitule beuveries.
3 tous les jours
4 10 300 litres
5 0041 0473 2680 2588
6 en france, on n'a jamais vraiment compris la différence entre la science et la techni
7 ronny -> r o n n y
8 43 900 livres
OK
ANNOTATEUR : felipe <-> APPEL : ./lien_alalin/App_r1070/r1070.cmp


Vocabulaire
zéro un une deux trois quatre cinq six sept huit neuf
dix onze douze treize quatorze
vingt trente quarante cinquante soixante septante
cent mille million et virgule apostrophe oui
é è ê à â ç û accent grave accent aigu
heure heures moins plus quart demi tiret Monsieur
Affichage Complet
Affichage Simplifié
Lisez le nombre (numéro d'identification)

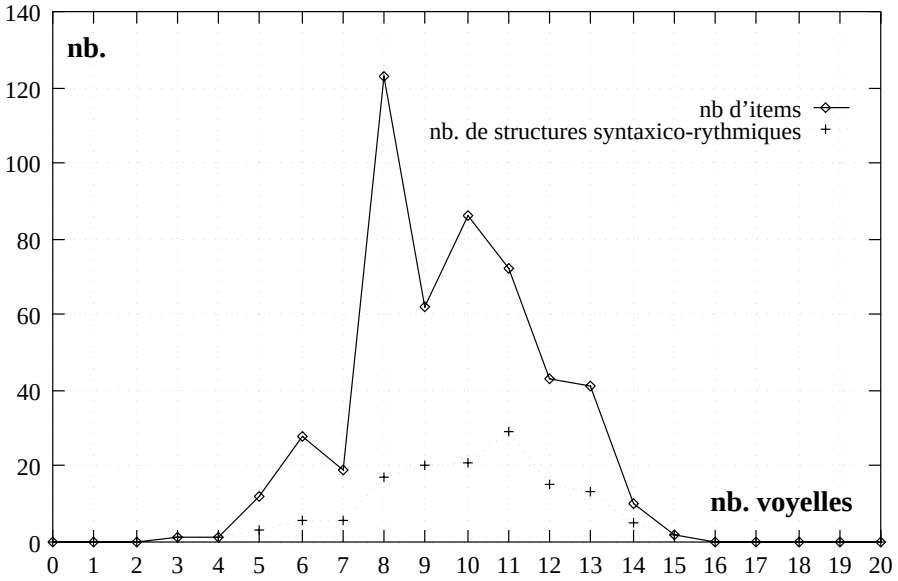


Figure 3.2: Nombre d'items et de structures syntaxico-rythmiques différentes du corpus PolyNombre en fonction du nombre de voyelles.

Le corpus de test

Ce corpus (désigné par la suite PolyNombreTest) contient 298 nombres prononcés par 35 locuteurs différents : quinze locutrices (dont quatre n'appartiennent pas aux locutrices de la base PolyNombre) et vingt locuteurs (dont six n'étaient pas présents parmi les locuteurs du corpus d'apprentissage). La table 3.2 rappelle le nombre de réalisations de chacun des locuteurs de la base. Des 298 nombres du corpus de test, seulement 115 sont différents dont 112 sont des nombres décimaux (soit à peu près les mêmes proportions que dans le corpus d'apprentissage). Il est intéressant de noter que parmi ces 115 nombres, 77 ne sont prononcés par aucun locuteur du corpus PolyNombre. À l'instar du corpus d'apprentissage nous indiquons en figure 3.3 les distributions des items du corpus de test en fonction de leur nombre de voyelles ainsi que la distribution des différentes structures "grammaticales".

Locuteur	occ.	locuteur	occ.	locuteur	occ.	locuteur	occ.	Locuteur	occ.
AC	19	AS	1	BM	1	CA	5	CM	35
CN	1	CZA	11	GM	4	GL	1	MJ	4
MA	2	NI	3	SB	6	WB	2	ZA	4
15 locutrices $\Longrightarrow$ 99 réalisations									

Locuteur	occ.	locuteur	occ.	locuteur	occ.	locuteur	occ.	Locuteur	occ.
AH	21	BR	6	BO	20	BS	4	CG	21
CJ	16	DL	1	DC	10	EJ	1	GD	1
JC	13	KT	1	LP	30	MG	11	ME	32
MK	1	MT	3	OO	5	SS	1	WB	1
20 locuteurs $\Longrightarrow$ 199 réalisations									

Table 3.2: Décompte des réalisations des différents locuteurs féminins puis masculins de la base PolyNombreTest.

3.1.3 PolyPhrase

Une autre base de parole continue sera également utilisée au chapitre 5. Comme pour les nombres, deux bases ont été extraites de la base PolyVar : une base réservée à l'apprentissage et une base destinée aux tests. Ces deux bases contiennent des phrases de structures grammaticales simples sélectionnées parmi l'ensemble des phrases déjà annotées orthographiquement ; la répartition des phrases entre les deux corpus s'est déroulée aléatoirement, l'intersection des deux corpus étant bien sûr vide en terme de réalisations[3].

[3]Ces corpus ont cependant de nombreux locuteurs en commun et la majorité des phrases retenues sont communes aux deux corpus

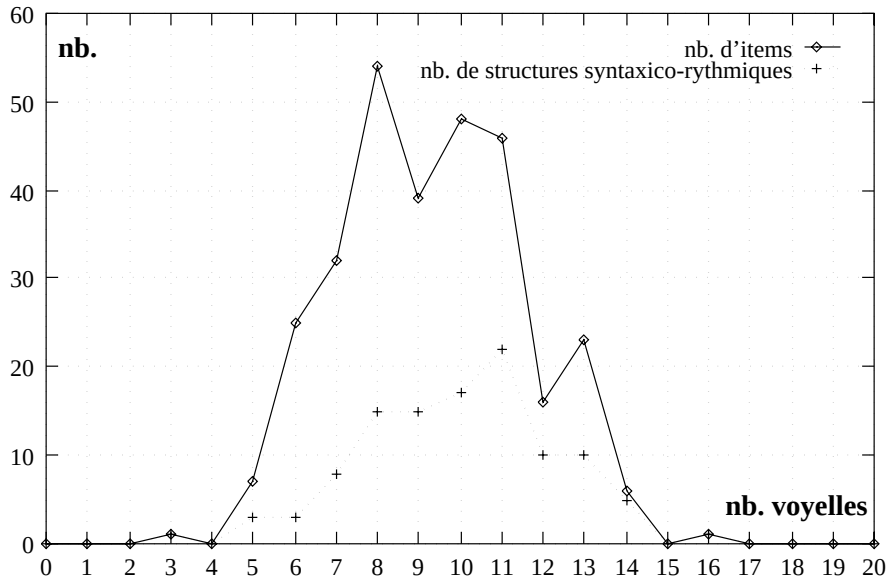


Figure 3.3: Nombre d'item et de structures syntaxico-rythmiques différentes du corpus PolyNombreTest en fonction du nombre de voyelles.

le corpus d'apprentissage

500 phrases ont été retenues pour les besoins de l'apprentissage. Les arbres syntaxiques nécessaires à la phase d'apprentissage ont été créés manuellement, aussi les phrases sélectionnées sont-elles de nature grammaticale simple et en nombre relativement réduit afin d'assurer le maximum de cohérence à cette opération somme toute assez fastidieuse. La liste des phrases retenues ainsi que les arbres syntaxiques les accompagnant sont fournis en annexe (pages 231 à 244), on peut cependant préciser que le nombre de phrases différentes du corpus PolyPhrase est de 80. Le corpus est prononcé par 50 locuteurs dont 21 sont des locutrices ; la table 3.3 reporte le nombre de réalisations de chacun d'eux. Parmi ces locuteurs 38 sont de langue maternelle française (18 suisses romands et 20 français) , 6 sont de langue maternelle allemande (ou suisse allemande) et 6 de langue maternelle diverses. Comme pour les corpus de nombres, seules les phrases ne possédant pas d'hésitation ou d'erreur de prononciation nuisant à leur intelligibilité ont été conservées. La figure 3.5 nous renseigne sur le nombre de phrases et de structures syntaxico-rythmiques du corpus d'apprentissage en fonction de leur nombre de voyelles.

Le corpus de test

Le corpus de test PolyPhraseTest est lui constitué de 348 phrases prononcées par 47 locuteurs : 22 femmes (dont 6 ne sont pas présentes dans la base d'apprentissage) et 25 hommes (parmi lesquels 8 n'étaient pas dans le corpus d'apprentissage) ; 34 d'entre eux sont de langue maternelle française (18 français et 16 suisses romands) et 8 sont de langue maternelle allemande ou suisse allemande. Le nombre de phrases prononcées par cha-

Locuteur	occ.	locuteur	occ.	locuteur	occ.	locuteur	occ.	Locuteur	occ.
AC	11	AS	11	BC	4	CA	5	CM	42
CN	7	CZA	10	CC	1	DD	1	GM	9
GML	4	MJ	5	LS	2	MA	2	NI	7
PI	1	SB	6	VM	1	WiB	2	WoB	3
ZA	8								
21 locutrices $\Longrightarrow$ 142 réalisations									

Locuteur	occ.	locuteur	occ.	locuteur	occ.	locuteur	occ.	Locuteur	occ.
AH	33	BR	8	BE	1	BaS	4	BO	31
BrS	13	CP	1	CG	36	CK	1	CJ	16
CCA	1	DC	15	EJ	1	GD	1	GC	1
HJ	2	HM	2	JC	8	KT	24	LP	48
LJ	2	MC	3	MG	26	ME	51	MT	13
OO	3	SS	1	VK	11	ZZ	1		
29 locuteurs $\Longrightarrow$ 358 réalisations									

Table 3.3: Décompte des réalisations des différents locuteurs féminins puis masculins de la base PolyPhrase.

cun est décrit dans la table 3.4. Au total 84 phrases différentes sont prononcées, cinq d'entre-elles n'appartenant pas au corpus d'apprentissage. Le nombre de phrases et de structures syntaxico-rythmiques différentes en fonction du nombre de voyelles est reporté sur la figure 3.5.

3.2 Les bases de mots isolés

3.2.1 AviLex

Cette base de parole a été conçue au Laboratoire Informatique d'Avignon et des Pays de Vaucluse (LIUAPV) pour la mise au point et les tests d'un module d'accès lexical [11] que nous emploierons par la suite. Elle est constituée de cinq locuteurs (4 hommes / 1 femme) qui prononcent isolément 500 mots tirés au hasard dans le lexique BdLex [117]. Deux de ces cinq locuteurs ont également prononcé une autre série de 700 mots également issus du même lexique. L'enregistrement de ces 3900 mots s'est déroulé dans des conditions dites de laboratoire (parole peu bruitée) via un micro connecté à une carte d'acquisition Oros (échantillonnage à 12,8 kHz). La liste des mots est présentée en annexe (pages 226 à 230) et la figure 3.6 précise les distributions des mots en fonction du nombre de voyelles pour les deux séries de 500 et 700 mots. Nous nommerons occasionnellement par la suite la série de 500 mots par AviLex1 et celle de 700 mots par AviLex2.

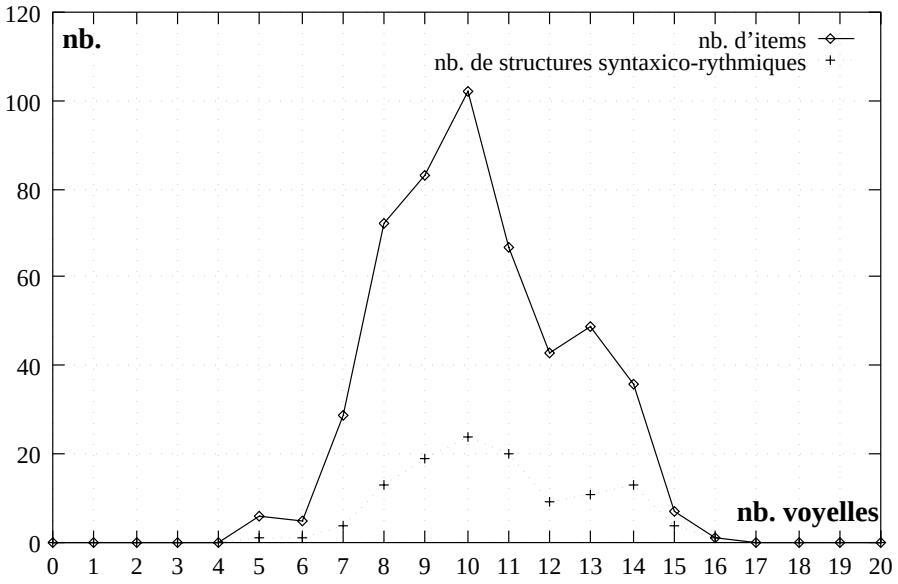


Figure 3.4: Nombre d'items et de structures syntaxico-rythmiques différentes du corpus PolyPhrase en fonction du nombre de voyelles.

Locuteur	occ.	locuteur	occ.	locuteur	occ.	locuteur	occ.	Locuteur	occ.
AC	10	AS	8	BN	1	BC	1	BV	1
CM	23	CN	4	CZA	6	DS	1	GM	9
GML	2	LQS	1	LS	1	MJ	7	MM	1
MA	1	NI	3	PI	2	SN	1	SI	3
SB	12	ZA	4						
22 locutrices $\Longrightarrow$ 102 réalisations									

Locuteur	occ.	locuteur	occ.	locuteur	occ.	locuteur	occ.	Locuteur	occ.
AH	25	BR	13	BaS	1	BF	1	BO	10
BrS	9	CG	34	CJ	12	CCA	2	DN	2
DC	16	EJ	1	GC	1	JC	3	KT	16
LP	31	MC	4	MG	14	ME	25	MK	2
MT	12	OO	2	SS	1	VK	8	VP	1
25 locuteurs $\Longrightarrow$ 246 réalisations									

Table 3.4: Décompte des réalisations des différents locuteurs féminins puis masculins de la base PolyPhraseTest.

3.2.2 PVM

Ce corpus est constitué des mots isolés de la base PolyVar qui étaient disponibles au moment de l'étude. Il est composé de 3364 répétitions de 116 mots différents prononcées par 37 locuteurs. Les occurrences de chaque mot sont hétérogènes : un sous-ensemble de 17 mots (prononcés en vue de la réalisation d'un serveur vocal interactif) est représenté 2329 fois dans le corpus PVM. La liste des mots est fournie en annexe (page 230).

3.2.3 AviTel

Afin de pouvoir comparer les analyses faites sur AviLex et PVM, nous avons collecté une base de parole téléphonique constituée de deux locuteurs qui ont prononcé une fois chacun les 500 mots isolés de la base AviLex1.

3.2.4 FeLex

Au cours de nos recherches — et plus particulièrement lors de l'étude des variations microprosodiques des paramètres de durée, de fréquence fondamentale et d'énergie — nous avons été amenés à enregistrer des échantillons de parole à des fins biens précises. Le corpus FeLex est constitué de 800 mots prononcés isolément par deux locuteurs sur ligne téléphonique. Ces mots — tous trisyllabiques — ont été sélectionnés de manière à obtenir une homogénéité du nombre des représentants des voyelles hautes, basses, moyennes et nasales dans les contextes consonantiques droits voisés et non voisés pour les trois positions syllabiques (initiale, médiane et finale de mot). La sélection a été effectuée par programmation à partir du lexique BdLex et le nombre de représentants des différents contextes (le détail est donné dans la table 3.5) avoisine la centaine. La liste des mots retenus est reportée en annexe (pages 223 à 225).

Voyelles	Positions					
	Initiale		Médiane		Finale	*toutes*
	V	NV	V	NV		
Hautes	100	100	101	101	184	586
Basses	100	100	100	100	201	601
Moyennes	100	100	101	102	207	610
Nasales	100	100	94	101	208	603
toutes	400	400	396	404	800	2400

Table 3.5: Répartition des cardinalités des différents contextes en fonction du contexte droit (voisé V ou non voisé NV) et de la position de la voyelle dans le mot

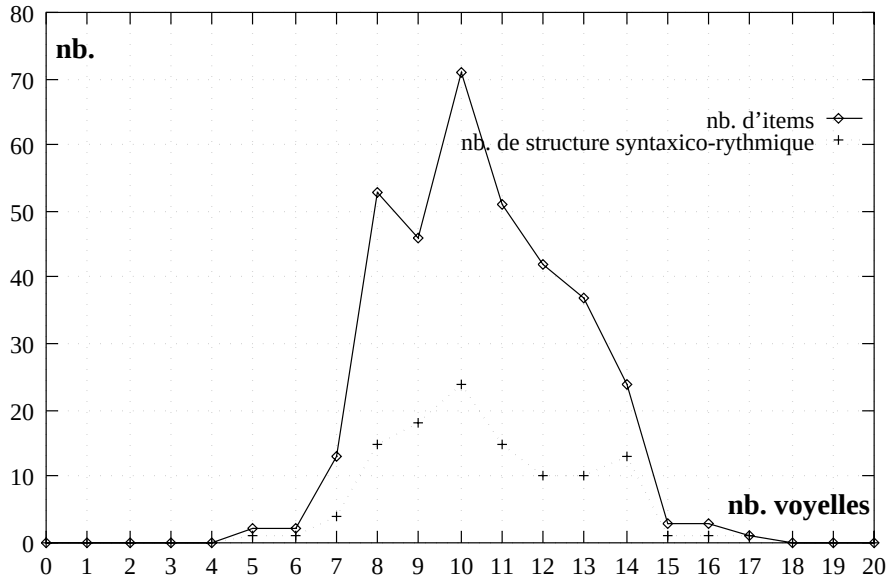


Figure 3.5: Nombre d'items et de structures syntaxico-rythmiques différentes du corpus PolyPhraseTest en fonction du nombre de voyelles.

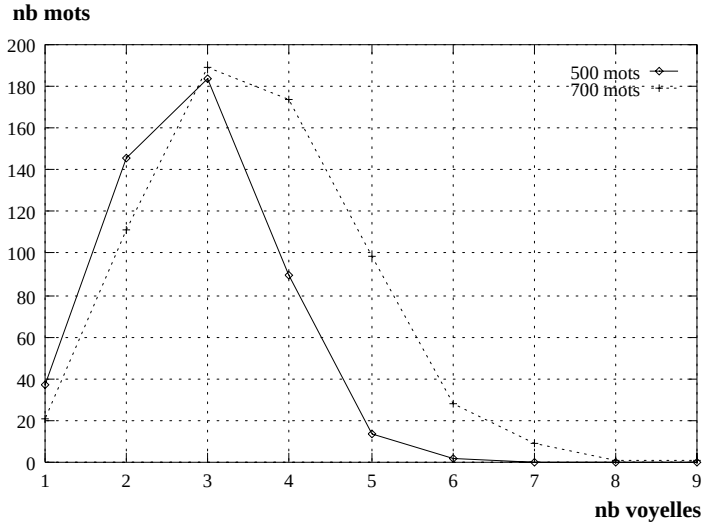


Figure 3.6: Distribution du nombre de mots des corpus AviLex1 et AviLex2 en fonction du nombre de voyelles.

Chapitre 4

Prosodie et filtrage lexical

De nombreuses études attestent le rôle essentiel de la prosodie dans le processus de communication et principalement dans les domaines émotionnel, pragmatique, sémantique et syntaxique [37]. Cela ne signifie cependant pas que la prosodie n'intervient pas à des niveaux inférieurs.

On a en effet souvent prêté à la prosodie un rôle non négligeable pour le décodage du flux de parole en unités segmentales [157]. Waibel rappelle par exemple des résultats obtenus très tôt par Blesser [14] qui précise que l'altération des informations spectrales ou de la durée des segments seule, n'entraîne pas une perte d'intelligibilité importante, mais qu'en revanche la perturbation des deux canaux simultanément dégrade la parole entraînant des taux de reconnaissance du message émis de l'ordre de 10%.

Di Cristo dans un document riche en informations a mis en relief les variations microprosodiques en français [40] que plusieurs études avaient déjà soulignées pour d'autres langues comme l'anglais[1] [80].

Nakatani et Schaffer [109] dans une étude maintenant classique, se sont employés à mettre en évidence la pertinence de l'information prosodique en langue anglaise pour accéder au lexique, en demandant à des sujets d'identifier des frontières de mots dans des portions de phrases — toujours composées de trois syllabes — où l'information phonétique a été supprimée par une technique"réitérante". Leur conclusion était alors que les sujets étaient capables d'identifier les frontières lexicales sur les seules informations prosodiques mieux qu'aléatoirement, et que le rythme était le facteur prédominant dans la décision du découpage lexical. Très récemment, dans une étude sur la segmentation de mots dissyllabiques contenant deux monosyllabes enchâssées, chacune prise isolément ayant un sens, Banel et Bacri ont démontré le rôle essentiel des schémas de durées pour la perception des limites des mots [8]. Elles notent que le schéma usuel iambic (court – long) entraîne le plus souvent une décision dissyllabique (indépendamment de la fréquence de la monosyllabe initiale) alors que le schéma trochaïque (long – court) augmente fortement le nombre de segmentations.

Dans une étude encore actuelle, Waibel [162] a démontré — pour l'anglais — qu'il était

[1] Voir [40] pour une liste détaillée des études microprosodiques existantes (pages 42, 180, 366–369 et 475–476).

possible de réduire de manière significative le nombre de mots candidats d'un processus d'accès lexical, par la prise en compte de modules de connaissances suprasegmentaux qui permettent de proposer des cohortes de mots autrement qu'en réalisant un décodage de la chaîne phonétique ; et ce aussi bien pour de la parole continue que pour une application en mots isolés. Il montre en particulier le rôle important joué par la durée (ratio de non voisement à l'intérieur de la syllabe, schémas rythmiques, ...). Chaque source de connaissance est pré-compilée sous forme d'un dictionnaire et seuls les mots qui satisfont les contraintes pré-codées sont retenus comme candidats potentiels.
Notre intime conviction est qu'une telle procédure — dans le cadre d'une application de mots prononcés isolément — ne peut-être mise en place avec des résultats aussi probants pour une langue comme le français. Notre langue possède en effet la caractéristique (commune à de nombreuses langues) de ne pas posséder d'accentuation fonctionnelle au niveau du mot comme c'est le cas en anglais (*ex.* : le verbe *per*mit et le mot *per*mit qui ne peuvent être différenciés que sur la base de la place de l'accent). Ce qui se traduit — dans notre langue — par un faible nombre d'études sur le filtrage lexical à l'aide d'indices prosodiques.
Il nous appartenait cependant de vérifier cette hypothèse en étudiant le potentiel d'indices prosodiques segmentaux et suprasegmentaux pour une tâche de filtrage lexical de mots énoncés isolément.

4.1 Objectifs

Le but principal de ce chapitre est de proposer une étude sur les possibilités et les limites de l'intégration d'informations prosodiques au sein d'un processus de filtrage lexical de mots prononcés isolément. Nous étudierons tout d'abord le potentiel filtrant des informations prosodiques relevant d'une organisation suprasegmentale en commentant les résultats de l'étude de Waibel [162] pour l'anglais puis en présentant quelques expériences simples visant à montrer la fragilité de ces informations dans un cadre aussi restreint que les mots isolés. Nous proposerons ensuite une étude détaillée des informations microprosodiques véhiculées par les paramètres de durée, d'intensité et de fréquence fondamentale. Nous dresserons pour cela l'inventaire des indices mis en évidence par nos prédécesseurs afin d'en mesurer la pertinence lorsqu'ils sont appréhendés par des techniques entièrement automatiques. Nous terminerons cette étude par la présentation des filtres prosodiques mis en place en précisant leur efficacité dans le processus de reconnaissance.

4.2 Le module d'accès lexical

Avant d'entreprendre l'exposé des expériences réalisées nous allons décrire brièvement le système d'accès lexical (SPEX) développé au LIUAPV [11]. Ce dernier s'inscrit dans une approche "fondée sur les connaissances" et se réclame indépendant des applications auxquelles on le destine. Il n'est donc pas tributaire d'une phase d'apprentissage spécifique souvent longue et délicate *a contrario* d'une grande partie des systèmes actuels

— appréciés entre autre pour leur résultats — qui mettent en œuvre des techniques statistiques (markoviennes le plus souvent, neuromimétiques également ou encore hybrides pour les plus récentes). Bien que fonctionnant de manière satisfaisante en mode multi-locuteur, SPEX dispose d'une phase d'apprentissage réduite pseudo-automatique consistant à recueillir les références spectrales — pour un locuteur donné — de chaque phonème de notre langue. Il est alors demandé au locuteur de prononcer une série de 34 mots qui ont été choisis afin de contenir l'ensemble des phonèmes du français — chacun en trois exemplaires — dans des contextes peu déformants. La référence d'un phonème est alors calculée comme la moyenne des trois occurrences retenues et est constituée de :

- un spectre représenté sur 24 canaux répartis sur une échelle de Mel,
- un spectre calculé sur 128 points,
- certaines valeurs mesurées lors de l'acquisition (énergie maximum, énergie minimum, *etc.*) afin de garantir l'homogénéité de l'apprentissage et de normaliser certaines valeurs de paramètres.

Nous pouvons décomposer le système d'accès lexical en deux étapes principales décrites maintenant que sont le décodage acoustico-phonétique et la composante lexicale.

4.2.1 Le niveau acoustico-phonétique

L'unité minimale de décodage est le phonème qui reste encore l'unité de base la plus utilisée dans les systèmes existants, principalement en raison de leur nombre réduit dans une langue donnée (entre 30 et 60) en comparaison à d'autres unités comme le diphone, la syllabe, le triphone ou encore le polyson qui permettent de s'affranchir — au moins partiellement pour certaines — des phénomènes co-articulatoires importants dans l'acte de parole.

Segmentation des zones vocaliques et consonantiques

Le module de DAP[2] [59] cherche à identifier des voyelles (respectivement des consonnes) sur les portions de signal étiquetées vocaliques (respectivement consonantiques). La méthode de segmentation employée consiste à observer l'évolution temporelle de la courbe d'énergie du signal de parole, qui permet généralement de distinguer les zones vocaliques — sur les maxima relatifs du paramètre — des zones consonantiques. Un paramètre spécial (EnDap) est utilisé pour cette tâche et permet de remédier aux principaux défauts d'une telle méthode à savoir :

- La présence de pics parasites pour certaines consonnes dans certaines bandes de fréquences.

[2]Nous ne présentons ici que la phase ascendante du module de décodage acoustico-phonétique développé au LIUAPV.

- Dans un contexte intervocalique fermé, une consonne sonnante ne se manifeste pas toujours par un abaissement de la courbe d'énergie.
- Des pics énergétiques peuvent se présenter à l'attaque du signal dans le cas d'une occlusive voisée par exemple.
- Dans des séquences de plusieurs voyelles, un seul maximum peu apparaître.

Le paramètre EnDap est constitué de la somme :

- de l'énergie totale du signal qui met en relief les noyaux vocaliques principaux,
- de l'énergie spectrale entre 200 et 3600 Hz, bande de fréquence qui comprend l'essentiel de l'énergie des voyelles,
- et de la moyenne des maxima spectraux dans les bandes 100-800 Hz et 600-3600 Hz qui renforce ainsi l'énergie autour du premier formant puis du second et troisième formant.

Ce paramètre est ensuite fortement lissé afin d'éliminer les pics parasites qui peuvent apparaître dans les zones transitoires ou consonantiques. Les outils de reconnaissance des formes [19] permettent alors la segmentation du signal en zone vocaliques (resp. consonantiques) par l'emploi d'un schéma de *colline* (resp. de *vallée*) qui reconnaît pour frontière toute variation positive (resp. négative) d'un point de la courbe.

Identification et valuation des voyelles

Les zones vocaliques précédemment détectées sont ensuite soumises à un éventuel redécoupage pour pallier l'éventualité où plusieurs phonèmes auraient donnés lieu à un unique maximum du paramètre EnDap. Chaque voyelle est alors identifiée à l'aide d'un paramètre résultant de la somme d'une fonction d'instabilité spectrale et de la distance au phonème de référence qui permet une localisation plus fine de ses limites. Une valuation leur associe un score de vraisemblance qui prend en compte le niveau énergétique de la voyelle ainsi que sa ressemblance globale au spectre de référence associé (mesurée par la moyenne — sur la durée totale de la voyelle — d'une distance directe des 24 canaux de la référence mise à niveau sur l'énergie moyenne de tout le spectre).

Identification et valuation des consonnes

Les consonnes sont localisées différemment selon qu'il s'agit d'occlusives sourdes ou pas. La limite gauche des premières est fixée par le paramètre EnDap qui marque le début de l'occlusion et la trame qui caractérise le mieux leur explosion les délimitent à droite. Pour les secondes, qui semblent difficiles à séparer correctement (si plusieurs se suivent), on ne retient que la zone centrée autour du minimum de distance au phonème recherché. Cette méthode à défaut d'être précise assure un traitement homogène des différentes consonnes. La valuation des consonnes ainsi localisées est identique à celle des voyelles.

Validation des hypothèses phonétiques

À ce stade du processus de décodage, nous disposons d'un treillis phonétique riche où de nombreuses consonnes sont valuées sur chaque zone consonantique de même que la presque totalité des voyelles sont notées sur chaque zone vocalique. Un premier filtre rejette les hypothèses trop mal notées et un second élimine les hypothèses irrecevables à l'aide d'un ensemble restreint de macro-traits robustes [96].

Cahier des charges

Un module de décodage acoustico-phonétique tout aussi perfectionné qu'il puisse être (qui plus est en phase ascendante) ne peut prétendre échapper aux phénomènes d'insertions, de suppressions et de substitutions dont doivent tenir compte les niveaux supérieurs. C'est pourquoi il est indispensable que ses limites soient formalisées ce qui est fait par l'ensemble de règles suivantes :

- pour chaque phonème localisé, le DAP propose un ensemble de phonèmes valués,
- chaque phonème voyelle réalisé doit être présent dans le treillis,
- les groupes consonantiques intervocaliques sont tous identifiés, à l'exception de certaines consonnes sonnantes placées dans un contexte vocalique fermé (ex: le phonème /*n*/ dans /*ini*/),
- la consonne la plus fermée d'un groupe consonantique réalisé identifiable est obligatoirement présente dans le treillis (ex: le phonème /*t*/ dans /*t*/).

Un exemple de treillis issu du processus décrit est proposé en figure 4.1.

4.2.2 Le niveau lexical

Le module lexical a pour fonction de faire correspondre les unités phonétiques du treillis avec les mots du lexique. Il s'inspire du modèle de cohorte [92] en proposant de réduire progressivement les hypothèses lexicales candidates à une cohorte aussi réduite que possible. Le cahier des charges du DAP n'autorise pas la manipulation directe des phonèmes du treillis et des décompositions phonétiques des entrées lexicales de manière efficace, aussi une représentation intermédiaire du treillis en zones dénommées *familles* permet de remédier aux insertions nombreuses de phonèmes et aux suppressions fréquentes des consonnes. Deux familles sont distinguées (voyelles et consonnes), chacune d'elle représente l'ensemble des unités de même type dont les noyaux stables déterminés lors de la phase de localisation du DAP se chevauchent. Les familles consonnes sont caractérisées par l'unique étiquette CO tandis que les familles voyelles possèdent plusieurs modalités qui sont déterminées à l'aide d'heuristiques sur divers paramètres (durée, note, énergie, ...) de leur unités constitutives :

VO Caractérise les familles qui couvrent une zone sur laquelle une voyelle a été réalisée de façon certaine.

EE Distingue les familles qui représentent un schwa non pleinement réalisé ou un simple appui vocalique dans un groupe consonantique.

VX Décrit les familles voyelles qui ne peuvent être répertoriées dans les deux classes précédentes.

Cette décomposition en famille du treillis est alors dérivée par un ensemble restreint de règles[3] qui se composent pour donner naissance à un ensemble d'hypothèses (appelées *images*) qui représentent les interprétations possibles du treillis phonétique quant à la structure du mot prononcé. Le dictionnaire est par ailleurs pré-compilé en tenant compte de quelques règles phonologiques comme l'élision du schwa ou les confusions vocaliques usuelles (ex: [] $\Longleftrightarrow$ [*o*]) ; chaque entrée lexicale pouvant donner lieu à plusieurs images directement comparables à celles obtenues à partir du treillis phonétique. C'est cette opération peu coûteuse, que réalise un premier filtre qui permet d'éliminer près de la moitié des entrées lexicales. Opèrent ensuite séquentiellement deux filtres descendants qui assurent la cohérence acoustico-phonétique des informations lexicales.

Filtre A

Ce filtre vérifie, pour une image donnée d'un mot , la présence — parmi les n meilleures propositions phonétiques de chaque famille — des phonèmes attendus (déduits de la décomposition phonétique réduite à l'image traitée) et élimine des candidats potentiels tout mot dont un phonème réduit viendrait à manquer dans une famille d'une de ses images.
Il est efficace car il ne fait aucunement état de l'indice temporel de chaque hypothèse et réduit le nombre de mots candidats à 15% du dictionnaire initial[4] avec un taux d'erreur très faible qui atteste — si besoin était ! — le bon fonctionnement du processus de décodage acoustico-phonétique.

Filtre B

Le second met à profit les notes des phonèmes réduits proposées par le DAP ainsi que leur recouvrement temporel pour réduire davantage encore la cohorte candidate en ne retenant qu'une seule image d'un même mot (pour les cas fréquents où un mot du lexique se dérive en plusieurs images) et en ne sélectionnant que les p meilleurs scores restants (p étant fixé en fonction de l'application et représente un compromis entre le nombre de mots dans la cohorte finale et le taux d'échec toléré).
Il reste à l'issu de ce filtre déjà plus coûteux, une cohorte réduite à environ 5% du dictionnaire initial, le taux d'erreur ajouté étant inférieur à 1%.

[3] Ces règles permettent de pallier les spécificités du décodage acoustico-phonétique.
[4] Taux moyen qui dépend bien évidemment du lexique initial.

Notation des hypothèses sélectionnées

Les phases de filtrage précédentes ont permis d'affiner progressivement les hypothèses lexicales en passant d'une représentation vocalique/consonantique à une décomposition phonétique réduite. Cette dernière étape complète ces décompositions phonétiques réduites afin de noter l'ensemble des phonèmes de la chaîne, y compris ceux qui pourraient ne pas être localisés par le processus de décodage acoustico-phonétique (voir le cahier des charges du DAP page 56). Les phonèmes absents du treillis sont replacés au mieux pour couvrir le signal de parole en localisant la zone centrée autour du minimum de distance au phonème de référence sur la zone où il est attendu.
Le système définit ensuite de nouvelles frontières pour couvrir au mieux le signal de parole avec la chaîne phonétique entièrement connue de chaque mot candidat. Ce re-cadrage se base sur une constante *intermin* qui correspond à la durée moyenne d'une transition entre deux phonèmes, permettant de limiter la notation des candidats sur les zones transitoires. Les espaces inter-phonétiques sont alors réduits aux *intermin* trames les plus instables de la zone correspondante et les phonèmes sont alors étendus en respectant quelques règles limitatives (un phonème ne peut être étendu sur une famille d'un autre type, une borne maximale permet d'éviter certaines erreurs de segmentations en familles, *etc.*). Les écarts inter-phonétiques qui restent supérieurs à *intermin* sont alors étudiés précisément afin de localiser d'éventuels schwa faiblement réalisés ou encore des appuis vocaliques qui peuvent survenir dans des séquences consonantiques (ex: /*b*/). Sont finalement retenus les p meilleurs scores représentant le score phonétique de chaque candidat.
Au regard de cette description du processus d'accès lexical, nous sommes en droit de nous poser plusieurs questions auxquelles nous allons tenter d'apporter des éléments de réponse dans la suite de l'exposé :

- Est-il possible que la prise en compte d'une information prosodique complémentaire à celles mises actuellement en œuvre puisse apporter une amélioration au processus décrit ?
- Dans l'affirmative, quelles sont précisément les informations pertinentes ?
- Présentent-elles une robustesse suffisante permettant leur intégration efficace dans notre module lexical ?

4.3 Étude macroprosodique

Nous emprunterons dans la suite de cet exposé le terme de macroprosodie à Di Cristo [40] par opposition aux phénomènes microprosodiques que nous décrivons plus loin et le considérons comme équivalent à celui plus répandu de suprasegmental. L'étude la plus complète de l'utilisation d'informations suprasegmentales pour l'amélioration d'une tâche de reconnaissance de mots isolés d'un grand vocabulaire est — à notre connaissance — la thèse de Waibel [162]. Dans sa revue des phénomènes macroprosodiques, l'auteur conclut à la difficulté d'une intégration efficace des informations prosodiques dans les systèmes de

reconnaissance en invoquant les complications de la prise en compte des erreurs induites par la mesure des paramètres, et du caractère pluridisciplinaire des différents niveaux de connaissance :

> "In summary, despite the rich knowledge we have acquired on the importance of prosody for human speech perception only a few systems use the prosodic information that is encoded by the speaker."

Sa contribution décrit alors le potentiel filtrant d'informations prosodiques automatiquement mesurées sur le signal de parole (les *sources de connaissances*) qui sont compilées en deux étapes :

- un système de synthèse à partir du texte [1], qui fournit de nombreuses informations : *f0* cibles, durées segmentales, marqueurs accentuels, syllabation,...,
- une deuxième étape complète les informations disponibles en dérivant différentes prononciations possibles d'une même entrée lexicale et en fournissant des limites syllabiques compatibles avec le détecteur de syllabes mis au point.

Les informations macroprosodiques pré-compilées qu'il utilise relèvent de la métrique, de l'intensité et de l'accentuation (qui joue un rôle fonctionnel à un niveau lexical en anglais). Nous allons rappeler brièvement les résultats qu'il a mis en évidence et faire le point sur nos propres analyses des différents paramètres suprasegmentaux pour l'amélioration du filtrage lexical.

4.3.1 La fréquence fondamentale

Waibel ne présente pas de résultat sur l'intégration directe de ce paramètre dans le cadre de la reconnaissance de mots isolés. Il l'utilise cependant lors de la détection des syllabes accentuées qui — en anglais — permettent un filtrage efficace des cohortes de mots. Dans une étude qui présente les différentes contributions possibles de la prosodie [157], Vaissière rappelle des résultats qu'elle a mis en évidence dans de précédents travaux sur de la parole continue ([153],[155]) :

- 60% des mots lexicaux en français sont marqués d'une fréquence fondamentale montante à l'initiale et 70% d'une *f0* finale descendante.
- 96% des chutes de *f0* sont localisées en finale de mot et jamais sur la syllabe initiale. 96% des montées se situent sur les syllabes initiales et finales des mots.

Son système élimine ainsi 33% des mots de plus de deux syllabes proposés par un module lexical à base de connaissances spectrales. Dans une étude très récente [164], Ward et Novick ont étudié le potentiel discriminant de la fréquence fondamentale pour distinguer la sémantique associée au mot "right" dans différentes situations de parole spontanée (contexte de réponse, d'acquiescement et d'indication de direction). Ils annoncent un

classement "sémantique" correcte des occurrences de ce mot dans 67% des cas, sur la seule information véhiculée par le fondamental et propose d'utiliser une telle méthode pour vérifier certaines hypothèses lexicales. Ces constatations sortent cependant du cadre limité de la reconnaissance de mots isolés que nous nous sommes assigné dans une première étape.
Pour notre part, nous nous contenterons de présenter la répartition des différents schémas de *f0* pour les mots de deux, trois et quatre voyelles du corpus AviLex en utilisant les différents codages du paramètre de fréquence fondamentale que nous décrivons maintenant :

Le codage 1234 : obtenu de manière classique par le découpage de la dynamique du fondamental en quatre bandes d'égale hauteur (le grave 1, le médium 2, l'infra-aigu 3 et l'aigu 4) ; la dynamique est celle de l'ensemble des réalisations d'un même locuteur (ce que G. Caelen [21] appelle le codage *Texte*). Nous distinguons ici deux types de codage — *Absolu* et *Relatif* — le premier attribuant à une voyelle un niveau en fonction de l'appartenance de sa *f0* (valeur aux deux-tiers) à l'une des quatre bandes précédemment décrites, le second imposant un écart fréquentiel d'au moins un quart de la dynamique du paramètre pour que deux voyelles successives se voient attribuer des niveaux différents.

Le codage duc : ce codage se déduit du précédent en codant *u* le passage d'un niveau à un niveau supérieur, *d* la suite de deux niveaux dont le premier est supérieur au second, et *c* code deux voyelles successives de même niveau (*ex.* 2331 $\Longrightarrow$ ucd).

Afin de ne pas alourdir inutilement cette description, nous ne retiendrons que deux locuteurs représentatifs de la base dont les dynamiques du paramètre de *f0* sont reportées sur la figure 4.2. Les figures 4.3, 4.4 et 4.5 présentent ces distributions pour le codage synthétique "duc" et sont reportées dans la table 4.1 en codage standard sur 4 niveaux (codage "1234"). On peut résumer ces observations aux constatations suivantes :

- seul un ensemble restreint de schémas sont observés,
- les mots de deux voyelles sont dotés d'un schéma descendant,
- les mots de trois voyelles, ont un schéma essentiellement descendant ; 10% ont un schéma montant puis descendant,
- les mots de quatre voyelles présentent des configurations descendantes ou montantes puis descendantes en quantité égales,
- les proportions varient sensiblement selon le type de codage considéré.

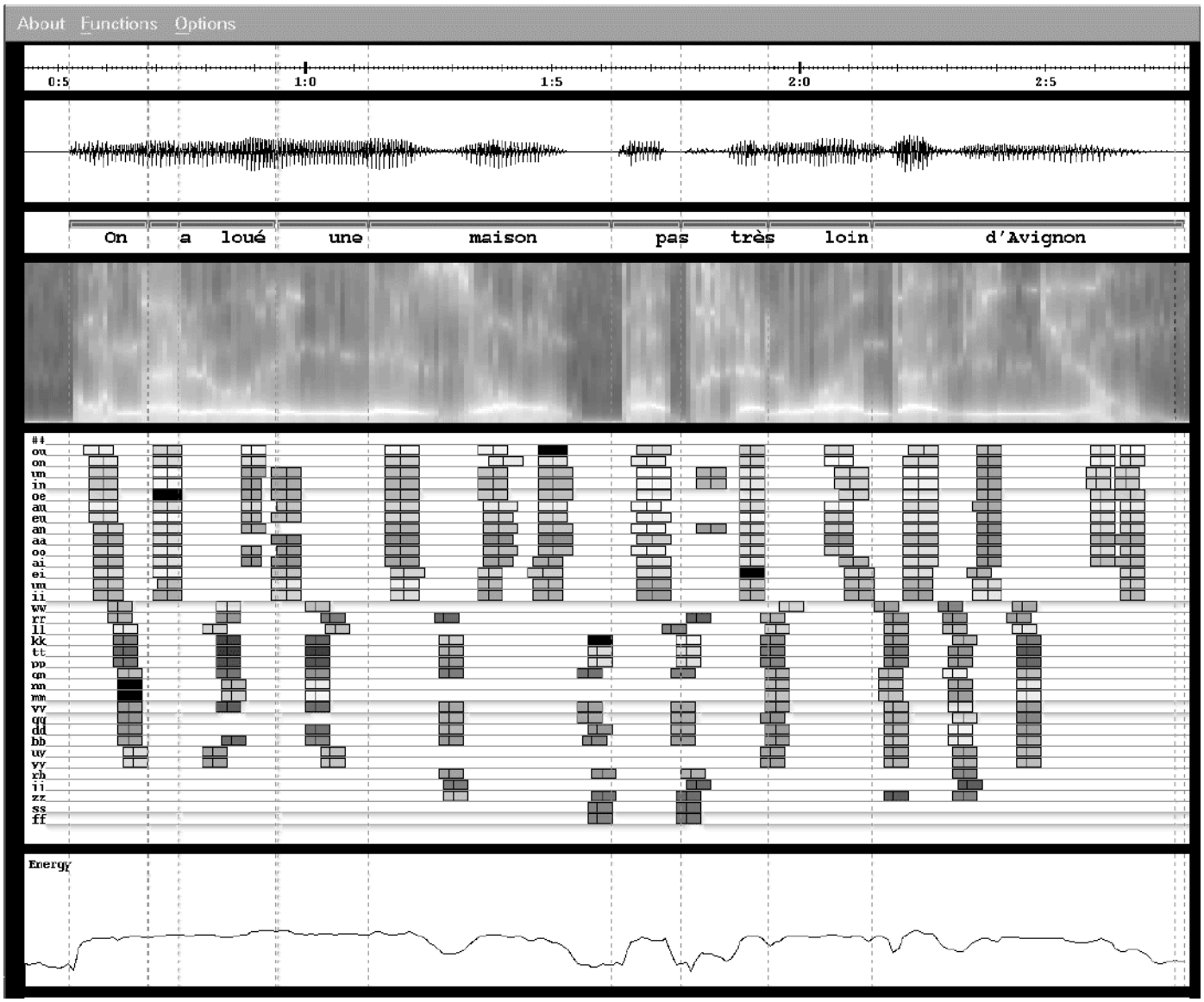


Figure 4.1: Treillis phonétique obtenu pour la phrase *On a loué une maison pas très loin d'Avignon.*

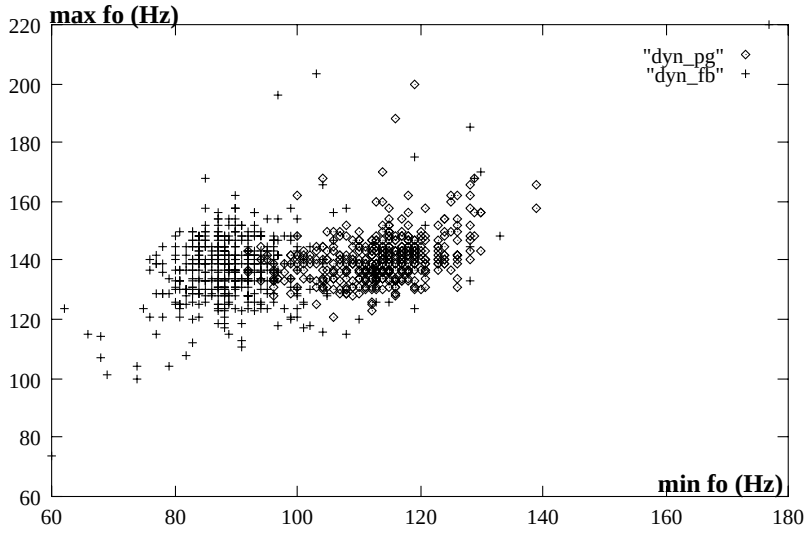


Figure 4.2: Dynamiques du paramètre de fréquence fondamentale pour les locuteurs *pg* et *fb* de la base AviLex. La dynamique de chaque item étudié est caractérisée par un point dont l'abscisse représente la valeur inférieure et l'ordonnée la valeur supérieure de *f0* sur l'ensemble de l'item. On remarque que le locuteur *fb* utilise une plus grande plage de variation du fondamental.

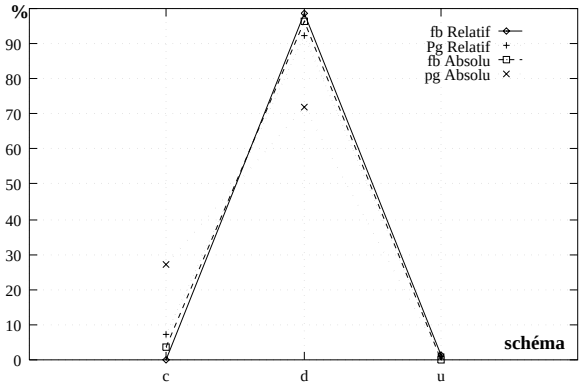


Figure 4.3: Distribution des schémas de *f0* au format "duc" (pour les codages *Absolu* et *Relatif*) des mots de 2 voyelles prononcés par les locuteurs *fb* et *pg* de la base AviLex1.

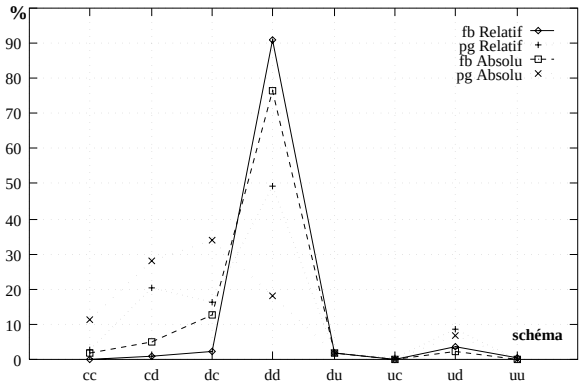


Figure 4.4: Distribution des schémas de *f0* au format "duc" (pour les codages *Absolu* et *Relatif*) des mots de 3 voyelles prononcés par les locuteurs *fb* et *pg* de la base AviLex1.

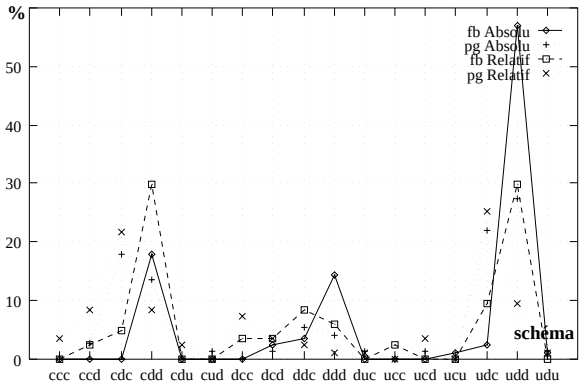


Figure 4.5: Distribution des schémas de *f0* au format "duc" (pour les codages *Absolu* et *Relatif*) des mots de 4 voyelles prononcés par les locuteurs *fb* et *pg* de la base AviLex1.

Schéma	Codage *Absolu*				Codage *Relatif*			
	fb		pg		fb		pg	
	nb.	%	nb.	%	nb.	%	nb.	%
111	2	1.1	-	-	-	-	5	2.9
121	3	1.7	1	0.6	1	0.6	8	4.7
122	-	-	-	-	-	-	1	0.6
123	-	-	-	-	1	0.6	-	-
132	-	-	-	-	3	1.7	1	0.6
211	-	-	2	1.1	4	2.3	28	16.4
212	-	-	-	-	-	-	1	0.6
213	-	-	-	-	3	1.7	-	-
221	1	0.6	10	5.6	2	1.1	35	20.5
222	-	-	17	9.6	-	-	-	-
231	-	-	1	0.6	2	1.1	6	3.5
232	-	-	8	4.5	-	-	-	-
311	9	5.1	4	2.3	-	-	-	-
312	-	-	-	-	-	-	2	1.2
321	38	21.7	21	11.9	159	90.9	84	49.1
322	1	0.6	47	26.6	-	-	-	-
324	-	-	1	0.6	-	-	-	-
331	4	2.3	1	0.6	-	-	-	-
332	2	1.1	39	22.0	-	-	-	-
333	-	-	3	1.7	-	-	-	-
341	1	0.6	-	-	-	-	-	-
342	-	-	1	0.6	-	-	-	-
343	-	-	1	0.6	-	-	-	-
411	7	4.0	-	-	-	-	-	-
412	-	-	1	0.6	-	-	-	-
421	77	44.0	2	1.1	-	-	-	-
422	5	2.9	5	2.8	-	-	-	-
423	-	-	1	0.6	-	-	-	-
424	3	1.7	-	-	-	-	-	-
431	14	8.0	-	-	-	-	-	-
432	5	2.9	9	5.1	-	-	-	-
433	-	-	2	1.1	-	-	-	-
441	2	1.1	-	-	-	-	-	-
444	1	0.6	-	-	-	-	-	-

Table 4.1: Distribution des différents schémas de *f0* (codage "1234" *Absolu* et *Relatif*) mesurés pour les réalisations des mots de trois voyelles des locuteurs *fb* et *pg* de la base AviLex1.

Dans l'expérience qui suit, nous allons nous servir de ces observations pour mesurer le potentiel filtrant des schémas mélodiques en tentant de vérifier que le schéma d'un mot inconnu prononcé par un locuteur dont on a stocké les réalisations permet de réduire sensiblement la cohorte de mots pouvant lui correspondre. Nous allons pour cela considérer dans la base AviLex2 les réalisations du locuteur *pg* — dont les différents schémas mélodiques ont été mesurés sur la base AviLex1 — puis observer le classement de chacune parmi le lexique de 20.000 mots associé. Pour chaque candidat du lexique décrit par sa décomposition phonétique "théorique", un schéma mélodique est mesuré sur les voyelles localisées grossièrement sur chaque portion voisée du signal de parole, et sa note est donnée par la probabilité d'occurrence du schéma mesuré en consultant les distributions recueillies précédemment (cf table 4.1). Afin de ne pas introduire de biais dans les taux de filtrage, seuls les mots du lexique qui sont conformes au voisement du mot candidat sont retenus, ce qui est réalisé par l'application du filtre présenté en section 4.4.4[5]. Enfin, seuls les mots d'au moins deux voyelles et d'au plus six voyelles ont été considérés dans ce test. Les résultats sont reportés dans la table 4.2 pour divers codages ; nous les commentons brièvement :

- La donnée seule du rang moyen des mots prononcés dans les cohortes (ligne 4) ne suffit pas à rendre compte de manière effective du taux de filtrage obtenu. Il est indispensable pour cela de prendre en compte le nombre d'ex æquo (ligne 5). Cette constatation est d'autant plus évidente que l'intervalle de notation est petit (dans le cas présent, il y a autant de notes que de schémas différents alors qu'une cohorte comporte en moyenne 4500 mots). Cette précaution n'ayant pas été prise dans l'étude de Waibel, nous supposons donc que peu ou pas de mots se voyaient attribuer une même note.

- Les classements sont meilleurs dans les réalisations pg, ce qui est normal puisque c'est d'elles que sont issus les schémas de référence. L'écart avec les réalisations de pg700 est cependant inférieur à 10% (ligne 5), et un mot est en moyenne classé dans la première moitié des mots du lexique (résultat qui varie faiblement avec le codage considéré).

- Classer n'est pas filtrer ! Il convient en effet pour cela de tenir compte du nombre de mots qui seraient au dessus d'un seuil donné (qui peut être fixé dynamiquement) et donc rejetés lors du filtrage. Les écarts-types (ligne 3) permettent d'évaluer grossièrement ce seuil : on peut dans notre cas remarquer que l'écart-type reste globalement pour tous les codages inférieur à la moitié du nombre de classes.

- Le type de codage ne semble pas très important et au vu des courbes des figures 4.6, 4.7 et 4.8, un filtrage de l'ordre de 10 à 20% seulement semble envisageable avec un taux d'erreur acceptable (ce qui est pour le moins éloigné du classement moyen du mot prononcé, proposé dans la ligne 4 !).

[5] Cette méthodologie nous garantit une mesure exacte du taux de filtrage des seuls schémas mélodiques, en ne considérant qu'une information orthogonale à celle mise à profit dans le filtre de voisement.

	Codage *Absolu*				Codage *Relatif*			
	1234		duc		1234		duc	
	pg	pg700	pg	pg700	pg	pg700	pg	pg700
nb. moy. de classes avant	3	4	9	4	3	4	3	4
nb. moy. de classes au total	12	9	12	11	13	13	12	12
écart-type	3	3	3	3	3	3	3	3
rang du mot	27%	29%	30%	47%	27%	41%	28%	42%
rang avec ex æquo	46%	45%	50%	62%	48%	56%	49%	57%

Table 4.2: Classements des mots de la base AviLex2 prononcés par le locuteur *pg* dans le lexique de 20.000 mots associé en utilisant les schémas de référence mesurés sur les mots du corpus AviLex1. La première ligne indique le nombre moyen de classes qui précède la classe du mot prononcé ; la deuxième ligne indique le nombre moyen de classes (*i.e.* de notes différentes) d'une cohorte ; la troisième ligne indique l'écart-type de ce classement par classes ; les deux dernières lignes reportent les rangs moyens des mots exprimés en pourcentage sans tenir compte d'éventuels ex æquo (ligne 4) puis en les considérant (ligne 5).

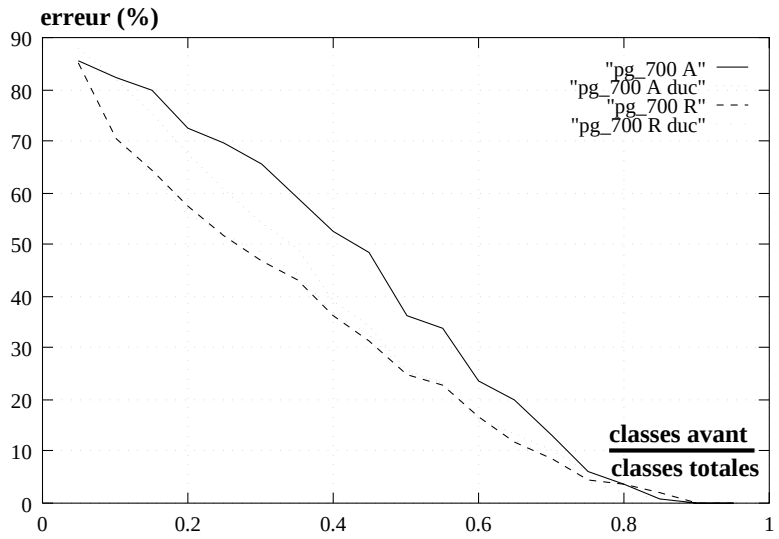


Figure 4.6: Taux d'erreur (sur l'axe des ordonnées en pourcentage) exprimés en fonction de la proportion de classes retenues dans une cohorte (cette proportion est reportée sur l'axe des abscisses, une valeur de 1 signifiant que l'on garde toutes les classes). A désigne le codage *Absolu*, R le codage *Relatif*.

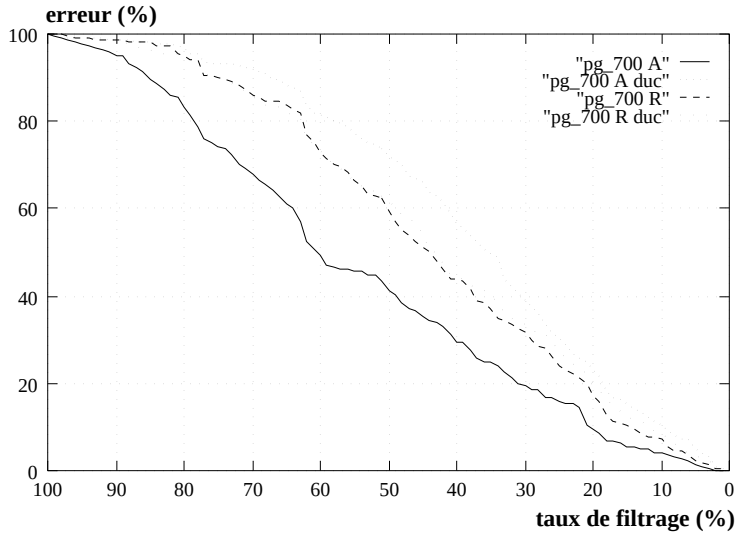


Figure 4.7: Taux d'erreur (exprimés en pourcentage sur l'axe des ordonnées) en fonction du pourcentage de mots filtrés par cohorte (seuil fixe). A désigne le codage *Absolu*, R le codage *Relatif*.

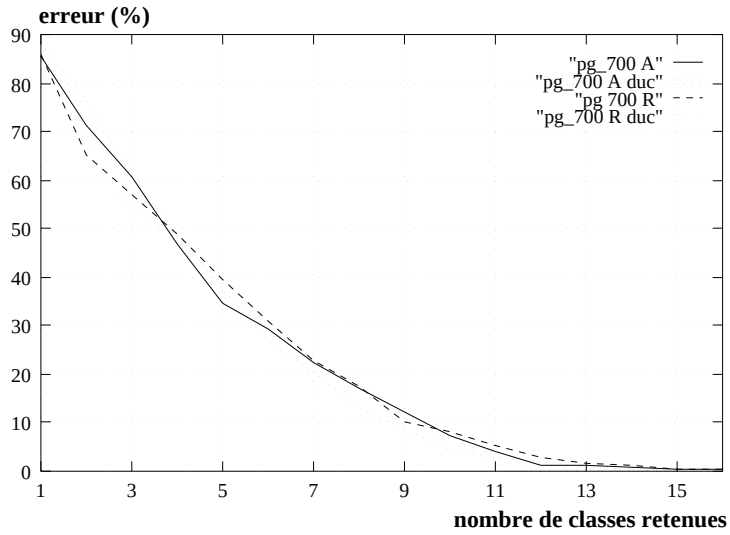


Figure 4.8: Taux d'erreur (exprimés sur l'axe des ordonnées en pourcentage) en fonction du nombre de classes (une classe étant définie par l'ensemble des mots qui obtiennent la même note) retenu pour chaque cohorte (ce nombre est fixé pour toutes les cohortes). A désigne le codage *Absolu*, R le codage *Relatif*.

4.3.2 L'intensité

Waibel tente pour l'intensité (qu'il définit dans le cas des mots prononcés isolément par la fonction de *loudness* de Mermelstein [99]), de vérifier si un ensemble de schémas extraits automatiquement autorise une réduction efficace de la cohorte de mots probables. Il emploie à cet effet une technique de classification hiérarchique avec divers taux de similarité qu'il applique à une base de 100 mots prononcés par un locuteur. Il montre alors que pour une valeur de similarité de 0.9, ces prototypes permettent de classer — en moyenne — le mot correct dans les 25 premiers candidats ; les tests étant réalisés sur les mêmes 100 mots qui ont servi à l'apprentissage des différents schémas et cela pour deux locuteurs (dont un est celui qui à servi pour l'apprentissage). Les résultats qu'il obtient sur cette même tâche limitée, dans le cas où aucune classification n'est opérée, sont légèrement moins bons, ce qui semble appuyer la position de Di Cristo qui considère que la microprosodie doit être effacée avant toute interprétation suprasegmentale[6]. Compte tenu des réserves que nous avons formulées dans la section précédente sur l'information véhiculée par un taux de classement seul, nous préférons — à l'instar de la fréquence fondamentale — étudier les divers schémas relevés sur le corpus AviLex et vérifier si la prise en compte de leur distribution, permet raisonnablement de diminuer le nombre de mots candidats d'une cohorte. Un seul codage — le codage *Absolu* — est ici retenu au regard du peu de différence sur les taux de filtrage obtenus précédemment. La table 4.3 fait état des différents schémas mesurés pour les réalisations du locuteur pg sur la base AviLex1 ; et l'on peut simplement remarquer que la majorité des schémas sont descendants, et que comme pour la fréquence fondamentale, on rencontre davantage de schémas montants puis descendants dans les mots plus longs.
Pour chaque mot de deux, trois ou quatre voyelles de la base AviLex2 on effectue un classement du lexique de la même manière que précédemment avec cette fois-ci les notes provenant des schémas d'intensité. Le taux de classement des mots prononcés est de 42% (57% en tenant compte des ex æquo), avec une variance de 24%. Le nombre de notes différentes attribuées aux mots de chaque cohorte (classe) est en moyenne de 13 et le mot prononcé est noté en 4ème position (avec une variance de 4.5). On observe sur la courbe 4.9 les taux d'erreur associés à un filtrage dynamique du nombre de classes dans les cohortes ; un filtrage de l'ordre de 30% du lexique peut être réalisé avec un taux d'erreur acceptable.

[6]Notons cependant que la différence de rendement obtenue pour les deux valeurs de similarités (0.9 et 1) est de l'ordre de 3%, ce qui est peu significatif au vu du nombre restreint de tests effectués.

schéma	nb	%	schéma	nb	%
Mots de 2 voyelles			Mots de 4 voyelles		
c	83	59.3	ccd	18	20.9
d	52	37.1	ccc	17	19.8
u	5	3.6	cud	10	11.6
			dud	10	11.6
			dcc	7	8.1
			udc	6	7.0
			ucd	4	4.7
			duc	3	3.5
Mots de 3 voyelles			ucc	2	2.3
cc	63	35.0	dcu	2	2.3
cd	39	21.7	cdd	1	1.2
dc	32	17.8	dcd	1	1.2
ud	29	16.1	udu	1	1.2
dd	5	2.8	cdc	1	1.2
du	5	2.8	ddu	1	1.2
uc	4	2.2	cuc	1	1.2
cu	3	1.7	ccu	1	1.2

Table 4.3: Schémas d'intensité pour les mots de 2, 3 et 4 voyelles recueillis sur les réalisations du locuteur pg de la base AviLex1.

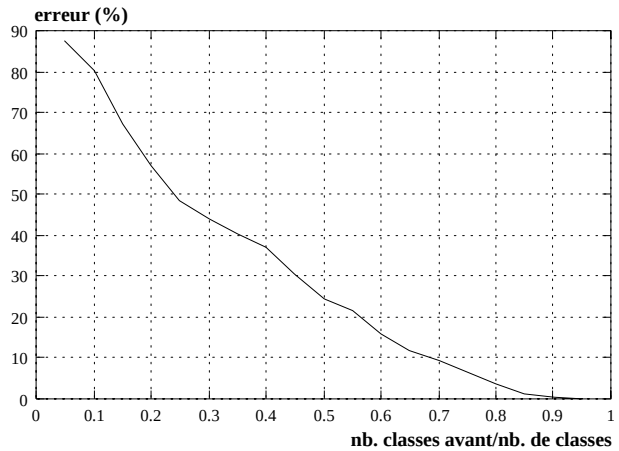


Figure 4.9: Taux d'erreur (sur l'axe des ordonnées en pourcentage) exprimés en fonction de la proportion de classes retenues dans une cohorte (cette proportion est reportée sur l'axe des abscisses, une valeur de 1 signifiant que l'on garde toutes les classes). Le rapport 0.8 correspond à un filtrage effectif de 29.8 %.

4.3.3 La durée

Waibel étudie la contribution de plusieurs sources de connaissances temporelles : le rythme (par normalisation puis par calcul d'une distance euclidienne entre les durées fournies par un détecteur syllabique et celles dérivées synthétiquement), le rapport de voisement de chaque syllabe, et les contributions de différentes classes (NASAL, L, R, FRONT, BACK) à la durée de chaque syllabe. Il teste alors ces filtres sur une série de 862 mots prononcés parmi ceux d'un lexique de 1478 entrées et obtient — par combinaison des huit filtres — l'arrivée du mot prononcé en $64^{ème}$ position (valeur moyenne tenant compte des mots monosyllabiques et polysyllabiques). Ces résultats nous laissent à la fois perplexes et admiratifs : ce classement nécessite d'une part, un algorithme de synthèse qui fournit des durées suffisamment fiables pour être utilisées comme durée de référence dans le processus de filtrage et, d'autre part, un algorithme de syllabation non moins efficace. Ne pouvant pas prétendre à une telle précision — *a fortiori* dans une phase ascendante — nous ne pensons pas être en mesure de proposer un filtrage aussi probant. L'auteur reconnaît cependant que 116 mots ont été éliminés du corpus de test initial (alors composé de 978 mots) soit pour avoir été mal enregistrés, soit pour avoir mis en défaut un des algorithmes de la chaîne de traitement. On peut également émettre quelques réserves quant au caractère prosodique de filtres portant sur des traits de nasalisation ou encore de postériorité/antériorité qu'il combine ici avec le filtre de rythme. Aussi nous proposons de vérifier le pouvoir discriminant des seuls schémas de durée en mesurant ceux du locuteur *pg* de la base AviLex1 puis en observant sur les réalisations de la base AviLex2 du même locuteur les taux de classement obtenus à partir de ces schémas. Ces derniers sont obtenus à partir des durées des voyelles d'un mot en associant à chacune d'elles le rang qu'elles auraient si on les classait par ordre croissant de durée puis en codant u, d ou c chaque voyelle qui est respectivement plus courte, plus longue ou de même durée que la voyelle qui la suit directement (*ex.* si un mot possède quatre voyelles de durées respectives 10, 7, 9, et 7 centisecondes alors le schéma associé transforme la séquence 3121 en le schéma *dud*).
Les schémas de durée mesurés pour le locuteur *pg* sont reportés pour les mots de trois et quatre voyelles sur les figures 4.10 et 4.11. Nous proposons donc de vérifier si un mot candidat — dont on connaît le schéma de durée — peut-être classé mieux qu'aléatoirement à l'aide des distributions de schémas recueillies. Comme le calcul d'un schéma nécessite la connaissance de la durée des voyelles, nous ne considérons dans cette expérience que les mots issus du processus de filtrage lexical qui constituent une cohorte d'environ une centaine de mots pour lesquels SPEX fournit un alignement phonétique. La note d'un mot candidat est la probabilité d'occurrence de son schéma de durée (donnée par les distributions recueillies précédemment). On désigne par *classe* l'ensemble des mots d'une cohorte qui possèdent la même note. La figure 4.12 permet d'observer les limites d'un filtrage basé sur la seule information rythmique :

- On remarque tout d'abord le nombre réduit de classes différentes par cohorte (qui est maximisé par le nombre de schémas de durée différents des mots de la cohorte) : en moyenne une cohorte est décomposée en 13 classes seulement.

- La deuxième distribution présente la position de la classe à laquelle appartient le mot réellement prononcé. En moyenne, le mot fait partie de la $7^{ème}$ classe d'une cohorte ce qui ne fait pas des schémas de durée mesurés sur les voyelles une information très pertinente. Un filtre sans perte peu cependant, au regard des distributions présentées, être réalisé en ne conservant d'une cohorte que les 13 premières classes, ce qui avouons-le est d'un rendement très faible (10% de mots éliminés pour les cohortes du locuteur *pg* de la base AviLex2 d'au moins 14 classes).

- Ce résultat n'a rien de surprenant puisque les mots proposés par le module de filtrage lexical sont assez proches et possèdent en général un découpage en voyelle/consonne homogène à l'intérieur d'une même cohorte.

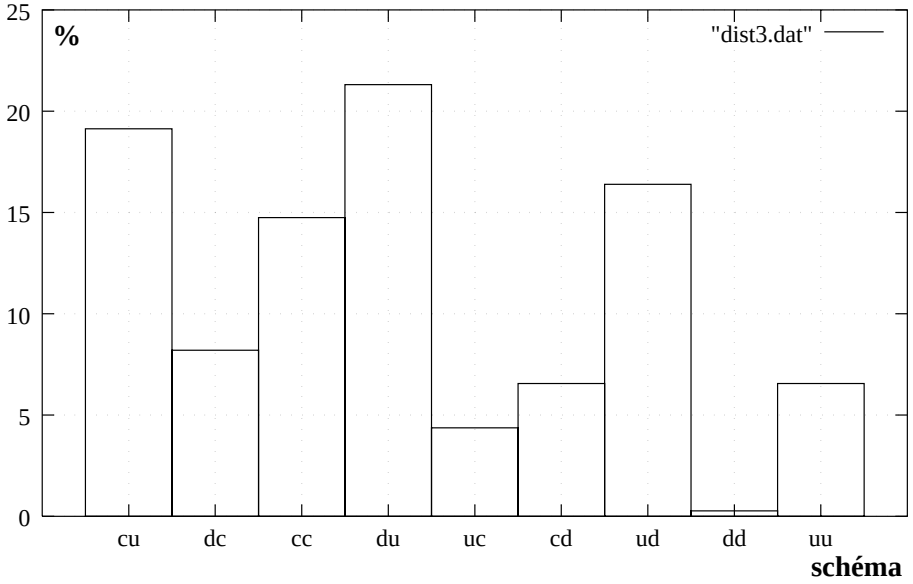


Figure 4.10: Distribution des schémas de durée du locuteur *pg* de la base AviLex1 pour les mots de 3 voyelles.

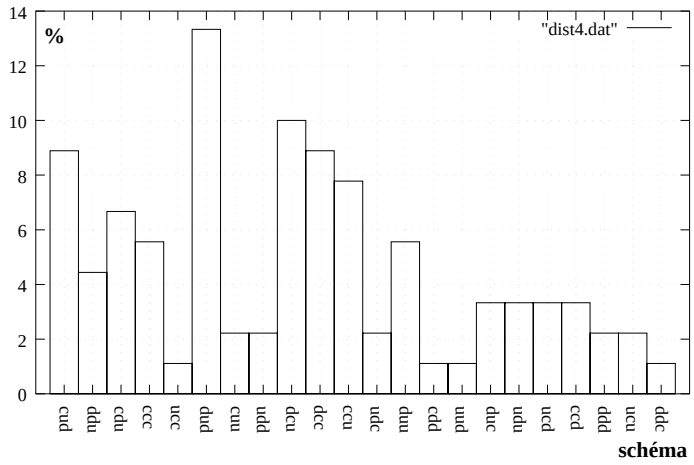


Figure 4.11: Distribution des schémas de durée du locuteur *pg* de la base AviLex1 pour les mots de 4 voyelles.

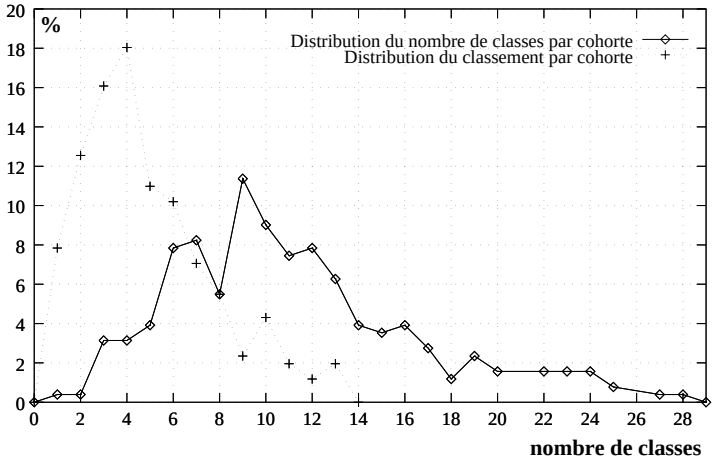


Figure 4.12: Distribution du nombre de classes (*i.e.* mots ayant la même note) pour les cohortes — proposées par SPEX — du locuteur *pg* de la base AviLex2 et distribution de la position de la classe du mot réellement prononcé.

4.4 Étude microprosodique

Après avoir présenté les possibilités et limites du filtrage lexical de mots prononcés isolément par des indices suprasegmentaux, nous allons maintenant entreprendre l'étude des phénomènes microprosodiques afin d'en mesurer le pouvoir discriminant dans notre module de filtrage lexical.

4.4.1 La durée

Nous n'allons pas faire ici un inventaire des nombreuses études consacrées aux variations temporelles de la parole mais simplement tenter de dégager les phénomènes qu'elles ont permis d'établir. Nous nous aiderons pour cela de la synthèse réalisée par Di Cristo dans [40]. À partir des phénomènes décrits, nous tenterons de vérifier par des analyses sur des corpus de mots, ceux qui sont pertinents pour une utilisation dans un système de reconnaissance de la parole.

Quelques résultats sur les phénomènes de durée

Nous savons — grâce aux nombreux travaux traitant de l'aspect temporel de la parole [26, 72] — que les manifestations acoustiques de ce paramètre sont régies par des facteurs multiples. Selon Klatt [71] il y a au moins sept facteurs qui sont responsables de la durée des divers segments : l'accent syllabique, l'emphase du mot auquel appartient le phonème, les phonèmes précédent et suivant, la position dans la phrase, position dans le mot et enfin la nature du segment.
Une étude rigoureuse de l'influence de ces facteurs — qui peuvent interagir entre eux — devrait emprunter une méthodologie capable de rendre compte de l'espace des variations multi-paramétriques [160]. Sans vouloir dévoiler la suite de cet exposé, nous verrons (hélas) très rapidement que dans le cadre de mesures automatiques de la durée des segments, de telles techniques ne sont pas (loin s'en faut) nécessaires .... En étudiant la microprosodie de mots prononcés isolément, nous éliminons déjà un bon nombre des facteurs régissant les variations de la durée des phonèmes de telle sorte que l'étude présentée ici traitera des "seules" informations véhiculées par les variations intrinsèques et co-intrinsèques de la durée vocalique.
Voici la liste des principaux facteurs régissant les variations intrinsèques de la durée des voyelles :

- Une première constatation à caractère "quasi-universelle" a très tôt mis en évidence la corrélation étroite entre l'aperture d'une voyelle et sa durée. En particulier, toutes les études attestent que la voyelle [*i*] est plus courte — toute chose égale par ailleurs — à la voyelle [*a*].

- Une seconde observation spécifique cette fois au français, fait état de l'écart intrinsèque moyen positif entre les voyelles nasales et les voyelles orales correspondantes.

- On a constaté également dans plusieurs langues dont le français que les noyaux vocaliques initiaux et finaux sont souvent plus longs [154].

De nombreux travaux ont également permis de mettre en relief l'influence du contexte consonantique sur la durée des voyelles adjacentes (les variations co-intrinsèques) :

- Une premiere observation, semble-t-il commune à toutes les langues fait état de l'importance du mode phonatoire de la consonne sur la durée de la voyelle qui précède. Plus précisément, on constate que les voyelles sont plus longues devant une consonne voisée que lorsqu'elles précèdent une consonne non voisée. Des études comparatives [26] mettent en évidence que cette différence de durée est sensiblement plus importante en anglais que dans les autres langues.

- Le mode articulatoire de la consonne — bien que dans des proportions moindres — affecte également la durée de la voyelle qui précède. Les voyelles suivies d'une constrictive sont généralement plus longues que celles suivies d'une occlusive. Ce résultat est considéré comme secondaire par rapport au premier par plusieurs auteurs.

- Enfin, peu de résultats permettent de conclure quant à l'influence de la consonne sur la durée du noyau vocalique qu'elle précède.

Forts de ces phénomènes, nous allons, par une série d'analyses, déterminer ceux qui peuvent être retenus dans une tâche de filtrage lexical. Nous sommes tout de suite confrontés à deux difficultés dont les solutions nous semblent receler une part d'"arbitraire" :

- Comment segmenter le continuum de parole en unités discrètes (dans notre cas les phonèmes) ?

- Quelle précision peut-on attendre de nos mesures ?

Une expérience simple permet de mettre en évidence des différences notables (de l'ordre de 20 ms) dans la segmentation en phonèmes du même continuum de parole par plusieurs experts phonéticiens à partir de la lecture de spectrogrammes. Ces différences sont d'autant plus importantes que la parole analysée contient beaucoup de consonnes vocaliques. Comme le fait remarquer Campbell [22], la détection précise des limites de phonèmes est probablement une gageure. On comprendra aisément — malgré les techniques modernes disponibles — que nous émettions quelques réserves quant à la réalisation de cette tâche de manière automatique et a fortiori sur les conclusions que nous pourrions tirer de ces mesures sur différents corpus. C'est pour cette raison que dans la suite de cet exposé nous allons considérer plusieurs méthodes automatiques — habituellement employées en reconnaissance de la parole — pour extraire la durée d'un phonème et les confronter à différents corpus de parole. Nous allons donc à travers des corpus de mots prononcés isolément par plusieurs locuteurs, tenter de dégager des phénomènes précédemment décrits, ceux qui se révèlent détectables automatiquement.

Durée mesurée par un système d'accès lexical

Notre première démarche est d'étudier les variations de la durée des voyelles en prenant comme mesure des durées celles fournies en sortie d'un module lexical. Notre choix s'est tout naturellement porté sur le module d'accès lexical développé au LIUAPV qui présente de bons résultats : un mot prononcé est reconnu parmi une cohorte de 50 mots dans 90% des cas.
Nous allons présenter dans cette section les observations faites sur le corpus AviLex, et tenter de vérifier si les phénomènes microprosodiques précédemment décrits sont mesurables par notre technique de mesure automatique de la durée.
Le tableau 4.4 (repris — par souci de clarté — pour deux locuteurs de la base AviLex dans la figure 4.13) résume les moyennes et écarts-types calculés pour chaque locuteur pour toutes les voyelles analysées de notre corpus, et la table 4.5 précise le nombre d'observations de chacune d'elles. La différence du nombre d'observations pour chaque locuteur s'explique simplement par le fait que nous ne considérons ici que les mots qui ont été reconnus par le module d'accès lexical (soit en moyenne 470 mots pour un locuteur).

Loc	*a*	*i*	*y*		*e*		*u*	*o*		~	~	~	*∂*
fb	7.5 1.7	7.8 2.1	7.7 3.3	7.1 1.4	7.6 1.9	8.5 2.8	8.2 1.9	7.7 1.7	7.8 1.4	9.0 3.0	10.2 2.5	9.4 3.4	6.6 1.6
pg	7.0 1.4	7.0 1.6	7.0 1.8	6.5 1.1	7.1 1.3	7.3 1.6	7.9 3.2	6.8 0.8	7.0 1.3	9.5 2.6	10.1 2.6	8.5 2.3	6.6 1.7
ts	8.0 2.8	7.2 2.0	7.7 3.0	7.0 1.7	7.0 1.4	7.6 2.1	7.4 2.2	7.4 1.9	7.0 0.9	8.3 2.7	8.3 2.3	7.7 2.4	7.3 1.9
hm	8.2 3.1	8.6 3.2	9.5 4.1	7.6 2.6	9.6 4.3	9.9 3.3	8.5 3.1	7.9 2.0	8.1 2.2	14.6 5.3	12.8 5.1	12.7 5.1	7.2 1.6
si	7.6 2.2	7.1 1.8	6.7 1.4	6.9 1.3	7.2 1.5	7.9 2.2	7.2 1.3	7.3 1.7	7.5 1.7	8.5 2.5	7.7 1.6	8.2 2.0	7.0 1.9
lc	7.7 2.2	7.4 2.5	7.8 2.5	7.6 2.0	9.1 3.3	8.3 3.3	7.8 2.3	7.1 1.5	7.5 1.5	15.0 4.7	12.6 5.7	15.7 4.7	7.6 2.2
si7	7.2 1.6	6.8 1.7	6.5 1.6	6.9 2.0	6.9 0.9	7.3 1.3	6.8 1.6	6.9 1.6	7.5 2.4	8.2 2.6	8.0 2.5	7.7 2.4	7.2 2.0
pg7	6.9 1.3	6.6 1.1	6.6 1.3	6.6 1.2	6.9 1.2	6.8 0.9	6.7 1.4	6.6 0.9	6.8 1.6	8.4 2.5	8.8 2.7	7.6 1.9	6.8 1.9
tous	7.5 2.1	7.2 2.1	7.2 2.5	7.1 1.8	7.4 2.2	7.9 2.4	7.4 2.2	7.1 1.5	7.3 1.8	9.9 4.2	9.2 3.4	9.7 4.2	7.0 1.8

Table 4.4: Récapitulatif des moyennes et écarts-types de chaque phonème étudié pour chaque locuteur de la base AviLex ; les durées étant fournies par le module d'accès lexical SPEX. Le terme *tous* désigne l'ensemble des locuteurs.

Loc	*a*	*i*	*y*		*e*		*u*	*o*		~	~	~	*∂*
fb	177	193	50	91	279	47	41	58	31	109	26	143	67
pg	181	187	49	103	281	42	41	50	45	105	27	151	48
ts	178	181	43	101	165	136	41	52	37	107	28	148	49
hm	166	162	45	98	189	109	35	54	40	98	29	142	76
si	175	193	49	105	204	112	42	57	40	106	28	154	47
lc	158	172	44	168	230	63	36	51	34	96	24	130	35
si7	268	301	102	100	476	209	70	193	72	136	108	115	51
pg7	280	321	112	79	624	88	80	183	80	141	112	118	63
tous	1583	1710	494	845	2448	806	386	698	379	898	382	1101	436

Table 4.5: Décompte des voyelles étudiées par locuteur tous contextes confondus pour le corpus AviLex.

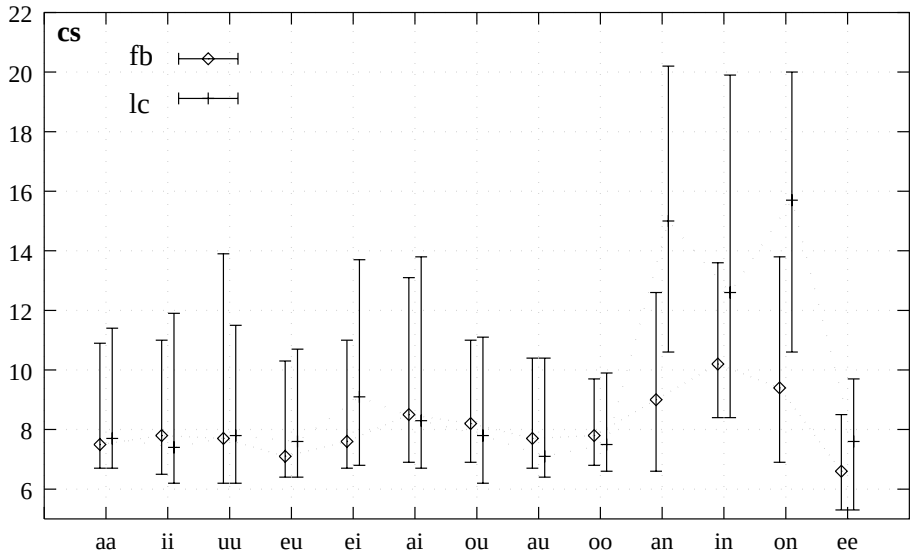


Figure 4.13: Durées moyennes (mesurées par SPEX) des différentes voyelles du français pour deux locuteurs de la base AviLex. Chaque phonème est décrit par une ligne verticale reliant les trois points suivants (par ordre décroissant de valeurs) : l'écart-type des valeurs supérieures à la moyenne, la moyenne des durées pour le phonème, et l'écart-type des durées inférieures à la moyenne.

Il ressort de cette première série de mesures deux informations principales :

- Les différences de durées entre les voyelles hautes (nominativement [*i*] et [*y*]) et la voyelle basse [*a*] ne semblent pas significatives. Notons que les réalisations des voyelles [*i*] et [*a*] des locuteurs *fb* et *hm* indiquent même une moyenne de durée supérieure pour la voyelle haute. Ceci peut être observé plus clairement pour deux locuteurs de la base sur la figure 4.13. Plusieurs facteurs qu'il convient d'analyser peuvent fournir une explication pour autant que nous ne remettions pas en cause la relation liant l'aperture d'une voyelle et sa durée. La première raison qui peut être invoquée serait l'inadéquation de notre système de mesure des durées, une seconde explication pourrait être le manque d'homogénéité des observations : il n'est en effet pas possible de savoir par ce tableau si la répartition des configurations (position dans le mot, nature des consonnes adjacentes, nombre de syllabes dans le mot,...) est uniforme. C'est pourquoi nous devons étudier les durées des voyelles dans des contextes plus contraints.

- Les voyelles nasales sont en moyenne plus longues que les voyelles orales. L'étude des rapports des durées des voyelles nasales sur les durées des voyelles orales "associées" atteste cette différence (voir le tableau 4.6). Il ne semble cependant pas assuré que cette différence de durée soit discriminante pour l'attribution du trait nasal/oral. Si l'on fait l'hypothèse que les distributions des durées mesurées, sur notre corpus de test, sont représentatives (approximation d'autant plus exacte que le nombre d'observations du corpus est grand) alors une décision bayésienne de la classe d'appartenance (orale ou nasale) d'une observation quelconque à partir de sa durée, pourra être prise par la distribution qui maximise sa probabilité d'être observée. On peut alors mesurer la probabilité d'erreur de notre décision par la surface commune des deux distributions. C'est cette mesure qui est reportée dans le tableau 4.6 en colonne de droite. Une probabilité d'erreur de 50% indiquerait que notre décision n'est pas meilleure qu'un choix aléatoire ! La figure 4.14 montre que si pour un locuteur dans notre base (le locuteur féminin *lc*) la décision s'avère efficace, il n'en va pas de même pour tous les autres (et en particulier pour le locuteur *si7*). Une décision partielle permet d'obtenir une probabilité d'erreur raisonnable : en ne prenant une décision pour le trait oral ou nasal d'une voyelle que dans les cas où sa durée n'est pas la valeur modale des deux distributions, alors l'erreur de l'estimateur tombe à 6% ; la contrepartie étant bien sûr le nombre réduit de cas sur lesquels notre décision sera proposée : environ 20% des observations de l'ensemble de notre corpus.

  Concernant les différences de longueurs des voyelles nasales, Di Cristo fait état d'un allongement légèrement supérieur de la nasale [˜] par rapport aux nasales [˜] et [˜] qui — toujours d'après l'auteur — ne présentent pas de différence notable. Il précise cependant que le rapport de durée bien que significatif n'excède pas 5%. Nos propres mesures sur le corpus AviLex ne nous permettent pas de conclure à une telle tendance de manière significative (cf table 4.4) pour l'ensemble des locuteurs.

Loc	˜/a		˜/		˜/		nasales/orales	
	rap %	err %	rap %	err %	rap %	err %	rap %	err %
fb	17	31	17	29	17	20	17	31
pg	26	25	18	34	28	23	21	30
ts	4	38	9	38	8	30	4	40
hm	44	20	36	23	23	33	34	28
si	11	34	9	36	3	46	6	39
lc	49	12	52	10	34	24	48	13
si7	12	42	2	40	9	43	9	45
pg7	18	33	11	39	23	29	17	34
tous	24	32	25	35	14	38	22	35

Table 4.6: Rapports des durées (mesurées par SPEX) des voyelles nasales et orales du corpus AviLex exprimés en pourcentage et taux d'erreur engendré lors d'une décision bayésienne de discrimination entre voyelle orale et nasale. La dernière colonne présente un rapport moyen des durées des nasales aux durées des voyelles orales.

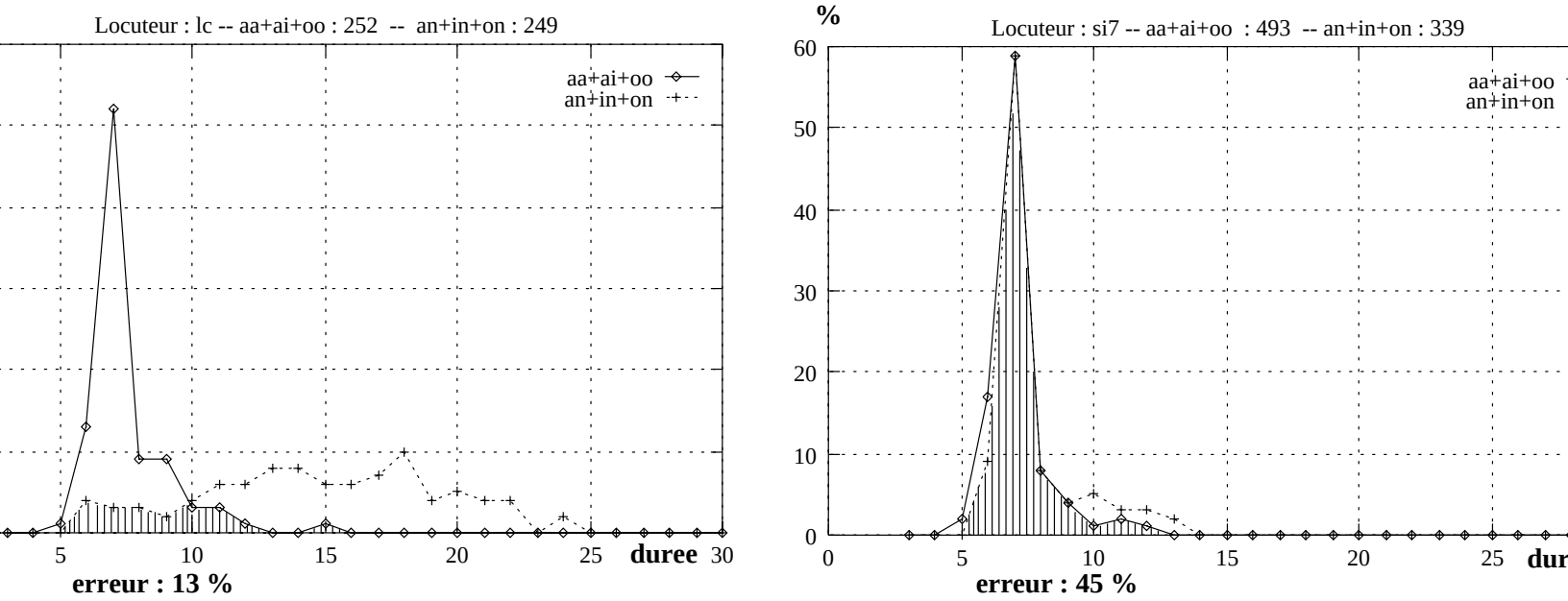


Figure 4.14: Distributions des durées des voyelles nasales et orales associées pour deux locuteurs. Dans le cas du locuteur *lc* une décision orale/nasale peut être envisagée avec efficacité, ce qui n'est pas du tout le cas pour le locuteur *si7*.

Au regard de nos précédentes mesures, nous avons constaté qu'il n'était pas évident de mettre en relation l'aperture d'une voyelle et sa durée. Il nous faut pour conclure définitivement étudier plus précisément la distribution des voyelles basses et hautes dans notre corpus. Le tableau 4.7 résume les moyennes et écarts-types mesurés sur ces mêmes voyelles en prenant soin de distinguer le contexte consonantique qui suit chaque voyelle. Plus précisément nous distinguons les consonnes par leur mode articulatoire[7] (occlusif/constrictif) et leur mode phonatoire (voisé/non voisé). La table 4.8 reporte le nombre d'observations pour chaque contexte. La lecture de ces mesures nous permet de constater que :

- Le nombre d'observations dans chaque contexte est relativement homogène pour chaque voyelle. On ne peut donc pas expliquer l'absence de corrélation nette entre l'aperture et la durée d'une voyelle par une hétérogénéité des contextes consonantiques. Notons que les voyelles [*i*] du locuteur *hm* de la base AviLex sont plus longues que les voyelles [*a*] uniquement dans le cas où elles sont suivies d'une consonne non voisée. Compte tenu du caractère isolé de cette observation, nous pouvons conclure prudemment à un artefact de prononciation du locuteur *hm*.

- L'effet du mode phonatoire de la consonne de droite, n'est là encore pas très marqué. On peut simplement remarquer pour les consonnes non voisées un écart-type généralement inférieur à celui mesuré pour les consonnes voisées. Ce qui en d'autres termes signifie que c'est dans le cas d'un contexte consonantique droit voisé que les observations s'écartent le plus de la moyenne (qui s'avère être le plus souvent la valeur modale de la distribution). Mais là encore une décision bayésienne sur le mode de voisement de la consonne subséquente ne pourrait être envisagée qu'avec une probabilité d'erreur avoisinant les 40% !

- L'effet allongeant des constrictives est également mis en relief au regard des données moyennes avec cependant un écart-type plus grand qui encore une fois entraînerait une décision occlusive/constrictive peu efficace sur la base de ces distributions.

[7] Bien que notre position soit de considérer que la production de toute consonne s'accompagne d'une constriction, nous emploierons la classification d'Alain Marchal dans [89] qui répertorie dans la classe constrictive les consonnes fricatives et les consonnes sonnantes.

Loc	a_v	a_nv	i_v	i_nv	y_v	y_nv	a_oc	a_co	i_oc	i_co	y_oc	y_co
fb	8 2.2	7 0.9	8 1.8	7 1.2	7 1.5	6 1.0	7 0.6	8 2.3	7 1.0	8 1.7	6 1.2	7 1.5
pg	7 1.3	7 1.4	7 1.2	7 1.0	7 1.2	6 0.9	6 1.4	7 1.3	7 0.7	7 1.2	6 1.0	7 1.1
hm	9 3.5	7 1.1	7 2.0	9 2.5	8 3.4	7 1.7	7 1.2	8 3.4	7 1.7	8 2.6	7 2.0	8 2.0
ts	9 3.7	7 1.4	7 2.0	7 1.3	8 2.9	7 2.6	7 1.1	8 3.8	6 1.0	7 1.8	8 3.1	7 2.9
lc	8 2.3	7 0.9	7 1.6	7 1.8	6 0.6	7 0.9	7 0.6	8 2.3	7 0.9	7 1.9	6 1.0	6 0.8
si	8 2.6	7 1.7	7 1.3	7 1.6	7 1.7	6 0.8	7 0.4	8 2.9	7 1.6	7 1.4	6 0.9	7 1.6
si7	7 1.5	7 1.3	7 2.2	7 0.9	6 1.7	7 1.8	7 1.8	7 0.9	7 1.6	7 1.9	6 0.9	7 2.3
pg7	7 1.4	7 1.2	6 1.3	7 1.0	6 1.4	7 1.1	7 1.0	7 1.5	6 0.9	7 1.3	6 0.8	7 1.6
tous	8 2.5	7 1.3	7 1.8	7 1.5	7 1.9	7 1.5	7 1.2	8 2.5	7 1.3	7 1.8	6 1.3	7 1.9

Table 4.7: Moyennes et écart-type des durées des voyelles [*a*], [*i*] et [*y*] du corpus AviLex mesurées par SPEX dans différents contextes consonantiques droits : *v* voisé, *nv* non voisé, *co* constrictif et *oc* occlusif.

Loc	a_v	a_nv	i_v	i_nv	y_v	y_nv	a_oc	a_co	i_oc	i_co	y_oc	y_co
fb	74	66	69	71	15	12	53	65	36	99	9	13
pg	76	68	72	65	14	12	56	66	34	97	9	13
hm	69	63	64	62	13	9	48	63	36	85	7	11
ts	76	63	65	65	14	10	50	64	33	92	7	13
lc	66	61	59	62	12	10	51	57	29	87	8	10
si	74	66	70	70	15	11	52	67	35	101	8	13
si7	103	79	106	89	52	29	78	87	76	104	29	36
pg7	112	81	115	90	52	30	81	89	79	113	27	41
tous	650	547	620	574	187	123	469	558	358	778	104	150

Table 4.8: Nombre d'observations des voyelles [*a*], [*i*] et [*y*] du corpus AviLex en fonction de leur contexte consonantique droit.

Nous achevons notre étude en analysant l'influence de la position de la voyelle dans le mot ainsi que l'incidence du nombre de voyelles qui le constituent sur la durée segmentale des voyelles. La figure 4.15 montre les valeurs moyennes des voyelles orales en fonction du nombre de voyelles des mots. Bien que la distribution des mots en terme de nombre de voyelles ne soit pas homogène (voir les distributions des mots du corpus AviLex en fonction de leur nombre de voyelles sur la figure 3.6 de la page 51) on constate la pente négative de toutes les courbes qui confirme une tendance déjà énoncée par de nombreux prédécesseurs : la durée vocalique est inversement proportionnelle au nombre de syllabes dans le mot. Un aperçu des distributions des durées de la figure 4.16, nous rappellera cependant — comme précédemment — qu'il est difficile d'exploiter de telles mesures avec fiabilité. La figure 4.17 montre enfin la durée moyenne des voyelles orales dans les mots de 3 et 4 voyelles afin de vérifier l'influence de la position de la voyelle dans le mot sur la durée. On remarquera simplement que la durée semble augmenter depuis la première syllabe jusqu'à la dernière. Les écarts-types — qui ne sont pas reportés — sont cependant suffisamment grands pour qu'aucune décision ne puisse être effectuée avec une probabilité d'erreur acceptable. Di Cristo [40, p. 431] signale que les voyelles sont sensiblement plus longues dans les syllabes médianes que dans les syllabes initiales. Cette tendance n'est pas décelée ici.

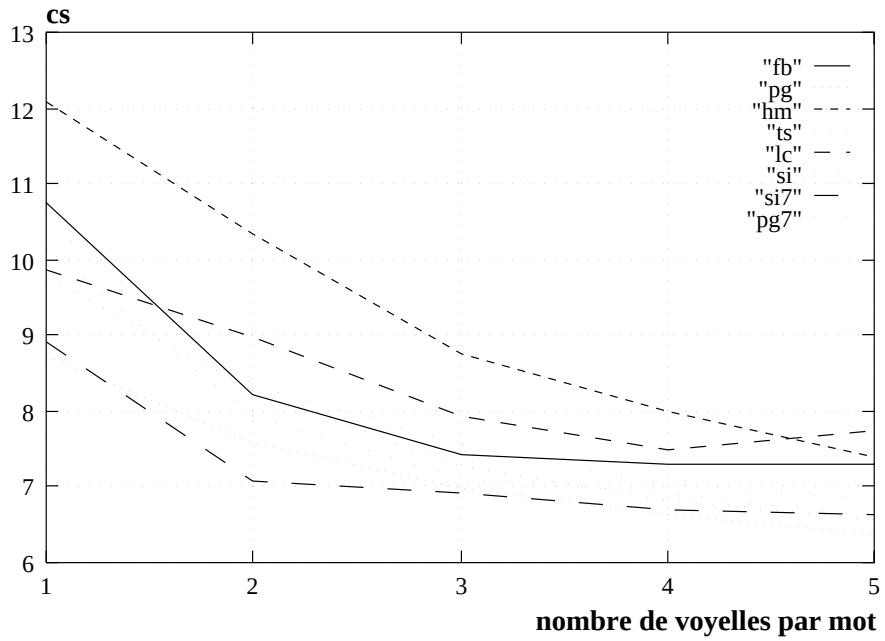


Figure 4.15: Moyennes des durées des voyelles orales du corpus AviLex mesurées par le système SPEX en fonction du nombre de voyelles dans le mot.

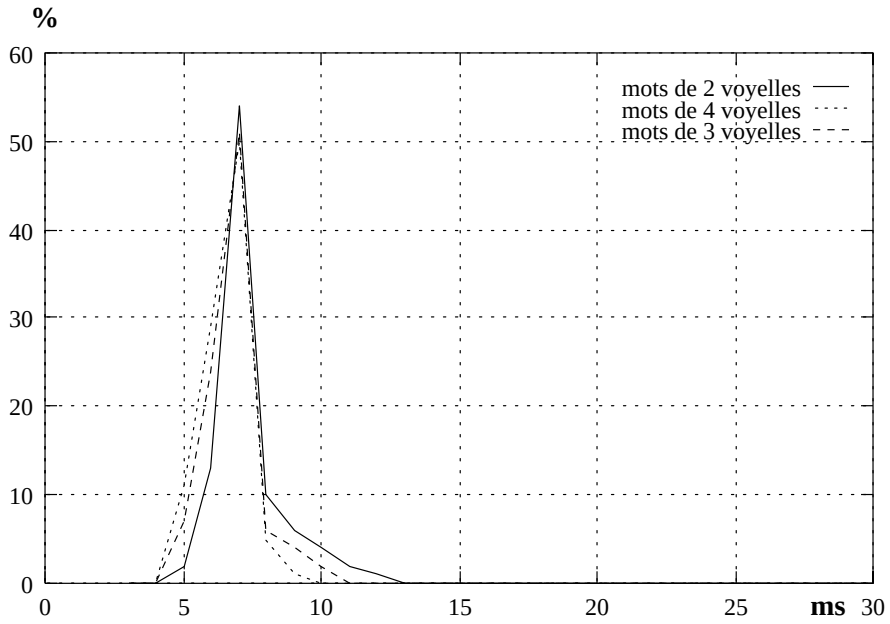


Figure 4.16: Distribution des durées des voyelles orales pour les mots de 2, 3 et 4 voyelles pour les réalisations d'un locuteur du corpus AviLex. Des courbes similaires sont obtenues pour les autres locuteurs.

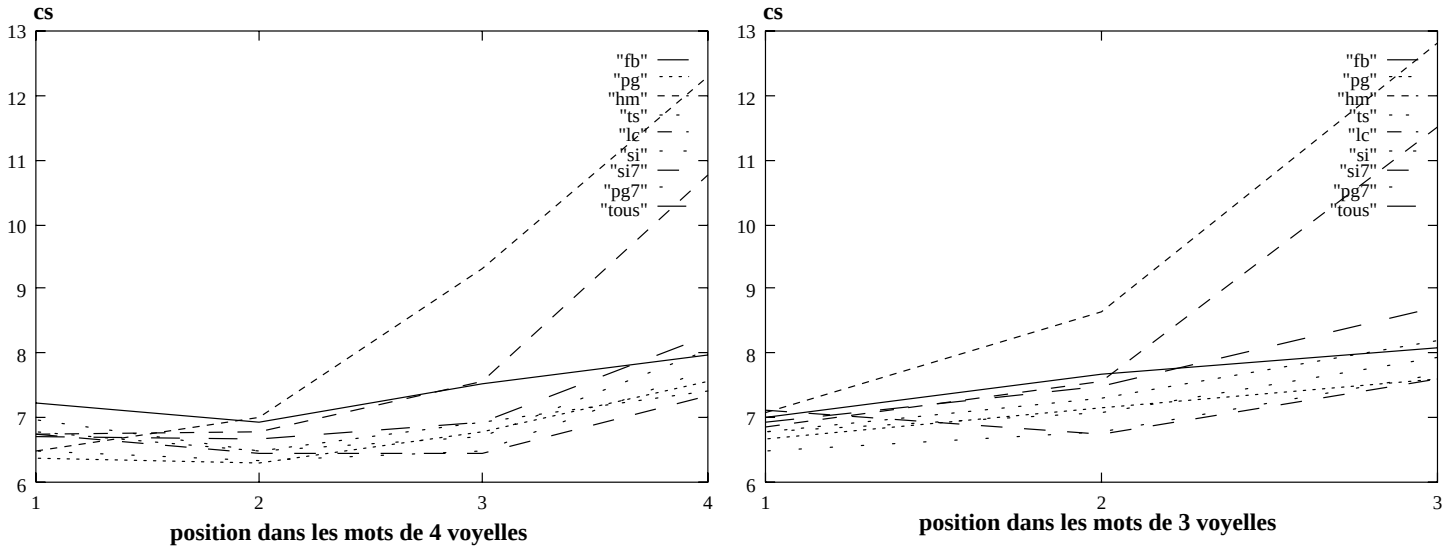


Figure 4.17: Moyennes des voyelles orales dans les mots de 3 et 4 voyelles dans toutes les positions.

Bien qu'étant conscient des limites de cette série de mesures, nous pouvons cependant nous risquer à quelques commentaires :

- Notre mode de calcul des durées ne permet pas toujours de mettre en évidence les observations faites par nos prédécesseurs. Lorsque c'est le cas pour des valeurs moyennes, il n'est pas pour autant assuré qu'une décision bayésienne prise à partir des distributions des diverses observations, puisse être suffisamment fiable. Ceci ne remet cependant aucunement en cause les études existantes sur le sujet mais renforce simplement les craintes que nous avions formulées plus haut sur la difficulté de mesurer les durées automatiquement avec précision.

- L'investigation des variations microprosodiques ne saurait être limitée à l'analyse du seul corpus AviLex qui — bien que composé de plusieurs locuteurs prononçant de nombreux mots — ne présente pas un nombre de représentants suffisant de tous les contextes vocaliques (nature de la voyelle, mode phonatoire et articulatoire de la consonne qui suit, position dans le mot, ...). Elle ne saurait non plus s'accommoder des seules mesures proposées par notre système d'accès lexical.

Nous allons donc maintenant réaliser des mesures de durée à l'aide des modèles de phonèmes présentés dans le chapitre 2 sur différents corpus de mots isolés.

Durée obtenue par un système stochastique

Pour obtenir des modèles de phonèmes de qualité, un grand nombre de données a été nécessaire. Cette contrainte nous a imposé d'entraîner nos modèles sur la base téléphonique PolyVar qui possède un grand nombre d'échantillons de parole. Aussi l'étude des durées que nous présentons décrit des corpus de qualité téléphonique : le corpus PVM extrait du corpus PolyVar et le corpus AviTel conçu spécialement pour cette étude.
La chaîne de phonèmes d'un mot, obtenue à l'aide du lexique BdLex est utilisée par l'algorithme de Viterbi pour effectuer un alignement de la suite de modèles phonétiques avec le signal de parole. Une analyse visuelle de la segmentation pour des mots contenant des consonnes voisées (plus difficiles à délimiter) nous a permis d'en vérifier la qualité (voir la figure 4.18).
Les analyses qui suivent sont limitées aux huit locuteurs les plus présents (en terme d'occurrences de mots prononcés) dans la base ; les valeurs moyennes calculées pour tous les locuteurs de la base seront cependant présentées (sous le libellé *tous*). Le tableau 4.9 présente les longueurs (exprimées en centiseconde) recueillies pour les différentes voyelles de notre corpus indépendamment de sa position dans le mot et de son contexte consonantique. Le tableau 4.10 indique le nombre d'observations de chaque voyelle.
On constate tout de suite une certaine disparité dans les cardinalités qui s'explique par le nombre réduit de mots différents dans la base PVM et par le nombre inégal de répétitions de chacun de ces mots. Nous éviterons donc dans la suite de cette analyse de tirer des conclusions hâtives quant aux valeurs obtenues pour les voyelles [], [u] ou encore [$\tilde{}$] qui possèdent peu de représentants. Une première lecture du tableau 4.9 nous permet de

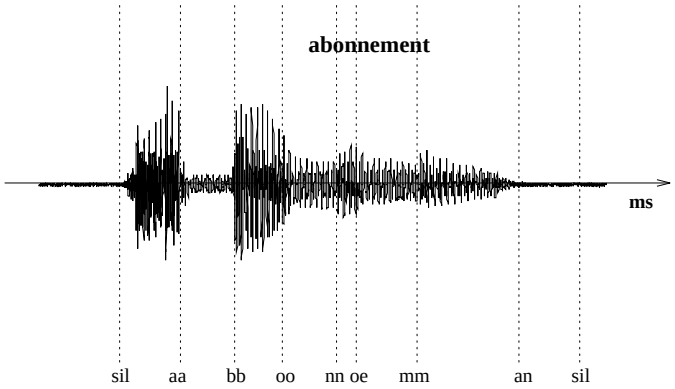


Figure 4.18: Exemple d'alignement de Viterbi obtenu à partir de nos modèles d'allophones pour le mot *abonnement*.

constater que la relation liant l'aperture d'une voyelle à sa durée n'est pas frappante (une vérification visuelle à l'aide de la figure 4.19 confirme ce constat pour deux locuteurs de la base PVM). La seule observation conforme aux études de nos prédécesseurs qui paraît corroborée par nos valeurs semble être l'allongement sensible des nasales comparées aux voyelles orales "associées". Cette constatation est confirmée par le tableau 4.11 qui fait état d'un rapport moyen des durées des voyelles nasales sur les durées des voyelles orales de l'ordre de 30%. La probabilité d'erreur d'une décision bayésienne sur le trait oral/nasal d'une voyelle quelconque prise à partir des distributions mesurées sur le corpus PVM est d'environ 27%.

Loc	*a*	*i*	*y*		*e*		*u*	*o*		~	~	~	*∂*
CG	12 6.6	12 5.2	12 2.4	15 8.5	15 6.3	13 6.7	18 5.5	15 7.8	15 7.1	20 4.4	17 0.0	18 5.7	11 6.8
CM	11 5.7	12 5.4	11 3.4	6 2.1	14 6.7	12 6.8	16 5.6	14 7.9	10 4.2	17 5.0	11 1.8	13 4.5	8 4.4
GM	10 6.0	11 4.8	11 2.9	6 2.4	11 5.2	12 4.0	11 0.0	13 7.3	10 5.5	14 3.7	8 0.0	14 5.1	5 2.7
ME	9 5.4	10 4.5	9 2.6	7 8.2	11 6.1	11 5.0	14 4.9	12 6.8	9 3.0	16 4.8	13 3.8	15 5.4	6 3.2
LP	11 5.2	12 4.9	10 3.1	9 8.2	14 6.2	13 6.2	17 4.8	15 7.8	13 6.2	18 4.2	14 3.2	16 4.1	7 3.6
CJL	11 6.2	12 5.5	8 1.8	7 0	13 7.7	13 6.4	19 6.4	10 6.5	9 1.2	17 4.9	0 0	18 5.9	4 0.9
VKR	11 6.0	12 5.8	10 2.4	9 6.7	14 6.6	13 6.3	15 7.0	13 7.6	11 3.5	17 4.0	15 5.0	15 3.6	6 2.3
AS	8 3.2	11 3.6	10 2.4	9 6.1	11 3.9	10 4.3	9 1.4	8 3.2	7 2.0	12 2.9	10 2.4	8 2.2	5 1.3
tous	11 5.8	11 5.3	10 3.5	9 7.2	13 6.7	13 6.4	15 6.2	13 7.6	11 5.1	17 5.1	14 4.3	16 5.5	7 4.2

Table 4.9: Moyennes et écarts-types des durées obtenues par nos modèles de phonèmes non contextuels pour les huit locuteurs les plus représentés du corpus PVM. Le libellé *tous* précise le nombre de voyelles considéré pour l'ensemble des locuteurs de la base.

Loc	*a*	*i*	*y*		*e*		*u*	*o*		˜	˜	˜	*∂*
CG	109	88	40	6	93	41	4	27	34	38	1	32	29
CM	138	86	43	7	87	56	8	31	32	60	4	34	25
GM	50	39	20	3	46	18	1	13	10	26	1	9	10
ME	215	141	56	8	143	87	12	63	39	68	8	60	37
LP	188	138	61	11	120	74	12	56	41	58	6	52	36
CJL	79	56	22	1	42	28	4	20	12	21	0	22	9
VKR	153	110	48	11	120	70	11	42	34	66	4	46	31
AS	95	58	31	5	57	36	5	29	25	28	3	29	13
tous	1872	1258	545	91	1197	781	113	479	383	659	62	530	346

Table 4.10: Nombre d'observations de chaque voyelle du corpus PVM pour les huit locuteurs les plus représentés de la base. Le libellé *tous* précise le nombre de voyelles considéré pour l'ensemble des locuteurs de la base.

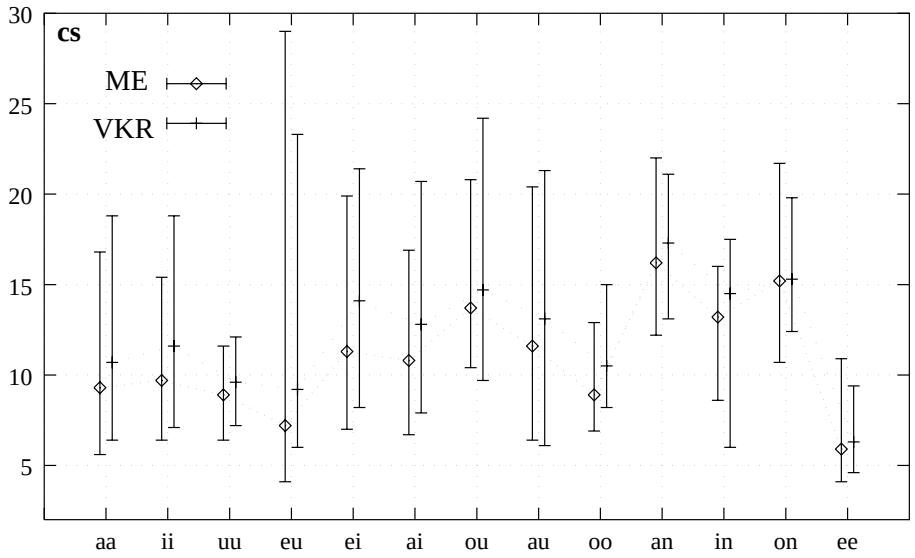


Figure 4.19: Durées moyennes des différentes voyelles du français pour deux locuteurs de la base PVM. Chaque phonème est décrit par une ligne verticale reliant les trois points suivants (par ordre décroissant de valeurs) : l'écart-type des valeurs supérieures à la moyenne, la moyenne des durées pour le phonème et l'écart-type des durées inférieures à la moyenne.

Loc	˜/a		˜/		Nasale/Orale	
	rap %	err %	rap %	err %	rap %	err %
CG	39	16.3	18	20.9	32	21.6
CM	35	20.4	26	22.4	30	25.4
GM	26	16.8	34	10	24	21.8
ME	43	18.7	41	15.3	38	21.6
LP	39	19.7	15	20.5	31	23.6
CJL	33	14.7	52	4.2	33	19.9
VKR	38	16.2	31	15.1	32	18.5
AS	35	21.1	16	31.2	18	31.2
tous	37	23.8	31	26.3	31	27.5

Table 4.11: Rapports des durées exprimés en pourcentage des voyelles nasales et orales associées du corpus PVM. Les rapports $[\tilde{\varepsilon}]/[\varepsilon]$ ne sont pas reportés car trop peu représentés dans ce corpus.

À l'instar de l'étude des durées fournies par notre module d'accès lexical, il convenait de vérifier l'influence du contexte consonantique droit sur la durée intrinsèque des voyelles. Les tableaux 4.12 et 4.13 résument les moyennes, écarts-types et cardinalités des observations recueillies à cet effet. Nous ne pouvons qu'observer une large supériorité numérique des contextes consonantiques droits voisés et ce aussi bien pour les voyelles hautes que basses (ce qui ne peut donc pas expliquer les valeurs peu bavardes du tableau 4.9). Ces données nous indiquent également que l'influence du mode phonatoire de la consonne subséquente n'est pas mis en relief ici, pas plus que l'importance de son mode articulatoire (on note même ici une moyenne de durée plus faible dans le cas de voyelles suivies de constrictives).
L'étude de l'influence de la position sur les valeurs intrinsèques est maintenant proposée : les figures 4.20 et 4.21 montrent d'une part, que les durées moyennes mesurées sont inversement proportionnelles à la longueur des mots prononcés et d'autre part, rappellent que pour une longueur de mot donnée les durées des voyelles dépendent de leur position dans le mot. Le tableau 4.14 reporte les probabilités d'erreur d'une prise de décision de la longueur (en nombre de voyelles) d'un mot en fonction de la durée d'une voyelle de ce mot et précise, sans surprise, que les distributions sont d'autant plus distinctes que la longueur des mots est éloignée : la plus petite probabilité d'erreur est obtenue pour une distinction entre les mots de une puis cinq voyelles (12%), la plus grande incombant à une discrimination entre les mots de 3 et 4 voyelles (40%).

Loc	a_v	a_nv	i_v	i_nv	y_v	y_nv	a_oc	a_co	i_oc	i_co	y_oc	y_co
CG	11 5.7	10 1.4	10 3.7	11 2.2	12 2.2	11 0.8	11 2.9	11 5.5	13 2.9	10 3.1	11 0.8	12 1.9
CM	10 4.6	10 2.2	10 3.7	11 3.9	10 2.1	11 1.4	11 1.9	10 4.6	13 4.2	10 3.4	11 1.4	10 2.0
GM	11 5.6	9 5.1	10 4.5	10 4.5	11 3.1	10 0.0	12 3.3	10 6.4	13 4.2	9 4.3	10 0.0	12 2.3
ME	9 4.8	7 2.2	9 3.5	8 2.5	9 2.4	10 0.4	9 2.7	8 4.8	11 3.0	8 2.7	10 0.4	10 1.8
LP	11 4.2	10 3.1	11 4.6	10 3.5	11 2.9	12 0.5	12 3.6	11 4.1	15 3.2	9 3.4	12 0.5	11 1.7
CJL	11 5.2	9 1.5	9 2.7	8 3.1	8 1.6	10 0.5	10 1.5	11 5.4	13 1.9	8 2.5	10 0.5	8 1.4
VKR	11 6.5	10 2.1	11 4.6	8 2.7	10 2.3	12 0.0	12 2.4	10 6.3	13 3.3	8 3.8	12 0.0	10 1.7
AS	8 3.5	7 1.1	11 3.9	10 3.0	10 2.3	12 1.0	12 3.4	7 2.5	12 2.5	9 3.3	12 1.0	10 2.1
tous	10 5.3	10 3.6	10 4.5	9 3.5	10 3.3	11 1.5	11 3.4	10 5.2	13 3.9	9 3.8	11 1.5	11 2.9

Table 4.12: Moyennes et écarts-types des voyelles [i], [y] et [a] des huit locuteurs les plus présents de la base PVM en fonction des différents contextes consonantiques droits : v voisé, nv non voisé, co constrictif et oc occlusif. Le libellé *tous* indique les valeurs moyennes pour l'ensemble des locuteurs de la base.

Loc	a_v	a_nv	i_v	i_nv	y_v	y_nv	a_oc	a_co	i_oc	i_co	y_oc	y_co
CG	55	20	43	26	32	3	11	57	16	52	3	24
CM	64	22	40	23	32	3	13	68	12	50	3	19
GM	24	5	21	7	14	1	6	18	6	20	1	11
ME	101	31	65	38	39	4	22	101	26	75	4	30
LP	86	32	65	41	45	2	15	95	29	73	2	32
CJL	38	10	26	11	14	3	10	34	7	28	3	10
VKR	71	23	50	34	37	1	13	75	23	58	1	30
AS	46	9	22	19	22	2	10	39	14	25	2	14
tous	869	300	576	366	399	29	177	914	236	673	29	275

Table 4.13: Cardinalités des voyelles [a], [i] et [y] des huit locuteurs les plus récents de la base PVM dans différents contextes consonantiques droits : v voisé, nv non voisé, co constrictif et oc occlusif.

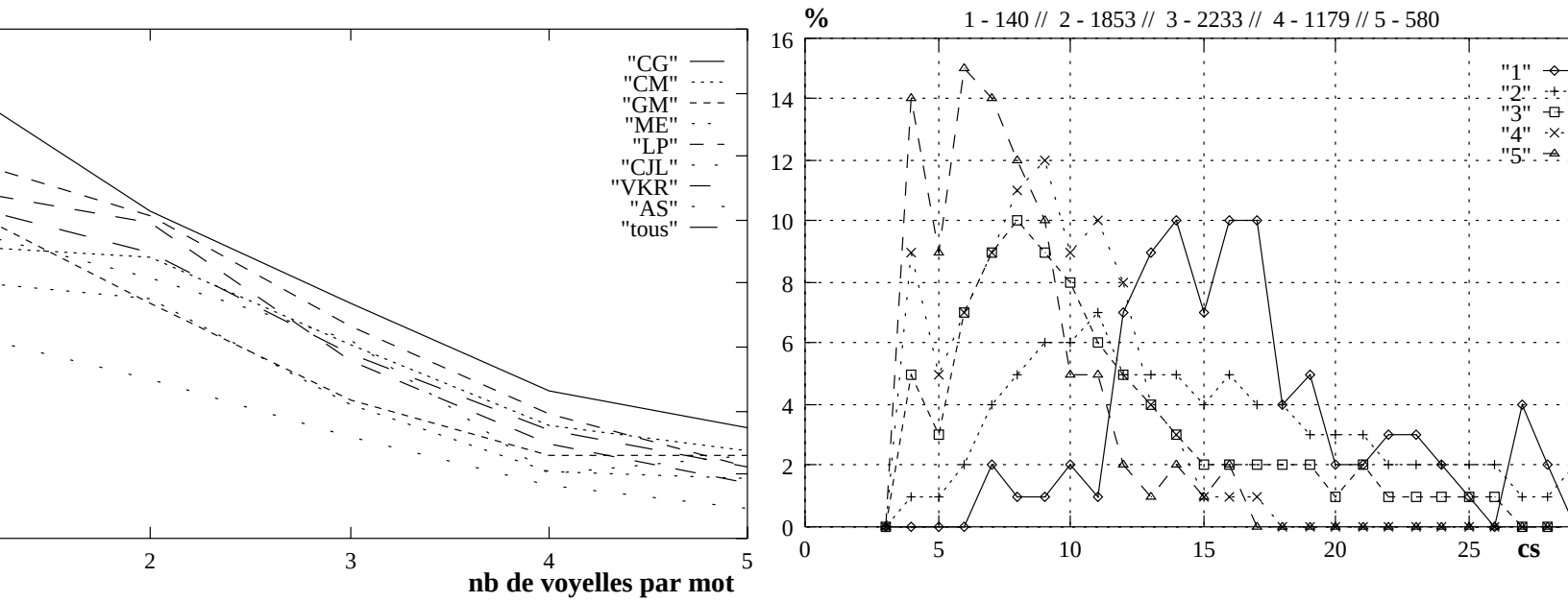


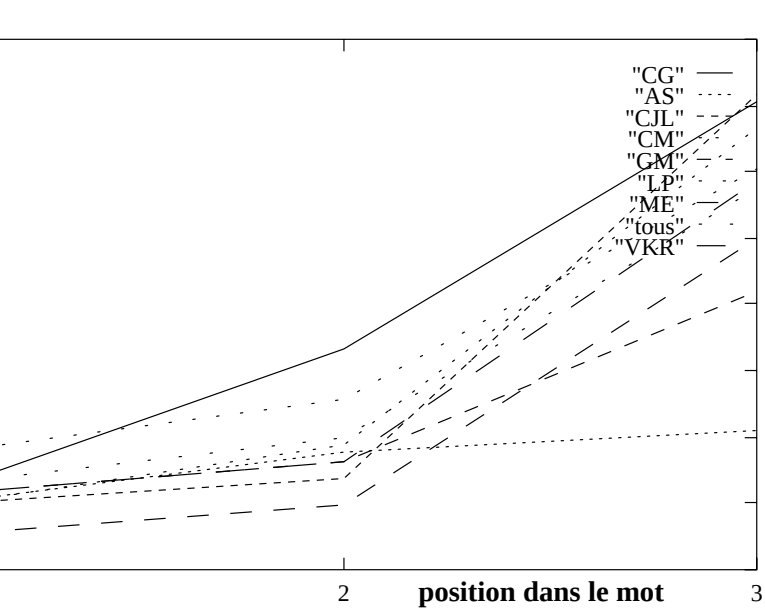
Figure 4.20: Moyennes des durées des voyelles orales du corpus PVM en fonction du nombre de voyelles par mot et distributions des observations associées pour l'ensemble des locuteurs de la bases. Des distributions semblables sont obtenues dans des contextes plus spécifiques (voyelles hautes, voyelles basses,...).

Nous allons maintenant compléter notre étude des variations microprosodiques de durée des voyelles en présentant deux séries de mesures :

- Afin de vérifier l'influence du choix d'un corpus sur nos mesures, nous allons étudier brièvement celles réalisées sur le corpus AviTel qui, rappelons-le, est composé des mêmes 500 mots de la base AviLex1 prononcés par deux locuteurs à travers une ligne téléphonique.

- À aucun moment de nos investigations sur les durées, nous n'avons été en mesure de fournir des observations prises dans les mêmes conditions (position dans le mot, longueur du mot, contexte consonantique précis,...). Nous avons à chaque fois privilégié l'étude d'un facteur en nous contentant de vérifier que la répartition des autres était suffisamment homogène pour ne pas produire un biais dans nos conclusions. Nous allons combler cette lacune en présentant les mesures réalisées sur le corpus FeLex qui a été spécialement conçu pour disposer d'une répartition parfaitement équitable des divers facteurs pouvant intervenir dans les variations des durées des voyelles.

Nous rappelons que ces deux derniers corpus étant enregistrés dans les mêmes conditions (ligne téléphonique) que le corpus PolyVar ayant servi à l'apprentissage de nos modèles



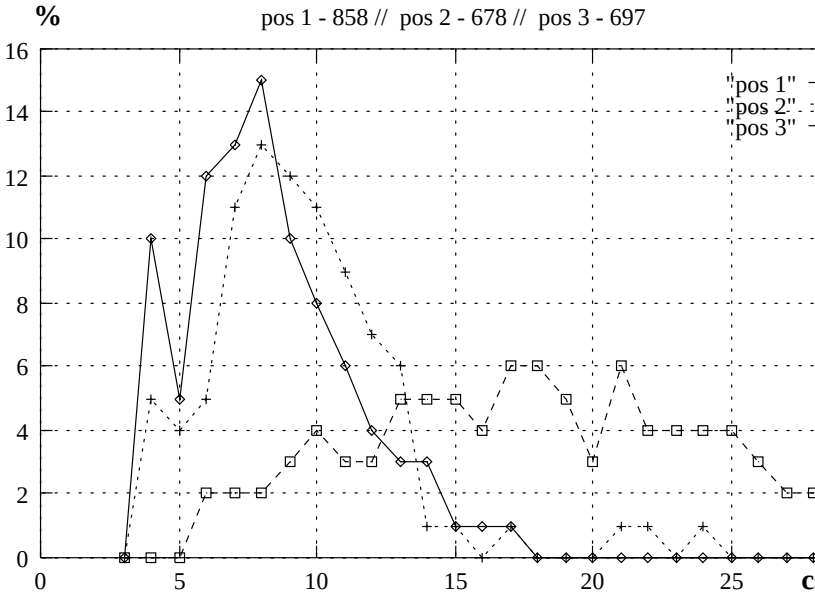


Figure 4.21: Moyennes des durées des voyelles orales dans les mots de 3 voyelles du corpus PVM en fonction de la position de la voyelle dans le mot et distributions des observations associées. Les probabilités d'erreur qu'engendrerait une décision prise à partir des distributions 1/2, 1/3 puis 2/3 sont respectivement de 39.7%, 17.2% puis 19.7%.

de phonèmes, les mesures présentées dans les tables et figures qui suivent sont également obtenues automatiquement à l'aide de ces modèles.
La figure 4.22 présente les mesures effectuées sur la base AviTel. On peut simplement retenir qu'à l'instar des corpus précédents, les différences de durées résultant de l'aperture des voyelles orales ne sont pas significatives, contrairement à la différence de durée des voyelles orales et nasales (détaillée en figure 4.23). On se contentera simplement ici de noter que la probabilité d'erreur d'une décision du trait oral/nasal d'une voyelle prise à partir des distributions des observations du corpus AviTel est proche de celle mesurée pour le corpus PVM qui était pour l'ensemble des locuteurs de la base de 27%.

longueur des mots	1	2	3	4	5
1	-	33.9 %	24.6 %	16.7 %	12.4 %
2	33.9 %	-	37.7 %	28.5 %	24.1 %
3	24.6 %	37.7 %	-	40.1 %	35.5 %
4	16.7 %	28.5 %	40.1 %	-	38.1 %
5	12.4 %	24.1 %	35.5 %	38.1 %	-

Table 4.14: Table présentant les probabilités d'erreur associées à la décision — qui serait prise à partir des distributions des observations mesurées sur le corpus PVM — du nombre de voyelles d'un mot avec comme seule information discriminante la durée d'une voyelle de ce mot.

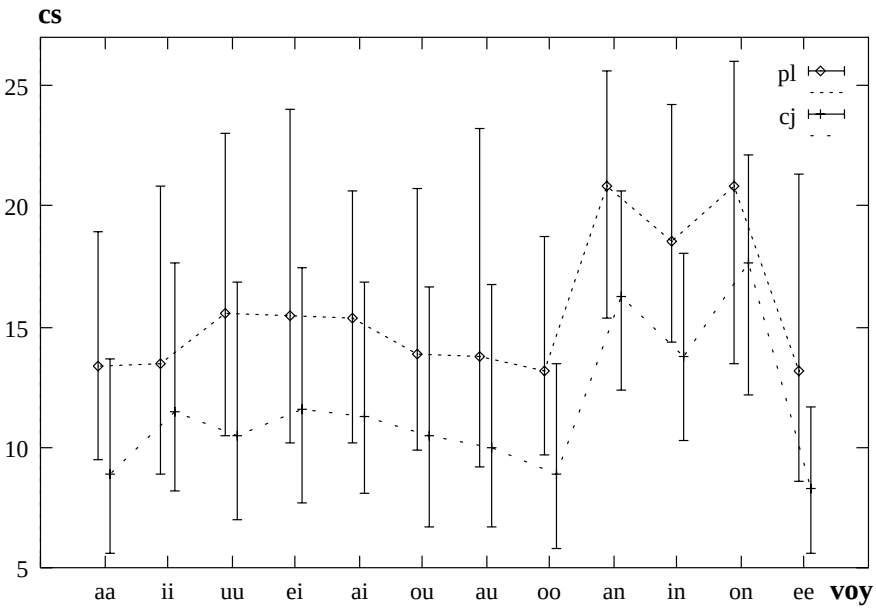


Figure 4.22: Durées moyennes des différentes voyelles du français pour les deux locuteurs de la base AviTel.

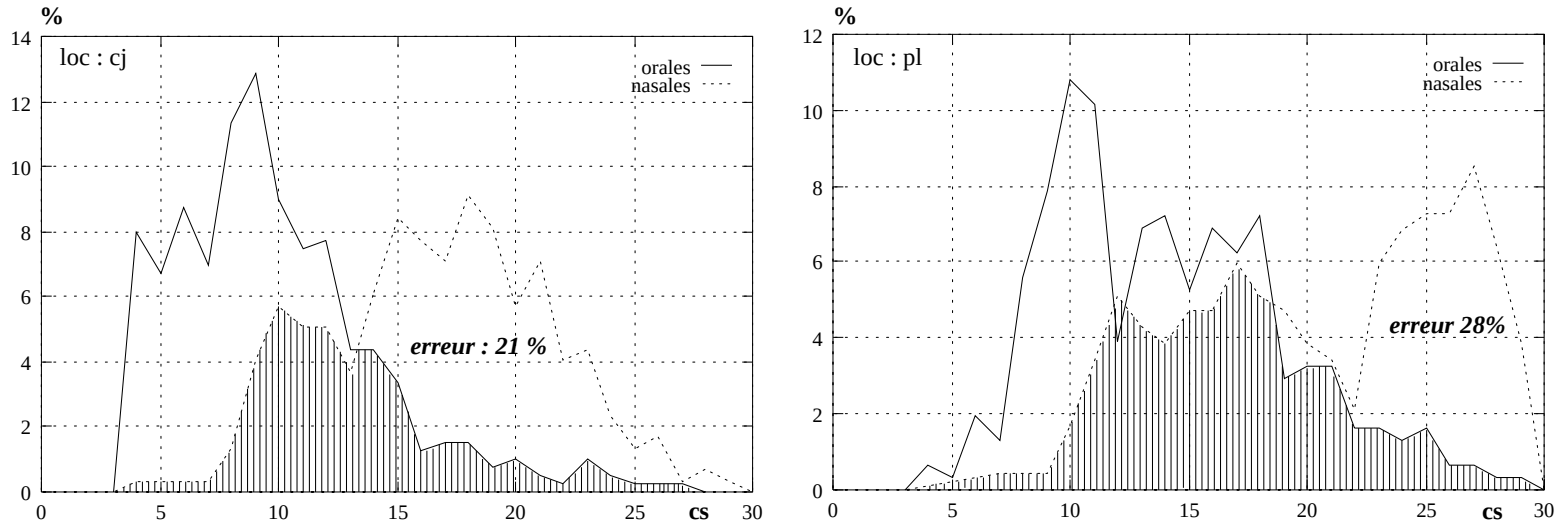


Figure 4.23: Distributions des voyelles orales et nasales pour les deux locuteurs de la base AviTel.

Les mesures réalisées sur le corpus FeLex sont reportées dans les tableaux 4.15, 4.16 et 4.17. Elles appellent quelques commentaires qui nous permettront de clore l'étude des durées des voyelles :

- Bien que les voyelles semblent légèrement plus longues (4%) au contact des consonnes subséquentes voisées (tendance plus marquée en position médiane de mot), nous ne pouvons que remarquer qu'il ne s'agit pas là d'une différence significative (les probabilités d'erreur mesurées sur les distributions de ces observations sont de l'ordre de 40%).

- Les voyelles hautes ne sont pas plus courtes que les voyelles basses aux vues du tableau 4.17 et cela quelque soit la position de la voyelle dans le mot (on remarque même une tendance inverse !).

	VH		VM		VB		Nasales	
Loc	V	NV	V	NV	V	NV	V	NV
PL	12.0 3.6	11.2 3.9	10.9 2.5	11.0 1.5	12.3 3.2	12.7 3.4	16.1 2.5	15.5 2.4
CJ	11.4 3.9	9.5 3.4	9.0 2.5	9.6 2.7	9.2 3.6	9.3 3.5	13.6 3.4	12.5 3.0
tous	11.7 3.8	10.4 3.8	10.0 2.7	10.3 2.3	10.8 3.8	11.0 3.8	14.9 3.3	14.1 3.1

Table 4.15: Durées moyennes des voyelles hautes (VH), moyennes (VM), basses (VB) et nasales prises *à l'initiale* des mots du corpus FeLex en distinguant les contextes consonantiques droits voisés (V) et non voisés (NV). Chaque case contient la durée moyenne et l'écart-type calculés à partir d'une centaine d'observations.

	VH		VM		VB		Nasales	
Loc	V	NV	V	NV	V	NV	V	NV
PL	12.0 2.9	10.4 2.7	10.3 2.4	9.4 2.2	10.4 1.5	10.0 1.9	15.7 2.3	14.8 2.4
CJ	10.0 2.9	9.6 2.2	7.6 2.1	7.8 1.5	6.5 1.9	7.1 1.4	12.2 2.1	11.6 2.1
tous	11.0 3.0	10.0 2.5	8.9 2.6	8.6 2.0	8.5 2.6	8.6 2.2	14.0 2.8	13.2 2.8

Table 4.16: Durées des voyelles hautes (VH), moyennes (VM), basses (VB) et nasales prises *en position médiane* des mots du corpus FeLex en prenant soin de distinguer les contextes consonantiques droits voisés (V) et non voisés (NV). Chaque case contient la durée moyenne et l'écart-type calculés pour une centaine d'observations.

	position initiale							
Loc	voy. hautes		voy. moyennes		voy. basses		nasales	
PL	11.6	3.8	11.0	2.1	12.5	3.3	15.8	2.5
CJ	10.5	3.8	9.3	2.6	9.2	3.6	13.1	3.2
tous	11.1	3.8	10.1	2.5	10.9	3.8	14.5	3.2

	position médiane							
Loc	voy. hautes		voy. moyennes		voy. basses		nasales	
PL	11.2	2.9	9.9	2.4	10.2	1.7	15.3	2.4
CJ	9.8	2.6	7.7	1.8	6.8	1.7	11.9	2.1
tous	10.5	2.8	8.8	2.4	8.5	2.4	13.6	2.8

	position finale							
Loc	voy. hautes		voy. moyennes		voy. basses		nasales	
PL	19.7	3.8	18.2	2.5	18.7	2.8	18.3	2.6
CJ	18.2	4.4	16.8	3.4	14.5	3.4	18.4	3.3
tous	18.9	4.2	17.5	3.0	16.6	3.7	18.3	3.0

Table 4.17: Moyennes et écart-type des durées des voyelles hautes, moyennes, basses et nasales du corpus FeLex observées pour les positions initiale, médiane et finale de mot.

Bilan de l'étude des durées

Un bilan de cette section sur la durée des voyelles s'impose maintenant. De tous les facteurs recensés pouvant intervenir sur les valeurs intrinsèques des durées des voyelles, il semble que peu sont observables — du moins par les techniques employées ici — au-delà de simples valeurs moyennes. On peut cependant retenir :

- qu'une distinction du trait oral/nasal d'une voyelle peut être envisageable à partir de sa durée mesurée par nos modèles de phonèmes ; l'étude de corpus AviLex avec les durées fournies par SPEX faisant ressortir une impression contraire ou du moins fortement dépendante du locuteur considéré,

- que la position de la voyelle influence fortement la longueur de celle-ci sans pour autant que les différences imputables aux différents facteurs ne soient plus facilement localisables de manière significative en finale de mot,

- que l'influence consonantique droite n'est pas facilement mesurable, loin s'en faut,

- qu'il n'est pas aisé de fournir de manière automatique une bonne mesure des durées, c'est pourquoi il ne nous semble pas réaliste — même si cela est très séduisant — de

proposer des coefficients de pondération robustes — dans le cadre d'une application automatique — de correction des effets microprosodiques,

- que la syllabe n'a fait l'objet d'aucune étude particulière. Il est en effet connu que la durée des voyelles dépend — entre autre facteur — de la nature de la syllabe dans laquelle elle est enchâssée. Plusieurs raisons nous ont fait ignorer ce fait dont la complication des mesures déjà fastidieuses qui ont été présentées, l'aspect non déterministe de la syllabation et donc non automatisable totalement (au moins pour le français) [8]. Enfin, nous avons vérifié la répartition uniforme du nombre de syllabes ouvertes et fermées dans nos différents corpus.

4.4.2 La fréquence fondamentale

L'étude des variations microprosodiques de la fréquence fondamentale a suscité l'intérêt de nombreux chercheurs depuis de longues années [152, 40, 42, 121]. On attribue habituellement au paramètre de fréquence fondamentale de nombreuses possibilités segmentales qui ne sont pas ou peu employées dans les systèmes de reconnaissance pour plusieurs raisons dont la principale est la fiabilité des mesures de *f0*. Vaissière précise par exemple [157] que la tâche de lecture de spectrogrammes est grandement facilitée par l'affichage de la courbe de la fréquence fondamentale du moins pour une langue comme le français. Plus particulièrement, elle décrit les principaux effets que la prosodie peut apporter à une tâche de segmentation :

- Le contour de fréquence fondamentale pourrait éliminer dans une phase de vérification certains cas de sur-segmentation ou plus rarement de sous-segmentation. Nous pensons que ceci est effectivement concevable dès lors que l'on travaille sur des unités supérieures au mot. Donnons simplement comme exemple la phrase *Il se garantira du froid avec un bon capuchon* où la courbe du fondamental pourrait être utilisée pour séparer le [*a*] terminal du mot *froid* du [*a*] initial du mot *avec*. La figure 4.24 illustre ceci pour deux prononciations de cette phrase via le canal téléphonique. Il est assez intéressant de noter la ressemblance frappante entre ces deux courbes, d'autant plus que les autres réalisations de cette même phrase — toutes du même locuteur — sont également très semblables.

- La présence de *f0* peut être utilisée en français pour distinguer les consonnes voisées des autres ; remarque à laquelle nous souscrivons totalement.

- La forme de la courbe *f0* peut fournir des indications pour la distinction des consonnes obstruentes et non obstruentes. Les consonnes obstruentes étant généralement accompagnées d'une configuration concave de la courbe du fondamental pouvant s'expliquer par une baisse de la pression sous-glottique. Nous montrerons dans la suite ce qu'il en est sur nos corpus de mots isolés.

[8] Mariani [91, p. 362] rappelle par exemple que la syllabation de phrases (dont on connait le texte) est parfois délicate particulièrement dans le cas de *∂* caduques non prononcés.

Nous allons donc, à l'instar des durées, faire l'inventaire des principaux phénomènes liés aux variations microprosodiques du fondamental [9] mis en relief par les travaux passés puis tenter de déterminer ceux qui peuvent servir notre tâche de filtrage lexical.

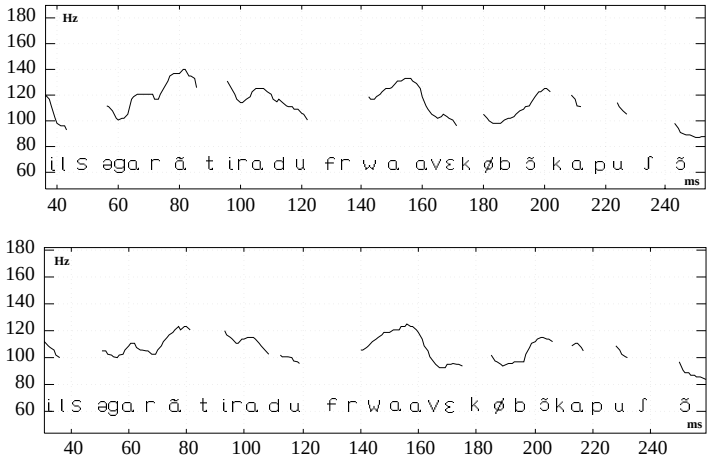


Figure 4.24: Courbes de fréquence fondamentale de deux réalisations de la phrase *il se garantira du froid avec un bon capuchon* pour un même locuteur.

Les facteurs responsables des variations intrinsèques et co-intrinsèques de la fréquence fondamentale des voyelles sont peu nombreux et peuvent se résumer au mode articulatoire de la voyelle (aperture et nasalité) et au caractère voisé ou non voisé des consonnes précédentes :

- Une tendance commune à l'ensemble des langues étudiées fait état d'une relation entre la hauteur d'une voyelle et la moyenne de la fréquence fondamentale mesurée lors de sa réalisation. Les valeurs de *f0* les plus fortes sont associées aux voyelles les plus hautes (comme [*i*] et [*y*]) alors que les valeurs les plus basses sont observées pour les voyelles basses (et plus particulièrement la voyelle [*a*]). L'écart intrinsèque des valeurs de *f0* pour les voyelles hautes et basses varie selon les études de 6% à 17%, différence pouvant s'expliquer par la diversité des corpus analysés, les différentes prises de mesure du paramètre et l'influence de chaque langue.

- Le trait de nasalité semble affecter la fréquence intrinsèque des sons voisés. Carré observe en effet dans une étude sur les voyelles nasales [25] que leur *f0* est généralement plus élevée que celle mesurée pour les voyelles orales dites associées. Di Cristo constate à ce sujet que la voyelle [˜] est affectée d'une *f0* intrinsèque supérieure à celle des autres nasales.

- Le voisement des consonnes entraîne une baisse sensible de la *f0* des voyelles subséquentes, l'effet inverse étant observé en l'absence de voisement. L'importance

[9] Le lecteur trouvera dans [40] une revue détaillée des études traitant ce thème pour différentes langues.

accordée à ce trait fluctue d'une étude à l'autre, mais toutes s'accordent à retenir ce facteur comme principal responsable des variations co-intrinsèques de la *f0* des voyelles adjacentes.

Nous décrivons maintenant notre propre analyse de ces phénomènes acoustiques afin de déterminer ceux qui par leur stabilité autoriseraient un traitement automatique dans le cadre d'une tâche d'accès lexical. Le choix du corpus s'est porté sur AviLex dont la richesse lexicale est supérieure à celle de PolyVar. Les courbes de fréquence fondamentale sont issues de l'algorithme d'*amdf* présenté dans le chapitre 2. Les points de mesure relevés lors des expériences qui suivent ont été multiples afin de constater et d'écarter les éventuels problèmes induits par le choix d'une méthode particulière :

fo_{moy} **:** la moyenne arithmétique des valeurs de *f0* sur la durée du phonème observé (durée qui nous est fournie par le module d'accès lexical).

$fo_{2/3}$ **:** la valeur relevée au deux tiers du segment vocalique. Ce point de mesure constitue une valeur perceptuelle critique que Rossi a mis en relief dans plusieurs études sur la perception des glissandos de *f0* [128, 130].

$fo_{1/2}$ **:** la valeur médiane de *f0* sur la durée du noyau vocalique observé.

Il faut cependant avouer — en dépit des remarques faites par nos prédécesseurs sur l'importance du choix d'un protocole de mesure [40, pp. 77–82] — qu'avant même de commencer notre analyse, les précautions prises quant à ces points de mesure nous paraissaient intuitivement peu influentes sur les résultats que nous pouvions obtenir. En effet si l'on considère — suite aux expériences réalisées précédemment sur la durée — que la valeur modale des distributions de durées des voyelles est de l'ordre de 7 à 8 trames, prendre la valeur d'un paramètre — dont l'évolution temporelle n'est pas accidentelle — aux deux tiers ou au milieu du segment vocalique ne doit guère entraîner de différences dans les mesures et leurs interprétations. Il doit également en aller de même avec le calcul de la moyenne au moins dans notre cas (mots énoncés isolément dans des conditions *normales* de diction).
Ces précisions expérimentales faites, nous pouvons présenter notre étude de la pertinence — sur nos données — des facteurs précédemment énoncés qui régissent les variations intrinsèques et co-intrinsèques de la fréquence fondamentale des voyelles. Nous avons veillé tout au long de nos mesures à ne mettre en correspondance que des données homogènes : La figure 4.25 illustre la tendance bien connue à la décroissance des valeurs de *f0* dans le temps que l'on nomme ligne de déclinaison[10]. La position de la voyelle dans le mot jouant un grand rôle sur la valeur représentative que nous mesurons, deux choix s'offraient alors à nous pour palier ce problème pratique :

[10]Il est intéressant de noter que les dynamiques du paramètres de *f0* séparent assez bien les deux locuteurs étudiés. Nous renvoyons le lecteur à une étude de Monaghan [105] qui traite du potentiel de cet indice pour la caractérisation d'un locuteur.

- Appliquer des coefficients pour *corriger* l'effet de la déclinaison. Cette approche bien qu'intéressante, car elle permet de ne pas diviser l'ensemble des observations, a été écartée pour des raisons évidentes de fiabilité.

- Diviser l'ensemble des observations en fonction de la position de la voyelle étudiée dans le mot. C'est cette méthodologie que nous appliquerons par la suite.

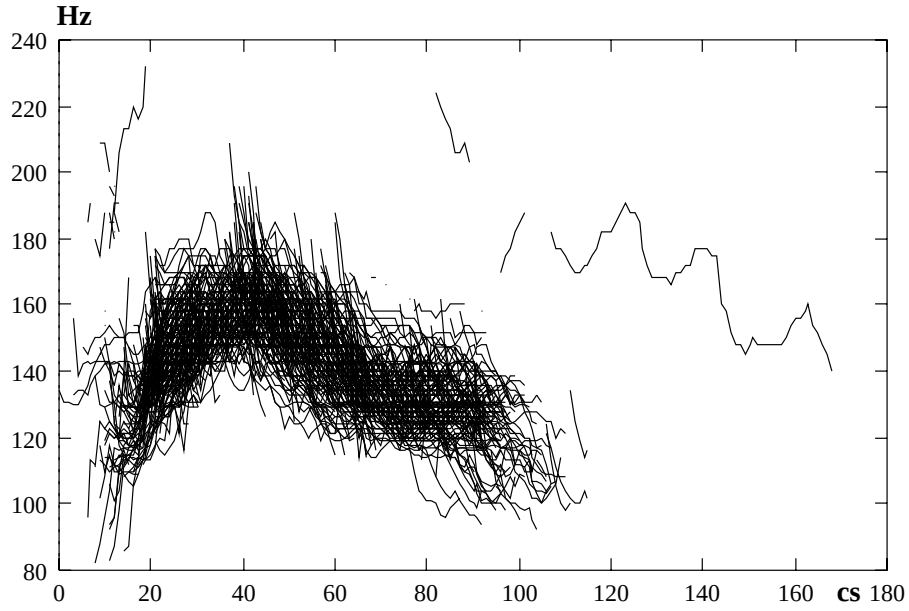


Figure 4.25: Courbes de fréquence fondamentale des mots de 4 voyelles (non terminés par un e-muet) de la base AviLex pour le locuteur PG.

Les tableaux 4.18, 4.19 et 4.20 présentent les moyennes de la fréquence fondamentale intrinsèque des voyelles hautes ([*i*] et [*y*]) et de la voyelle basse [*a*]. La première constatation vient confirmer notre intuition quant à la précision des mesures : il n'y a pas de différence sensible entre les trois méthodes de calcul de la valeur représentative de la *f0* d'une voyelle, aussi, pour la suite de notre analyse, nous contenterons-nous de ne donner que les valeurs de $fo_{2/3}$. Il ressort également que la *f0* moyenne des voyelles hautes est plus élevée que la *f0* moyenne de la voyelle basse [*a*]. L'écart moyen exprimé en pourcentage est d'environ 6% pour l'ensemble des locuteurs ce qui est tout à fait conforme aux observations faites par Di Cristo [40] qui constate un écart de 6.5%. On remarquera cependant que cet écart varie sensiblement d'un locuteur à l'autre et que malgré le nombre trop faible d'observations, une décision bayésienne prise à partir des distributions de la fréquence fondamentale pour une position donnée engendrerait une probabilité d'erreur élevée.

Loc	i+y				a				$(i$+$y)/a$		
	$fo_{2/3}$	fo_{moy}	$fo_{1/2}$	nb	$fo_{2/3}$	fo_{moy}	$fo_{1/2}$	nb	$R_{2/3}$	R_{moy}	$R_{1/2}$
fb	133	132	135	58	131	129	132	104	1.5	2.3	2.2
pg	139	139	140	57	131	131	132	112	5.8	5.8	5.7
ts	164	162	162	49	148	148	148	94	9.8	8.6	8.6
hm	167	167	169	48	149	148	149	100	10.8	11.4	11.8
si	138	137	139	60	134	132	134	107	2.9	3.6	3.6
lc	289	285	284	54	249	248	245	93	13.8	13.0	13.7
si7	141	140	142	44	128	128	129	24	9.2	8.6	9.2
pg7	154	152	152	50	142	142	142	26	7.8	6.6	6.6
tous	165	164	165	420	153	153	153	660	7.3	6.7	7.3

Table 4.18: Valeurs moyennes (Hz) de la *f0* des voyelles hautes ([i] et [y]) et de la voyelle basse [a] observées en *début* de mot sur le corpus AviLex. La dernière colonne exprime en pourcentage le rapport des deux premières colonnes. Le libellé *nb* indique le nombre d'observations de chaque voyelle.

Loc	i+y				a				$(i$+$y)/a$		
	$fo_{2/3}$	fo_{moy}	$fo_{1/2}$	nb	$fo_{2/3}$	fo_{moy}	$fo_{1/2}$	nb	$R_{2/3}$	R_{moy}	$R_{1/2}$
fb	116	117	118	137	117	117	118	58	-0.9	0.0	0.0
pg	134	134	135	132	127	128	129	50	5.2	4.5	4.4
ts	155	155	156	119	151	151	152	48	2.6	2.6	2.6
hm	156	157	159	127	148	148	150	54	5.1	5.7	5.7
si	133	134	135	131	133	132	134	51	0.0	1.5	0.7
lc	277	277	278	130	256	258	258	52	7.6	6.9	7.2
si7	131	131	133	276	124	124	126	196	5.3	5.3	5.3
pg7	152	152	152	296	148	149	150	195	2.6	2.0	1.3
tous	153	153	154	1348	144	145	146	704	5.9	5.2	5.2

Table 4.19: Valeurs moyennes (Hz) de la *f0* des voyelles hautes ([i] et [y]) et de la voyelle basse [a] observées en *milieu* de mot sur le corpus AviLex. La dernière colonne exprime en pourcentage le rapport des deux premières colonnes. Le libellé *nb* indique le nombre d'observations de chaque voyelle.

Loc	i+y				a				$(i$+$y)/a$		
	$fo_{2/3}$	fo_{moy}	$fo_{1/2}$	nb	$fo_{2/3}$	fo_{moy}	$fo_{1/2}$	nb	$R_{2/3}$	R_{moy}	$R_{1/2}$
fb	98	98	99	48	105	104	107	21	-7.1	-6.1	-8.1
pg	124	123	124	53	122	122	122	26	1.6	0.8	1.6
ts	177	177	180	45	164	165	166	22	7.3	6.8	7.8
hm	138	138	139	36	128	128	129	15	7.2	7.2	7.2
si	121	121	122	55	120	120	121	24	0.8	0.8	0.8
lc	252	252	253	31	219	221	220	10	13.1	12.3	13.0
si7	108	108	109	76	103	104	104	47	4.6	3.7	4.6
pg7	132	132	131	85	128	128	128	59	3.0	3.0	2.3
tous	135	135	136	429	127	127	127	224	5.9	5.9	6.6

Table 4.20: Valeurs moyennes (Hz) de la *f0* des voyelles hautes ([i] et [y]) et de la voyelle basse [a] observées en *finale* de mot sur le corpus AviLex. La dernière colonne exprime en pourcentage le rapport des deux premières colonnes. Le libellé *nb* indique le nombre d'observations de chaque voyelle.

À la suite de cette première série de mesures, nous sommes en droit de nous interroger sur l'importance des effets relatifs au contexte consonantique et de nous demander s'ils n'interfèrent pas sur les résultats de notre analyse. Nous allons à cet effet mesurer les fréquences fondamentales moyennes des voyelles hautes [*i*] et [*y*], puis de la voyelle basse [*a*], dans des contextes consonantiques gauches voisés puis non voisés et cela pour les trois positions : à l'initiale, au milieu et en finale de mot. Ces mesures sont reportées dans les tableaux 4.21 et 4.22. Le moins que l'on puisse dire en analysant ces tables est que l'influence du contexte consonantique est loin d'être marquée ; les écarts sont de plus très inégaux d'un locuteur à l'autre. Il semble donc difficile au regard de ces données de conclure à l'importance du contexte consonantique. On attribuera cependant les valeurs fantaisistes en finale de mots mesurées pour la voyelle [*a*] au faible nombre d'observations effectuées ainsi qu'à une erreur locale de fonctionnement de notre détecteur de *f0*. L'analyse de toutes les voyelles confondues du corpus AviLex fait apparaître la même conclusion : l'influence du mode de voisement de la consonne précédente sur la valeur moyenne de la fréquence fondamentale n'est pas mesurable de manière significative. Ces deux tables faisant apparaître un léger déséquilibre des voyelles dans les différents contextes (notamment en position initiale), et malgré le peu d'influence du contexte consonantique gauche sur la fréquence fondamentale des voyelles, nous reportons tout de même en table 4.23 les écarts intrinsèques mesurés pour les voyelles hautes ([*i*] et [*y*]) et la voyelle basse [*a*] en prenant soin de différencier le contexte consonantique gauche selon le diacritique voisé/non voisé et la position de la voyelle dans le mot.

Loc	Initiale					Médiane					Finale				
	voisée		non voisée		R	voisée		non voisée		R	voisée		non voisée		R
	$fo_{2/3}$	*nb*	$fo_{2/3}$	*nb*	%	$fo_{2/3}$	*nb*	$fo_{2/3}$	*nb*	%	$fo_{2/3}$	*nb*	$fo_{2/3}$	*nb*	%
fb	139	18	127	18	-9.4	119	68	113	48	-5.3	96	20	96	16	0.0
pg	140	18	138	17	-1.4	133	68	135	44	1.5	124	22	121	17	-2.5
ts	177	15	168	12	-5.4	151	63	163	39	7.4	166	16	194	15	14.4
hm	169	17	167	12	-1.2	159	62	155	46	-2.6	136	17	136	12	0.0
si	140	19	139	18	-0.7	135	65	131	45	-3.1	122	24	120	16	-1.7
lc	279	19	317	16	12.0	274	64	281	44	2.5	248	17	258	7	3.9
si7	140	18	145	13	3.4	129	150	132	89	2.3	107	39	108	29	0.9
pg7	149	18	161	13	7.5	150	172	157	90	4.5	130	41	132	33	1.5
tous	167	142	170	119	1.8	151	712	155	445	2.6	134	196	134	145	0.0

Table 4.21: Valeurs moyennes de $fo_{2/3}$ pour les voyelles [*i*] et [*y*] du corpus AviLex dans les contextes consonantiques gauches voisés puis non voisés, *nb* indique le nombre d'observations de ces voyelles, R précise le rapport exprimé en pourcentage des moyennes obtenues dans un contexte non voisé sur celles mesurées dans un contexte gauche voisé.

Loc	Initiale					Médiane					Finale				
	voisée		non voisée		R	voisée		non voisée		R	voisée		non voisée		R
	$fo_{2/3}$	nb	$fo_{2/3}$	nb	%	$fo_{2/3}$	nb	$fo_{2/3}$	nb	%	$fo_{2/3}$	nb	$fo_{2/3}$	nb	%
fb	128	18	129	21	0.8	120	26	116	17	-3.4	96	5	114	2	15.8
pg	135	19	129	24	-4.7	128	24	127	15	-0.8	122	9	119	2	-2.5
ts	166	16	150	22	-10.7	144	23	161	14	10.6	165	9	118	2	-39.8
hm	155	14	152	21	-2.0	152	22	148	15	-2.7	128	6	118	2	-8.5
si	135	18	135	24	0.0	132	24	130	14	-1.5	121	5	125	3	3.2
lc	249	16	263	20	5.3	257	23	264	13	2.7	220	3	125	3	-76.0
si7	132	5	140	5	5.7	125	84	123	57	-1.6	106	20	100	11	-6.0
pg7	140	6	150	5	6.7	147	84	150	58	2.0	127	25	127	14	0.0
tous	157	112	157	142	0.0	144	310	144	203	0.0	127	82	116	34	-9.5

Table 4.22: Valeurs moyennes de $fo_{2/3}$ des voyelles basses [a] du corpus AviLex dans les contextes consonantiques gauches voisés puis non voisés, nb indique le nombre d'observations de ces voyelles, R précise le rapport exprimé en pourcentage des moyennes obtenues dans un contexte non voisé sur celles mesurées dans un contexte gauche voisé.

Loc	CG. Voisé			CG. Non Voisé		
	I	M	F	I	M	F
fb	8.6	-0.8	0.0	-1.6	-2.6	-15.8
pg	3.7	3.9	1.6	7.0	6.3	1.7
ts	6.6	4.9	0.6	12.0	1.2	64.4
hm	9.0	4.6	6.2	9.9	4.7	15.3
si	3.7	2.3	0.8	3.0	0.8	-4.0
lc	12.0	6.6	12.7	20.5	6.4	106.4
si7	6.1	3.2	0.9	3.6	7.3	8.0
pg7	6.4	2.0	2.4	7.3	4.7	3.9
tous	6.4	4.9	5.5	8.3	7.6	15.5

Table 4.23: Écarts intrinsèques (exprimés en pourcentage) des voyelles hautes [i] et [y] à la voyelle basse [a] du corpus AviLex dans les contextes consonantiques gauches voisés puis non voisés pour les trois positions : à l'initiale au milieu et en finale de mot.

En conséquence, il convient de considérer avec prudence les valeurs moyennes de la fréquence fondamentale intrinsèque et co-intrinsèque des voyelles du français. Notre position sera de ne pas employer ces variations microprosodiques pour notre tâche de filtrage lexical. Si nous tentons cependant de donner une explication à ce constat somme toute contradictoire avec d'autres études, nous pouvons invoquer le mode expérimental résolument différent de notre analyse. Dans son étude de la variation de la fréquence fondamentale des voyelles, Di Cristo considère principalement des mots monosyllabiques qu'il enchâsse dans des phrases de type P $\rightarrow$ SN + SV, où le mot fait office de nom dans le syntagme nominal composé d'un article, d'un nom et d'un complément de nom. Une seconde raison qui peut être invoquée est la fiabilité des mesures ; nous avons cependant vérifié notre détecteur sur une base où la *f0* était incluse dans la description. Les mots qui ont servi de test, ont subi de plus un contrôle visuel qui n'a permis de mettre en évidence que quelques problèmes locaux peu nombreux.
Pour achever ce constat pessimiste (quant aux perspectives d'utilisation dans notre tâche de filtrage lexical) nous nous devions d'étudier la fréquence fondamentale intrinsèque des consonnes voisées. Il existe parmi toutes les études sur le sujet une observation largement décrite ([15, 40, 157],...) qui présente un indice de segmentation semble-t-il robuste : la tenue des obstruantes intervocaliques s'accompagne d'une pente négative qui relie le décrochage d'implosion à celui de la détente. Avec moins de précision, on observe plus simplement la configuration concave (triangulaire pour certains) qui accompagne les obstruentes intervocaliques voisées, alors que les non obstruentes voisées ([m], [n] et [ℓ]) semblent s'insérer dans le continuum mélodique avec un contour plat. Les données de Di Cristo sur le sujet, ne sont une fois de plus pas avares de précisions. Il mesure la rupture de *f0* des occlusives voisées et relève des valeurs de 9% pour un locuteur masculin et 15% pour les voix féminines. Cette rupture dans le cas des constrictives voisées s'élève à 13% pour un homme et 16.5% pour une femme [40, pp. 349 et 353]. Ne prétendant pas atteindre une telle finesse, nous allons comme précédemment tenter de mesurer la probabilité d'erreur qu'engendrerait une décision bayésienne prise sur l'amplitude de la rupture de *f0* dans la consonne (pour autant que celle-ci soit facilement mesurable). Nous avons donc tenté, dans une première expérience de donner une mesure de la configuration creusée des consonnes afin de constater les différences de distribution de cette valeur dans des contextes intervocaliques pour toutes les consonnes. Nous avons pour cela retenu deux mesures de la configuration concave que nous précisons maintenant. Soient x_g la dernière trame de la voyelle qui précède la consonne étudiée et x_d la première trame de la voyelle qui suit la consonne (ces valeurs sont fournies par le processus de filtrage lexical) et soit $\hat{f0}$ la courbe de *f0* obtenue par interpolation linéaire entre les points (x_g,$f0(x_g)$) et (x_d,$f0(x_d)$) ; on nomme alors $Rf0$ la hauteur de la concavité et S l'aire de la surface définie par l'intersection des contours de *f0* et de $\hat{f0}$ (voir la figure 4.26).

$$Rf0 = Min(\hat{f0}(x) - f0(x)) \text{ pour tout } x \ / \ x \in \ ]x_g, x_d[\text{ et } \hat{f0}(x) - f0(x) > 0$$

$$S = \sum_{x=x_g+1}^{x_d-1} \hat{f0}(x) - f0(x)$$

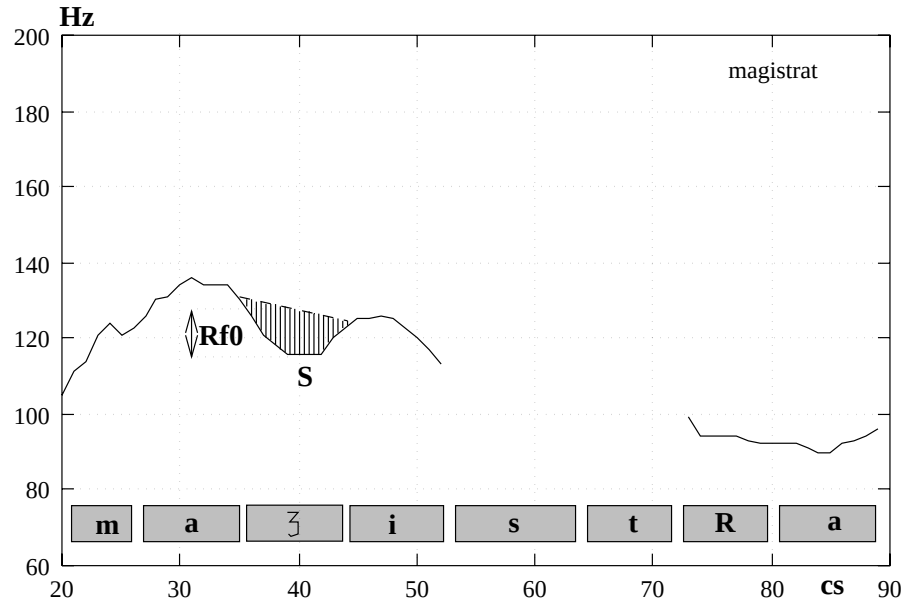


Figure 4.26: Configuration de la fricative [] dans le mot *magistrat*.

Le tableau 4.24 résume les moyennes de ces deux mesures pour la plupart des consonnes voisées du français. Cela nous permet de confirmer qu'en moyenne les occlusives et fricatives sont plus creusées que les consonnes liquides et nasales [ℓ], [m] et [n]. Une deuxième observation, moins nette, semble également confirmer l'observation faite par Di Cristo qui tend à affirmer que les constrictives voisées (ici réduites aux fricatives) présentent une rupture plus marquée que dans le cas d'une occlusive voisée. La plupart des réalisations de la liquide [] se manifeste par un contour creusé qui diffère peu des constrictives voisées. Enfin, la semi-voyelle [j] semble caractérisée par une configuration intermédiaire. La figure 4.27 présente les distributions des ruptures et de l'aire de la concavité mesurées pour les consonnes voisées intervocaliques liquides+nasales (L+N), occlusives et fricatives. Il apparaît que les liquides+nasales possèdent des valeurs plus faibles de $Rf0$ et S ce qui confirme la tendance de la fréquence fondamentale de ces consonnes à s'insérer dans le continuum contrairement aux consonnes intervocaliques fricatives et occlusives. La table 4.25 indique les probabilités d'erreur associées aux diverses distinctions sur la nature des consonnes à l'aide de ces distributions. Il en ressort qu'une distinction des consonnes liquides+nasales est réalisable avec une probabilité d'erreur de l'ordre de 20% (une décision partielle pouvant diminuer ce taux) alors qu'une distinction entre les fricatives et les occlusives ne peut se faire qu'avec une probabilité d'erreur proche de 40% !

Loc	Occlusives						Fricatives						Nasales				Liquides					
	b		*d*		*g*				*z*				*m*		*n*		*ℓ*				*j*	
	R	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*
fb	7	29	6	23	9	43	7	32	8	51	9	50	1	4	0	3	1	3	7	37	4	21
pg	1	7	2	9	1	4	1	6	1	8	4	19	0	0	0	0	0	2	2	8	1	7
hm	4	20	4	19	6	43	5	40	6	51	6	42	0	0	0	1	1	5	4	30	1	16
si	3	13	4	17	5	29	4	20	5	29	7	35	0	0	0	0	1	4	6	28	3	16
lc	5	27	6	26	7	33	5	30	6	40	5	35	1	6	1	5	1	5	5	23	1	7
si7	4	15	5	22	6	30	4	16	6	30	6	33	0	0	0	1	0	2	4	16	2	8
pg7	4	17	4	16	4	18	3	14	3	17	3	14	2	5	2	4	2	5	4	13	0	1
tous	4	17	4	18	5	25	4	20	4	28	5	29	1	2	0	2	1	4	5	23	2	10

Table 4.24: Moyennes des valeurs de $Rf0$ (R exprimé en pourcentage) et de l'aire de la concavité (S) de la *f0* des consonnes voisées fricatives, occlusives, liquides, nasales ainsi que pour la semi-voyelle [j] du corpus AviLex.

	L+N/Fri	L+N/Occ	Fri/Occ	$(b{+}d{+}g{+}{+}z{+})/(m{+}n{+}\ell)$
S	19.7	21.5	39.6	20.7
$Rf0$	22.8	22.4	46.4	22.7

Table 4.25: Probabilités d'erreur (exprimées en poucentage) associées à l'affectation d'une classe consonantique (liquide+nasale (L+N), occlusive, fricative) à l'aide des distributions mesurées sur le corpus AviLex.

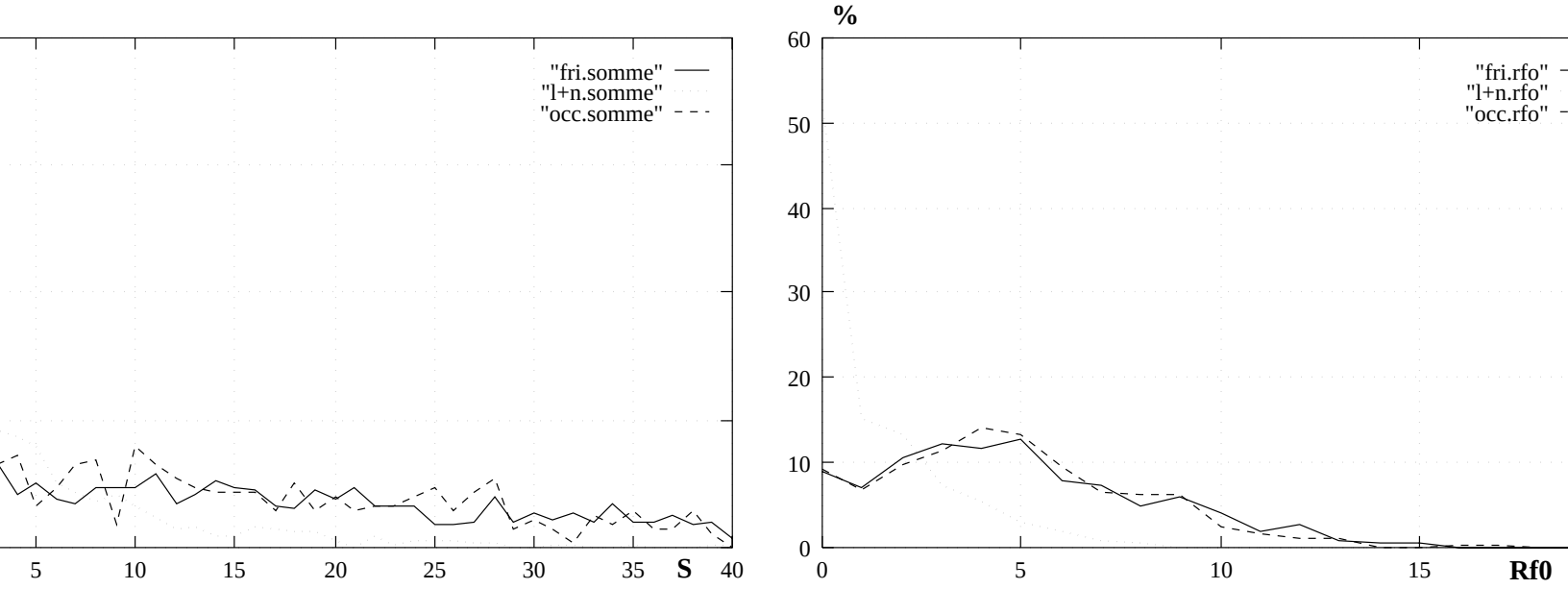


Figure 4.27: Distributions des valeurs de $Rf0$ et de S pour les trois classes de consonnes voisées : occlusives, fricatives et liquides+nasales (LN) étudiées dans le corpus AviLex dans des situations intervocaliques.

Afin de vérifier que les mesures précédentes ne sont pas dues aux spécificités du corpus AviLex, nous avons réalisé quelques contrôles sur le corpus FeLex. Une première vérification confirme que la *f0* d'une voyelle dépend grandement de sa position dans le mot (figure 4.28). On mesure en effet un abaissement de la *f0* des voyelles finales (comparées aux valeurs mesurées sur la première voyelle) de l'ordre de 11%, alors que la différence entre les valeurs de *f0* à l'initiale et en position médiane semble peu significative. Nous avons également vérifié les valeurs intrinsèques de la fréquence fondamentale des voyelles hautes, basses et nasales en prenant soin d'isoler leur contexte consonantique gauche et ce pour chaque position des voyelles dans le mot : initiale (table 4.26), médiane (table 4.27) et finale (table 4.28). Le nombre d'observations sur lesquelles ces tables sont fondées ne sont pas reportés pour des raisons de clarté mais le minimum de représentants de chaque classe (*i.e.* case dans une table) est de 70. Il ressort de cette série de mesures quelques faits brièvement commentés :

- On vérifie bien que le mode phonatoire de la consonne qui précède une voyelle influence la *f0* de cette dernière. Plus précisément, la fréquence fondamentale d'une voyelle est en moyenne plus élevée lorsqu'elle est précédée d'une consonne non voisée que lorsqu'elle suit une consonne voisée. Les rapports moyens rapportés montrent cependant, que ce phénomène — qui dépend du locuteur, de la nature des voyelles et de la position des voyelles étudiées — peut être considéré comme mineur (voir la figure 4.29). Cette observation confirme donc l'appel à la prudence que nous avons émis après l'analyse des données du corpus AviLex sur l'influence consonantique.

- On constate également que les voyelles hautes [*i*] et [*y*] sont en moyenne supérieures à la voyelle basse [*a*]. Des rapports intrinsèques de l'ordre de 5% pour les voyelles en positions initiale et médiane sont mesurés alors que des rapports nuls ou négatifs sont observés en finale de mot. Une décision de la nature haute ou basse de la voyelle prise à partir des distributions recueillies en position initiale, engendrerait une probabilité d'erreur de l'ordre de 25 % , cette valeur pouvant diminuer fortement dans le cas d'une décision partielle (5%). Un rapport moyen de 6% avait été mesuré dans le cas des voyelles du corpus AviLex avec cependant moins de régularité d'un locuteur à l'autre de la base.

Enfin nous présentons — à l'instar du corpus AviLex — l'étude de la fréquence fondamentale des consonnes intervocaliques du corpus FeLex. Sur la figure 4.30 sont présentées les distributions des mesures de rupture $Rf0$ et d'aire S pour les consonnes voisées intervocaliques du corpus FeLex. Si ces distributions sont ressemblantes à celles obtenues sur le précédent corpus, la probabilité d'erreur d'une distinction constrictive / liquide+nasale est ici plus élevée (voir la table 4.29). Une décision prise partiellement à partir de ces distributions pourra cependant être envisagée.

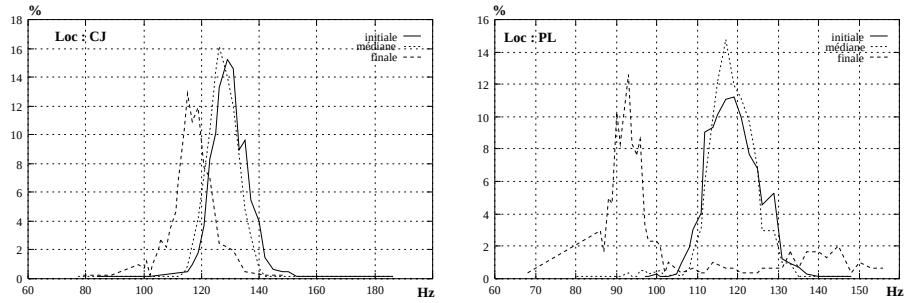


Figure 4.28: Distributions de $fo_{2/3}$ mesurées sur les voyelles de FeLex en fonction de leur position dans le mot.

	[i]+[y]			[a]			[˜]+[˜]+[˜]			[i]+[y] / [a]	
Loc	V	NV	R	V	NV	R	V	NV	R	V	NV
CJ	131	135	3.0	125	127	1.6	125	126	0.8	4.6	5.9
PL	120	126	4.8	115	116	0.9	122	123	0.8	4.2	7.9
tous	125	129	3.1	120	121	0.8	123	124	0.8	4.0	6.2

Table 4.26: Moyenne des valeurs de $fo_{2/3}$ observées pour les voyelles (hautes, basses et nasales) du corpus FeLex à l'*initiale* de mot dans les contextes consonantiques gauches voisés (V) puis non voisés (NV). R indique le rapport — exprimé en pourcentage — des mesures effectuées dans les contextes non voisés sur celles des contextes voisés. Les deux dernières colonnes reportent les rapports (%) des mesures des voyelles hautes sur celles des voyelles basses en fonction du contexte consonantique gauche.

Loc	[i]+[y]			[a]			[˜]+[˜]+[˜]			[i]+[y] / [a]	
	V	NV	R	V	NV	R	V	NV	R	V	NV
CJ	127	130	2.3	123	127	3.1	127	127	0.0	3.1	2.3
PL	121	122	0.8	115	116	0.9	117	118	0.8	5.0	4.9
tous	124	127	2.4	119	121	1.7	122	122	0.0	4.0	4.7

Table 4.27: Moyenne des valeurs de $fo_{2/3}$ observées pour les voyelles (hautes, basses et nasales) du corpus FeLex en *milieu* de mot dans les contextes consonantiques gauches voisés (V) puis non voisés (NV). R indique le rapport — exprimé en pourcentage — des mesures effectuées dans les contextes non voisés sur celles des contextes voisés. Les deux dernières colonnes reportent les rapports (%) des mesures des voyelles hautes sur celles des voyelles basses en fonction du contexte consonantique gauche.

Loc	[i]+[y]			[a]			[˜]+[˜]+[˜]			[i]+[y] / [a]	
	V	NV	R	V	NV	R	V	NV	R	V	NV
CJ	117	119	1.7	117	116	-0.9	116	116	0.0	0.0	2.5
PL	95	103	7.8	94	106	11.3	92	100	8.0	1.1	-2.9
tous	107	111	3.6	108	113	4.4	106	111	4.5	-0.9	-1.8

Table 4.28: Moyenne des valeurs de $fo_{2/3}$ observées pour les voyelles (hautes, basses et nasales) du corpus FeLex en *finale* de mot dans les contextes consonantiques gauches voisés (V) puis non voisés (NV). R indique le rapport — exprimé en pourcentage — des mesures effectuées dans les contextes non voisés sur celles des contextes voisés. Les deux dernières colonnes reportent les rapports (%) des mesures des voyelles hautes sur celles des voyelles basses en fonction du contexte consonantique gauche.

	L+N/Fri	L+N/Occ	Fri/Occ	$(b+d+g++z+)/(m+n+\ell)$
S	28.6	30.6	36.6	29.9
$Rf0$	31.6	31.7	41.0	31.9

Table 4.29: Probabilités d'erreur des décisions bayésiennes associées (exprimées en pourcentage).

Bilan de l'étude de la fréquence fondamentale

Il est temps de conclure sur cette étude des variations microprosodiques de la fréquence fondamentale des voyelles et des consonnes. Si, comme pour l'étude de la durée, peu de phénomènes recensés sont ici observables de manière significative, on peut tout de même dresser une liste des quelques points mis en évidence par cette étude :

- Les valeurs intrinsèques de la fréquence fondamentale des voyelles sont très dépendantes de leur position dans le mot et nous renvoyons le lecteur à l'étude des différents schémas de *f0* mesurés dans les mots du corpus AviLex.

- Les voyelles hautes ont — en moyenne — une fréquence fondamentale plus élevée que celle des voyelles basses — toutes choses égales par ailleurs — et abstraction faite de l'écart nul mesuré en finale de mot. La différence relevée entre les deux classes de voyelles semble cependant fragile pour une discrimination fonctionnelle. Elle est en effet dépendante du corpus étudié : le corpus AviLex semble indiquer des variations importantes selon le locuteur observé, alors que les mesures réalisées sur le corpus FeLex (composé de seulement deux locuteurs) semblent plus stables.

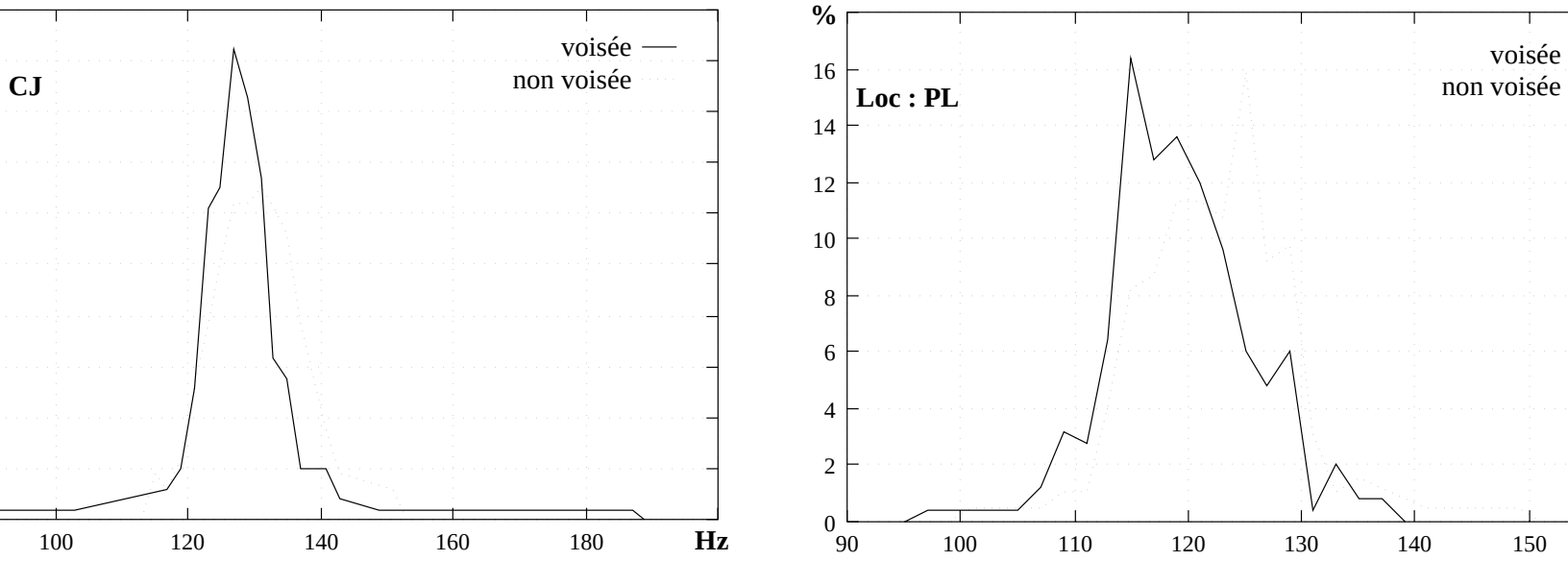


Figure 4.29: Distributions de $fo_{2/3}$ pour l'ensemble des voyelles du corpus FeLex mesurées en position initiale de mot en fonction du caractère voisé ou pas de la consonne de gauche.

- L'influence du contexte consonantique n'a pas — loin s'en faut — été mis en évidence sur les mesures de nos deux corpus. On peut simplement remarquer qu'en moyenne la fréquence fondamentale des voyelles est plus élevée lorsqu'elles suivent une consonne non voisée qu'une consonne voisée, sans pour autant que la différence — dépendante du locuteur observé — puisse donner lieu à une exploitation dans le cadre du filtrage lexical.

- On a relevé cependant un indice qui pourrait être utilisé à des fins de filtrage lexical : la distinction des consonnes [ℓ], [m] et [n] des autres consonnes voisées lorsqu'elles sont entourées de voyelles. Les mesures proposées dans cette étude, de la concavité de la courbe du fondamental généralement observée pour les obstruentes, autorisent une décision exacte dans 70% à 80% des cas (selon le corpus étudié). Bien que meilleure qu'un choix aléatoire, il reste cependant à vérifier l'apport bénéfique d'une telle décision dans un processus de filtrage, ce que nous nous proposons de faire dans la section 4.5.

4.4.3 L'intensité

L'intensité est sans aucun doute le paramètre le plus négligé de la recherche prosodique, bien qu'étant de loin le plus facile à extraire du signal. Cette lacune vient certainement en grande partie du fait que l'intensité a longtemps été considérée comme une co-variable de la fréquence fondamentale, les deux paramètres étant étroitement liés aux variations de la pression sous-glottique. Di Cristo [40, pp. 474–475] souligne le caractère erroné de cet

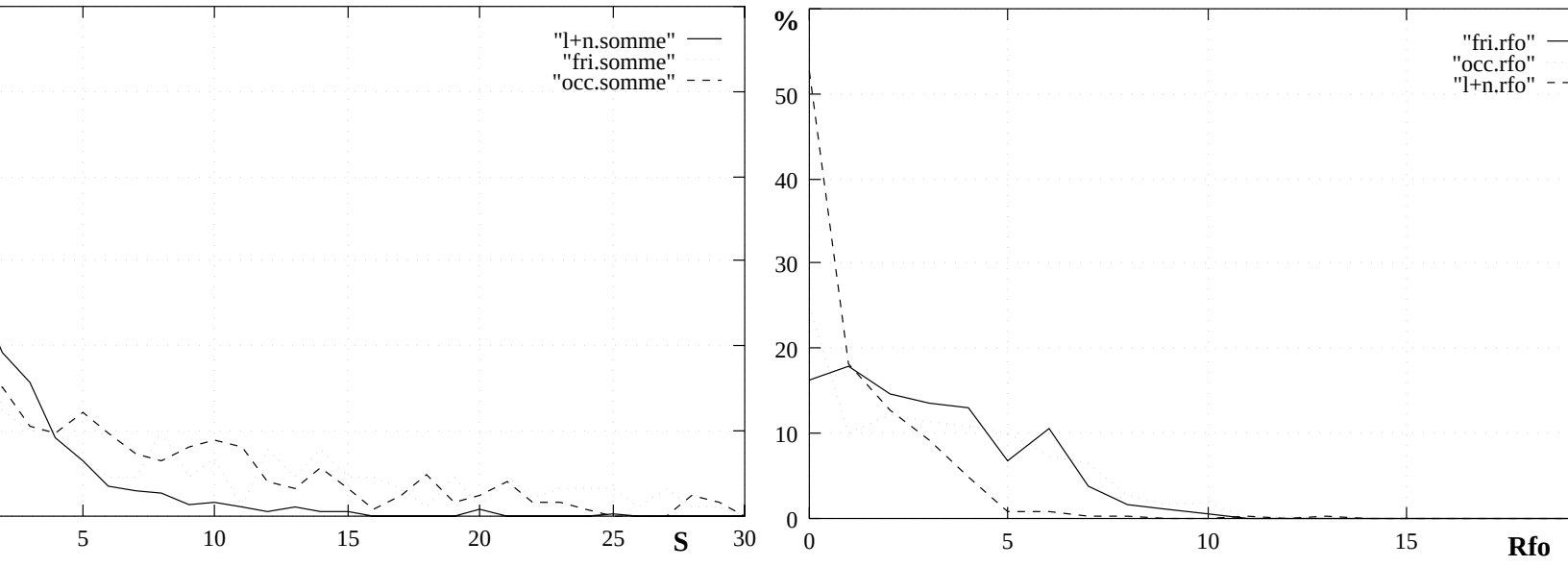


Figure 4.30: Distribution des valeurs de $Rf0$ et de S pour les trois classes de consonnes : occlusives, fricatives et liquides+nasales (L+N) étudiées dans le corpus FeLex dans les situations intervocaliques.

argument et rappelle à ce propos les résultats obtenus par Rossi [131] qui attestent que les variations d'intensité affectent la reconnaissance auditive des contours mélodiques. En conséquence, nous disposons de peu d'études sur ce paramètre desquelles on peut cependant retenir les résultats suivants :

- Les voyelles basses sont généralement affectées d'une intensité spécifique supérieure à celle des voyelles hautes, avec un minimum pour la voyelle [i] et des valeurs maximales pour les voyelles [a] et []. Cette constatation est en partie rejetée par les expériences de Di Cristo [40] qui note que les voyelles [] et [] sont les plus intenses et par les travaux de Rossi [127] qui affirme que l'intensité spécifique ne dépend pas de l'aperture des voyelles.

- Di Cristo relève également que les voyelles nasales [˜] et [˜] possèdent une intensité intrinsèque supérieure à la nasale [˜].

- L'intensité globale des voyelles semble généralement plus faible lorsque celles-ci sont précédées de consonnes non voisées ou occlusives. Di Cristo mesure une tendance inverse pour l'influence du mode articulatoire de la consonne et confirme un écart intrinsèque positif en faveur des consonnes voisées qu'il qualifie cependant de minime. Aucune influence post-vocalique n'est relevée dans ces études.

Nous allons dans la suite présenter — à l'instar de la durée et de la fréquence fondamentale — les mesures réalisées sur le corpus FeLex afin de vérifier si l'information microprosodique

véhiculée par le paramètre d'intensité peut s'avérer d'un quelconque soutien dans une tâche de filtrage lexical. Nous n'étudierons cependant que les variations du paramètre sur les voyelles ; l'information localisée sur les consonnes ayant déjà fait l'objet de soins attentifs lors de la phase de décodage acoustico-phonétique utilisée par le module d'accès lexical [59]. Nous limiterons nos investigations au plan acoustique ce qui est discutable puisque nous savons que deux syllabes de même intensité objective, peuvent être perçues non isophones si elles diffèrent sensiblement par leur durée [127] ; nous rappelons cependant que Di Cristo — dans son étude de l'intensité [40, p. 517]— conclut à l'inadéquation partielle de l'intégration temporelle des sons. Des points de mesures multiples ont été relevés (milieu, point de phonie [132], valeur maximale,...) sans grand changement quant à leur interprétation, aussi ne reportons-nous ici qu'une série de mesures prises au centre des segments vocaliques analysés.

On peut tout d'abord noter d'après la figure 4.31 que les deux locuteurs de la base FeLex se comportent de manière analogue ; aussi nous contenterons-nous par la suite de cumuler les observations de chacun.

De manière très générale, on peut remarquer que les valeurs les plus faibles de l'intensité s'observent naturellement en finale de mot ; ce qui a déjà été vu en détail en section 4.3.2 et que l'on peut observer pour les voyelles [i] et [a] dont les mesures ont été reportées en figure 4.32. Notons simplement qu'il ne s'agit pas là d'une constante et que les réalisations de la voyelle [] — par exemple — sont plus intenses en finale de mot qu'à l'initiale.

Nous vérifions également sur la figure 4.31 que les voyelles basses sont plus intenses que les voyelles hautes [i] et [y], et constatons comme nos prédécesseurs que la voyelle [y] possède une intensité supérieure à la voyelle [i] ; cette différence n'étant cependant pas hautement significatives pour une décision bayésienne (voir la figure 4.33). Il semble cependant possible que l'intensité puisse être utilisée à des fin de filtrage dans certaines situations particulières, faute d'observations en quantité suffisante, nous bornerons notre remarque aux voyelles les plus représentées du corpus FeLex, [a] et [i], pour lesquelles une distinction peut être proposée avec une probabilité d'erreur de moins de 15% en position initiale et médiane de mot.

La courbe 4.31c) nous permet de vérifier la prédominance des voyelles orales [] et [] en finale de mot, ce qui est conforme aux observations de Di Cristo [40, p. 499], le nombre de représentants (voir le décompte en table 4.30) de ces deux voyelles n'étant cependant pas suffisant pour conclure quant à la régularité de cette constatation.

Nous terminons rapidement cette étude en mesurant l'influence du mode phonatoire et/ou articulatoire de la consonne qui précède la voyelle. Nous relevons pour les deux locuteurs de la base FeLex (voir table 4.31) que les voyelles sont plus intenses — en moyenne — au contact de consonnes voisées qu'au contact des consonnes non voisées et que le mode articulatoire de la consonne précédente influence également l'intensité de la voyelle : ces dernières possédant une intensité supérieure lorsqu'elles suivent une constrictive. Ces différences ne sont cependant pas susceptibles d'être exploitées de manière fiable pour une classification de la consonne précédente (cf. figure 4.34).

		i	*y*	*u*	*a*	~		*e*	~		*o*	~		œ
cj	initiale	114	50	9	176	103	61	78	50	23	6	36	1	0
pl	initiale	128	61	10	196	107	83	72	50	25	8	42	0	1
cj	médiane	141	47	18	195	132	41	46	10	89	9	45	3	0
pl	médiane	137	44	19	200	131	41	41	12	82	16	49	2	1
cj	finale	102	42	5	161	46	18	114	9	5	9	47	0	0
pl	finale	89	41	2	151	27	24	93	9	5	10	12	0	0

Table 4.30: Nombre d'observations des voyelles du corpus FeLex en fonction de la position de la voyelle dans le mot.

loc	CG. Voisé			CG. non Voisé			CG. CO			CG. OC		
	moy.	ec	nb	moy.	ec	nb	moy.	ec	nb	moy	ec	nb
cj	49	5.1	380	46	5.5	632	48	5.6	704	46	5.0	363
pl	52	6.4	370	49	7.1	630	51	7.1	698	48	6.4	347
tous	51	6.1	750	47	6.5	1262	49	6.6	1042	47	5.8	710

Table 4.31: Moyenne et écart-type de l'intensité des voyelles du corpus FeLex observées dans des contextes consonantiques gauches divers : voisé/non voisé, occlusif(OC)/constrictif(CO). *nb* indique le nombre d'observations considérées.

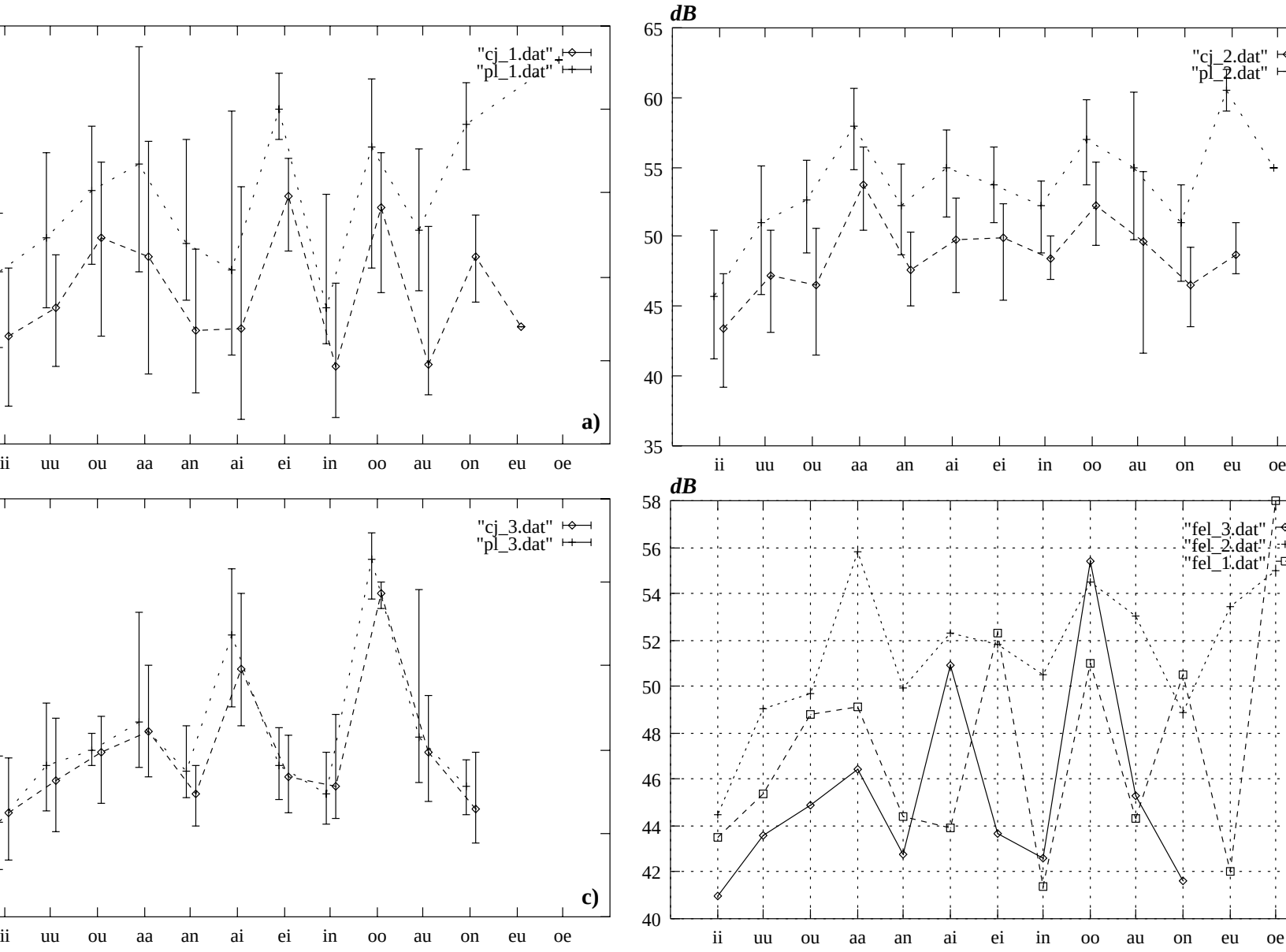


Figure 4.31: Moyennes et écarts-types de l'intensité des différentes voyelles du corpus FeLex en position initiale (a), médiane (b) , finale (c) puis toute position confondue (d)pour les deux locuteurs *pl* et *cj* de la base.

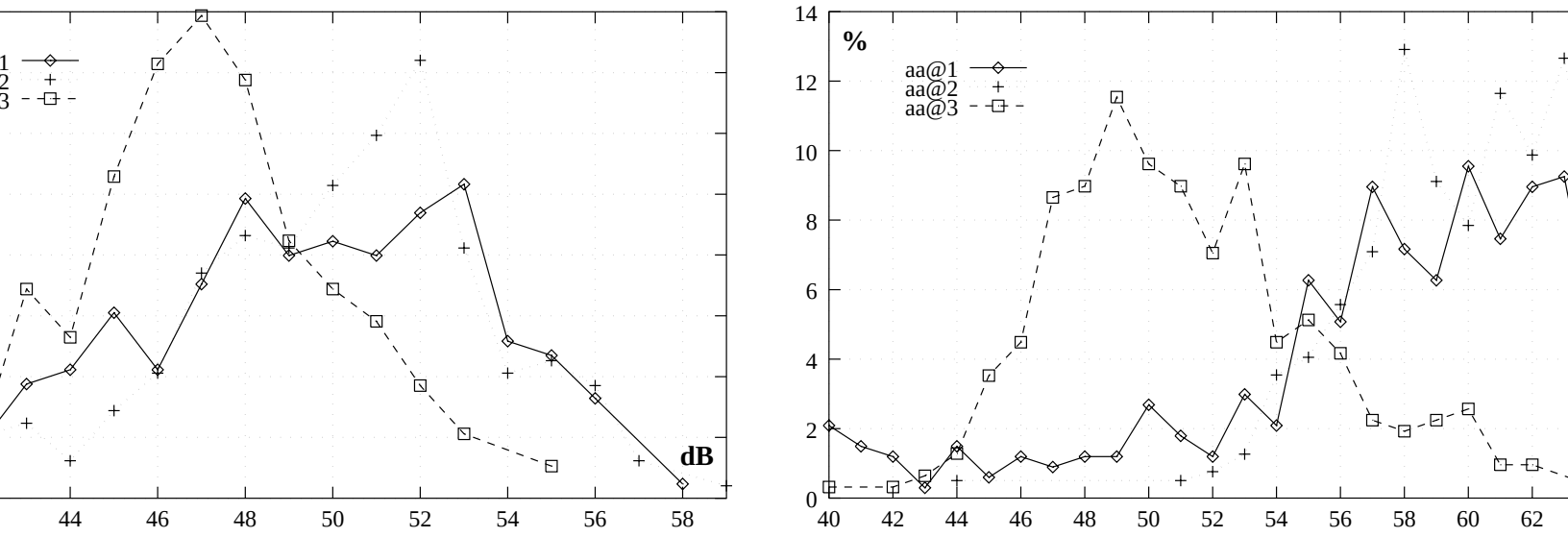


Figure 4.32: Distributions de l'intensité des voyelles [*i*] et [*a*] du corpus FeLex en fonction de la position de la voyelle dans le mot.

4.4.4 Discrimination par décision voisée/non voisée

Le paramètre de fréquence fondamentale permet — par l'absence et/ou la présence de ses valeurs — de distinguer les consonnes voisées des autres [157]. Nous savons cependant (voir la sous-section 2.1.1) qu'une détection voisée/non voisée depuis le signal de parole est difficile à obtenir dans toutes les conditions. Aussi allons-nous maintenant mesurer l'apport de notre décision de voisement pour la tâche de filtrage lexical qui nous préoccupe.

Position du problème

Concrètement, nous décomposons le problème en deux étapes :

- Un premier filtre doit opérer en amont du processus de reconnaissance. Il dispose pour cela d'une chaîne phonémique associée à chaque mot du lexique, représentative d'une prononciation "usuelle". Il convient alors d'éliminer — autant que faire se peut — les entrées dont le schéma de voisement "théorique" ne peut correspondre à celui que l'on mesure sur le signal de parole à reconnaître.

- Une deuxième étape consiste alors à filtrer parmi une cohorte issue du processus de filtrage les mots dont le schéma de voisement ne peut correspondre à celui du signal avec cette fois-ci la connaissance de la chaîne phonémique effectivement prononcée et son alignement temporel.

Attribution d'un schéma de voisement

Pour autant que nous sachions réaliser une décision voisée/non voisée à partir du signal, il reste un obstacle à la réalisation de ces deux filtres : le mode de calcul d'un schéma de voisement d'une chaîne phonétique donnée. Une première méthode consiste à mettre bout à bout les diacritiques de voisement (N_V et V) de chaque phonème de la chaîne — que l'on considère isolément — en supprimant les répétitions consécutives des symboles identiques et en élimant les éventuels diacritiques N_V en début et fin de mot (cette dernière clause étant spécifique au premier filtre où la distinction des consonnes non voisées avec le silence n'est pas réalisable).

Ex. $[sa\ell is\tilde{}]$ (salissons) $\Longrightarrow N_V.V.V.V.N_V.V \Rightarrow N_V.V.N_V.V \Rightarrow V.N_V.V$

Cette méthode — qui peut convenir pour le premier filtre — fait abstraction de plusieurs phénomènes décrits par la phonétique combinatoire [89], [91, pp. 75–77], dont notamment l'assimilation et l'élision, qui sont fréquents dans notre langue.
Nous négligeons volontairement l'assimilation vocalique de sourdité (désonorisation des voyelles hautes précédées ou entourées de consonnes sourdes) qui — bien que répandue en français québécois — est peu fréquente dans les prononciations "franco-françaises". L'assimilation consonantique retiendra en revanche toute notre attention. Nous distinguons deux cas rencontrés tous les deux dans la même situation où deux consonnes sont en contact, l'une étant sourde l'autre pas :

- L'assimilation de voisement qui se caractérise par un transfert de voisement de la deuxième consonne sur la première ; on parle alors d'assimilation régressive (droite ← gauche).

  Ex. $[\kappa ze\kappa o]$ (ex æquo) :assimilation de voisement régressive ($[\kappa] \leftarrow [b]$) avec le voisement du phonème $[\kappa]$ (pouvant donner lieu à une prononciation $[gze\kappa o]$).

- L'assimilation de sourdité qui se caractérise par un dévoisement de la consonne sonore au contact de la consonne sourde ; l'assimilation pouvant donc être aussi bien progressive (droite → gauche) que régressive.

  Ex. $[\rho\underline{r}ize]$ (prisé) : assimilation de sourdité progressive ($[\rho] \rightarrow [r]$) avec le dévoisement du phonème $[r]$.

  Ex. $[\underline{d}\kappa a]$ (vodka) : assimilation de sourdité régressive ($[d] \leftarrow [\kappa]$) avec dévoisement du phonème $[d]$ (pouvant entraîner la prononciation $[t\kappa a]$).

L'élision du $[\partial]$ très fréquente dans notre langue, peut également être responsable d'erreurs au cours de la détermination du schéma de voisement. La règle la plus générale que l'on puisse faire concernant cette chute est la suivante : la voyelle $[\partial]$ peut tomber si son élision ne fait apparaître aucun groupe consonantique difficile à prononcer.

Ex. $[a\rho idm\tilde{}]$ (rapidement).

Cette règle possède bien sûr de nombreuses exceptions dont il est difficile de dresser un inventaire exhaustif. On connaît par exemple la tendance dialectale du sud de la France à prononcer tous les [∂] même en finale de mot.

Quelques chiffres sur les lexiques

Avant d'entreprendre toute expérience, nous avons effectué quelques mesures sur le lexique BdLex qui nous ont servi par la suite et que nous reportons maintenant (cf. table 4.32).

Quelques mesures concernant BdLex	Nombre	%
entrées	273842	100
mots qui contiennent au moins 2 consonnes consécutives	99 114	36.2
mots qui contiennent au moins 3 consonnes consécutives	1503	0.5
mots qui contiennent au moins 4 consonnes consécutives	2	–

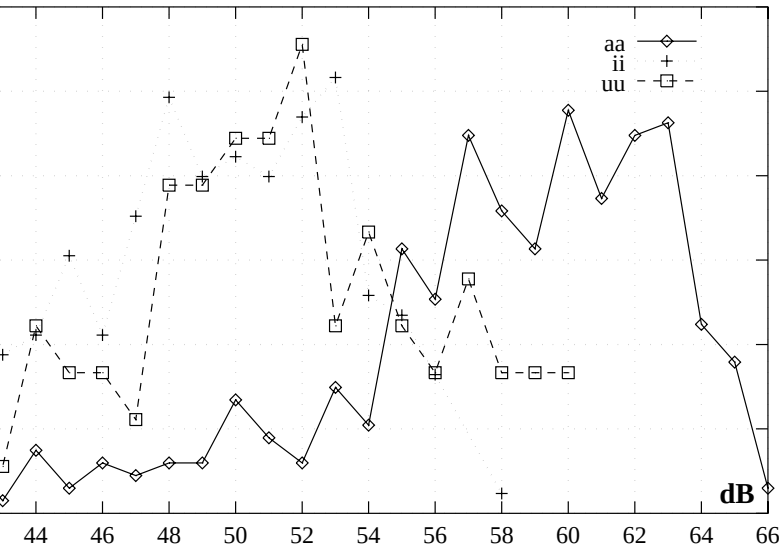
Schémas de voisement rencontrés dans BdLex	Nombre	%
$V.N_V.V$	123 875	45.2
V	99 741	36.4
$V.N_V.V.N_V.V$	44 499	16.2
$V.N_V.V.N_V.V.N_V.V$	5 521	2
$V.N_V.V.N_V.V.N_V.V.N_V.V$	194	0.1

Table 4.32: Cardinalités des mots et classes de mots de BdLex et nombre de schémas de voisement différents de ce lexique (sans prise en compte des phénomènes d'élision et d'assimilation).

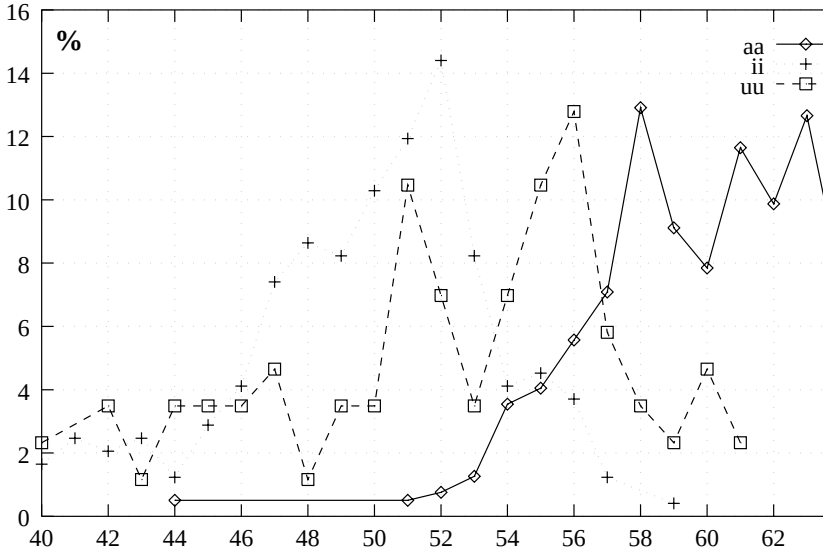
Le nombre de mots contenant deux consonnes consécutives se divise en trois parties :

- Les deux consonnes sont en même temps voisées ou non voisées (18%) : il n'y a pas d'assimilation possible.
- La première est voisée, l'autre sourde (14%) : il y a possibilité d'assimilation de sourdité régressive.
- La première est sourde, l'autre sonnante (4%) : il y a possibilité d'assimilation de voisement et de sourdité progressive.

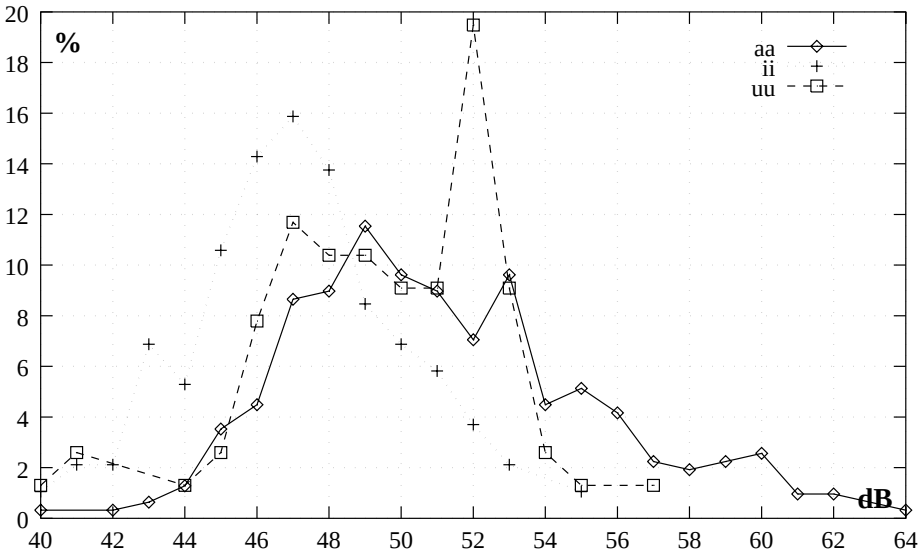
S'il paraît surprenant de rencontrer deux mots qui contiennent quatre consonnes consécutives, il convient de remarquer qu'il s'agit en fait du même mot d'origine anglaise proposé au singulier puis au pluriel (cold-cream, cold-creams) !



position initiale : 15 %

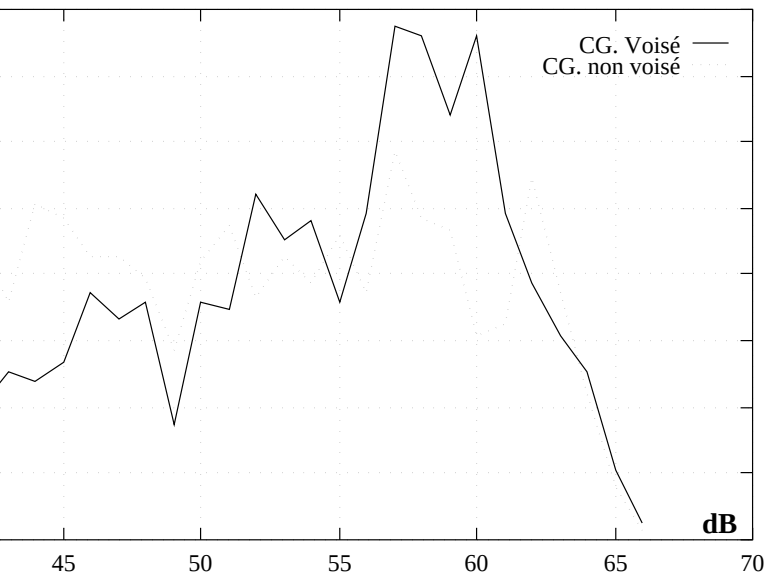


position médiane : 8 %



position finale : 28 %

Figure 4.33: Comparaison des distributions de l'intensité des voyelles $[i]$, $[y]$ et $[a]$ du corpus FeLex pour les voyelles initiales, médianes puis finales de mot. La probabilité d'erreur de la distinction des voyelles $[a]$ et $[i]$ à partir des distributions mesurées est indiquée à côté de chaque position.



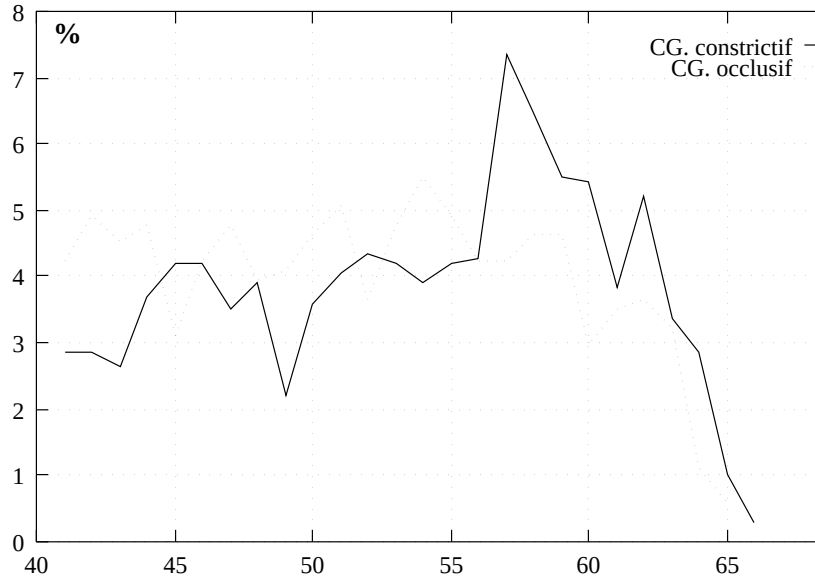


Figure 4.34: Distributions de l'intensité des voyelles de FeLex dans les contextes consonantiques gauches voisés, non voisés, occlusif et constrictif.

Filtre 1

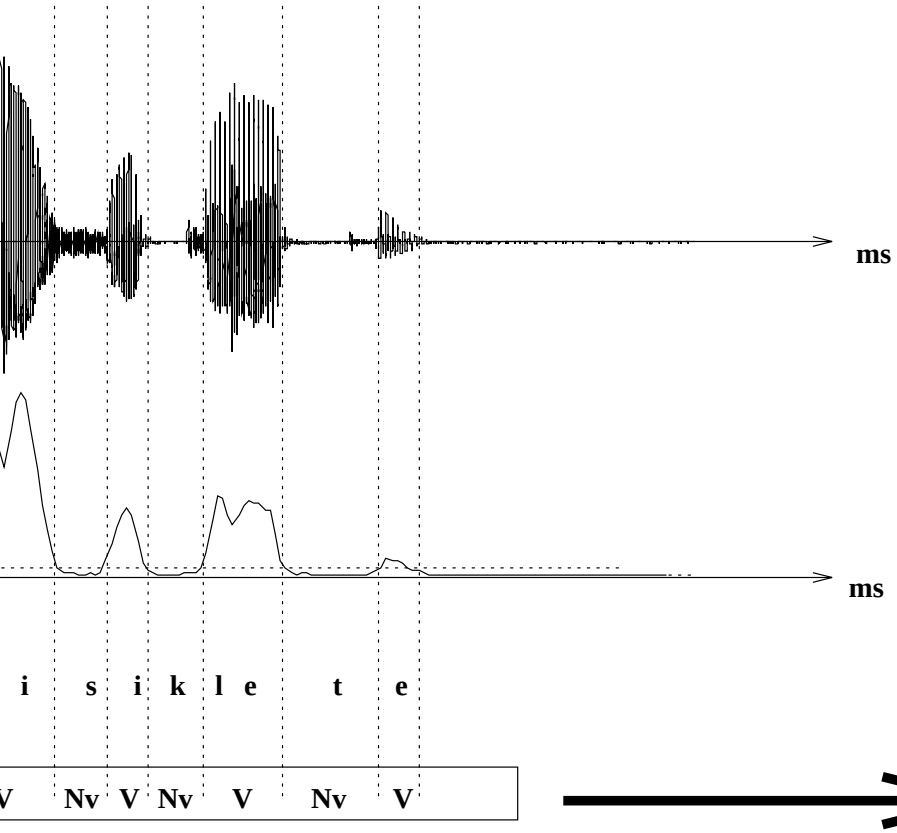
Le schéma de voisement du signal de parole à reconnaître est obtenu à partir de la courbe de voisement calculée lors de la détection de la fréquence fondamentale (voir la figure 4.35). La pré-compilation des schémas de voisement des entrées du lexique est réalisée sans tenir compte des phénomènes d'assimilation et d'élision du [∂] de la manière suivante :

- Peu de mots contiennent une suite de trois consonnes, nous négligeons donc les assimilations consonantiques potentielles.
- Dans le cas de deux consonnes consécutives, nous ne prenons pas en compte les assimilations consonantiques de sourdité. L'alignement temporel n'étant encore pas connu, un tel phénomène ne doit pas entraîner de perturbation du schéma de voisement (Ex: [$a\rho\tilde{}d$] $\rightarrow V.N_V.V$ / [$a\rho_{\circ}\tilde{}d$] $\rightarrow V.N_V.V$).
- Nous considérons les éventuelles assimilations consonantiques de voisement qui peuvent modifier le schéma de voisement, notamment pour les mots se terminant par le suffixe [-ism] où il est fréquent que le trait de voisement de [m] soit transféré sur le [s] (Ex: [$\tilde{}da\ell ism$] $\rightarrow$ [$\tilde{}da\ell izm$]).
- Nous prenons enfin en considération la chute du [∂] en finale de mot qui pourrait être responsable d'un mauvais rendement de notre premier filtre ; les prononciations du [∂] final sont fréquentes dans notre base AviLex ce qui a pour effet de modifier le schéma de voisement (Ex: (attente) [$at\tilde{}t$] $\rightarrow V.N_V.V$ / [$at\tilde{}t\partial$] $\rightarrow V.N_V.V.N_V.V$).

Nous avons testé l'efficacité de notre filtre sur la base AviLex qui a été utilisée lors des tests du module d'accès lexical, avec les mêmes sous-lexiques de respectivement 15 000 et 20 000 mots [11]. Les résultats sont reportés dans la table 4.33 et font apparaître un taux de filtrage de l'ordre de 60% pour une erreur de moins de 3%. Ces résultats sont satisfaisants tant en terme de réduction des entrées lexicales qu'en simplicité de mise en œuvre. Notons également que ces résultats sont obtenus par un algorithme indépendant du locuteur, un taux d'erreur de moins de 1% — pour un taux de filtrage identique — est obtenu, lorsque nous spécifions un seuil de voisement par locuteur. Bien que nous pensons qu'il est possible de déterminer automatiquement ce seuil, lors de la phase d'apprentissage des références spectrales, nous n'avons pas désiré pousser plus loin nos investigations, nous contentant par conséquent de l'algorithme global dont les performances — au regard des résultats présentés dans la section 4.5) — s'avèrent satisfaisantes.

Loc	taux de filtrage	erreur	lexique
fb	60%	2.9%	lexi_15000
hm	60%	2.1%	lexi_15000
ts	59%	3.4%	lexi_15000
pg	59%	1.2%	lexi_15000
lc	59%	3.7%	lexi_15000
si	59%	2.7%	lexi_15000
pg7	62%	3.1%	lexi_20000
si7	61%	4%	lexi_20000
tous	59.9%	2.89%	—

Table 4.33: Taux de filtrage et d'erreur obtenu par le premier filtre sur la base AviLex avec les lexiques de 15 000 et 20 000 mots qui ont été employés lors de la phase d'évaluation du module d'accès lexical.



Lexique

entrée	schéma	emuet
affectueuse	V.Nv.V.Nv.V	
affirmation	V.Nv.V.Nv.V	
agricultrice	V.Nv.V.Nv.V	x
agriculture	V.Nv.V.Nv.V	
agriculteur	V.Nv.V.Nv.V	
association	V.Nv.V.Nv.V.Nv.V	
ainsi	V.Nv.V	

Figure 4.35: Le premier filtre. La courbe de voisement calculée pendant la détection de la fréquence fondamentale, permet de déterminer un schéma de voisement qui sera comparé aux schémas des entrées lexicales qui sont pré-compilés.

Filtre 2

Nous avons expérimenté dans ce deuxième filtre l'introduction de deux nouvelles familles dans la phase de décomposition d'un treillis en familles (voir la section 4.2.2 page 56 pour un rappel des familles considérées) : la famille CN pour les consonnes non voisées et la famille CV pour les consonnes voisées. En pratique, les consonnes voisées génèrent une famille CV, à l'exception de la consonne [] sur laquelle aucune décision n'est prise (famille CO) ; les consonnes non voisées sont étiquetées CN. Les entrées lexicales sont alors précompilées avec ces nouvelles familles, autorisant ainsi un accès lexical plus sélectif : les mots les plus courants ayant une décomposition VO/CO simple, on arrive fréquemment à la situation où un schéma va pratiquement correspondre à la moitié des mots du lexique (voir la table 4.32 de la page 121). Nous reportons dans la table 4.34 les vingt schémas les plus fréquemment rencontrés dans le lexique de 15 000 mots que nous utilisons dans nos tests pour le codage habituel VO/CO puis pour le nouveau codage VO/CN-CV-CO. introduit dans ce filtre. On comprend alors aisément l'intérêt d'un tel filtre.

rg	schéma (VO/CO)	nb	%	schéma (VO/CV-CN-CO)	nb	%
1	VOCOVOCOVO	6996	46.8	VOCVVO	2674	17.9
2	VOCOVO	5589	37.4	VOCNVO	2572	17.2
3	VOCOVOTRVO	3676	24.6	VOCNVOTRVO	1837	12.3
4	VOCOVOCOVOTRVO	3119	20.9	VOCVVOTRVO	1699	11.4
5	VOCOVOCOVOCOVO	2880	19.3	VOCNVOCVVO	1640	11.0
6	VOCOVOCOCOVO	2080	13.9	VOTRVO	1346	9.0
7	VOCOCOVO	2035	13.6	VOCVVOCVVO	1345	9.0
8	VOCOCOVOCOVO	1860	12.4	VO	1260	8.4
9	VOTRVOCOVO	1539	10.3	VOCNVOCNVO	1210	8.1
10	VOTRVO	1346	9.0	VOCVVOCNVO	1156	7.7
11	VOCOVOTRVOTRVO	1302	8.7	VOCNVOCOVO	997	6.7
12	VO	1260	8.4	VOTRVOTRVO	987	6.6
13	VOCOVOTRVOCOVO	1154	7.7	VOCNVOCVVOTRVO	865	5.8
14	VOCOCOVOTRVO	1046	7.0	VOCVVOCOVO	783	5.2
15	VOTRVOTRVO	987	6.6	VOCVVOCVVOTRVO	696	4.7
16	VOCOCOVOCOVOTRVO	820	5.5	VOCNVOTRVOTRVO	681	4.6
17	VOCOCOVOCOVOCOVO	733	4.9	VOCNVOCNVOTRVO	670	4.5
18	VOTRVOCOVOCOVO	611	4.1	VOTRVOCVVO	602	4.0
19	VOCOVOCOCOVOTRVO	600	4.0	VOCOVO	600	4.0
20	VOCOVOCOCOVOCOVO	584	3.9	VOCVVOTRVOTRVO	574	3.8
total	336			2059		

Table 4.34: Occurrence et pourcentage des 20 schémas précodés les plus courants du lexique de 15 000 mots (codage VO/CO et VO/CV-CN-CO).

L'apport de ce filtre est reporté dans la table 4.35 pour le corpus AviLex1 sur le lexique de 15 000 entrées : 25 % des mots du lexique sont éliminés avec une erreur de l'ordre de

1%. Ce filtre est d'autant plus intéressant que son coût est infime ; les entrées du lexique étant pré-compilées.

locuteur	taux de filtrage	erreur
fb	25 %	1.2 %
pg	24 %	0.6 %
ts	26 %	1.60 %
lc	28 %	2.40 %
si	25 %	0.40 %
moyenne	25 %	1.2 %

Table 4.35: Taux de filtrage et d'erreur associé du filtre 2 à l'issue de la phase d'accès au lexique (15 000 entrées) sur le corpus AviLex1.

Filtre 3

Ce filtre opère en amont du filtrage lexical dont la sortie est constituée d'une cohorte valuée de 50 à 150 mots parmi lesquels doit normalement se trouver le mot prononcé. A contrario du premier filtre, nous disposons de la suite de phonèmes reconnus ainsi que leur alignement dans le temps. Il nous est alors possible de filtrer plus finement certains mots en intégrant les règles d'assimilation consonantiques et d'élision du [∂]. La table 4.36 présente les résultats obtenus pour l'ensemble des locuteurs de la base AviLex. On constate qu'il est possible d'obtenir un filtrage moyen de 20% sur les cohortes avec un taux d'erreur de moins de 3% qui peut s'expliquer soit par une décision de voisement erronée (30% des cas erratiques), ou par un défaut dans l'alignement proposé par le niveau lexical qui propose un phonème non voisé sur une zone voisée ou réciproquement (le restant des cas). Le gain moyen — tous locuteurs confondus — est de trois places dans les situations où le mot prononcé n'est pas le premier reconnu. Un exemple de filtrage qu'il n'était pas possible d'effectuer par les filtres 1 et 2 (faute de précision sur la durée) est présenté ici ; les mots précédés d'une croix sont ceux rejetés par le filtre, la première ligne indique le voisement mesuré (v pour voisé, n pour non voisé et i pour indécis) ; le mot prononcé est le mot *bicyclette* ici classé en tête.

```
  iivvvvvvvvvvvvvvinnnnnnnnnvvvvvvviinnnnnnnnniivvvvvvvvvvvvviinnnnnnnnnnnnnnnniivvii
  .bb    ...ii..    ..ss.     ..ii..  ..kk.      .ll ...ai..     ......tt.....
       .rr.ei..     ..ss.     ..ii..   ..pp.     .rr ...oo..     ......kk.....
  .pp  .rr.ei..     ..ss.     ..ii..   ..pp.       ....ii...     ......tt.....
  .kk    ...on..    ..ss.     ..ei..  ..kk.       ....an....     ......ss.....
         ...ei..    ..ff.     ..ii..  ..kk.       ....aa....     ......ss.....
x .dd    ...ii..    ..ss.     ..ii..   ..pp.   .rr...........on.............
x .dd    ...ii..    ..ss.     ..ii..   ..pp.       ............an...........
x .dd    ...ii..    ..ss.     ..ii..   ..pp.   ..............on.............
x      .rr.ei..     ..ss.     ..ii..   ..pp.     .yy ...........an...........
x  .ss   ...uu..    ..ss.     ..ei..  .pptt.       ....ii...     ....bb...       .ll
  .kk    ...on..    ..ff.     ..ii..   ..tt.       ....uu...     ......rr.....
x .dd    ...ii..    ..ss.     ..ii..   ..pp.     .rr ...........ai...........
         ...in..    ..sspp    ..ei..  .kktt.       ....oe...     ......rr.....
x      .rr.ei..     ..ss.     ..ii..   ..tt.   .rr...........on.............
x .kk    ...on..    ..ss.     ..an..   ..tt.     .rr ...........an...........
x .dd    ...ii..    ..sspp    ..uu..   ..tt.   .rr...........on.............
x .dd    ...ii..    ..sskk    ..uu..   ..tt.   .rr...........on.............
x .dd    ...ii..    ..ss.     ..ii..   ..pp.       ............oe...........
x .dd    ...ii..    ..ss.     ..ii..   ..pp.       ............ai...........
x      .rr.ei..     ..ss.     ..ii..   ..tt.       ............an...........
x      .rr.ei..     ..ss.     ..ii..   ..tt.   ..............on.............
x        ...in..    ..sspp    ..ei..  .kktt.     .rr ...........an...........
x        ...in..    ..ss.     ..ei.rr  ..tt.       ............in...........
x .kk    ...on..    ..ss.     ..uu.ll  ..tt.   .rr...........on.............
x .dd    ...ii..    ..sspp    ..uu..   ..tt.       ............an...........
x .dd    ...ii..    ..sskk    ..uu..   ..tt.       ............an...........
x .dd    ...ii..    ..sspp    ..uu..   ..tt.   ..............on.............
```

Locuteur	taux de filtrage	taux d'erreur	gain moyen
pg	20.9%	2.1%	3.5
fb	17.5%	2.2%	2.7
ts	21%	2.7%	3.7
si	16.9%	2.2%	3
lc	17.8%	1.5%	2
tous	20%	3%	3

Table 4.36: Résultats du filtre 3 par locuteur puis tous locuteurs confondus *tous* avec les taux de filtrage et d'erreur exprimés en pourcentage et le gain moyen (exprimé en place) pour tous les mots non reconnus en première position.

4.5 Bilan

Nous résumons brièvement les tentatives d'intégration de filtres prosodiques dans notre système de filtrage lexical :

Les indices macroprosodiques : le caractère très spécifique des expériences menées dans le cadre du filtrage macroprosodique (tests réalisés sur un seul locuteur) et leur faible rendement pour obtenir des taux d'erreur acceptables nous ont amenés à ne pas les retenir dans notre processus de filtrage. Nous rappelons tout de même que parmi les indices de durée, de fréquence fondamentale et d'intensité, les résultats les plus intéressants ont été mesurés pour le dernier paramètre. Nous retiendrons de cette étude — et conformément à nos convictions initiales — qu'il ne semble pas réalisable pour le français, d'utiliser à l'instar de Waibel [162] pour l'anglais, des informations suprasegmentales à des fins de filtrage lexical de mots prononcés isolément.

Les indices microprosodiques : il faut bien reconnaître que peu d'indices microprosodiques ressortent de notre étude ! Il semblerait tout d'abord qu'une distinction des voyelles orales et nasales soit réalisable à partir des durées fournies par nos modèles de phonèmes. L'intégration de cette information nécessiterait cependant une remise en cause des choix méthodologiques retenus pour notre module d'accès lexical sans garantie véritable d'une amélioration significative. Des indices de fréquence fondamentale, il s'avère que seule une distinction des consonnes [ℓ], [m] et [n] des autres consonnes voisées soit réalisable avec des taux d'erreur variables suivant les différents corpus étudiés. Nous avons réalisé un filtre mettant en place cette distinction de manière partielle — à partir des distributions présentées sur la figure 4.26 — et s'appliquant aux mots proposés en sortie de SPEX. Le faible nombre de cas faisant intervenir la décision et le gain moyen en place engendré (inférieur à l'unité pour l'ensemble des cohortes de AviLex1 !) nous a amenés à négliger cet indice. L'étude des indices d'intensité a fait ressortir la possibilité d'une distinction efficace des voyelles [i] et [a] en position initiale et médiane de mot (15% d'erreur à l'initiale et 8% en position médiane des mots de trois voyelles du corpus FeLex). De piètres performances (en terme de taux de filtrage) ont été obtenues par un filtre basé sur cet indice avec les mots du corpus AviLex, ce qui signifie entre autre chose que le système SPEX fonctionne bien et qu'il est à même de réaliser finement une telle distinction. Ainsi il semble que l'indice de voisement soit le seul qui autorise un filtrage efficace avec une perte acceptable. Nous reportons sur la figure 4.36 les taux de classement obtenus avec les différents filtres de voisement tous locuteurs confondus : il est intéressant de noter qu'en plus d'un filtrage performant, le nombre de mots classés dans les dix premières positions peut augmenter avec la prise en compte de l'indice de voisement (différence qui ne semble cependant pas hautement significative).

Nous conclurons ce chapitre en remarquant que nos résultats tendent à confirmer ceux d'une étude récente de Dumouchel [50] qui mesure l'apport d'une composante micro-

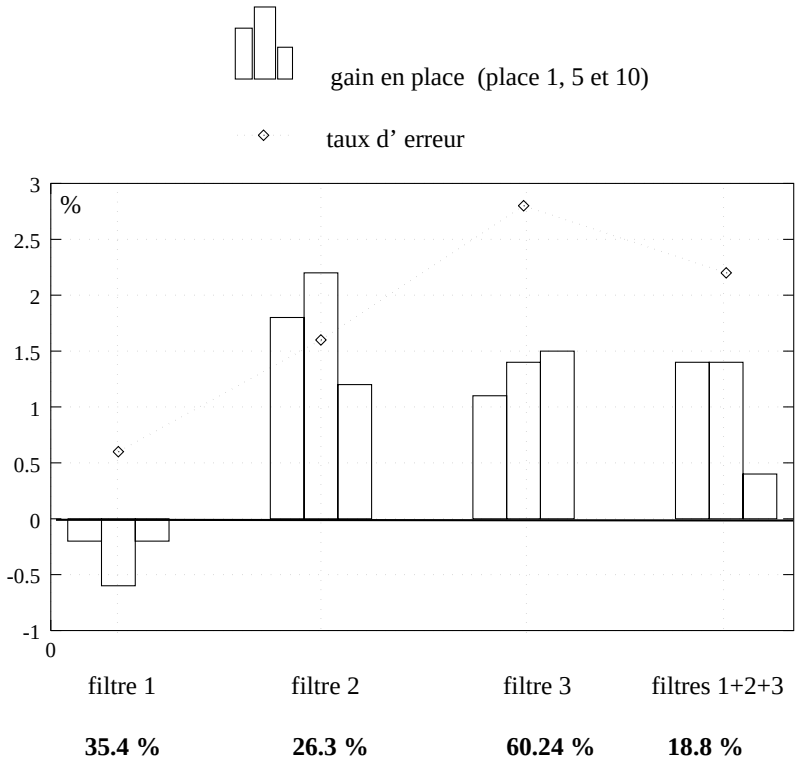


Figure 4.36: Cette figure indique l'amélioration apportée par chaque filtre de voisement au taux de classement des mots en tête, dans les 5 premières positions, puis dans les 10 premières positions. La courbe indique le pourcentage d'erreur engendré, et la taille du lexique (exprimée en pourcentage) restant après filtrage à la fin de l'étape d'accès au lexique (*i.e.* avant que n'intervienne le filtre A du processus de filtrage). Le filtre 3 intervenant seulement sur les cohortes issues du filtrage lexical, le taux de 60.24 % indique le pourcentage du lexique qu'il reste sans application des filtres de voisement.

prosodique (obtenue automatiquement à l'aide d'un classificateur bayésien à distributions multi-gaussiennes) dans un système de reconnaissance markovien et qui, en dépit du fait qu'une décision voisée/non voisée semble pertinente, conclut à une amélioration non significative des résultats.

Chapitre 5

Organisation suprasegmentale

Au cours du précédent chapitre, nous avons fait le point sur l'utilisation d'indices prosodiques (et essentiellement microprosodiques) dans un processus de filtrage lexical de mots prononcés isolément. Conscient de l'étroitesse de cette étude, nous nous devions d'explorer l'organisation suprasegmentale des paramètres prosodiques afin d'en mesurer l'apport dans un processus général de reconnaissance automatique de la parole continue. Dans la suite de ce chapitre, le terme de prosodie référera — sauf précision contraire — à sa seule composante suprasegmentale ; les éléments suprasegmentaux se définissant classiquement par leur domaine de réalisation qui s'étend au-delà d'un simple segment (ou phonème) et dont l'existence ne tient qu'à l'étude de leur agencement en séquence[1].

5.1 Objectifs

Avant de présenter nos objectifs, nous aimerions préciser que, bien que traitant de l'information prosodique utilisable dans un processus de reconnaissance de la parole, ce chapitre n'a pas pour fonction de proposer un ensemble de règles décrivant nos connaissances prosodiques que nous validerions par l'amélioration des résultats de reconnaissance apportée à un système particulier. Nous allons au contraire exposer les limites d'une telle approche et préciser les raisons qui nous font penser qu'un apprentissage à partir d'exemples est une solution satisfaisante à l'intégration de la prosodie en reconnaissance de la parole mais également à l'analyse prosodique à but exploratoire. Une partie de ce chapitre sera dédiée à la description d'un système capable de réaliser automatiquement une étude corrélative entre des indices prosodiques et différents niveaux d'organisation du message. Nous montrerons ensuite sur deux applications (reconnaissance de nombres et reconnaissance de phrases) le rôle prédictif que peut assurer notre processus au sein d'un système de reconnaissance de la parole.

[1] Voir [83] pour une définition plus complète des informations suprasegmentales.

5.2 Quelques points concernant la prosodie

Comme nous l'avons déjà précisé dans le premier chapitre, la prosodie intervient à tous les niveaux du processus de communication (y compris au niveau phonétique abordé lors du précédent chapitre). Les schémas prosodiques que nous mettons en œuvre durant l'acte de parole ne sont pas le fruit du hasard mais répondent au contraire à des impératifs qui assurent à la prosodie un statut linguistique. Nous disposons cependant d'un degré de liberté assurant à l'acte de parole son caractère vivant. Au delà de la fonction intégratrice de la prosodie qui permet d'accorder à une suite de mots grammaticalement incorrecte un statut linguistique (très fréquent dans les situations naturelles de dialogue), il n'est cependant pas interdit de penser que la prosodie joue un rôle moins important que celui assumé par les unités segmentales. Plus exactement on peut très bien concevoir un mode de communication privé de toute prosodie, il suffit pour s'en convaincre — au moins partiellement — de fixer le paramètre du fondamental d'un système de synthèse à une valeur donnée et d'écouter les réalisations ainsi produites pour s'apercevoir qu'il nous est alors toujours possible d'accéder au sens du message avec cependant une attention plus soutenue. G. Caelen [118] remarquait à ce propos :

> "Quand par exemple, on dit d'un énoncé que sa prosodie est inacceptable cela veut dire que le message est décodé et sa prosodie analysée puis après confrontation (qui peut commencer en début de décodage) jugée inappropriée au contenu. Autrement dit, il y a eu des traitements très largement indépendants du message vocal et de sa prosodie."

Ceci ne signifie bien sûr pas qu'il faille négliger la prosodie dans le processus de communication : si la parole est par nature redondante, c'est que sa fonction principale est d'assurer la transmission d'une information d'un locuteur à un auditeur ; et c'est cette redondance qui facilite la tâche de ce dernier. Nous pouvons reprendre la classification des faits prosodiques en trois catégories proposées par Fujisaki [57] et dont les limites ne sont pas toujours nettement définies [157] :

Le niveau linguistique qui concerne les informations symboliques des niveaux lexical, syntaxique, sémantique et pragmatique. Parmi les fonctions relevant du domaine linguistique, la fonction de groupement (fonctions identificatrice et hiérarchisante) — qui permet d'associer entre eux des mots sémantiquement liés — est certainement celle qui est la plus convoitée pour une utilisation dans un processus de traitement automatique [157].

Le niveau para-linguistique qui relève des variations contrôlées par le locuteur permettant de transmettre à la personne qui l'écoute des informations précisant son attitude par rapport au message qu'il transmet. Ces indices sont particulièrement importants dans des situations naturelles de dialogue où le locuteur a souvent recours à des schémas prosodiques pour indiquer son état d'esprit face au contenu du message énoncé [74].

Le niveau non-linguistique qui permet à un auditeur de disposer d'informations sur le locuteur, comme son état de santé, son accent ou encore son état émotionnel. Ces informations ne sont pas le fruit d'une planification consciente de la part du locuteur (bien qu'elles puissent être reproduite) et sont souvent qualifiées de variables non contrôlées.

Pour ce premier contact avec la prosodie, nous allons nous limiter à l'étude de la fonction de segmentation d'un message en groupes de mots sémantiquement liés (comme un nom avec l'article qui le précède ou bien un adjectif qui le précède ou le suit).

Quels sont les problèmes soulevés ?

Réaliser une étude de la structuration prosodique d'un énoncé impose que l'on pose un ensemble de questions telles que :

- Quelle est la place accordée aux informations prosodiques dans le système linguistique ?
- Est-on capable d'ériger un ensemble de règles prédictives de tout ou partie de ces informations ?
- À l'inverse, peut-on déceler dans le signal de parole des indices prosodiques (acoustiques ou perceptuels) qui seraient révélateurs de la structuration constituante d'un message ?
- Existe-t-il des invariants ? Si oui quels sont-ils ?
- Quelles données d'observation doit-on étudier pour répondre à ces questions ?

Il est presque inutile de préciser qu'aucune de ces questions ne possède actuellement de réponse qui satisfasse l'ensemble de la communauté des linguistes, même si les deux premières — qui relèvent de préoccupations plus théoriques — ont donné naissance à des paradigmes accordant partiellement l'ensemble des chercheurs. Nous n'avons bien sûr pas dans la suite de ce mémoire la prétention de répondre de manière définitive à ces questions, mais simplement de préciser nos vues quant à ces points et de proposer un outil à vocation exploratoire et prédictive : le système ProStat (pour **Pro**sodie et **Stat**istique).

Quelques éléments de réponse ...

Le statut de la prosodie a considérablement évolué en l'espace d'une trentaine d'années et s'il est un point qui rassemble de nos jours tous les chercheurs en prosodie, c'est bien son appartenance au code linguistique[2].

[2]Nous renvoyons le lecteur à [40] pour une discussion sur l'évolution du statut prosodique à travers les années.

Les désaccords interviennent lorsque l'on tente de décrire les liens entre l'organisation prosodique et les autres niveaux de structuration du langage. Il semble raisonnable de dire que le problème restera ouvert tant que nous ne serons pas en mesure d'expliquer pleinement les mécanismes de planification de l'acte de parole. On peut toutefois s'accommoder de systèmes explicatifs élaborés par des experts et dont la complexité dépend grandement du nombre de paramètres qu'ils tentent de modéliser. Notre propos n'étant pas de les décrire[3] nous renvoyons le lecteur aux travaux de thèse de [144, pp. 6–8] qui réalise une revue des principaux modèles existants. Nous pouvons simplement remarquer avec Vaissière [158] qu'aucun de ces modèles ne peut répondre aux attentes multiples de leurs utilisateurs potentiels.

En schématisant le problème on peut considérer que la prosodie est à la fois gouvernée par les organisations syntaxique, sémantique et pragmatique des énoncés[4] et qu'elle répond également à un ensemble de contraintes rythmiques. La priorité contextuelle à accorder à ces différents niveaux de structuration — souvent conflictuels — étant un des principaux obstacles à la modélisation des informations prosodiques [157, 45].

Dès lors, toute tentative d'explication possède son lot de partisans et de détracteurs. En premier lieu, l'influence de la syntaxe sur l'organisation prosodique peut-être discutée, ainsi très récemment Monaghan [104] écrit-il en décrivant son système de création de dialogues oralisés :

> "An important point is that BRIDGE's intonation rules make no use of syntactic structure : our working hypothesis is that syntax is not relevant to determining intonation."

Pour autant que l'on accepte que la prosodie et la syntaxe entretiennent des relations particulières, il n'en reste pas moins compliqué de les décrire. On peut tout d'abord s'interroger sur la profondeur de ces relations ; il semble à la lumière des nombreux travaux traitant de ce problème que le débat sur la congruence de ces deux structurations est maintenant obsolète et que la notion de points de *rendez-vous* entre les deux soit largement acceptée au sein de la communauté prosodique (voir par exemple les exposés de Martin et Rossi dans [126]). Une des préoccupations des chercheurs en prosodie est alors de quantifier ces points de rendez-vous afin d'en rechercher d'éventuels invariants dans des buts divers ; en synthèse d'abord où la qualité de la parole générée est grandement dépendante de ces connaissances et également en reconnaissance de la parole dont les processus pourraient s'appuyer sur de tels points pour réduire leur espace de recherche des solutions.

En plus de la syntaxe, nous savons que la prosodie entretient des liens étroits avec l'organisation sémantique. Dans son étude sur les stratégies prosodiques des locuteurs dans des situations de lecture obéissant à des consignes particulières, G. Caelen montre très

[3] Voir un exposé intéressant de Martin dans [126, pp. 235–236] pour une discussion du statut de ces différents systèmes.

[4] Voir la thèse d'état de Mariani [91] pour une définition et une description des différentes approches des domaines syntaxique (pages 84 à 99), sémantique (pages 101 à 112) et pragmatique (pages 113 à 127).

clairement que les comportements prosodiques sont variés et font intervenir des niveaux de structuration syntaxique certes, mais également sémantique et pragmatique [21]. On trouvera dans son mémoire une synthèse des approches syntaxiques et sémantiques (pages 45–56) ainsi qu'un descriptif des deux modèles syntaxiques, des trois modèles sémantiques et du modèle pragmatique qu'elle considère dans son étude. De même Rossi [134, 135] dans son modèle fait une large place aux informations sémantiques pour décrire les diverses configurations prosodiques.

Enfin, la prosodie répond aussi à des contraintes rythmiques. Le rôle de la pause et de l'allongement final a suscité l'intérêt de nombreux chercheurs dont notamment Grosjean [32] qui utilise les pauses pour dériver ses structures de performance ou encore Emerard qui définit le rythme comme la structuration de l'énoncé par les pauses [34]. Tous démontrent que les pauses et les allongements finaux sont des éléments pertinents du découpage de la parole en unités linguistiques. Le rythme abrite également deux notions largement étudiées [31] que sont l'*isosyllabicité* (*i.e.* les durées des segments sont compressées ou au contraire étendues afin de garder constante la durée d'une syllabe) et l'*isochronie* (*i.e.* la tendance à garder constante la durée entre deux syllabes accentuées). Ces deux principes allant à l'encontre de la notion de phase temporelle mise en évidence pour le français par Pasdeloup [116] et qui se caractérise par un mouvement de ralentissement progressif du débit à l'intérieur d'une même phase, par la présence d'une voyelle allongée terminale de phase, ainsi que par un phénomène de ré-initialisation de la première syllabe qui est ramenée à une valeur brève à peu près constante pour toutes les phases d'une même phrase. Compte tenu des réserves que nous avons émises au cours du chapitre traitant des informations microprosodiques sur la fiabilité des mesures automatiques de la durée, nous pouvons considérer que, dans le cadre d'un traitement automatique, il ne semble pas réalisable d'apprécier les réductions progressives du débit dans les phases temporelles *a fortiori* pour de la parole réelle[5] où des variations intrinsèques et co-intrinsèques viennent perturber ces observations. De manière plus globale, il semble que les contraintes rythmiques sont relativement indépendantes des autres organisations linguistiques et qu'elles peuvent entrer en conflit avec ces dernières [5, 157, 45].

Ainsi le chercheur qui tente de modéliser les informations prosodiques est-il confronté à la multiplicité des facteurs intervenant (syntaxique, sémantique et rythmique pour l'essentiel). Pour mener à bien cette tâche il aura recours — en plus de son expérience — à des données d'observation lui permettant de vérifier ou au contraire d'infirmer certaines règles. On peut dès lors s'interroger sur la nature et la quantité des données d'observation nécessaires à la bonne réalisation de cette tâche. Plusieurs remarques peuvent être formulées à ce sujet :

- Le choix de la nature des données à étudier dépend bien entendu de celle des travaux envisagés et de présupposés de l'expérimentateur (parole de type lue ou bien spontanée, choix de la phrase comme unité de réalisation ou choix du texte, *etc.*).

[5] Pasdeloup utilise de la parole réitérée en "ma-ma-ma" pour mettre en évidence l'existence des phases temporelles.

- Un premier biais à éviter est celui qui consiste à n'étudier les réalisations que d'un seul locuteur, situation non souhaitable car l'expérimentateur peut à tort attribuer une valeur linguistique à de simples artefacts spécifiques au locuteur étudié [24] ; en particulier si le locuteur est l'expérimentateur, auquel cas, l'objectivité de ses réalisations face au propos de l'étude n'est pas nécessairement garantie.

- En second lieu, le choix de la grandeur des corpus d'observations n'est pas simple. Tout d'abord, il faut bien constater que la taille du corpus n'est pas toujours garante de son adéquation au propos de l'étude. Ainsi un corpus de petite taille bien choisi peut-il couvrir au mieux les besoins de l'étude *a contrario* de corpus de plus grandes tailles qui seraient mal sélectionnés. Pour ne citer que cet exemple emprunté à Shih et Ao [139] lors d'une étude récente des durées pour un système de synthèse du chinois mandarin à partir du texte, les auteurs rappellent que la sélection de 424 phrases à l'aide d'un "greedy" algorithme[6] leur a permis d'obtenir une couverture totale des facteurs dont ils souhaitaient faire l'étude alors qu'un choix aléatoire du même nombre de phrases n'aboutissait qu'à une couverture partielle de 74% ; 42 phrases seulement sélectionnées par cet algorithme auraient alors suffi pour obtenir une telle couverture. Si la grandeur d'un corpus n'est pas forcément un gage de qualité, il n'en reste pas moins surprenant de constater certains choix comme celui fait par G. Caelen lors de son étude sur les stratégies des locuteurs en réponse à des consignes particulières de lecture [21] : trois phrases seulement lui ont servi de corpus de référence prononcées par 12 locuteurs dans le cadre de trois consignes de lecture (notons cependant que selon l'auteur et pour ces seules réalisations, pas moins de 40 000 étiquettes ont été apposées sur le signal, de manière semi-automatique, durant une période de 6 ans !). Nous pouvons retenir de tout cela, qu'un corpus bien choisi est préférable à un grand corpus, ce qui implique cependant que l'expérimentateur sache au moment de son choix énumérer les différents facteurs sur lesquels il souhaite porter son analyse.

- Enfin et pour clore cette discussion sur les choix des données de référence, il est évident que des considérations d'ordre pratique peuvent influencer les choix faits à ce niveau. L'utilisation de corpus déjà existants est tentante quand on connaît le travail qu'impose la définition, l'enregistrement et l'étiquetage (orthographique, phonétique, prosodique, *etc.*) d'un corpus[7]. Encore faut-il faire remarquer que l'amélioration des techniques actuelles permet indéniablement de s'affranchir d'une partie des travaux coûteux d'étiquetage et par ce fait de pouvoir traiter davantage de données que par le passé [79, 5, 148, 166].

Ainsi, une des premières difficultés du chercheur est de faire le choix d'un corpus d'analyse répondant au mieux à ses préoccupations, problème qui s'avère non trivial compte tenu

[6] L'algorithme tire son nom du fait que chaque item sélectionné a nécessité l'analyse de la totalité des items du corpus.

[7] Voir par exemple le rapport interne décrivant ces travaux pour la partie suisse-romande de la base PolyPhon enregistrée à l'Idiap [77].

des différents points énoncés. À ce titre, nous rappelons une tentative qui s'est déroulée aux Journées d'Études sur la Parole de Bruxelles (1992) où une réunion satellite avait été spécialement organisée autour de ce thème avec l'échec (prévisible) qu'on lui a connu et qui traduit bien l'ampleur de la difficulté.
Un dernier point que nous désirons aborder rapidement ici est celui des méthodologies du chercheur et de leurs évolutions actuelles. Nous venons de voir que les techniques existantes se sont améliorées et qu'elles permettent maintenant d'obtenir automatiquement des informations qu'il était encore fastidieux de réunir il y a seulement quelques années. Il est bien évident que ces techniques ont des limites et qu'elles ne permettent pas de répondre aux besoins spécifiques de chaque étude prosodique, il n'en reste pas moins qu'elles ont instauré une dynamique constructive entre les chercheurs qui proposent des modèles et les autres qui tentent de les utiliser dans des systèmes (de synthèse ou de reconnaissance de la parole). On assiste même actuellement à une tendance très nette à remplacer les composantes obtenues par des experts par des composantes générées à l'aide de techniques statistiques [150, 161, 111, 140, 75, 73, 69, 68, 63, 114]. Plusieurs raisons peuvent expliquer cette mutation :

- En tout premier lieu, l'usage massif de techniques statistiques comme les modèles de Markov ou encore les réseaux neuro-mimétiques a permis d'obtenir rapidement des outils performants pour extraire du signal de parole des informations de bas niveau. On peut raisonnablement penser que ces techniques, dont les résultats annoncés sont souvent de très bonne qualité (pour la reconnaissance phonétique par exemple), ont atteint leur plafond et que les "luttes" pour le gain de quelques dixièmes supplémentaires relèvent davantage d'un exercice de style que d'une réelle nécessité. En plus de la progression des niveaux inférieurs (en référence à une organisation classique d'un système de reconnaissance), les méthodes statistiques proposent des solutions à des niveaux comme la syntaxe ou même encore la sémantique avec un large recours aux modèles de langage probabilistes [52]. Malgré tous ces efforts, si des systèmes aux prétentions limitées peuvent maintenant voir le jour (comme des systèmes de réservation ou de renseignement automatique [74]), il n'en reste pas moins que le problème général de la reconnaissance de la parole n'est pas encore — loin s'en faut — résolu. Ainsi la tendance actuelle est-elle de se tourner vers des sources de connaissances encore peu exploitées dans les systèmes de reconnaissance ; la prosodie fait bien sûr partie de ce type d'informations. On assiste donc actuellement à une arrivée de toutes ces techniques dans le domaine prosodique non pas nécessairement à des fins explicatives mais certainement dans un but évident de performance.

- Une autre raison qui peut expliquer l'engouement pour ces techniques tient également à la méthodologie du chercheur face à la multiplicité des facteurs qu'il doit prendre en considération pour modéliser correctement la structuration prosodique du langage. Une méthode courante et difficilement contournable est alors de considérer comme fixés certains paramètres afin d'analyser l'évolution des autres qui sont alors en nombre raisonnable. Les méthodes dites statistiques permettent de s'affranchir

de cette problématique en réalisant non plus une optimisation locale de quelques facteurs mais au contraire en autorisant une analyse globale de la totalité des facteurs intervenants.

- Ces méthodes nécessitent généralement moins de données manuelles qu'une expertise classique et sont donc d'une mise en route assez aisée, ce qui constitue un argument supplémentaire en leur faveur.

Il faut cependant bien reconnaître que de telles techniques possèdent également des défauts qu'il ne faut pas oublier. Le premier d'entre eux qui est davantage une frustration est celui de ne pas pouvoir utiliser ces techniques à des fins explicatives ; leur sorties étant assez rapidement illisibles et donc difficilement exploitables en tant que connaissance. Un autre problème lié à ces algorithmes est le besoin de corpus d'apprentissage importants qui ne sont pas nécessairement disponibles. Quand bien même ce serait le cas, les temps de calcul liés à ces techniques demeurent encore un problème à prendre en compte. Plus fondamentalement, si ces dernières sont habiles à modéliser les données d'apprentissage, elles le sont nettement moins lorsqu'on leur demande de se prononcer sur des données dont elles n'ont pas disposé lors de la phase d'apprentissage. Les problèmes de lissage qui en résultent sont loin d'être réglés ce qui nous donne à penser que le recours à la connaissance est encore (heureusement !) un passage obligé. Nous renvoyons le lecteur à la récente étude de Santen [159] pour une discussion plus complète sur l'utilisation de l'outil statistique dans le cadre de la synthèse de la parole à partir du texte. Une méthodologie idéale serait donc de mélanger les approches statistique et analytique afin de remédier aux défauts de l'une par les avantages de l'autre. C'est dans ce sens qu'a été conçu le système ProStat que nous allons maintenant présenter.
Rappelons simplement que c'est aussi l'approche qu'a employée Aubergé [4] dans son système de constitution semi-automatique d'un module de génération de l'intonation pour un système de synthèse : son système permet de formaliser — après analyse — un lexique hiérarchisé de formes intonatives globales (les contours moyens). Elle utilise pour cela des paires minimales d'attributs à différents niveaux linguistiques : la phrase (déclarative, interrogative ou impérative), la proposition (position absolue, indépendance, dépendance relative) et le groupe (nature, fonction, position absolue, position relative). Son système permet ainsi de modéliser l'intonation d'une phrase dont elle connaît les différents attributs par superposition des contours moyens appris semi-automatiquement sur une base spécifique de phrases isolées. L'auteur conclut après analyse de sa base à l'existence de formes intonatives caractéristiques des différents niveaux modélisés (phrase, proposition, groupe et sous-groupe) ainsi qu'à leur dépendance hiérarchique.

5.3 Qu'est-ce que ProStat ?

Initialement conçu à des fins exploratoires, ce système permet de mesurer la corrélation entre des indices prosodiques, des contraintes rythmiques et des niveaux d'organisation linguistique particuliers. Sa caractéristique principale est que notre système ne possède

aucun *a priori* que ce soit sur une hypothétique hiérarchie de ces différentes contraintes régissant l'organisation prosodique ou encore sur des entités telles que l'*accent* ou des marques intonatives particulières. Cette stratégie ne relève aucunement d'un goût prononcé pour la difficulté mais répond bien au contraire à des impératifs liés au traitement automatique. Nous aurions aimé dans cette étude manipuler des entités d'un niveau supérieur aux simples traits acoustiques mesurés depuis le signal ne serait-ce qu'en raison de leur nombre restreint. En particulier, si la notion d'accent nous paraît plaisante[8], il faut bien reconnaître qu'elle n'en reste pas moins à nos yeux ambiguë ; et ce bien au-delà de simples conflits terminologiques dont nous avons fait part dans l'introduction. Ainsi pour ne prendre qu'un exemple, lors d'une étude récente sur la distribution accentuelle dans un corpus de phrases lues, Delais [45] indique un algorithme de détection des syllabes accentuées basé principalement sur les paramètres de fréquence fondamentale et de durée[9] : une montée graduelle de la *f0* sur la syllabe, un mouvement descendant de la courbe de *f0* accompagnée d'un allongement syllabique ou enfin la présence d'un maximum de *f0* avec un léger allongement syllabique sont les caractéristiques acoustiques de l'accentuation. Elle distingue alors deux types d'accents en fonction de leur rôle et de leur réalisation :

- l'accent régulateur rythmique (faible montée mélodique et allongement syllabique non significatif) qui assure une fonction rythmique et qu'elle rapproche des notions d'ictus (défini par Rossi [133, 134] comme l'un des trois accentèmes) ou encore d'accent secondaire [115],
- l'accent démarcatif (mouvement ample montant ou descendant de la fréquence fondamentale accompagné d'un allongement significatif de la syllabe porteuse) assurant une fonction linguistique que l'on pourrait rapprocher des intonèmes continuatifs (CT et ct) et conclusif (CC) définis par ROSSI [134] (la table 1.1 de la page 5 de l'introduction résume la description de ces intonèmes).

Nous n'avons pas souhaité dans cette étude — même si nous avons été tenté initialement par une telle approche — entretenir des distinctions de ce type qui présupposent, d'une part, d'avoir connaissance du rôle des accents distingués — ce qui pour nous est une prise de position trop forte — et, d'autre part, d'être capable de réaliser automatiquement la distinction à partir de traits acoustiques de la nature de l'accent. Delais reconnaît d'ailleurs qu'il y a des cas où il lui est difficile de décider du caractère rythmique ou démarcatif d'un accent. On trouvait déjà très tôt à ce sujet dans une étude de Lehiste [83, p. 233] la phrase suivante :

> "There is no one-to-one correspondance between stress and any single acoustic parameter. Thus, there is also no automatic way to identify stressed syllabes."

[8] Voir la récente étude de Astesano [2] pour une synthèse des principales positions quant à l'accentuation du français ainsi qu'une étude des distributions acoustiques (durée et *f0*) de différentes classes accentuelles dans des discours.

[9] L'auteur précise que le calcul de durée s'appuie principalement sur celle du noyau vocalique.

Les unités manipulées par le système seront donc les seules étiquettes prosodiques décrites lors du premier chapitre (pages 35 à 37) ; ce qui ne signifie pas pour autant que ProStat ne sera pas amené à s'appuyer sur des configurations paramétriques particulières (combinaison des étiquettes) lors de prises de décisions, mais simplement que ces dernières ne seront en aucune façon nommées ni décrites par notre système. Le rôle de ProStat est donc de fournir à son utilisateur un moyen statistique et visuel pour tenter de dégager certaines régularités organisationnelles qui pourraient être exploitées notamment dans le cadre de la reconnaissance de la parole. Son principe de fonctionnement est d'envisager toutes les combinaisons de contraintes qu'il peut déduire à partir d'un corpus d'apprentissage en offrant à chacune une mesure de l'adéquation aux paramètres prosodiques. Il s'est ensuite avéré que le système ProStat doté d'une métrique pouvait assumer un rôle prédictif directement utilisable en reconnaissance de la parole. Bien que réducteur dans les choix faits à l'occasion de cette première étude, notre système revendique donc les points suivants :

- permettre à un utilisateur de disposer rapidement d'un ensemble d'informations l'autorisant à étudier le comportement des paramètres prosodiques en des points (connus ou pas) de la structuration linguistique et/ou rythmique d'un énoncé,
- de proposer pour un énoncé donné — et à partir d'un apprentissage préalable — un ensemble d'hypothèses valuées utilisables en reconnaissance de la parole s'appuyant non pas sur des règles décrivant des points précis de la structuration mais au contraire sur une prise en compte de l'énoncé dans sa globalité,
- de vérifier des propositions issues de divers modules (agent lexical, agent syntaxique, *etc.*).

Nous allons dans la section suivante préciser les choix initiaux de notre système et décrire son fonctionnement. Nous en montrerons ensuite des utilisations possibles dans deux tâches de reconnaissance de la parole.

5.4 Description du système ProStat

5.4.1 Les entrées

Deux types de fichiers sont fournis en entrée du système ProStat : les fichiers descriptifs des énoncés que l'on désire employer lors de la phase d'apprentissage et les fichiers prosodiques associés qui contiennent sous forme textuelle les treillis prosodiques calculés par les méthodes décrites au chapitre 2. Pour chacun d'eux, un formalisme particulier est défini que nous rappelons brièvement[10] afin de préciser la nature exacte des informations fournies au système.

[10]La typographie employée dans la suite différencie les symboles en caractère **gras** qui sont les symboles terminaux du langage décrit des symboles en *italique* qui indiquent les non-terminaux. Par convention le symbole $\longrightarrow$ est utilisé comme symbole de ré-écriture, le symbole ";" marque la fin d'une règle, "." sépare les éléments successifs d'une énumération (facultatif) et le symbole "|" sépare les différents choix possibles d'une règle (commodité de notation).

Formalisme des fichiers descriptifs

description	$\longrightarrow$	;
description	$\longrightarrow$	*syntaxe* . *alignement* . *description* ;
description	$\longrightarrow$	*commentaires* . *description*;
syntaxe	$\longrightarrow$	**décomposition_grammaticale**(”*nom_de_fichier*”,*arbre*)->; ;
alignement	$\longrightarrow$	;
alignement	$\longrightarrow$	**étiquetage**(”*nom_de_fichier*”,*numéro_du_mot_dans_la_phrase*,*phonèmes*)->; . *alignement* ;
phonèmes	$\longrightarrow$	**nil** ;
phonèmes	$\longrightarrow$	<”*pho*”,<*début*,*fin*>>. . *phonèmes* ;
arbre	$\longrightarrow$	**nil** ;
arbre	$\longrightarrow$	*symbole*(*sous-arbre*) . *arbre* ;
sous-arbre	$\longrightarrow$	<”*mot*”,*numéro_du_mot_dans_la_phrase*> ;
sous-arbre	$\longrightarrow$	*arbre* ;

Nous ne décrivons pas les règles telles que *commentaire* ou *nom_de_fichier* qui sont suffisamment explicites de par leur nom. Voici un extrait d'un fichier descripteur qui vérifie la précédente grammaire et constitue donc une entrée valide du système ProStat.

```
Phrase: une grenouille saute sur les nénuphars
*******
décomposition_grammaticale("12111846.cmp.item22",
  PH( SS( GN( ART(<"une",1>).NC(<"grenouille",2>).ADJ(<"verte",3>).nil ).nil ).
      SV( VB(<"saute",4>).
          CIRC( PREP(<"sur",5>).
                GN( ART(<"les",6>).NC(<"nénuphars",7>).nil ).
                nil
              ).
          nil
        ).
      nil
    ).nil)->;


étiquetage("12111846.cmp.item22",1,<"uu",<31,43>>.<"nn",<43,45>>.nil)->;
étiquetage("12111846.cmp.item22",2,<"gg",<45,47>>.<"rr",<47,53>>.<"ee",<53,62>>.
           <"nn",<62,68>>.<"ou",<68,77>>.<"yy",<77,85>>.nil)->;
étiquetage("12111846.cmp.item22",3,<"vv",<85,89>>.<"ai",<89,100>>.<"rr",<100,111>>.
           <"tt",<111,127>>.nil)->;
étiquetage("12111846.cmp.item22",4,<"ss",<127,130>>.<"au",<130,147>>.
           <"tt",<147,153>>.nil)->;
étiquetage("12111846.cmp.item22",5,<"ss",<153,164>>.<"uu",<164,168>>.
           <"rr",<168,170>>.nil)->;
```

```
étiquetage("12111846.cmp.item22",6,<"ll",<170,174>>.<"ai",<174,180>>.nil)->;
étiquetage("12111846.cmp.item22",7,<"nn",<180,185>>.<"ei",<185,192>>.
           <"nn",<192,197>>.<"uu",<197,204>>.<"ff",<204,216>>.
           <"aa",<216,236>>.<"rr",<236,253>>.nil)->;
```

La figure 5.1 reprend cet exemple de descripteur sous une forme visuelle. Nous remarquons avec cet exemple, qu'une phrase (si l'on travaille sur des phrases) est décrite par son alignement temporel et par sa décomposition syntaxique sous forme arborescente. Tout autre niveau de structuration (sémantique, pragmatique, *etc.*) peut remplacer ou s'ajouter à la description précédente pour autant qu'il puisse être représenté par un arbre[11]. Cependant, dans la suite de cet exposé seules les décompositions syntaxiques seront fournies au système ProStat ce qui ne signifie nullement — loin s'en faut — que nous accordons à la sémantique (par exemple) un rôle secondaire ; mais si fournir une décomposition syntaxique pour des phrases de structures relativement simples ne nous semblait pas une tâche insurmontable, il n'en allait pas de même quant à la structuration sémantique surtout si l'on garde à l'esprit l'aspect automatisable du processus. Remarquons enfin qu'en guise d'étude rythmique, le système ProStat se propose d'analyser la longueur (en nombre de voyelles) des divers regroupements linguistiques analysés.

Formalisme des fichiers de treillis prosodiques

treillis ⟶ ;
treillis ⟶ <<< *début*,*fin*>,<"*pho*",>>,<*liste_valeurs*>> . *treillis*
liste_valeurs ⟶ **nil**;
liste_valeurs ⟶ *entier* . *liste_valeurs*;

Voici en complément de l'exemple précédent un extrait du fichier treillis accompagnant le fichier descripteur vérifiant le formalisme qui vient d'être énoncé.

```
< <<31,43>,<"AL1">>, <nil> >
< <<31,43>,<"-">>, <-6909.nil> >
< <<68,77>,<"NIVA4">>, <nil> >
< <<68,77>,<"EFO2">>, <nil> >
< <<68,77>,<"-">>, <-16286.nil> >
< <<89,100>,<"INFO_FO">>,<120.130.152.152.32.nil> >
< <<130,147>,<"NIVA4">>, <nil> >
< <<130,147>,<"AL1">>, <nil> >
< <<130,147>,<"EFO1'">>, <nil> >
< <<185,192>,<"+">>, <22417.nil> >
< <<197,204>,<"EFO1'">>, <nil> >
```

[11] Une définition tout à fait générale d'un arbre est ici retenue : un arbre est un graphe non orienté connexe sans boucle.

En résumé, une entrée valide de ProStat est une observation dotée de sa décomposition syntaxique, de son alignement temporel ainsi que de sa caractérisation prosodique.

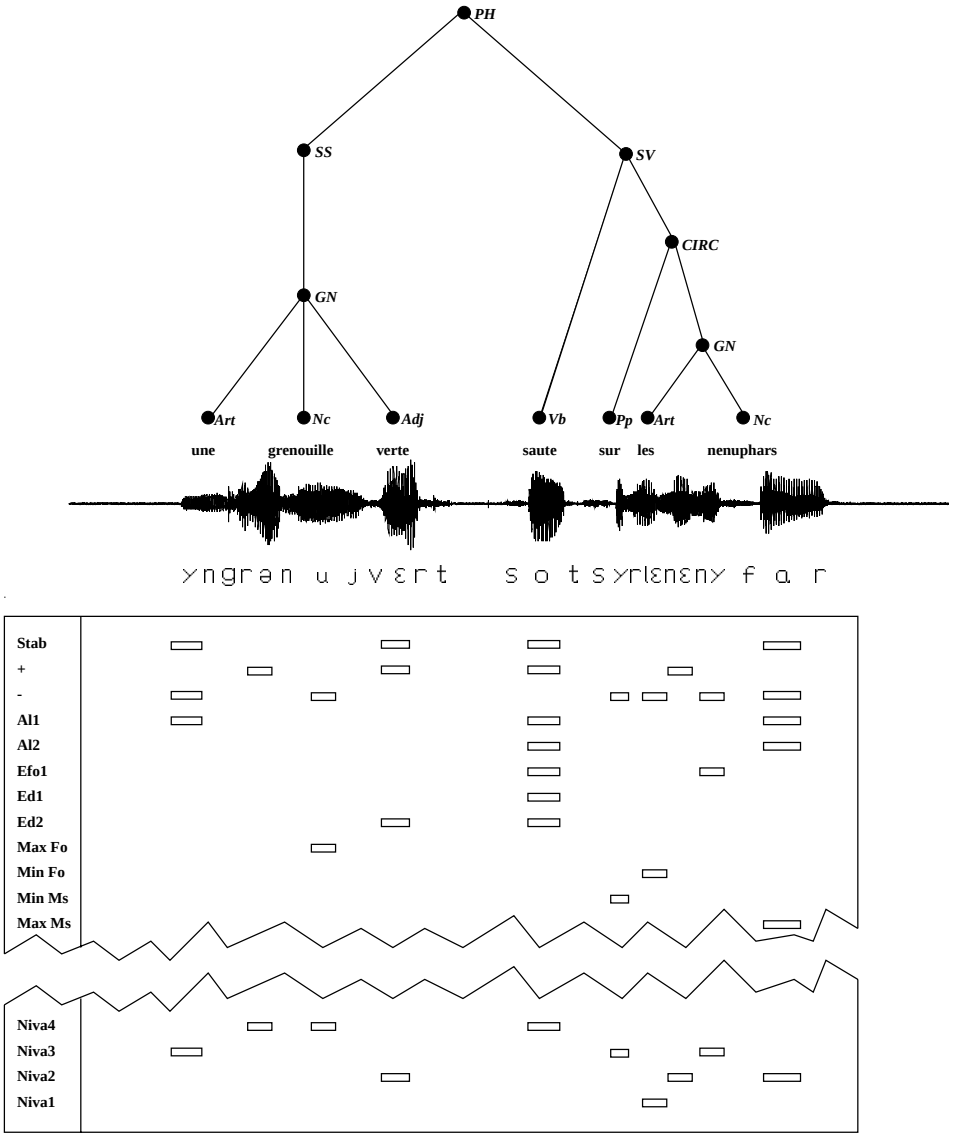


Figure 5.1: Représentation d'une entrée fournie au système ProStat.

5.4.2 L'apprentissage

Nous venons de présenter les entrées du système, nous allons maintenant décrire de quelle façon elles sont utilisées. La structure interne de données de ProStat peut être définie comme un graphe orienté non connexe. À chaque arc est associée une contrainte structurelle (que ce soit à un niveau syntaxique, rythmique ou autre) tenant compte du chemin déjà parcouru dans le graphe. Ainsi, plus on progresse dans le graphe, plus l'information contenue dans un nœud est spécifique.

Informations contenues dans un graphe de données de ProStat

Un nœud N du graphe est appelé un *P-nœud* ; il contient les informations suivantes :

- le nombre d'observations passées par le nœud considéré (qui décroît plus on avance dans le graphe),
- le décompte des diverses étiquettes prosodiques stockées dans le nœud,
- la structure syntaxico-rythmique $Crit(N)$ — également désignée *critère* ou *SR-structure* dans la suite — modélisée par le nœud.

Une SR-structure est représentée par un arbre (A) dont les nœuds n — dénommés *SR-nœuds* — sont définis par :

- un symbole $Symb(n)$ appartenant à l'ensemble des symboles définis par l'utilisateur,
- par $Voy(n)$, le nombre de voyelles qu'il modélise (qui peut être instancié ou non),
- ainsi que les SR-nœuds fils n_i ; le nombre de fils de n étant noté $Nbf(n)$.

On désigne par $Deg(A)$ le degré d'instanciation de A défini par le nombre de niveaux consécutifs (en partant de la racine) où tous les SR-noeuds sont instanciés. La profondeur de l'arborescence A est notée $Prof(A)$ dans la suite. Dans l'exemple de la figure 5.3, l'entrée représentée est un critère de profondeur 4 et de degré d'instanciation 4. Notons également que par définition, une observation d'apprentissage est une SR-structure de degré d'instanciation maximal égal à la profondeur de l'arborescence de cette observation. À chaque feuille de A sont associées les étiquettes prosodiques des observations unifiables avec A ; seules les étiquettes localisées sur les voyelles initiales et finales du groupe décrit par chaque feuille sont actuellement prises en compte. Les observations unifiables à un critère A sont toutes celles pour lesquelles A est un critère géniteur : on dit que A est géniteur de B (ce que nous noterons $A \Re B$) ssi :

$$\left\{ \begin{array}{rcl} Symb(A) & = & Symb(B)\,, \\ Voy(A) = Voy(B) & ou & Voy(A)\, libre\,, \\ \forall i \in [1, Nbf(A)], & si & A_i \exists,\, alors \left\{ \begin{array}{l} B_i \exists \\ A_i \Re B_i \end{array} \right. \end{array} \right.$$

La figure 5.2 propose deux exemples de cette relation.

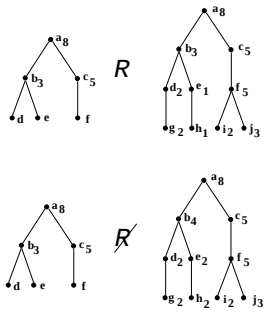


Figure 5.2: Illustration de la relation $\Re$. (a,b,d,e,f,g,h,i,j) appartiennent à l'ensemble des symboles *utilisateur*. Les indices associés à ces symboles correspondent au nombre de voyelles du SR-nœud décrit.

Accroissement du graphe

Notons $C_{i,j}$ le critère géniteur de C de degré d'instanciation j et de profondeur i.
L'ensemble ∂_C des critères géniteurs d'un critère C est alors défini par :
$\partial_C = \{C_{i,j} \,/\, i \,et\, j \in [0, Prof(C)] \,avec\, j \leq i\}$
L'accroissement du graphe ProStat se fait en parcourant, pour chaque observation O (de pronfondeur $Prof(O) = p_o$) du corpus d'apprentissage, les P-nœuds du graphe dont la structure syntaxico-rythmique appartient à l'ensemble des géniteurs de O (∂_O). Chaque P-nœud est alors mis à jour avec les informations prosodiques fournies en entrée et ceux qui sont absents de ProStat sont alors insérés dans le graphe. Le parcours suit l'algorithme suivant (où N est un P-nœud du graphe, O l'observation, p la profondeur d'analyse et j le degré d'instanciation) :

$parcours(N, O, p, j)$
  $si\ (p < p_o)$
    $si\ \nexists N'$: P-nœud $/\ Crit(N') = O_{p+1,j}$
      $alors$ créer N' fils de N
      $sinon$ mettre à jour N'
    $parcours(N', O, p+1, j)$
  $si\ (j < p_o)$
    $si\ \nexists N'$: P-nœud $/\ Crit(N') = O_{p,j}$
      $alors$ créer N' fils de N
      $sinon$ mettre à jour N'
    $parcours(N', O, p, j+1)$

Le parcours est lancé pour une observation O par : $parcours(racine, O, 1, 0)$. Ainsi, une observation de profondeur p peut générer jusqu'à $\frac{p \times (p+3)}{2}$ P-nœuds. En pratique, le nombre de recouvrements (dépendant de l'application) permet de freiner la croissance du graphe.

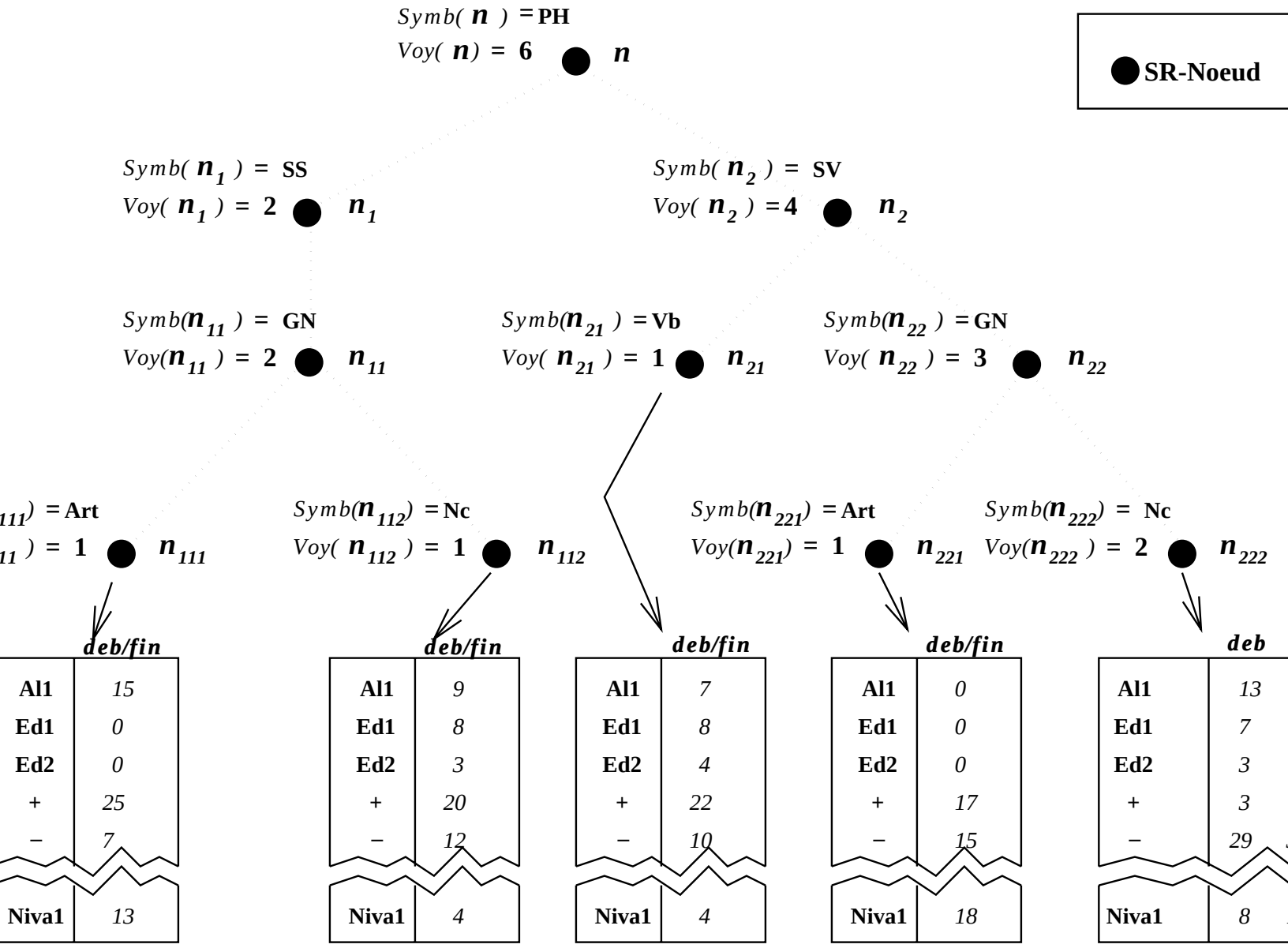


Figure 5.3: Exemple d'une SR-structure de profondeur et de degré d'instanciation 4. À chaque feuille de la structure sont associés les indices prosodiques localisés sur les voyelles initiale et finale.

Les figures 5.10 et 5.21 (pages 161 et 178) montrent — pour deux applications particulières — l'évolution du nombre de P-nœuds en fonction du nombre d'observations présentées au système.
La figure 5.5 représente l'état du graphe ProStat après apprentissage des observations A et B dont les géniteurs sont détaillés en table 5.1. On constate sur cet exemple la mise en facteur de nombreux P-nœuds.

i/j	1	2	3	4
0	Ph	Ph(Ss.Sv)	Ph(Ss(Gn).Sv(Vb.Gn))	Ph(Ss(Gn(Art.Nc.Adj)).Sv(Vb.Gn(Art.Nc)))
1	Ph(9)	Ph(Ss.Sv,9)	Ph(Ss(Gn).Sv(Vb.Gn),9)	Ph(Ss(Gn(Art.Nc.Adj)).Sv(Vb.Gn(Art.Nc)),9)
2		Ph(Ss(5).Sv(4),9)	Ph(Ss(Gn,5).Sv(Vb.Gn,4),9)	Ph(Ss(Gn(Art.Nc.Adj),5).Sv(Vb.Gn(Art.Nc),4),9)
3			Ph(Ss(Gn(5),5).Sv(Vb(2).Gn(2),4),9)	Ph(Ss(Gn(Art.Nc.Adj,5),5).Sv(Vb(2).Gn(Art.Nc,2),4),9)
4				Ph(Ss(Gn(Art(1).Nc(2).Adj(2),5),5).Sv(Vb(2).Gn(Art(1).Nc(1),2),4),9)

i/j	1	2	3	4
0	Ph	Ph(Ss.Sv)	Ph(Ss(Gn).Sv(Vb.Gn))	Ph(Ss(Gn(Art.Nc)).Sv(Vb.Gn(Art.Adj.Nc)))
1	Ph(9)	Ph(Ss.Sv,9)	Ph(Ss(Gn).Sv(Vb.Gn),9)	Ph(Ss(Gn(Art.Nc)).Sv(Vb.Gn(Art.Adj.Nc)),9)
2		Ph(Ss(5).Sv(4),9)	Ph(Ss(Gn,5).Sv(Vb.Gn,4),9)	Ph(Ss(Gn(Art.Nc),5).Sv(Vb.Gn(Art.Adj.Nc),4),9)
3			Ph(Ss(Gn(5),5).Sv(Vb(1).Gn(3),4),9)	Ph(Ss(Gn(Art.Nc,5),5).Sv(Vb(1).Gn(Art.Adj.Nc,3),4),9)
4				Ph(Ss(Gn(Art(2).Nc(3),5),5).Sv(Vb(1).Gn(Art(1).Adj(1).Nc(1),3),4),9)

Table 5.1: Description de l'ensemble des géniteurs des observations A ($\equiv A_{4,4}$) et B ($\equiv B_{4,4}$) respectivement. j indique la profondeur et i le degré d'instanciation.

Notons dès à présent que les arcs qui partent d'un même nœud ne sont pas nécessairement exclusifs.

5.4.3 Les sorties

Comme nous l'avons déjà précisé, ProStat est avant tout un système destiné à étudier de manière conviviale les corrélations entre divers niveaux de représentation linguistique. Il est donc doté d'une interface permettant à un utilisateur d'accéder rapidement à une information particulière et de disposer d'une vue synthétique des corrélations les plus marquées. Un opérateur peut consulter différentes informations concernant une SR-structure particulière que nous décrivons maintenant.

- Il peut tout d'abord visualiser les contours des divers paramètres acoustiques qui correspondent à la SR-structure donnée en entrée. La figure 5.4 montre un exemple d'affichage fourni par ProStat en réponse à deux requêtes. Les contours sont obtenus le plus finement à partir des valeurs initiales, médianes et finales du paramètre pour les voyelles de début et de fin de groupe. Il est bien évidemment possible de ne considérer qu'une partie de cette information afin d'obtenir des contours encore plus stylisés. Nous nous empressons de préciser que cette stylisation pour le moins grossière n'est le fruit d'aucun test perceptif. Nous avons déjà dans l'introduction mentionné les difficultés d'une telle entreprise de modélisation des paramètres prosodiques. S'il existe un grand nombre de méthodes de stylisation différentes, on peut cependant rappeler celle développée par Hirst [66, 65] qui présente un avantage important à nos yeux qui est d'être automatisable et la

fameuse, et non moins fastidieuse, "méthode IPO" d'équivalence perceptive [147] qui inspire la nouvelle méthode proposée par d'Alessandro et Al. [43] ainsi que la toute récente étude de Bosch [149] sur la classification automatique des mouvements du fondamental. Récemment Tournemire [44] proposait une étude perceptive sur la stylisation extrême (*i.e.* maximale) des courbes mélodiques de phrases isolées. Nous reportons brièvement quelques points de conclusion de ce travail qui nous permettent de penser que les contours que nous modélisons pourraient être utilisés à des fins de synthèse :

  - la présence ou l'absence de microprosodie des consonnes voisées n'est pas perceptible par un auditeur naïf,
  - la stylisation par le milieu des voyelles est la moins dégradée,
  - la stylisation par mot (3 valeurs fixes de *f0* par mot) semble tout à fait acceptable comme approche initiale.

- L'utilisateur dispose également d'une matrice qui lui permet d'apprécier rapidement la caractérisation prosodique d'une SR-structure donnée. La table 5.2 est un exemple de matrice fournie par ProStat. Dans la suite de l'exposé nous emploierons la variable Nbi pour désigner le nombre d'indices prosodiques différents calculés (nous rappelons qu'actuellement nous prenons en considération 39 indices prosodiques dont le descriptif est donné dans la section 2.2 pages 35 à 37).

- Enfin, le système ProStat peut dresser à la demande de l'utilisateur un bilan synthétique des points de rendez-vous (entre les indices prosodiques et l'organisation syntaxico-rythmique des observations) les plus marqués ; l'utilisateur ayant alors à sa charge de fournir au système des paramètres globaux comme par exemple le nombre minimal d'observations devant être consultées pour une éventuelle prise en compte de l'information. Remarquons simplement que cet outil d'analyse n'est en aucun cas un "extracteur"automatique de règles mais simplement une aide supplémentaire offerte à l'utilisateur. Voici à titre indicatif un extrait de synthèse fournie par le système (pour une application sur des nombres décimaux ici) à la demande de l'utilisateur :

```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
<ED2> : 83.3% d'être sur la voyelle de la feuille 3
<EEN1> : 75.0% d'être sur la voyelle de la feuille 6
<MAX_MS> : 60.0% d'être sur la voyelle de la feuille 7

OBS 30:NB(N1000_999999(MOT().MILLE().MOT()).VIRG().N100_999(MOT().
```

```
CENT().MOT()),8)
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
<ED2> : 72.5% d'être en dernière position de la feuille 1
<ENIVR2> : 65.1% d'être en première position de la feuille 3
<PAUSE> : 68.5% d'être en dernière position de la feuille 3

OBS 84:NB(N1000_999999(3).VIRG(2).N100_999(3),8)
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
```

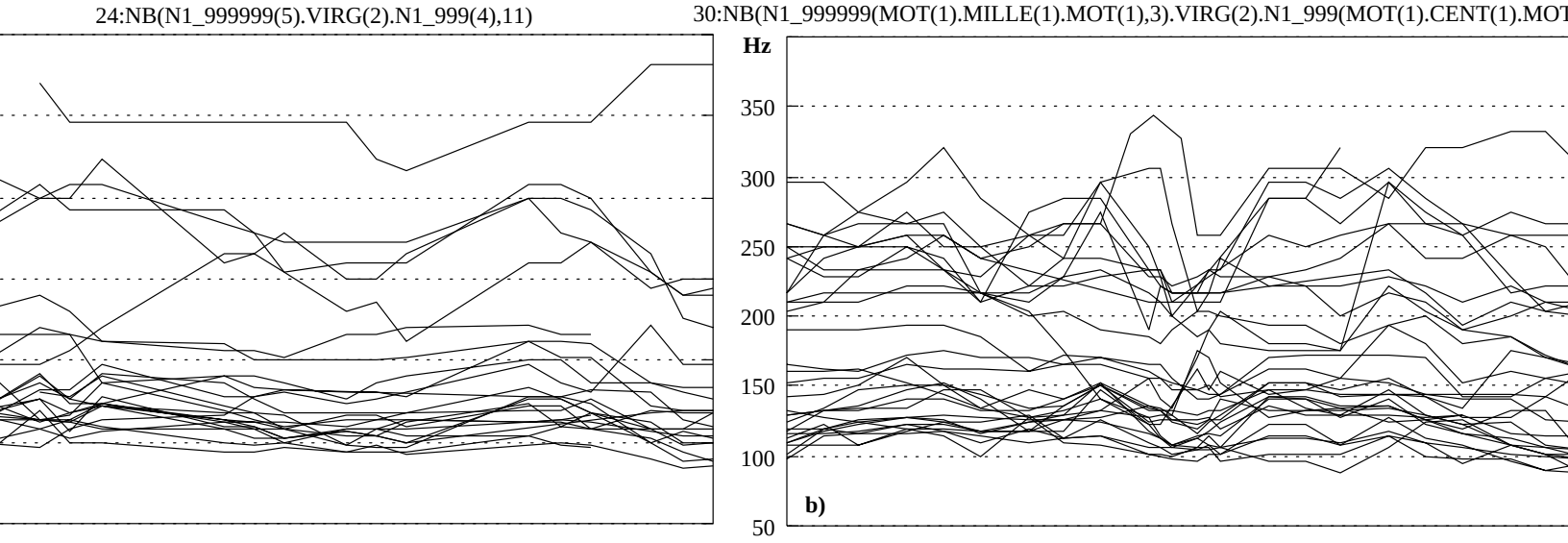


Figure 5.4: Contours du paramètre *f0* modélisés pour les nombres vérifiant chacun une contrainte structurelle particulière. Les contours présentés ici ont été obtenus à partir des valeurs du paramètre pris sur trois points (début, milieu et fin) des voyelles initiale et finale de chaque groupe terminal de la structure considérée. Ces valeurs sont ici simplement reliées par des segments de droite sans aucune prise en compte de la durée de chaque groupe.

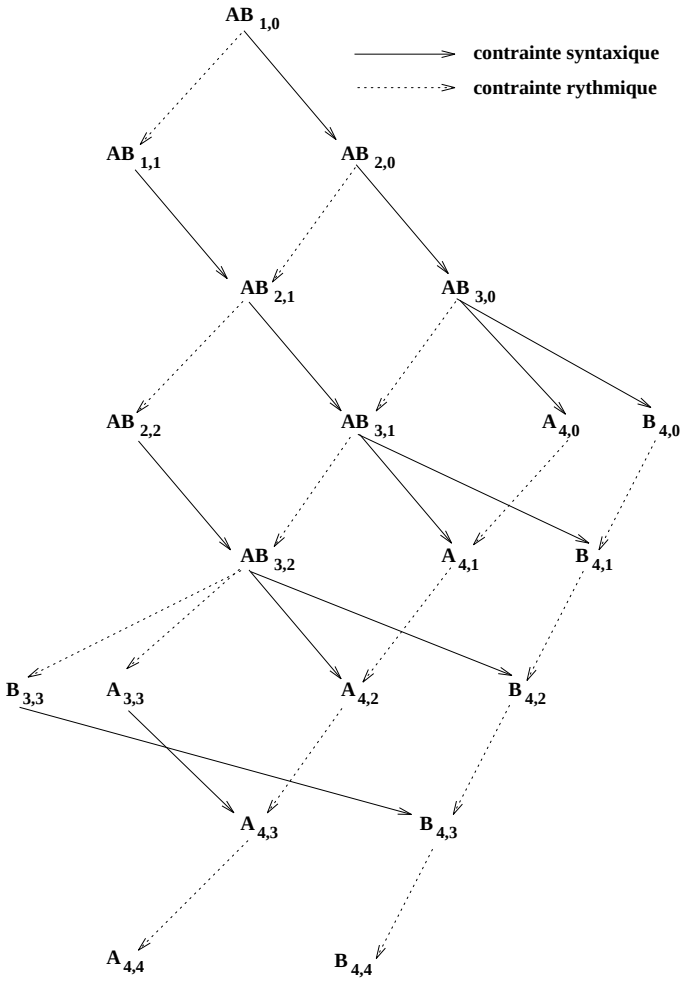


Figure 5.5: État du graphe ProStat après apprentissage des observations A et B présentées en table 5.1; les P-nœuds sont ici symbolisés par la SR-structure qu'ils contiennent.

5.4.4 Reconnaissance

Nous venons de faire l'inventaire des sorties proposées par le système ProStat, nous allons maintenant montrer comment l'information prosodique peut être utilisée à des fins purement applicatives.
Nous formulons l'hypothèse classique que les indices prosodiques — pour autant qu'ils soient pertinents — ne sont pas distribués de manière aléatoire, mais se réalisent bien au contraire en des endroits clés du message qui correspondent le plus souvent à des points particuliers des différents niveaux de structuration linguistique. Il semble donc tout naturel d'utiliser la "régularité" de ces rendez-vous pour proposer un faisceau d'hypothèses valuées sur tout ou partie de la structuration du message à reconnaître. Nous allons par la suite décrire de quelle façon nous utilisons l'information contenue dans le graphe de données de ProStat pour proposer un ensemble d'hypothèses valuées.
Comme nous l'avons déjà vu, chaque P-nœud du graphe mémorise une structure syntaxico-rythmique dont chaque feuille contient les étiquettes prosodiques rencontrées en leur début et fin ; et, plus profond est le P-nœud dans le graphe, plus précise est l'information qu'il contient. Les arcs partant d'un P-nœud donné ne sont pas exclusifs, aussi — mais également pour des raisons de simplicité — n'utiliserons-nous pas par la suite de techniques usuelles de classification[12]. La méthode que nous retenons est simplement fondée sur le calcul d'une distance entre deux matrices : une matrice d'observation et la matrice dérivée du critère associé au P-nœud en cours d'évaluation.

Matrice associée à un P-nœud du graphe

Elle est obtenue par un parcours *en profondeur d'abord* de la SR-structure associée ; sa largeur l_p est donc au plus de $2f$ (une valeur à l'initiale et en finale de groupe), f étant le nombre de feuilles de la SR-structure (dans l'exemple reporté en table 5.2, la SR-structure contient 7 feuilles et la largeur de la matrice associée est 8). Chacune des lignes de la matrice est alors normalisée de manière à ne considérer non plus de simples comptages d'indices prosodiques, mais leur probabilité d'occurrence :

$$M_p^{'}(i,j) \rightarrow \frac{M_p(i,j)}{\sum_{t=1}^{l_p} M_p(i,t)} \text{ , avec } j \in [1,l_p] \text{ et } i \in [1,Nbi]$$

Nbi désigne le nombre d'indices prosodiques (actuellement égal à 39) pris en considération.

Matrice d'observation

La matrice d'observation est directement mesurée depuis le signal de parole et sa largeur l_o correspond au nombre de voyelles contenues dans l'échantillon de parole analysé.

[12]Pour tous les problèmes liés à l'apprentissage des structures prosodiques, nous renvoyons le lecteur aux travaux plus spécifiques de Schneider [138]

Indice	Nb.	MOT DF	MILLE DF	MOT DF	VIRG D	VIRG F	MOT DF	CENT DF	MOT DF
NIVA1	63	5	5	6	11	9	3	6	18
NIVA2	45	4	2	3	10	9	5	9	3
NIVA3	77	15	8	9	8	9	15	10	3
NIVA4	47	6	14	11	1	1	6	4	4
+	75	19	4	22	3	13	5	4	5
-	147	11	25	5	25	14	23	25	19
=	12	0	1	3	2	1	1	0	4
ED1	47	0	7	21	4	9	0	6	0
ED2	30	0	0	25	4	1	0	0	0
EFO1	51	0	12	11	3	5	16	4	0
EFO1'	55	0	14	9	3	6	17	6	0
EFO2	33	0	0	16	1	1	15	0	0
EFO2'	36	0	0	15	2	1	18	0	0
ENIVA1	24	0	5	4	1	2	9	3	0
ENIVA2	22	0	0	10	1	1	10	0	0
ENIVR1	16	0	3	1	2	1	7	2	0
ENIVR2	14	0	0	4	1	1	8	0	0
EERO1	59	0	2	16	12	2	16	11	0
EERO1'	50	0	1	19	7	1	12	10	0
EERO2	37	0	0	9	18	2	8	0	0
EERO2'	34	0	0	14	10	2	8	0	0
EEN1	4	0	0	0	0	0	1	3	0
EEN2	5	0	0	1	2	0	2	0	0
MIN_FO	30	6	1	0	1	2	2	1	17
MAX_FO	30	2	8	10	1	1	2	4	2
MIN_MS	30	10	1	1	3	8	6	0	1
MAX_MS	30	0	1	8	0	1	0	2	18
MAX_ERO	30	16	0	7	5	1	1	0	0
MIN_ERO	30	3	0	1	0	1	5	5	15
STAB	128	12	29	23	12	10	0	13	29
PAUSE	18	0	0	5	0	0	1	0	12
AL1	86	3	14	23	4	1	0	16	25
AL2	64	0	12	18	2	1	0	7	24
VO	240	30	30	30	30	30	30	30	30

Table 5.2: Matrice des indices prosodiques relevés sur les nombres dont la structure est unifiable à la contrainte : NB(N1-999999(MOT(1).MILLE(1).MOT(1),3).VIRG(2).N1-999(MOT(1).CENT(1).MOT(1),3),8). D indique la voyelle initiale de groupe, F la voyelle finale.

Opération de réduction

La réduction d'une matrice par un critère (de f feuilles) donné consiste à déterminer un ensemble de f couples de colonnes (deb_i, fin_i) qui correspondent à des frontières possibles — en termes de voyelles — des constituants linguistiques modélisés par les feuilles de la SR-structure considérée. Une réduction est une opération qui doit nécessairement satisfaire les contraintes rythmiques imposées par la SR-structure et vérifier l'ensemble de règles suivant :

$$\forall\, i \in [1, f] \quad \begin{cases} deb_{i+1} & = & fin_i + 1 \\ fin_i & \geq & deb_i \\ deb_1 & = & 1 \\ fin_{l_p} & = & l_o \end{cases}$$

La figure 5.6 illustre cette opération simple pour une SR-structure donnée. Il est bien évident que certains découpages sont peu probables bien qu'envisagés ; nous considérons qu'il est du ressort de la métrique retenue de les éliminer ou de les déclasser au cours du processus de reconnaissance. La matrice résultante est appelée matrice *réduite*.

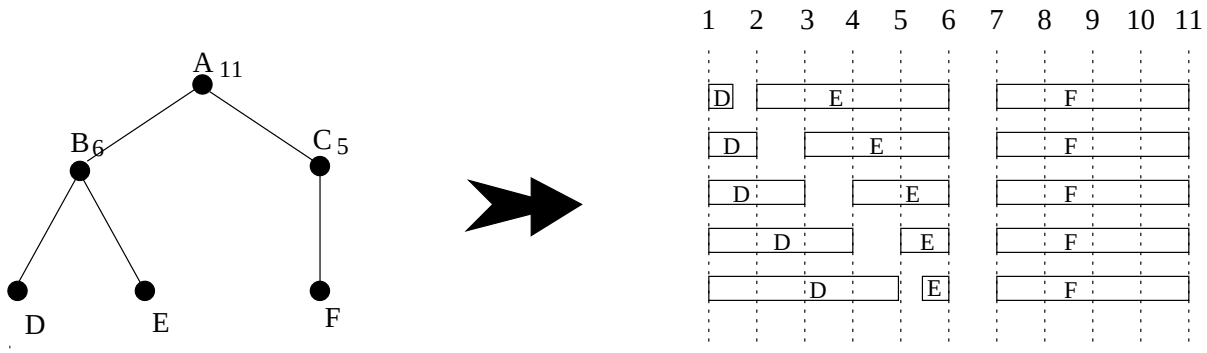


Figure 5.6: Illustration de l'opération de réduction. Ici 5 découpages sont envisagés pour la SR-structure décrite : les lettres A,B,C,D et E symbolisent des entités définies par l'utilisateur ; lorsqu'un nombre les accompagne, cela signifie que le nombre de voyelles de l'entité décrite est fixé à cette valeur (la SR-structure est ici de profondeur 3 et de degré d'instanciation 2).

Distance entre deux matrices

Après réduction de la matrice d'observations, nous sommes donc à même d'appliquer une mesure de ressemblance entre la matrice déduite de la SR-structure $M_p^{'}$ et chacune des matrices réduites M_o, toutes étant de même largeur l_p.

$$\begin{cases} d(M_o, M_p^{'}) & = \dfrac{\sum_{i=1}^{Nbi}\left(\alpha_i \sum_{j=1}^{lp} \delta_{ij}\right)}{Nbi} \\ \text{avec } \delta_{ij} & = \begin{cases} M_p^{'}(i,j) & si \quad M_o(i,j) = 1 \\ 0 & sinon \end{cases} \\ \text{et } \alpha_i & \text{coefficients de pondération} \end{cases}$$

Notation

Par cette mesure de similarité,[13] nous disposons d'un moyen de proposer et de noter, pour une observation donnée, un ensemble d'hypothèses syntaxiques et/ou rythmiques. Une première méthode simple qui consiste à classer les scores des différents P-nœuds considérés peut déjà s'avérer efficace si un ensemble représentatif d'exemples a été fourni lors de la phase d'apprentissage. Dans le cas contraire, les décisions peuvent être prises sur trop peu d'informations (cas de P-nœuds suffisamment profonds dans le graphe par exemple) ou en trop peu de points (essentiellement pour les P-nœuds proches de la racine). Les corpus d'apprentissage que nous employons dans nos expériences n'étant pas nécessairement de taille importante, nous adoptons une stratégie plus adéquate qui prend en compte non seulement l'information d'un P-nœud donné, mais également celle contenue dans l'ensemble des P-nœuds activés entre la racine et lui. Ainsi la note finale associée à un P-nœud est la note maximale obtenue pour tout chemin (de la racine jusqu'à ce P-nœud) dans le graphe de ProStat. La note d'un chemin donné étant simplement la moyenne des mesures de similarité des différents P-nœuds du chemin :

Soit $C_p : \{(n_1, ..., n_n)$P-nœuds $/ n_1 \equiv racine, n_n \equiv P$ et $\forall i \in [2, n], n_i \in Fils(n_{i-1})\}$
l'ensemble des chemins de la *racine* au nœud P,
Alors $Note(P) = \max_{C_p} \frac{\sum_{i=1}^{n} d(n_i)}{n}$
Où $d(n_i)$ est la mesure de similarité entre l'observation réduite et $Crit(n_i)$

5.5 Utilisation en reconnaissance de la parole

Maintenant que viennent d'être exposés les principes du système ProStat, nous allons nous attacher à montrer son utilité dans deux tâches de reconnaissance de la parole. Il s'agit ici davantage de démontrer que la prise en compte d'informations prosodiques dans un traitement automatique peut lui être bénéfique, plutôt que de présenter un système performant aux possibilités figées. Nous avons par exemple déjà précisé que tous les indices

[13] Les coefficients de pondération α_i sont actuellement tous égaux (à l'unité) ce qui signifie que chaque indice prosodique contribue à part égale à l'attribution de la note d'un P-nœud.

prosodiques considérés ici ne sont pas forcément tous pertinents, de même que certains autres pourraient l'être davantage (comme par exemple le passage par la "baseline" exposé par Vaissière [158]). Ceci étant précisé, nous pouvons décrire les expériences que nous avons réalisées dans ce but.

5.5.1 Tâche 1 : les nombres

Dans cette première tâche, nous allons nous intéresser à la base PolyNombre qui rappelons le est constituée de nombres décimaux prononcés via un canal téléphonique par près de 50 locuteurs, items sélectionnés par des critères de clareté (pas d'incident particulier de prononciation) et de disponibilité (appels ayant déjà fait l'objet d'une transcription orthographique). Le choix de cette application nous a paru répondre à un premier critère de simplicité (il est effectivement assez rapide de mettre en place un système de reconnaissance de nombres) puis à un second plus spécifique à la nature limitée et relativement régulière des faits prosodiques observés sur les nombres ; ceci devant nous permettre de tester la validité du système ProStat.

Les entrées

Dans cette première expérience à caractère exploratoire, nous avons réalisé un apprentissage à partir des 500 nombres de la base PolyNombre. Les entrées du système (alignement phonétique et arbre syntaxique) sont automatiquement obtenus depuis la seule annotation orthographique accompagnant chaque item de la base :

- Un réseau de reconnaissance phonétique est généré qui prend en compte les prononciations les plus courantes des nombres. Nous laissons le soin à un algorithme de Viterbi de sélectionner la prononciation la plus probable qui est celle dont l'alignement avec le signal paramétrisé a obtenu le plus haut score. Voici à titre indicatif un exemple de réseau automatiquement généré pour le nombre 5218,917[14] :

```
( [sil]
  ss in [kk [ee | sil]]
  mm ii ll [ee] [sil]
  dd eu [sil]
  ss an [sil]
  dd ii (zz | (ss [ee | sil]))
  uy ii tt [ee] [sil]
  vv ii rr gg uu ll [ee] [sil]
  nn oe ff [ee] [sil]
  ss an [sil]
```

[14]Le formalisme de HTK est ici employé [168] : [x] indique le caractère optionnel de x (*i.e.* 0 ou une fois), | sépare les différentes alternatives, (y) délimite un élément particulier du réseau (associativité) et enfin les symboles phonétiques reportés sont ceux réellement employés dans l'application.

```
dd ii [[ss [ee]] sil]
ss ai tt [ee]
[sil] )
```

- Une grammaire des nombres permet de disposer automatiquement de l'arbre grammatical d'une observation. Précisons que les choix faits à ce niveau conditionnent fortement les corrélations que l'on désire mettre en évidence et exploiter ultérieurement. Ainsi la grammaire que nous utilisons dans cette tâche et que nous reportons plus bas est-elle le reflet de quelques a priori. Nous différencions par exemple parmi le vocabulaire constitutif des nombres, certains mots (comme *cent*, *mille*, *virgule*) que nous suspectons de porter plus que les autres des marques prosodiques. Notons cependant que si cela s'avérait faux, la distinction faite serait alors caduque et ne gênerait en rien les traitements ultérieurs. De même, la structure arborescente choisie et les symboles retenus sont discutables, aussi nous décrivons les choix faits à ce niveau dans les figures 5.7 et 5.8. La figure 5.9 présente un arbre obtenu à l'aide de cette grammaire. Les graphes étant d'une lecture aisée, nous formulerons seulement quelques commentaires succincts sur cette grammaire. En tout premier lieu, il convient de remarquer que les groupements de mots sont tout à fait arbitraires et ne répondent pas nécessairement à des contraintes d'organisation prosodique. Rien ne garantit de plus que le vocabulaire du méta-langage descriptif (MOT,CENT,MILLE,VIRGULE,N100-999,N1000-999999) soit pertinent et/ou suffisant. Rien ne s'oppose en cas d'échec à la modification de tout ou partie de la grammaire des nombres afin de mettre au mieux en évidence les corrélations "syntaxico-prosodiques".

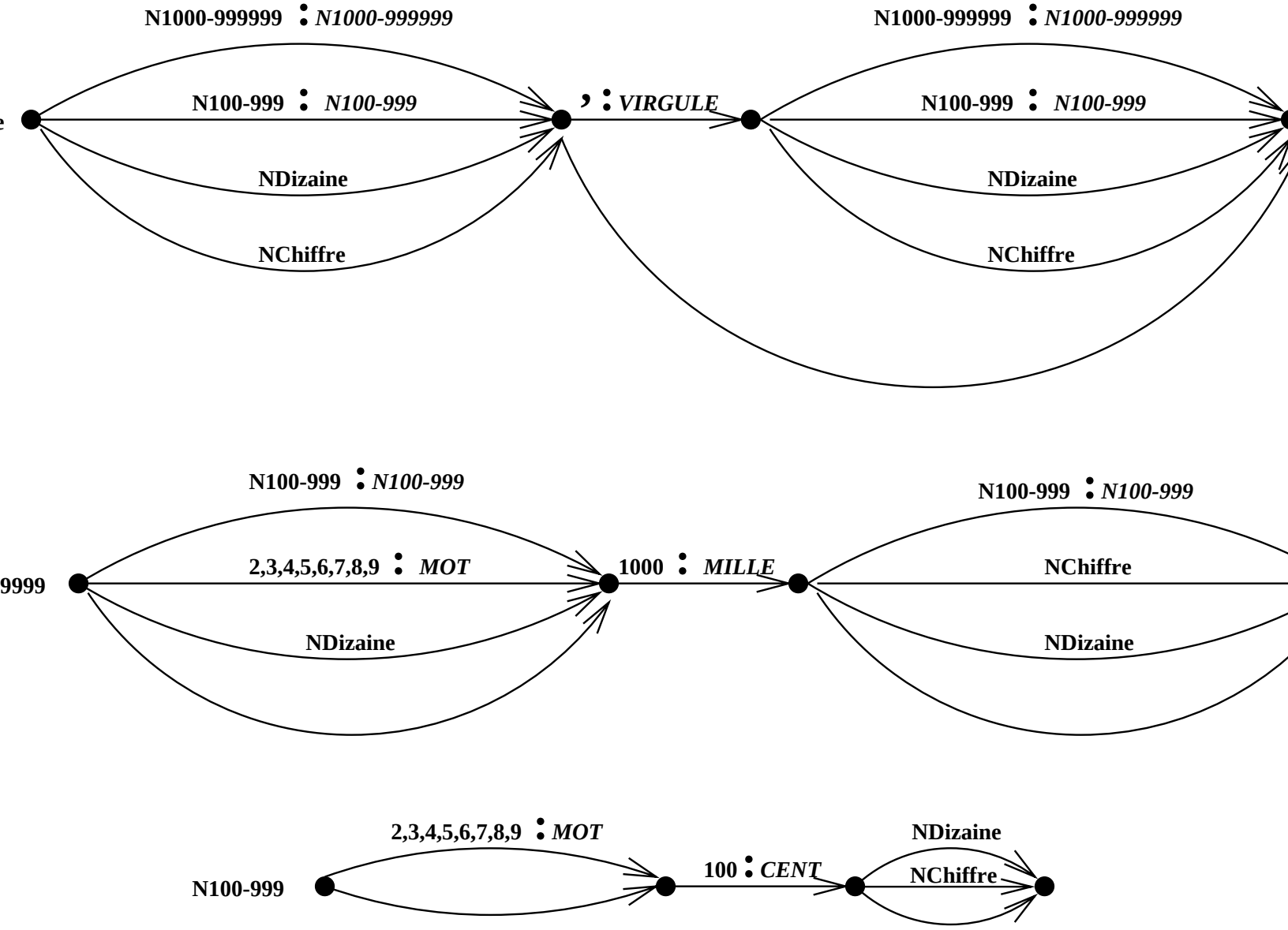
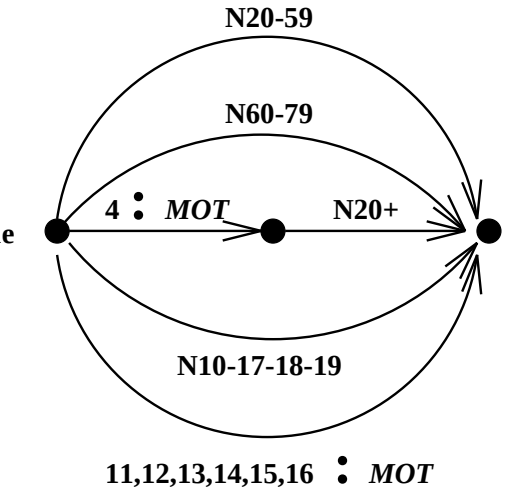
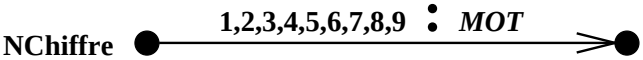
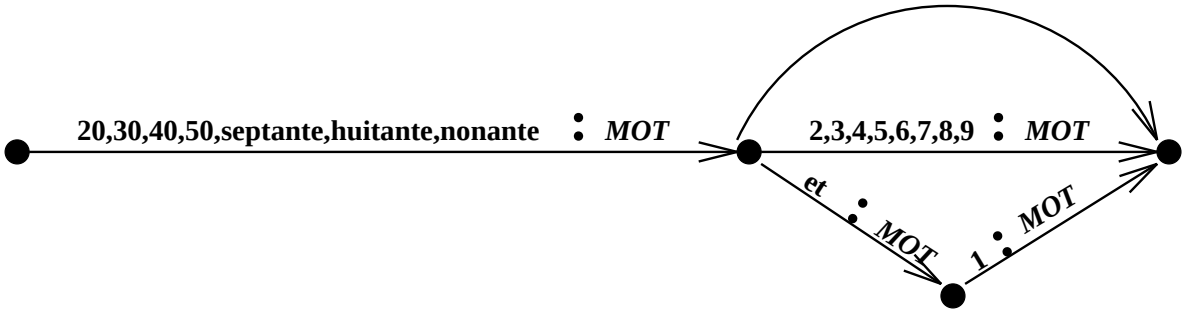
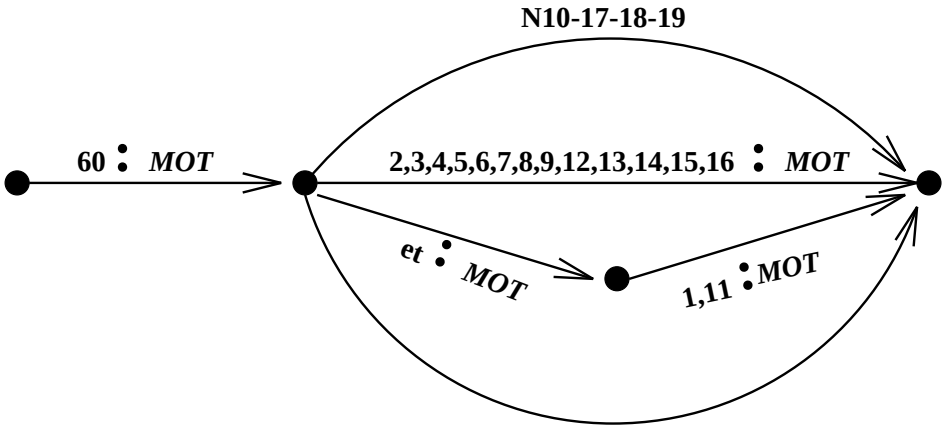


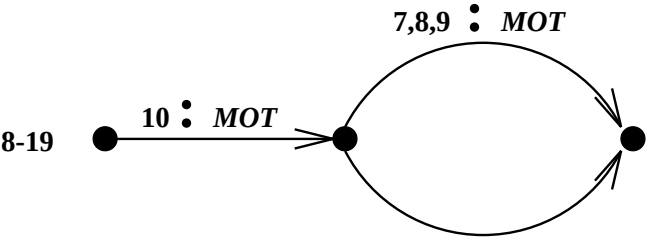
Figure 5.7: Représentation de la grammaire des nombres utilisée. Pour des raisons de lisibilité, cette représentation n'est pas LL1, bien qu'étant codée sous forme LL1 dans l'application. La symbolique utilisée ici est la suivante : chaque arc d'un automate est une règle dont les symboles associés sont soit les mots effacés par la règle, soit une autre tête de règle ; si $:x$ suit une liste de symboles, il y a émission du symbole syntaxique x.

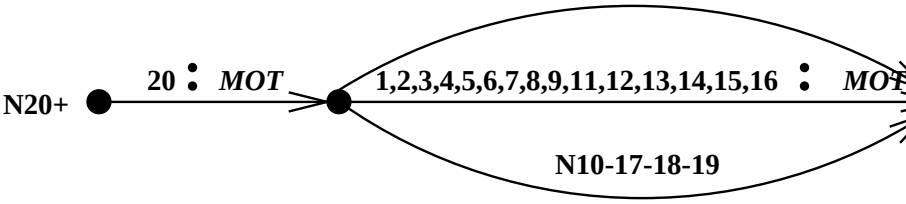

N20-59
N60-79
4 : MOT
N20+
N10-17-18-19
11,12,13,14,15,16 : MOT


NChiffre
1,2,3,4,5,6,7,8,9 : MOT


20,30,40,50,septante,huitante,nonante : MOT
2,3,4,5,6,7,8,9 : MOT
et : MOT
1 : MOT


N10-17-18-19
60 : MOT
2,3,4,5,6,7,8,9,12,13,14,15,16 : MOT
et : MOT
1,11 : MOT


7,8,9 : MOT
10 : MOT


N20+
20 : MOT
1,2,3,4,5,6,7,8,9,11,12,13,14,15,16 : MOT
N10-17-18-19

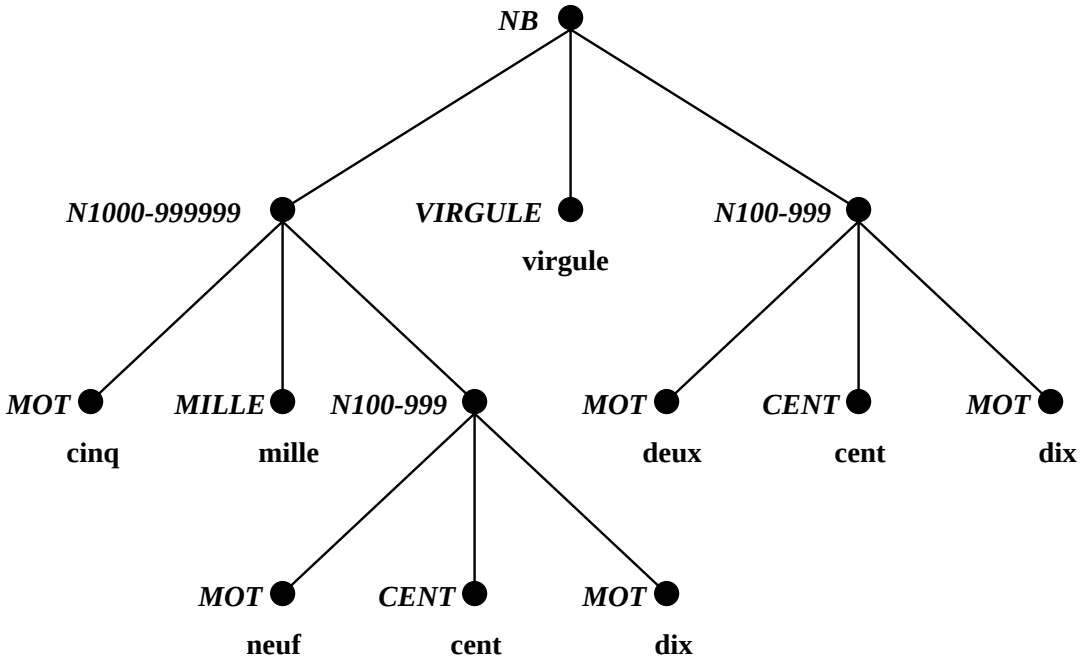


Figure 5.9: Arbre grammatical obtenu pour le nombre 5910,210.

L'apprentissage

Nous reportons quelques renseignements liés à la phase d'apprentissage des 500 nombres de la base PolyNombre ; en particulier le nombre de P-nœuds créés et le nombre de feuilles différentes (figure 5.10) ainsi que le décompte des diverses étiquettes prosodiques prises actuellement en compte au sein du système ProStat (table 5.3). Ces données appellent quelques commentaires sommaires :

- On constate en tout premier lieu que le nombre de P-nœuds du graphe est une fonction croissante du nombre d'observations même si le nombre de P-nœuds réellement créés est très inférieur au nombre maximum théorique de P-nœuds possibles (125 750 pour 500 observations).
- Le nombre de feuilles différentes nous indique un facteur moyen de regroupement des observations de l'ordre de 4, c'est-à-dire qu'en moyenne, une structure syntaxico-rythmique complète est représentée 4 fois dans le graphe.
- La table 5.3 confirme une constatation que nous avions déjà formulée au chapitre 4 : il y a peu de différence entre les étiquettes calculées à partir des valeurs moyennes et des valeurs prises au deux-tiers (EERO1/EERO1', EFO1/EFO1', *etc.*).
- On remarquera également que certaines étiquettes sont faiblement représentées dans la base, nous discuterons de leur pertinence dans la section suivante.

- Enfin, une analyse très grossière des cardinalités de chaque étiquette prosodique nous permet de dégager quelques caractéristiques globales du corpus PolyNombre :
  - un nombre de la base d'apprentissage possède en moyenne 9 à 10 voyelles,
  - il y a deux fois plus de voyelles qui s'inscrivent dans une pente descendante de la courbe de fréquence fondamentale que dans une pente montante,
  - près d'une voyelle sur trois semble porter une marque d'allongement (étiquette AL1) alors qu'un peu plus d'une voyelle sur cinq émerge de ces voisines directes par sa valeur de *f0* (étiquette EFO1).

Après ces quelques commentaires pour le moins généraux, nous proposons une analyse plus détaillée des faits prosodiques mesurés sur notre base de nombres.

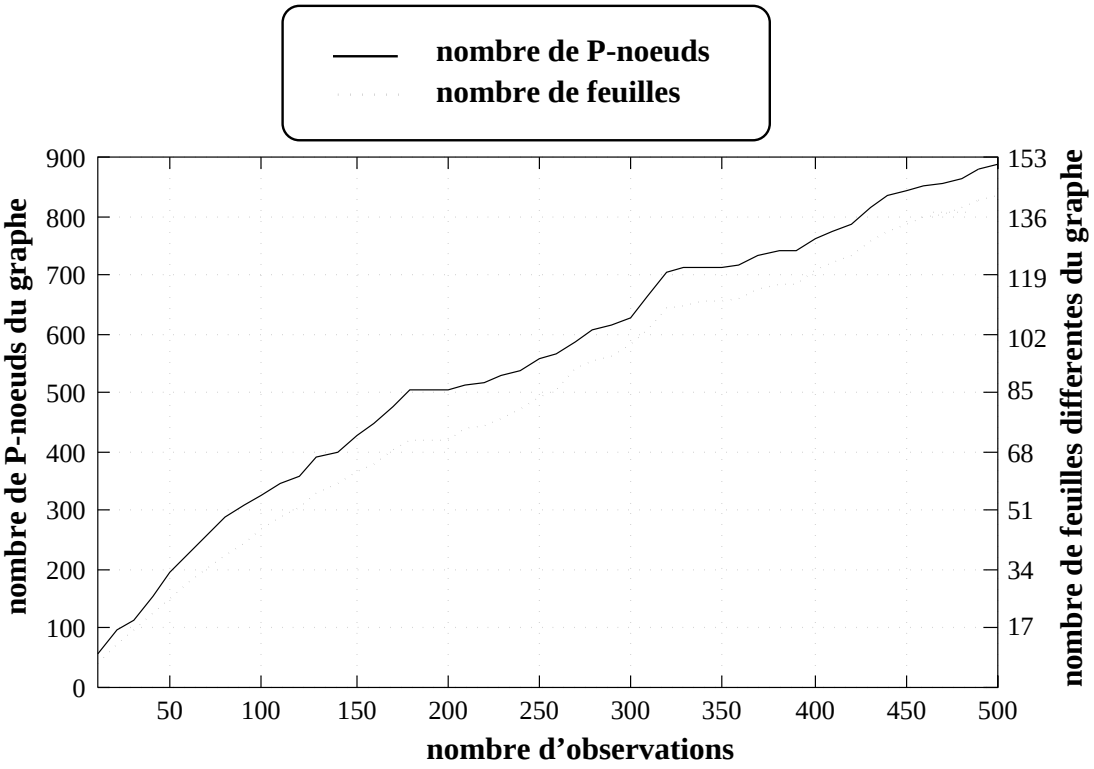


Figure 5.10: Nombre de P-nœuds et de feuilles différentes modélisés durant la phase d'apprentissage des 500 nombres de la base PolyNombre.

Une courte analyse

Nous pourrions dans les pages qui suivent nous livrer à une analyse méticuleuse des faits prosodiques se manifestant sur notre base PolyNombre afin de dégager un ensemble consistant de règles pouvant rendre compte de la plupart des configurations rencontrées. Nous

Étiquette	Nb.	Étiquette	Nb.	Étiquette	Nb.	Étiquette	Nb.
NIVA1	938	NIVA2	1278	NIVA3	1462	NIVA4	948
AL1	1666	AL2	1095	ED1	1101	ED2	703
EFO1	990	EFO1'	1070	EFO2	724	EFO2'	766
ENIVA1	504	ENIVA2	417	ENIVR1	261	ENIVR2	256
EERO1	1094	EERO1'	1106	EERO2	795	EERO2'	790
+	1361	-	2953	=	343		
EEN1	84	EEN2	65				
MAX_FO	500	MAX_ERO	500	MAX_MS	500		
MIN_FO	500	MIN_ERO	500	MIN_MS	500		
STAB	2310	PAUSE	199	VO	4781		

Table 5.3: Décompte des principales étiquettes prosodiques automatiquement apposées pour les 500 nombres de la base PolyNombre.

avons cependant déjà eu l'occasion d'aborder les problèmes liés à cette méthodologie ; en particulier, il est difficile à un expert de prendre en considération lors de son analyse la totalité des facteurs régissant les diverses manifestations prosodiques — pour autant que ces facteurs soient clairement définis — ce qui aboutit inévitablement à une prise en compte locale de quelques facteurs seulement (ce que van Santen appelle "piecemeal optimization" [159]). Nous allons illustrer des obstacles concrets qui se présentent alors avec une telle méthodologie tout en tentant de dégager quelques régularités que nous espérons suffisamment robustes pour attester la validité de notre système ProStat.

- Une première analyse auditive confirmée par une analyse visuelle des nombres de notre corpus d'apprentissage permet rapidement de constater la présence "régulière" d'une frontière — que nous appellerons majeure — à la fin du dernier groupe de mots précédant le mot *virgule*. Cette frontière se caractérise principalement par un allongement significatif de la dernière voyelle pleine de la partie entière des nombres, ainsi que par une augmentation sensible de sa fréquence fondamentale. Les schémas de la figure 5.13 proposent quelques exemples de courbes de fréquence fondamentale pour des nombres vérifiant la structure syntaxico-rythmique : NB(N1000_999999().VIRG().N100_999()). On y remarque bien sûr que les mouvements de la courbe de fréquence fondamentale sont plus ou moins prononcés selon les réalisations. En tout état de cause, une étude méticuleuse de nombreux exemples couvrant au mieux les différents contextes de réalisation doit impérativement précéder toute phase de construction d'un système de règles. Pour illustrer notre propos, nous avons reporté dans la table 5.5 le décompte des étiquettes prosodiques apposées sur la dernière voyelle pleine (v) de la partie entière des nombres décimaux de la base PolyNombre. Les observations sont réparties en huit colonnes en fonction de la position (comptée en voyelles) de v ; chaque colonne reportant alors deux informations : **no** qui comptabilise le nombre de fois où une

étiquette prosodique e a été apposée à v et **ne** le nombre de voyelles étiquetées e pour l'ensemble des voyelles des observations concernées[15].

Les vues 5.11 et 5.12 obtenues à partir de la table 5.5 compensent le manque de lisibilité de cette dernière et permettent conformément aux premières impressions de dégager quelques régularités. Il semble globalement que les indices de durée soient assez révélateurs de la terminaison de la partie entière d'un nombre et plus particulièrement les indices **MAX_MS** et **ED2**. Pour ne prendre qu'un seul exemple, 70% des étiquettes **ED2** apposées sur des signaux correspondant à des nombres dont la partie entière est constituée d'exactement 3 voyelles sont localisées à la fin de la partie entière de ces nombres. Les indices de fréquence fondamentale (émergence d'un niveau ou d'une valeur précise de *f0*), de manière moins régulière, viennent également confirmer l'augmentation de *f0* en position terminale de partie entière. Notons également que la faible probabilité mesurée pour l'indice "-" est également un indicateur fiable de la fin de la partie entière, surtout si l'on se rappelle le nombre important de voyelles étiquetées descendantes dans le corpus PolyNombre (près de trois fois le nombre d'étiquettes montantes). On remarque sur les deux figures que ces régularités sont plus ou moins marquées selon le nombre de voyelles que contient la partie entière. On devrait en tout état de cause étudier d'autres facteurs (comme la nature "syntaxique" de la partie entière, ou encore l'influence éventuelle de la partie décimale) avant d'énoncer une règle robuste décrivant les réalisations prosodiques se localisant en finale de partie entière. Ces études ne devraient pas être dissociées les unes des autres afin de vérifier qu'une configuration prosodique n'est pas attribuée à tort à un mauvais facteur. On le comprend bien vite, deux solutions s'offrent à nous ; la première étant d'utiliser les compétences d'un expert dont on peut espérer que ses connaissances seront suffisamment fiables pour réduire efficacement l'espace d'étude à un nombre raisonnable de facteurs, la deuxième étant une approche statistique que nous préconisons — au moins comme aide à l'expertise — qui n'impose pas d'*a priori* particulier quant à la nature des faits observés et ne se focalise pas en un point précis de la chaîne à traiter mais au contraire étudie les distributions des différents indices prosodiques globalement. Ainsi dans notre étude des phénomènes prosodiques caractérisant la dernière voyelle de la partie entière des nombres, notre système dispose-t-il non seulement des probabilités d'occurrences de chaque étiquette mais également de leur distribution sur le reste du nombre.

- Une deuxième analyse infirme nos convictions concernant le mot *mille* qui ne prend pas de marque prosodique particulière mais semble au contraire s'inscrire dans le continuum rendant ainsi caduque la distinction faite sur ce mot dans nos arbres syntaxiques. Après avoir étudié séparément les configurations prosodiques du mot *mille* dans des contextes variés, nous reportons dans la table 5.4 les résultats globaux obtenus tous contextes confondus sur 140 observations (seul le mot *mille* de la partie

[15]Ex: Sur les 171 observations des nombres dont la partie entière est constituée de 3 voyelles, 121 étiquettes ED2 ont été apposées sur la 3ème voyelle ; sur les 1378 voyelles que contiennent les 171 nombres seulement 172 ont été étiquetées ED2.

entière est ici considéré). On peut simplement remarquer qu'aucun indice prosodique particulier ne semble se produire de manière privilégiée excepté peut-être l'absence d'étiquettes d'extrema (local ou pas) *ENIVR2, MIN_ERO, MAX_ERO, MAX_FO* qui confirme l'hypothèse d'une neutralité des contours prosodiques sur le mot *mille*.

- De la même façon, notons également que contrairement à nos attentes le mot *cent* ne fait l'objet d'aucune caractérisation prosodique forte. On peut simplement observer que des études très contextuelles (voir 5.6) permettent de formuler quelques commentaires épars qui soulignent la nécessité d'un apprentissage automatique.

- Enfin et sans grande originalité, un allongement final est souvent mesuré dans notre corpus de nombres.

De cette courte analyse de la prosodie des nombres, on peut retenir que l'information prosodique la plus discriminante est celle qui est localisée en finale de partie entière, ce qui est bien sûr spécifique aux nombres décimaux. Certains phénomènes très locaux peuvent également donner lieu à quelques règles spécifiques (comme la caractérisation du mot *cent*) qui nous donne à penser qu'ils devraient être avantageusement traités par une approche automatique globale. Nous allons vérifier maintenant que le système ProStat présente les aptitudes requises pour capter non seulement les régularités que nous venons partiellement de décrire mais surtout toutes celles qui ont échappé à notre analyse — certes très rapide — et proposer ainsi des hypothèses valuées sur la structure syntaxico-rythmique des observations de notre corpus de test.

étiquette	nb.	tot.	%	étiquette	nb.	tot.	%	étiquette	nb.	tot.	%
NIVA1	22	271	8.12	NIVA2	38	442	8.60	NIVA3	47	491	9.57
NIVA4	28	284	9.86	AL1	64	540	11.85	AL2	39	340	11.47
ED1	51	382	13.35	ED2	43	235	18.30	EFO1	34	323	10.53
EFO1'	35	351	9.97	EFO2	17	250	6.80	EFO2'	20	264	7.58
ENIVA1	17	165	10.30	ENIVA2	6	145	4.14	ENIVR1	2	73	2.74
ENIVR2	0	81	0.00	EERO1	16	356	4.49	EERO1'	36	383	9.40
EERO2	11	272	4.04	EERO2'	35	268	13.06	EEN1	0	28	0.00
EEN2	0	23	0.00	+	30	390	7.69	-	95	993	9.57
=	12	116	10.34	MIN_FO	11	140	7.86	MAX_FO	4	140	2.86
MIN_MS	4	140	2.86	MAX_MS	14	140	10.00	MAX_ERO	0	140	0.00
MIN_ERO	0	140	0.00	STAB	151	747	20.21	PAUSE	1	61	1.64

Table 5.4: Caractérisation prosodique du mot *mille* (dans la partie entière uniquement) tous contextes confondus. *nb.* indique le nombre d'étiquettes prosodiques apposées au mot mille, *tot.* précise le nombre total d'étiquettes apposées sur l'ensemble des voyelles des observations ; la troisième colonne exprime la probabilité (exprimée en pourcentage) qu'une étiquette donnée corresponde au mot *mille*. Les données reportées dans cette table concernent un total de 140 observations (*i.e.* 140 nombres).

	Nombre de voyelles précédent le mot *virgule*															
	2		3		4		5		6		7		8		9	
étiquette	**no**	**ne**	**no**	**ne**	**no**	**ne**	**no**	**ne**	**no**	**ne**	**no**	**ne**	**no**	**ne**	**no**	**ne**
NIVA1	2	35	23	294	4	107	13	249	2	125	3	80	0	22	0	13
NIVA2	2	31	24	349	10	131	22	373	15	220	7	94	3	37	2	25
NIVA3	7	42	52	400	25	162	53	401	27	244	17	138	3	40	3	21
NIVA4	14	36	71	280	18	109	36	246	20	146	6	82	3	23	0	10
AL1	15	51	108	476	45	191	99	448	35	255	25	155	5	46	5	23
AL2	10	34	85	312	40	138	88	290	26	164	20	102	5	27	5	14
ED1	15	24	100	280	42	120	95	316	34	189	19	104	6	37	5	21
ED2	0	5	121	172	39	79	96	219	35	119	18	74	4	18	5	12
EFO1	11	24	87	275	31	115	76	273	38	154	21	94	8	31	3	15
EFO1'	12	27	75	286	30	122	70	291	46	183	23	101	8	34	3	15
EFO2	0	7	67	203	14	74	44	197	34	134	6	67	5	28	2	12
EFO2'	0	8	65	202	12	80	44	225	28	137	8	69	4	28	3	14
ENIVA1	6	10	56	151	14	59	53	136	30	80	13	41	4	13	0	8
ENIVA2	0	6	39	121	5	38	33	121	18	77	4	35	2	10	2	8
ENIVR1	4	7	37	96	13	31	32	62	16	34	4	18	2	8	0	2
ENIVR2	0	4	20	75	3	22	22	71	11	52	3	16	2	9	2	6
EERO1	9	32	65	311	28	117	57	307	35	176	8	90	4	35	1	15
EERO1'	11	27	76	284	20	120	40	316	31	194	8	104	6	29	2	20
EERO2	0	12	48	194	25	95	62	219	39	152	13	80	7	24	3	17
EERO2'	0	10	85	204	23	96	42	215	37	144	8	77	5	28	1	12
EEN1	0	6	4	21	0	10	3	20	4	15	1	8	0	2	0	0
EEN2	0	1	3	19	2	10	5	19	2	8	0	6	0	1	0	1
+	17	56	132	446	45	170	68	333	33	193	17	94	5	30	2	18
-	7	76	23	797	10	319	41	830	26	490	13	273	3	82	3	48
=	1	12	16	85	2	24	15	117	5	56	3	30	1	14	0	3
MIN_FO	2	25	11	171	3	57	6	124	5	64	2	33	0	10	0	5
MAX_FO	11	25	46	171	10	57	19	124	5	64	4	33	1	10	0	5
MIN_MS	1	25	4	171	3	57	0	124	2	64	1	33	0	10	0	5
MAX_MS	7	25	50	171	28	57	53	124	17	64	12	33	3	10	3	5
MAX_ERO	7	25	29	171	14	57	10	124	15	64	2	33	1	10	0	5
MIN_ERO	6	25	15	171	15	57	27	124	4	64	6	33	1	10	0	5
STAB	20	83	139	668	45	277	118	628	40	329	30	196	7	68	3	27
PAUSE	2	15	14	74	2	23	4	43	1	21	3	15	1	4	0	2
Nb. obs.	25		171		57		124		64		33		10		5	
Nb. voy.	146		1378		526		1306		754		405		132		71	

Table 5.5: Table récapitulative des indices prosodiques localisés sur la dernière voyelle pleine du groupe qui précède le mot *virgule*. Dans une même colonne sont regroupées toutes les observations dont la partie entière possède le même nombre de voyelles ; chacune d'elles étant divisée à son tour en deux colonnes : **no** indique le nombre de fois où une étiquette est située sur la voyelle terminale et **ne** indique le nombre de fois où la même étiquette à été attribuée pour toutes les voyelles de ces mêmes observations. Les deux lignes inférieures reportent respectivement le nombre d'observations puis le nombre de voyelles total de ces observations en fonction du nombre de voyelles de la partie entière.

étiquette	nb.	N1000_999999(6)						VIRG(2)		N100_999(5)				
		MOT(2)		MOT(1)	MILLE(1)	MOT(1)	MOT(1)	VIRG(2)		MOT(1)	CENT(1)	MOT(2)		MOT(1)
		d	f					d	f			d	f	
NIVA1	14	0	0	1	1	2	0	2	2	1	0	0	2	3
NIVA2	31	2	2	3	1	3	2	3	3	1	2	4	3	2
NIVA3	35	3	2	3	4	3	4	3	2	1	4	2	2	2
NIVA4	19	2	4	1	2	0	2	0	0	4	2	2	0	0
AL1	40	4	4	2	6	5	2	1	0	1	4	2	3	6
AL2	22	2	3	1	2	4	0	1	0	0	1	2	1	5
ED1	23	0	2	1	4	3	0	1	4	1	3	2	2	0
ED2	15	0	0	1	3	6	1	1	0	0	3	0	0	0
EFO1	26	0	5	0	5	0	6	1	1	4	1	3	0	0
EFO1'	27	0	5	0	6	0	6	1	1	3	3	2	0	0
EFO2	23	0	0	0	1	1	5	2	0	5	7	2	0	0
EFO2'	20	0	0	0	0	1	3	2	0	5	7	2	0	0
ENIVA1	12	0	1	0	2	0	5	0	0	3	1	0	0	0
ENIVA2	14	0	0	0	0	1	2	0	0	5	5	1	0	0
ENIVR1	10	0	1	0	1	0	3	1	0	3	0	1	0	0
ENIVR2	10	0	0	0	0	0	1	1	0	3	3	2	0	0
EERO1	24	0	2	4	0	1	6	0	1	0	5	3	2	0
EERO1'	22	0	2	4	0	0	8	0	0	0	5	1	2	0
EERO2	28	0	0	6	0	1	8	2	0	0	7	4	0	0
EERO2'	24	0	0	6	0	1	8	1	0	0	6	2	0	0
EEN1	2	0	0	0	0	0	2	0	0	0	0	0	0	0
EEN2	1	0	0	0	0	0	0	0	0	0	1	0	0	0
+	23	0	0	2	3	3	6	4	2	1	0	1	1	0
-	64	6	7	3	2	3	2	4	5	6	8	7	5	6
=	12	1	1	3	3	2	0	0	0	0	0	0	1	1
MIN_FO	8	0	0	0	0	1	1	0	1	1	0	0	1	3
MAX_FO	8	2	2	1	0	0	0	0	0	0	0	2	0	1
MIN_MS	8	0	0	1	0	0	1	2	1	2	0	1	0	0
MAX_MS	8	2	1	1	2	1	0	1	0	0	0	0	0	0
MAX_ERO	8	0	1	1	0	0	6	0	0	0	0	0	0	0
MIN_ERO	8	1	0	0	0	0	0	0	1	0	0	1	3	2
STAB	49	3	4	6	8	6	4	4	3	1	2	1	3	4
PAUSE	6	1	0	0	0	1	0	0	0	0	0	0	1	3
VO	104	8	8	8	8	8	8	8	8	8	8	8	8	8

Table 5.6: Caractérisation prosodique de huit observations du corpus PolyNombre possédant toutes la même structure syntaxico-rythmique. *d* et *f* désignent respectivement la voyelle initiale et finale du groupe décrit. On remarque en plus de la prédominance de nombreuses étiquettes sur le dernier mot de la partie entière, la majorité de certaines autres sur le mot *cent* notamment des étiquettes d'émergence de *f0* ou bien d'énergie.

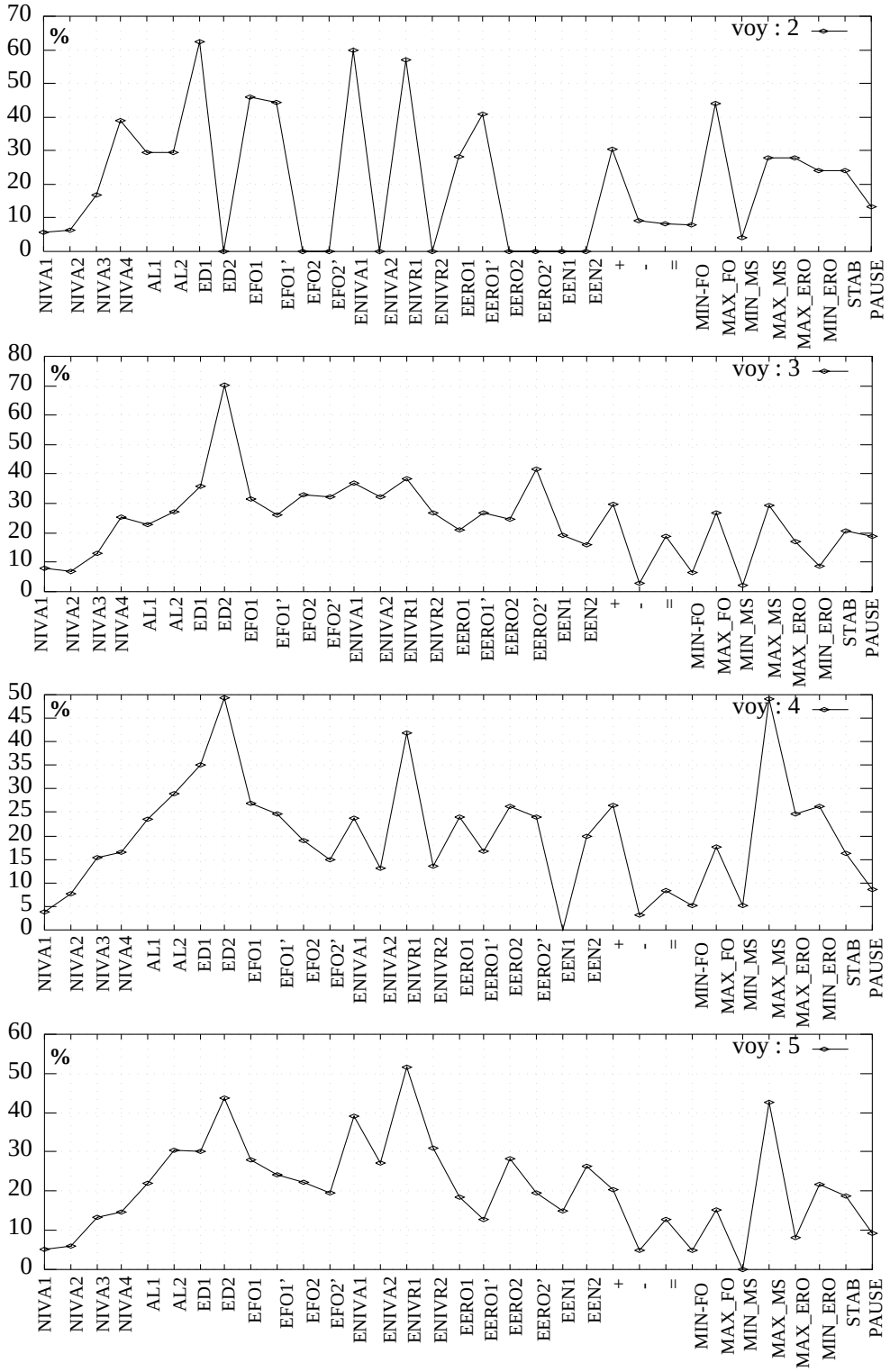


Figure 5.11: Probabilités (exprimées en pourcentage) qu'une étiquette prosodique donnée indique la dernière voyelle pleine de la partie entière des nombres de PolyNombre. Le nombre de voyelles de la partie entière est indiqué dans le coin supérieur droit de chaque courbe.

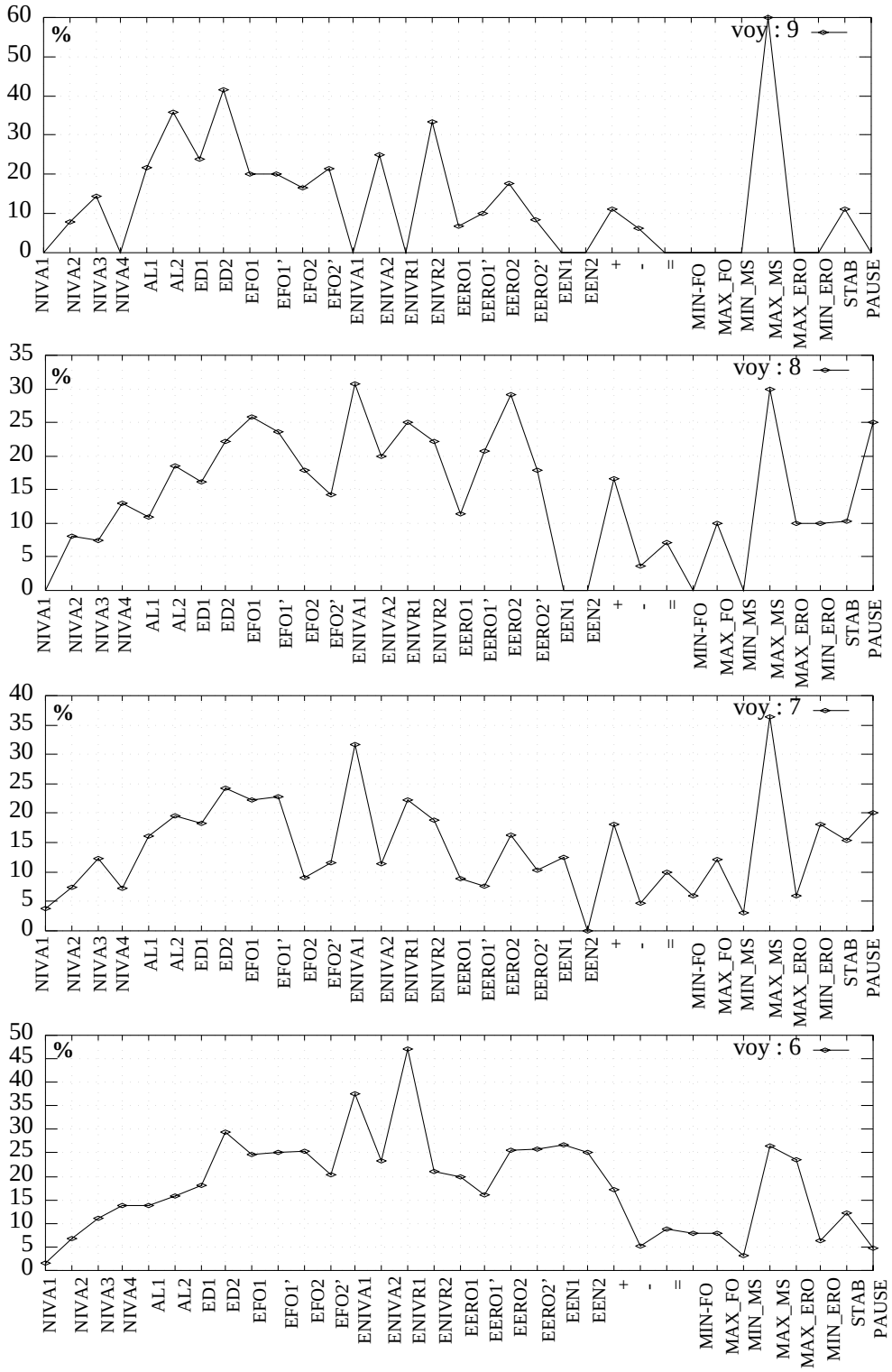
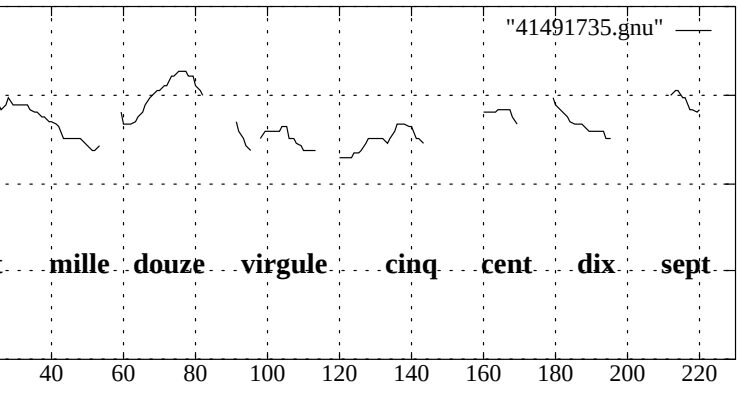
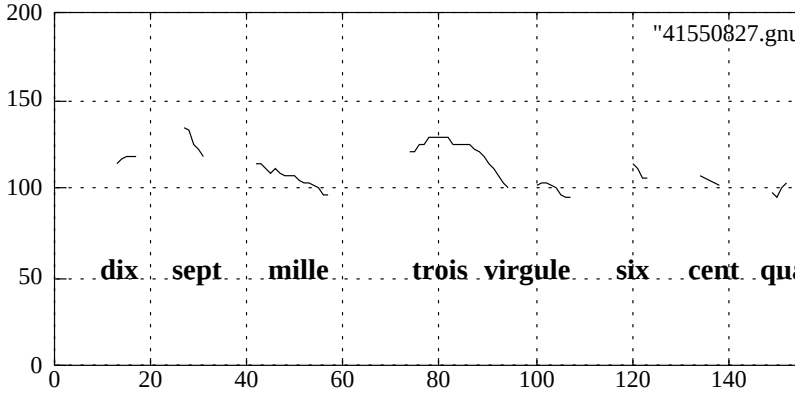


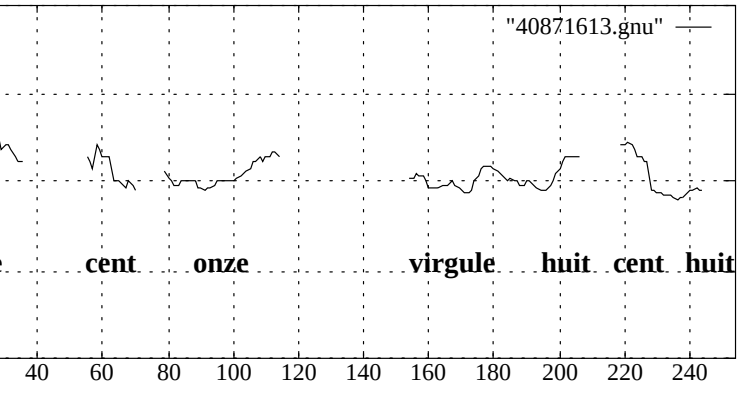
Figure 5.12: Probabilités (exprimées en pourcentage) qu'une étiquette prosodique donnée indique la dernière voyelle pleine de la partie entière des nombres de PolyNombre. Le nombre de voyelles de la partie entière est indiqué dans le coin supérieur droit de chaque courbe.



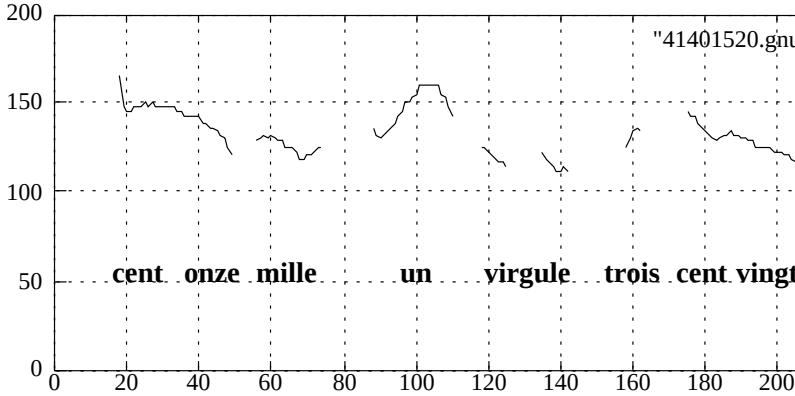
a) 2012,517 (loc: JLC)



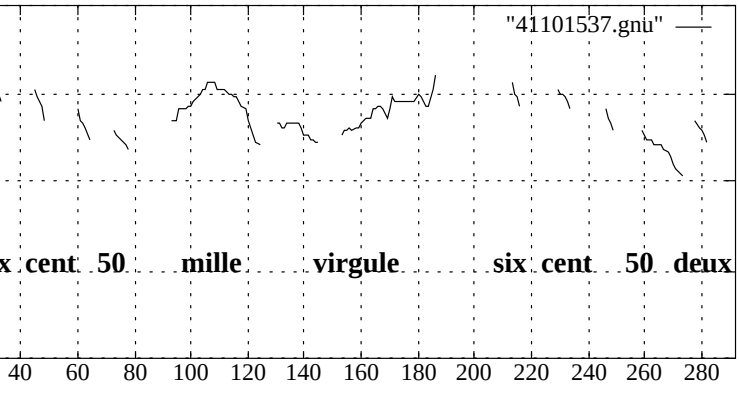
b) 17003,642 (loc: CD)



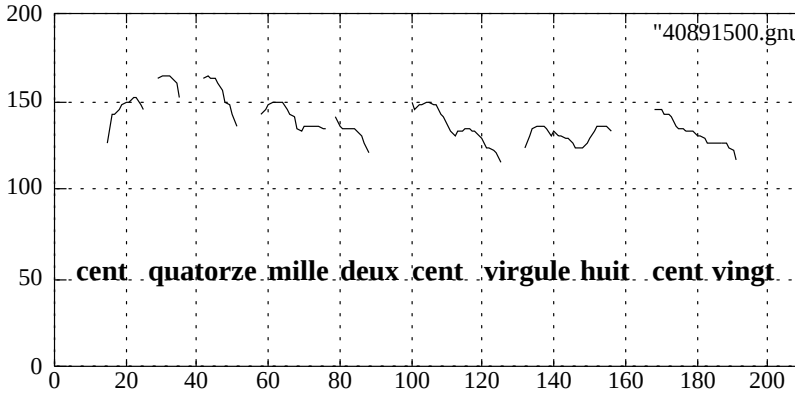
c) 1111,808 (loc: GC)



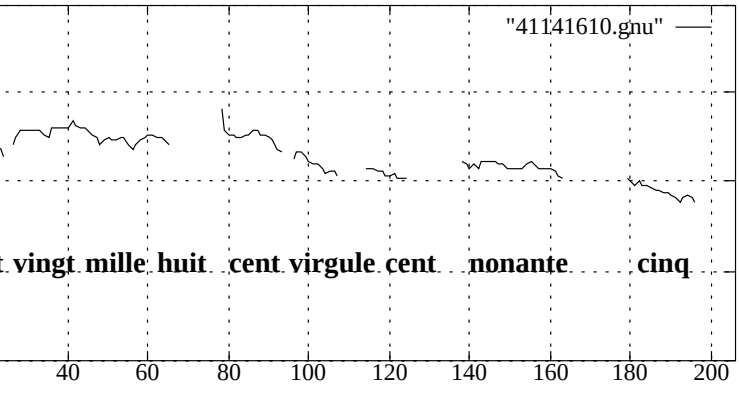
d) 111001,322 (loc: IMC)



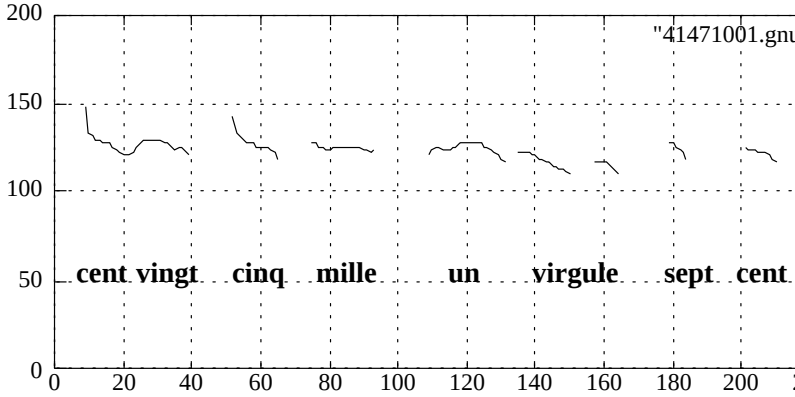
e) 250000,652 (JLC)



f) 114200,823 (loc: PL)



g) 120800,195 (loc: MM)



h) 125001,707 (loc: HA)

Figure 5.13: Exemple de courbes de *f0* mesurées pour un sous-ensemble de nombres de la

Les résultats

Une des prétentions du système ProStat est d'offrir en plus d'une aide à l'analyse de situations particulières, la possibilité d'utiliser le graphe généré lors de l'apprentissage pour proposer un ensemble d'hypothèses valuées sur les contraintes organisationnelles des niveaux linguistiques modélisés. Nous allons reporter dans les quelques lignes qui suivent un ensemble de résultats qui attestent la validité du système ProStat en tant que système prédictif.
Afin de nous assurer du fonctionnement global du système, nous avons tout d'abord consulté le classement des hypothèses syntaxico-rythmiques fournies par ProStat pour les données ayant servies à l'apprentissage (*i.e.* les 500 nombres de PolyNombre). Nous avons alors demandé à notre système de proposer son faisceau d'hypothèses parmi les seules feuilles du graphe d'apprentissage ; en d'autres termes, seules les structures syntaxiques complètes et pleinement instanciées les feuilles des SR-structures sont des mots dont le nombre de voyelles est spécifié) des signaux présentés en entrée du système ont été considérées et classées. Les données brutes ainsi que les résultats cumulés sont reportés sur la figure 5.14[16]. On remarque que le taux de propositions classées en tête est normalement élevé (*i.e.* supérieur à 80%) ; l'existence d'observations non classées en tête laisse toutefois présager une dégradation sensible lors des résultats sur les données de test. On peut expliquer ceci par le manque d'à propos de certaines étiquettes prosodiques ou encore par une inadéquation partielle du découpage syntaxique proposé à l'apprentissage. Une analyse plus précise des observations non classées en tête semble confirmer la deuxième hypothèse.

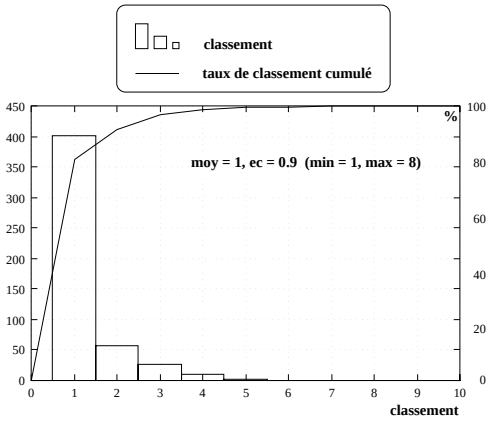


Figure 5.14: Taux de classement des 500 observations du corpus d'apprentissage. Seules les informations localisées dans les feuilles du graphe d'apprentissage sont ici en concurrence (leur nombre moyen étant de 17). On observe que 400 nombres du corpus PolyNombre (soit plus de 80% du corpus) sont classés en première position.

[16]Notons conformément à la remarque concernant les taux de classement formulée lors du chapitre 4 que la notation employée ici, ne fait pas apparaître d'ex æquo.

La figure 5.16 présente le classement obtenu dans les mêmes conditions avec cette fois-ci les données du corpus de test dont l'intersection avec le corpus d'apprentissage est nulle. Remarquons cependant que seules les observations dont la structure syntaxico-rythmique a été présentée au moins une fois lors de la phase d'apprentissage sont prises en compte dans ces résultats (soit en pratique seulement 148 observations sur les 298 que contient le corpus PolyNombreTest). La figure 5.15 indique la distribution du nombre de voyelles de ces 148 observations et rappelle le nombre de feuilles différentes dans le graphe ProStat en fonction du nombre de voyelles ; ceci afin de mesurer le nombre moyen de classes pouvant être affectées à une observation qui est ici proche de 15 (14.86).

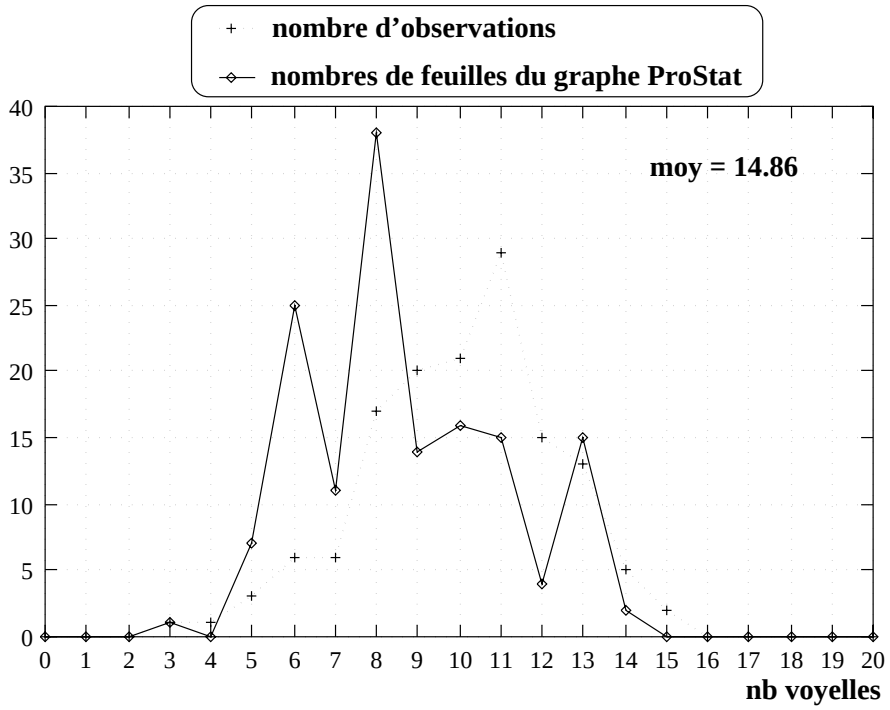


Figure 5.15: La courbe en pointillé indique le nombre de feuilles différentes du graphe d'apprentissage ProStat en fonction du nombre de voyelles. La courbe en trait plein indique quant à elle la distribution des 148 observations en fonction de leur nombre de voyelles. La moyenne pondérée affichée indique le nombre moyen de rangs d'un classement.

Comme on pouvait bien sûr s'y attendre, le classement fourni est de qualité inférieure à celui obtenu sur les données d'apprentissage. Il reste cependant très satisfaisant puisque près de la moitié des observations considérées sont classées en tête ; d'autant plus qu'à l'instar du test précédent, une analyse des observations non classées en tête permet de vérifier que les choix syntaxiques présentés plus haut sont parfois non pertinents et sont à l'origine d'une dispersion des résultats présentés comme l'illustrent les deux SR-structures suivantes qui possèdent toutes les deux le même nombre de mots, le même nombre de voyelles et possèdent de surcroît la même partie entière :

$$\left[\begin{array}{l} \text{N1000\_999999}_4 \left[\begin{array}{l} \text{MOT}_1 \\ \text{MILLE}_1 \\ \text{N100\_999}_2 \left[\begin{array}{l} \text{CENT}_1 \\ \text{MOT}_1 \end{array}\right] \end{array}\right] \\ \text{VIRG}_2 \\ \text{N100\_999}_4 \left[\begin{array}{l} \text{MOT}_1 \\ \text{CENT}_1 \\ \text{MOT}_1 \\ \text{MOT}_1 \end{array}\right] \end{array}\right] \quad \text{et} \quad \left[\begin{array}{l} \text{N1000\_999999}_4 \left[\begin{array}{l} \text{MOT}_1 \\ \text{MILLE}_1 \\ \text{N100\_999}_2 \left[\begin{array}{l} \text{CENT}_1 \\ \text{MOT}_1 \end{array}\right] \end{array}\right] \\ \text{VIRG}_2 \\ \text{MOT}_1 \\ \text{MOT}_1 \\ \text{MOT}_1 \\ \text{MOT}_1 \end{array}\right]$$

Nous avons cependant précisé au début de la présentation du système ProStat que notre but n'était pas ici de fournir des résultats optimaux pour une tâche précise, mais davantage de valider notre démarche, aussi n'allons-nous pas nous livrer maintenant à une amélioration des résultats qu'imposerait une implémentation réelle d'un système de reconnaissance de nombres. Nous formulerons cependant quelques commentaires dans la section suivante sur les possibilités à disposition pour réaliser proprement cette amélioration. Les résultats collectés ici permettent — en l'état — de conclure à la pertinence de l'information prosodique pour la prédiction de l'organisation syntaxico-rythmique des nombres décimaux. Pour étayer ceci nous proposons le report du classement moyen que l'on obtiendrait dans les mêmes conditions en attribuant à un P-nœud du graphe non plus une note obtenue par l'application de la méthodologie exposée en sous-section 5.4.4 mais une note aléatoire. Ce sont ces résultats qui sont reportés sur la figure 5.17. La figure 5.18 présente une comparaison des taux de classement qui permet d'apprécier l'apport de la composante prosodique ; on mesure ainsi une amélioration du classement en tête de l'ordre de 40%.

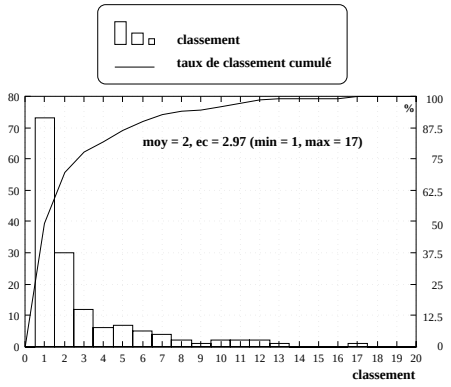


Figure 5.16: Taux de classement des 148 observations du corpus de test dont les structures syntaxico-rythmiques sont présentes dans le graphe ProStat pour un nombre moyen de classes voisin de 15.

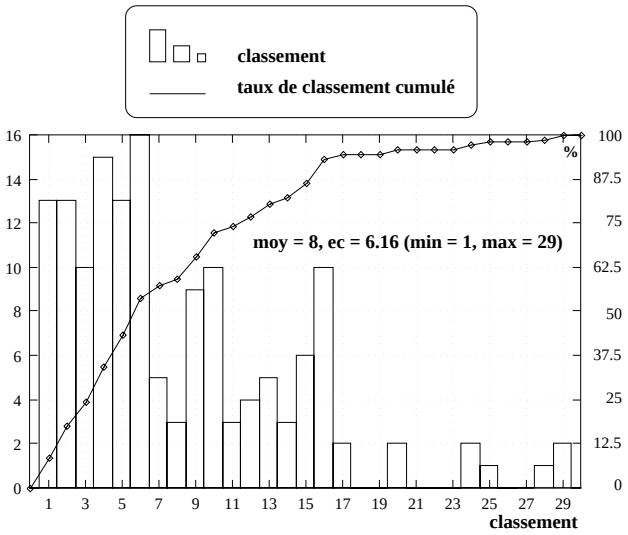


Figure 5.17: Taux de classement aléatoire des 148 observations du corpus de test dont les structures syntaxico-rythmiques sont présentes dans le graphe ProStat. Le nombre moyen de classes est proche de 15.

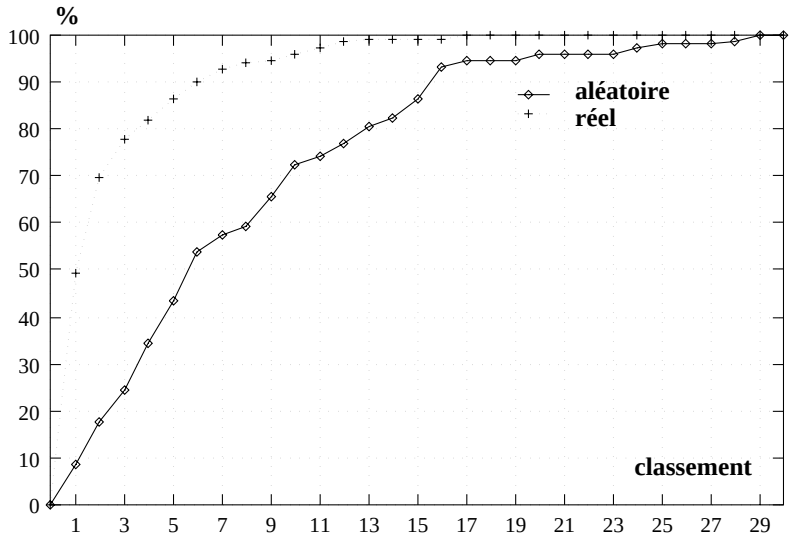


Figure 5.18: Comparaison du classement effectué sur les 148 observations du corpus de test dont les structures syntaxico-rythmiques sont présentes dans le graphe d'apprentissage ProStat. La courbe en pointillé indique les taux obtenus effectivement, alors que la courbe représentée par une ligne pleine indique le classement obtenu avec une notation aléatoire.

Nous avons jusqu'ici restreint notre champ d'étude aux seules observations dont la structure syntaxico-rythmique était présente dans le graphe ProStat et ainsi démontré la pertinence de l'information prosodique pour la prédiction structurelle de ces observations. Dans l'expérience qui suit, le système attribut à tous les nombres du corpus de test des hypothèses structurelles, qu'ils aient ou non été modélisés pendant la phase d'apprentissage. S'ajoutent donc aux 148 observations précédentes les 150 autres du corpus PolyNombreTest dont la structure syntaxico-rythmique n'existe pas — de manière complète — dans le graphe. La figure 5.19 reporte le classement cumulé obtenu pour l'ensemble du corpus de test puis pour sa restriction aux seuls nombres non entièrement modélisés dans le graphe d'apprentissage ; la ligne pointillée correspond à un classement des hypothèses par l'attribution d'une note aléatoire alors que la ligne pleine correspond au classement obtenu par application du calcul exposé plus haut. Le nombre moyen d'hypothèses formulées par le système pour un nombre est voisin de 400. Deux observations ressortent de ces deux schémas :

- On observe une amélioration nette du classement (30% de plus dans les 50 premières hypothèses) lorsque l'information prosodique est effectivement prise en compte dans la notation ; ce qui confirme fort heureusement les conclusions de la série d'expériences précédentes.

- On peut cependant s'étonner du classement peu sélectif des hypothèses formulées : pour que 90% des nombres du corpus de test soient classés, il faut retenir un faisceau d'environ 100 hypothèses ; le nombre de ces dernières étant en moyenne de 300 pour le corpus PolyNombreTest. L'explication de ce modeste classement tient — après analyse des résultats — à la nature arbitraire des groupements structurels proposés par les arbres "syntaxiques" des observations et qui ne semble donc pas refléter fidèlement un niveau d'organisation prosodique. La différenciation non pertinente de mots du vocabulaire (comme *mille*) contribue également à la dispersion des résultats. Afin d'illustrer ceci, la figure 5.20 présente le classement des hypothèses précédentes pour la prédiction du mot *virgule* dans les nombres du corpus PolyNombreTest. On voit ainsi très nettement que la première hypothèse que fournit le système ProStat — si elle n'est pas nécessairement la bonne — localise cependant correctement dans le signal le mot *virgule*. En d'autres termes, il semble que l'information prosodique — telle qu'elle est utilisée dans ProStat — semble pertinente pour la prédiction des structures de surface c'est-à-dire pour les groupements de plus haut niveau (ex : NB(N1000_999999(4).VIRG(2).N100_999(4),10)) alors qu'elle intervient dans des proportions moindres à l'intérieur des groupes ainsi délimités ; ce qui correspond finalement bien à nos intuitions initiales.

En tout état de cause, l'utilisation de ces résultats dans un système de reconnaissance des nombres nécessiterait un nouvel apprentissage avec des décompositions "grammaticales" simplifiées. La section 5.6 se propose ultérieurement de discuter des améliorations facilement réalisables ; nous nous contentons de rappeler qu'en l'état, notre système permet de classer des nombres dont la structure syntaxico-rythmique est déjà modélisée dans

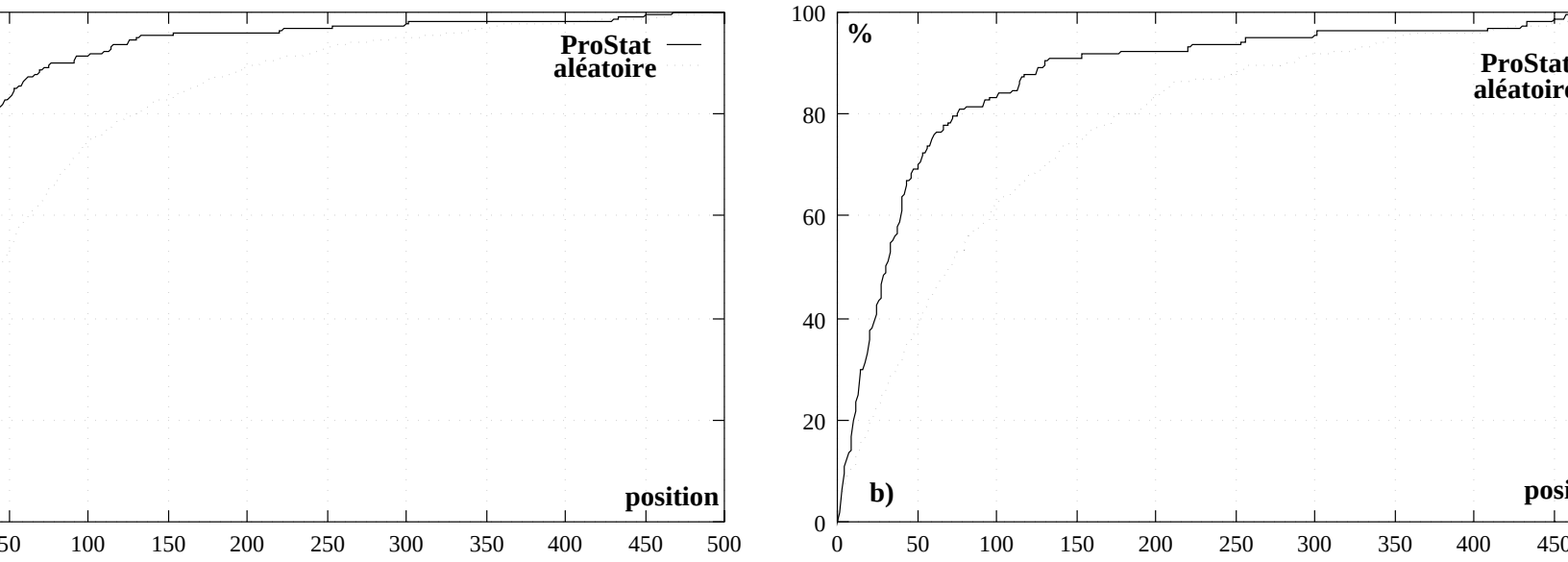


Figure 5.19: Pourcentage d'observations classées en fonction du rang par le système ProStat (ligne pleine) puis par une notation aléatoire (ligne pointillée). La figure a) consigne l'ensemble des observations du corpus de test PolyNombreTest ; la figure b) ne concerne que les observations du même corpus dont la structure syntaxico-rythmique n'est pas modélisée dans le graphe d'apprentissage (soit 150 observations).

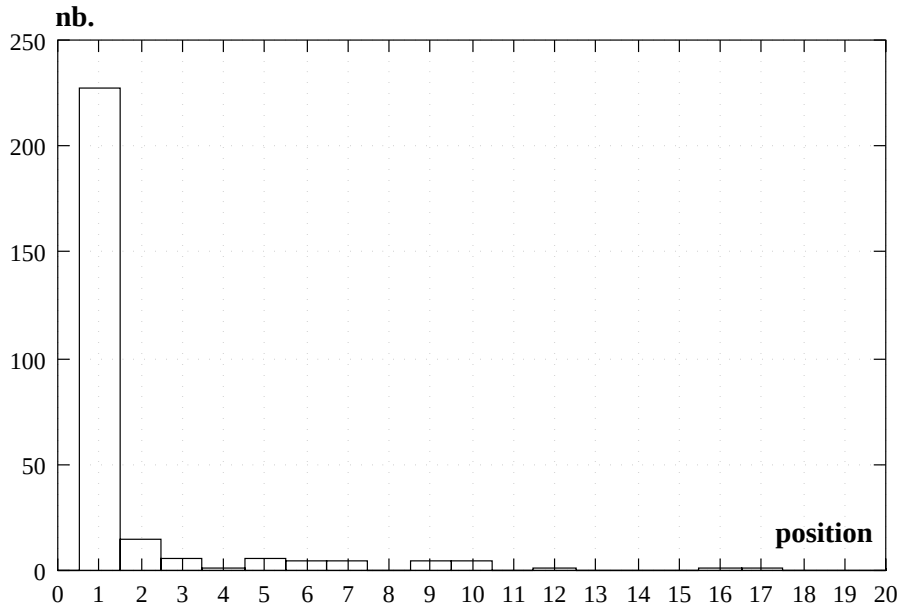


Figure 5.20: Classement des hypothèses fournies par le système ProStat pour les nombres du corpus PolyNombreTest en considérant uniquement l'exactitude de la position du mot *virgule* dans la chaîne.

le graphe d'apprentissage avec un bon taux de réussite, et qu'il permet avec fiabilité de déterminer la structure de surface des nombres du corpus de test.

5.5.2 Tâche 2 : les phrases

Après avoir testé la validité de notre système sur une première base de nombres décimaux, nous allons maintenant décrire nos travaux sur des corpus de phrases isolées. À l'instar de la précédente étude, nous rappelons brièvement les entrées de notre système ; nous précisons ensuite quelques données relatives à la phase d'apprentissage puis après avoir réalisé une courte étude de la prosodie des phrases de notre corpus, nous reportons enfin les résultats des diverses expériences auxquelles nous nous sommes livrées.

Les entrées

Cette tâche a pour but principal de montrer l'utilité de la prosodie dans un système de reconnaissance de la parole en continu. Plus précisément nous aimerions montrer la pertinence des informations prosodiques pour prédire tout ou partie de la structuration syntaxico-rythmique de phrases isolées. Comme nous avons déjà eu l'occasion de le préciser, nous sommes conscients des choix réducteurs que nous avons été amenés à faire dans cette étude :

- la restriction de l'étude à des phrases isolées nous place dans des conditions particulières où les schémas prosodiques sont relativement pauvres,
- le choix des structurations syntaxiques et rythmiques comme unique champ d'investigation pourra sembler limité à certains bien qu'il nous paraisse raisonnable dans le cadre d'une première étude, surtout pour des phrases isolées de type lues où la sémantique semble secondaire,
- enfin la sélection des corpus de phrases répond davantage à des contraintes pratiques que théoriques.

Nous aimerions simplement noter que le cadre de cette étude — aussi réducteur soit-il — correspond bien au conditions réalistes d'un environnement de reconnaissance de la parole : les systèmes actuels sont encore effectivement loin de pouvoir traiter de la parole dite naturelle au tout venant, les composantes syntaxiques de ces systèmes possèdent de nombreuses limites et si plusieurs modèles sémantiques et pragmatiques voient actuellement le jour [134, 21], il n'en reste pas moins qu'une restriction à un domaine très spécifique est nécessaire si l'on souhaite en tenir compte dans des systèmes automatiques (systèmes de réservation, de renseignements, *etc.*). En d'autres mots nous pensons qu'il n'est pas de notre ressort d'imposer à notre système des compétences que les autres composantes d'un système de reconnaissance ne sont pas ou peu capables de traiter actuellement. De plus l'étude que nous présentons maintenant ne nous place pas dans les conditions les plus favorables : nous travaillons sur des données "multi-locuteur" de qualité téléphonique ;

un nombre non négligeable de phrases est prononcé dans des conditions très bruitées et une partie d'entre-elles le sont par des personnes ayant des difficultés à s'exprimer en français[17].
Ces précisions étant faites, les 500 phrases de la base PolyPhrase (prononcées par 50 locuteurs) ont donc été utilisées pour la phase d'apprentissage. Rappelons simplement qu'elles sont des répétitions en nombre inégal de 80 phrases de base pour lesquelles des arbres syntaxiques ont été manuellement déterminés (voir les annexes pages 231 à 244). L'alignement phonétique quant à lui est obtenu de manière automatique : un phonétiseur réalisant un accès au lexique BdLex [117] et prenant en compte les principales règles phonologiques (effacement des /ə/, prise en compte des liaisons, *etc.*) permet d'obtenir un réseau de reconnaissance phonétique qui est alors fourni en entrée d'un algorithme de Viterbi qui sélectionne le meilleur chemin dans ce graphe à l'aide des modèles phonétiques décrits dans les pages 26 à 29 de ce mémoire. Voici un exemple d'un tel réseau[18] pour la phrase : *Le tapis était élimé sur le bord.*

```
(
  [sil]
  ll ee [sil]
  tt aa pp ii [sil]
  (ei | ai) tt (ei | ai) [tt | sil]
  (ei | ai) ll ii mm (ei | ai) [sil]
  ss uu rr [sil]
  ll [ee [sil]]
  bb oo rr [sil]
)
```

L'apprentissage

Nous reportons simplement ici quelques données relatives à la phase d'apprentissage, notamment le nombre de P-nœuds et de feuilles différents créés (figure 5.21) ainsi que le nombre total d'étiquettes prosodiques considérées par le système ProStat pour l'ensemble des données d'observation (table 5.7) et formulons quelques remarques sommaires sur cette phase d'apprentissage :

- Le nombre feuilles différentes du graphe d'apprentissage indique un facteur moyen de regroupement des observations de l'ordre de 10 ; ce qui devient raisonnable en comparaison du facteur moyen de regroupement mesuré après l'apprentissage des nombres.

[17] Sans aucune moquerie (...), j'ai envie d'ajouter qu'entre autre difficulté, l'accent suisse-romand est assez différent de l'accent français...

[18] Exemple décrit à l'aide du formalisme de HTK [168] et des symboles phonétiques réellement utilisés dans l'application présentée.

- On remarque encore un fois un écart faible des décomptes des étiquettes obtenues soit à partir des valeurs moyennes, soit à partir des valeurs prises au deux-tiers des paramètres prosodiques.
- Les voyelles s'inscrivant dans une pente montante de la courbe de fréquence fondamentale sont environ deux fois moins nombreuses que les voyelles affectées d'une pente descendante, ce qui signifie qu'en proportion, les voyelles montantes sont plus nombreuses que dans le corpus d'apprentissage des nombres. Ceci nous laisse à penser qu'un recours à des variations plus marquées du paramètre de *f0* caractérise la lecture des phrases (en comparaison à la lecture des nombres).
- Globalement, il semble que les décomptes des différentes étiquettes prosodiques soient voisins pour les corpus de phrases et de nombres décimaux ; la moyenne du nombre de voyelles par phrase étant voisin de 10.

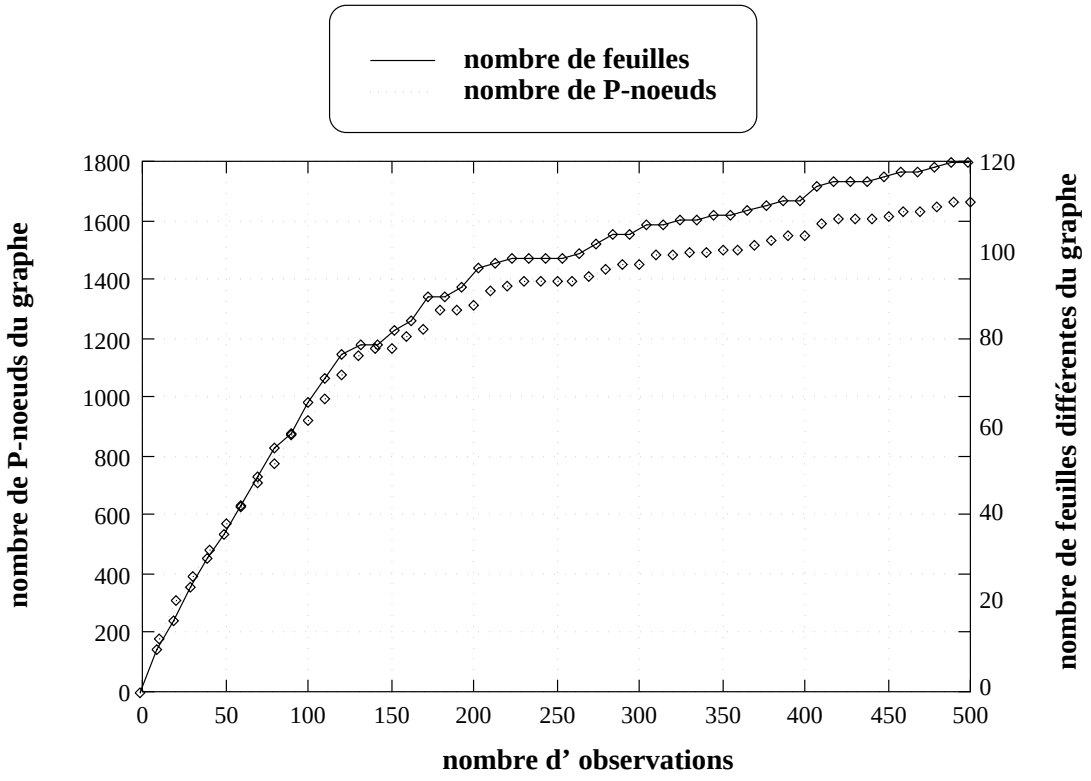


Figure 5.21: Nombre de P-nœuds et de feuilles différentes modélisés durant la phase d'apprentissage des 500 phrases de la base PolyPhrase.

Une courte analyse

Les études sur la prosodie des phrases isolées sont nombreuses [126, 31, 94, 23, 20, 110, 16, 102, 115, 45, 41, 158] et traduisent bien l'intérêt que l'on porte à la fonction de

Étiquette	Nb.	Étiquette	Nb.	Étiquette	Nb.	Étiquette	Nb.
NIVA1	1052	NIVA2	1507	NIVA3	1408	NIVA4	859
AL1	1638	AL2	987	ED1	1055	ED2	836
EFO1	1242	EFO1'	1306	EFO2	963	EFO2'	1029
ENIVA1	715	ENIVA2	552	ENIVR1	392	ENIVR2	362
EERO1	1281	EERO1'	1330	EERO2	836	EERO2'	952
+	1689	-	2830	=	335		
EEN1	94	EEN2	47				
MAX_FO	500	MAX_ERO	500	MAX_MS	500		
MIN_ERO	500	MIN_FO	500	MIN_MS	500		
VO	5018	STAB	1979	PAUSE	149		

Table 5.7: Décompte des principales étiquettes prosodiques automatiquement apposées sur les 500 phrases de la base PolyPhrase.

groupement de mots "sémantiquement" liés assurée par la prosodie ; que ce soit à des fins descriptives ou de traitement automatique. La démarche générale de la plupart de ces études consiste à proposer des unités autorisant un découpage de l'intonation en unités plus petites (groupe intonatif, unité intonative, mot prosodique, *etc.*) puis à en fournir une description prosodique (en terme de traits phonologiques, de contours, *etc.*) capable de rendre compte de leurs configurations possibles à la lumière de corpus d'observations de taille et de complexité variables. Pourtant, aucun de ces systèmes n'est en accord total ce qui — au-delà de différences dans les motivations et les méthodologies de chacun — s'explique très bien par le fait que les paramètres prosodiques sont à la fois gouvernés par les organisations syntaxique, sémantique mais aussi rythmique. Nous préconisons une approche plus globale n'imposant pas l'existence *a priori* d'unités particulières qu'il nous faudrait alors décrire[19] ; aussi n'allons-nous pas dans les pages qui suivent énoncer un système de règles qui viendrait uniquement alourdir l'existant. Nous allons simplement présenter des observations qui corroborent ou au contraire montrent les limites de règles consensuelles que l'on peut trouver dans ces nombreux travaux en espérant que notre approche globale guidée par les observations saura mieux que nous capter l'information résultante des différents niveaux de contraintes.
Afin d'alléger au plus l'analyse, nous décrirons nos observations avec un ensemble restreint d'étiquettes prosodiques en introduisant les symboles suivants :

- AL qui réunit les étiquettes d'allongement AL1 et AL2,
- ED qui rassemble les étiquettes d'émergence de durée ED1 et ED2,
- FO qui regroupe les symboles EF01, EF01', EF02 puis EF02',

[19]Notre propos n'étant cependant pas ici de remettre en cause la validité de ces différentes unités.

- EFO qui désigne les étiquettes d'émergence de niveau de *f0* (absolue ou relative),
- ERO qui réfère aux symboles EER01, EER01', EER02 et EER02',
- RO qui indique enfin une émergence de l'énergie (EEN1 et EEN2) ; les autres symboles restant inchangés.

Nous étudions dans un premier temps les phrases dont la structure de surface est composée de trois syntagmes comme dans l'exemple : (Aujourd'hui)$_{circ}$ (chaque village)$_{sujet}$ (a sa chapelle)$_{gv}$. Le décompte des étiquettes apposées sur les phrases du corpus PolyPhrase qui vérifient la contrainte structurelle énoncée est reporté en table 5.8 que nous nous empressons de commenter :

- Seules des étiquettes comme le minimum de fréquence fondamentale ou encore le maximum de durée ont des probabilités d'occurrence nettement marquées ; près de la moitié de ces étiquettes est effectivement localisée en finale de phrase (dernière voyelle pleine).
- Les autres étiquettes bien que non aléatoirement distribuées sont davantage réparties : si près de 60 % des étiquettes d'allongement correspondent à des fins de syntagmes de surface (une des trois positions F), plus de 15 % sont localisées à l'initiale des groupes ; 45 % des étiquettes d'émergence d'un niveau de *f0* sont localisées en finale d'un des deux premiers groupes.
- La prise en compte de configurations prosodiques plus complexes (combinaison booléenne des étiquettes) permettraient de dégager des régularités plus nettes, mais multiplierait la combinatoire des situations à considérer.

Cette première observation ne saurait tenir lieu d'analyse mais permet de représenter un élément d'information (celui associé à un P-nœud particulier du graphe d'apprentissage) dont dispose le système ProStat pour émettre des hypothèses structurelles d'une phrase. En tout état de cause il nous faut réaliser des analyses conjointes de l'influence de paramètres comme la nature des syntagmes ou leur nombre de voyelles. C'est ce que nous nous proposons de faire maintenant avec cependant des prétentions très modestes puisque nous ne ferons qu'illustrer quelques différences locales ; une analyse méticuleuse nécessiterait bien plus de données d'observations que nous n'en possédons pour couvrir correctement les divers contextes à étudier. De plus, la complexité d'une telle entreprise nous fait baisser les bras et nous conforte dans notre approche statistique.
Toujours dans le cadre des phrases constituées de trois syntagmes, nous nous sommes intéressés à observer les différences de localisation des indices prosodiques en comparant diverses combinaisons des positions des trois syntagmes (sujet, verbal et complément). La figure 5.22 résume cette analyse qui nous permet de dégager quelques points qui montrent l'influence de la nature des syntagmes sur l'organisation prosodique :

- La fin du syntagme sujet est très nettement marquée du maximum d'énergie dans la phrase lorsqu'il est suivi d'un complément circonstanciel (plus de 65% des étiquettes MAX_ERO) qui le sépare du syntagme verbal.

- De manière moins prononcée, les étiquettes d'émergence (FO, EFO, RO, et ERO) mais aussi le maximum du paramètre de fréquence fondamentale sont localisés plus fréquemment en final du syntagme sujet lorsque celui-ci est suivi d'un complément circonstanciel.

- Il est également intéressant de noter la forte régularité avec laquelle le minimum d'énergie (MIN_ERO) est localisé sur la première voyelle de la phrase lorsque celle-ci débute par le syntagme circonstanciel (plus de 85% des cas).

- Enfin nous pouvons aussi remarquer la forte présence de pauses en finale du premier syntagme lorsque le complément circonstanciel précède ou suit le syntagme sujet.

		Groupe 1						Groupe2						Groupe3			
		D		F		B		D		F		B		D		F	
étiquette	total	nb	%	nb	%	nb	%	nb	%	nb	%	nb	%	nb	%	nb	%
AL	1575	68	4.3	273	17.3	52	3.3	52	3.3	209	13.3	58	3.7	127	8.1	442	28.1
ED	1205	0	0.0	278	23.1	0	0.0	47	3.9	261	21.7	53	4.4	124	10.3	0	0.0
FO	2831	0	0.0	449	15.9	0	0.0	194	6.9	517	18.3	97	3.4	208	7.3	0	0.0
EFO	1233	0	0.0	255	20.7	0	0.0	68	5.5	242	19.6	48	3.9	96	7.8	0	0.0
RO	2751	0	0.0	268	9.7	0	0.0	226	8.2	373	13.6	23	0.8	241	8.8	0	0.0
ERO	83	0	0.0	7	8.4	0	0.0	7	8.4	4	4.8	0	0.0	8	9.6	0	0.0
+	1019	71	7.0	132	13.0	18	1.8	62	6.1	149	14.6	17	1.7	43	4.2	45	4.4
-	1743	142	8.1	76	4.4	38	2.2	143	8.2	59	3.4	31	1.8	178	10.2	190	10.9
=	227	14	6.2	15	6.6	4	1.8	9	4.0	27	11.9	5	2.2	20	8.8	22	9.7
MIN_FO	289	43	14.9	5	1.7	6	2.1	11	3.8	6	2.1	2	0.7	23	8.0	121	41.9
MAX_FO	289	23	8.0	74	25.6	2	0.7	16	5.5	29	10.0	15	5.2	16	5.5	21	7.3
MIN_MS	289	24	8.3	13	4.5	2	0.7	41	14.2	6	2.1	0	0.0	32	11.1	4	1.4
MAX_MS	289	7	2.4	61	21.1	8	2.8	5	1.7	29	10.0	12	4.2	11	3.8	137	47.4
MAX_ERO	289	20	6.9	43	14.9	2	0.7	45	15.6	19	6.6	1	0.3	9	3.1	4	1.4
MIN_ERO	289	74	25.6	16	5.5	44	15.2	13	4.5	1	0.3	9	3.1	33	11.4	35	12.1
PAUSE	104	0	0.0	28	26.9	2	1.9	3	2.9	9	8.7	3	2.9	2	1.9	28	26.9

Table 5.8: Décompte des étiquettes prosodiques dans les phrases du corpus PolyPhrase constituées de trois syntagmes de surface et probabilité d'occurrence (exprimée en pourcentage) des étiquettes en position initiale (D) et finale (F) de syntagme. La position (B) désigne les groupes constitués d'une seule voyelle.

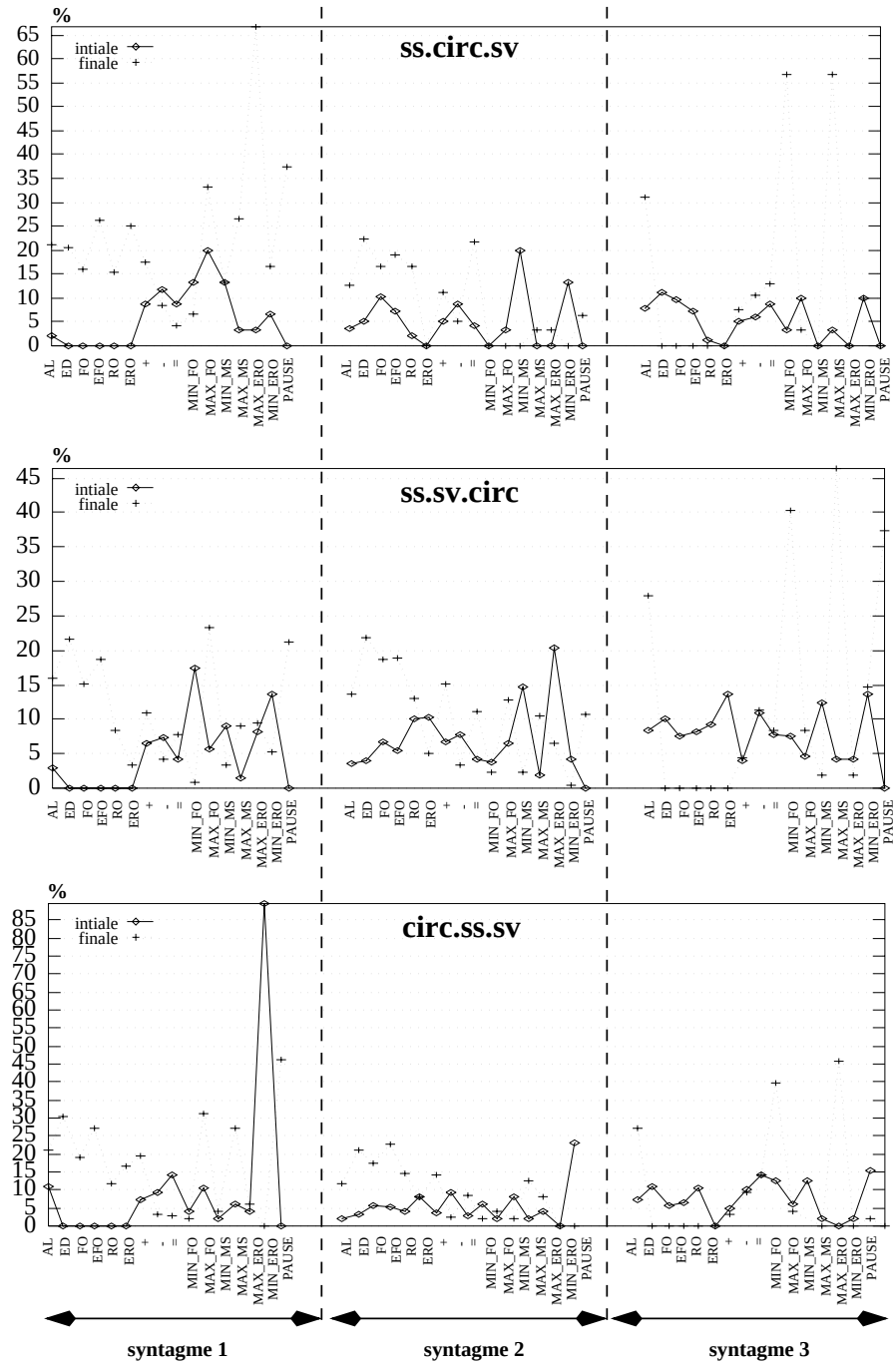


Figure 5.22: Comparaison du pourcentage d'étiquettes prosodiques localisées à l'initiale et en finale de syntagme pour les phrases du corpus PolyPhrase composées de trois syntagmes (un syntagme sujet *ss*, un syntagme verbal *sv* et un syntagme circonstanciel *circ*). Les pourcentages d'occurrence des étiquettes prosodiques ont été mesurés à partir de l'étude d'environ 300 phrases.

Nous avons ensuite voulu observer les variations prosodiques engendrées par la modification du nombre de voyelles à l'intérieur des constituants syntaxiques. Pour cette étude, le choix de phrases composées d'un syntagme sujet suivi d'un syntagme verbal s'est imposé à nous pour des raisons simples de quantité d'échantillons disponibles. Les tables 5.9 et 5.10 résument les configurations prosodiques mesurées pour les phrases du corpus PolyPhrase qui vérifient la contrainte syntaxique énoncée et dont le nombre de réalisations est supérieur à la dizaine. Malgré le faible nombre d'observations considérées ici, nous pouvons cependant formuler prudemment quelques commentaires qu'il nous faudrait vérifier sur des corpus spécifiques de plus grande taille. On peut vérifier l'importance du rythme sur la distribution des différentes étiquettes prosodiques dans la phrase et constater à ce titre que bien souvent la recherche d'un "équilibre rythmique" entre en conflit avec le niveau d'organisation syntaxique. Une comparaison des tableaux G et H de la table 5.10 permet d'illustrer nos propos. On observe en effet dans le cas des phrases parfaitement équilibrées (H) un marquage prosodique très net de la fin du syntagme sujet : près de la moitié des étiquettes d'émergence bilatérale de durée et de fréquence fondamentale (ED et EFO) ainsi que des maxima des paramètres de *f0* et de durée sont localisés sur la dernière voyelle du syntagme. Au contraire, les phrases présentant un déséquilibre du nombre de voyelles des syntagmes sujet et verbal présentent une répartition plus homogène des étiquettes prosodiques : la frontière syntaxique est prosodiquement moins claire et le nombre d'indices prosodiques localisés en milieu du syntagme est sensiblement plus élevé que pour les phrases équilibrées. De manière moins prononcée, les mêmes remarques peuvent être extraites des tableaux A, B et C de la table 5.9. Une autre illustration nous est proposée par le tableau E qui montre que le syntagme sujet mono-vocalique des phrases de six voyelles est très peu marqué (à l'exception des minimum de *f0* et d'énergie) alors que l'initiale du groupe verbal l'est davantage. Le marquage prosodique du début du second syntagme s'explique simplement par le fait que dans notre corpus le groupe verbal de ces phrases débute le plus souvent par un verbe contenant une seule voyelle pleine qui reçoit alors les diverses marques prosodiques. Le peu de données disponibles ne nous autorise pas — par exemple — à vérifier si dans le cas d'un verbe comportant deux voyelles, on observerait un report des marques prosodiques sur la deuxième voyelle du groupe verbal.

	A	PH(9)								B	PH(9)							
		SS(2)				SV(7)					SS(3)				SV(6)			
		D		F		D		F			D		F		D		F	
étiquette	total	nb	%	nb	%	nb	%	nb	%	total	nb	%	nb	%	nb	%	nb	%
AL	53	4	8	10	19	10	19	13	25	52	0	0	17	33	5	10	15	29
ED	32	0	0	6	19	8	25	0	0	38	0	0	19	50	4	11	0	0
FO	85	0	0	7	8	22	26	0	0	104	0	0	29	28	5	5	0	0
EFO	47	0	0	3	6	11	23	0	0	59	0	0	26	44	1	2	0	0
RO	63	0	0	10	16	1	2	0	0	109	0	0	25	23	4	4	0	0
ERO	4	0	0	1	25	0	0	0	0	6	0	0	2	33	0	0	0	0
MINFO	10	2	20	1	10	0	0	3	30	13	2	15	2	15	0	0	7	54
MAXFO	10	1	10	2	20	3	30	1	10	13	0	0	5	39	1	8	0	0
MINMS	10	0	0	0	0	0	0	0	0	13	3	23	1	8	0	0	0	0
MAXMS	10	1	10	4	40	0	0	3	30	13	0	0	6	46	0	0	4	31
MAXERO	10	3	30	4	40	1	10	0	0	13	3	23	2	15	0	0	0	0
MINERO	10	4	40	1	10	1	10	0	0	13	2	15	1	8	0	0	4	31

	C	PH(9)								D	PH(10)							
		SS(4)				SV(5)					SS(4)				SV(6)			
		D		F		D		F			D		F		D		F	
étiquette	total	nb	%	nb	%	nb	%	nb	%	total	nb	%	nb	%	nb	%	nb	%
AL	46	1	2	7	15	4	9	20	44	72	17	24	7	10	3	4	17	24
ED	36	0	0	7	19	7	19	0	0	43	0	0	8	19	1	2	0	0
FO	100	0	0	20	20	1	1	0	0	123	0	0	21	17	24	20	0	0
EFO	43	0	0	8	19	0	0	0	0	54	0	0	7	13	15	28	0	0
RO	87	0	0	32	37	3	3	0	0	113	0	0	29	26	7	6	0	0
ERO	7	0	0	4	57	0	0	0	0	0	0	0	0	0	0	0	0	0
MINFO	11	2	18	0	0	0	0	6	55	13	3	23	0	0	0	0	3	23
MAXFO	11	0	0	3	27	0	0	1	9	13	0	0	1	8	4	31	0	0
MINMS	11	2	18	0	0	0	0	0	0	13	1	8	1	8	6	46	0	0
MAXMS	11	0	0	1	9	0	0	8	73	13	4	31	0	0	0	0	3	23
MAXERO	11	3	27	4	36	0	0	0	0	13	4	31	1	8	0	0	0	0
MINERO	11	3	27	1	9	3	27	1	9	13	7	54	0	0	0	0	3	23

Table 5.9: Étude de l'influence du nombre de voyelles sur le comportement prosodique des phrases du corpus PolyPhrase composées de deux syntagmes de surface (un groupe sujet suivi d'un groupe verbal).

	E	PH(6)						F	PH(7)						
		SS(1)		SV(6)					SS(2)				SV(5)			
		B		D		F			D		F		D		F	
étiquette	total	nb	%	nb	%	nb	%	total	nb	%	nb	%	nb	%	nb	%
AL	44	3	7	2	5	23	52	47	0	0	15	32	5	11	16	34
ED	22	0	0	3	14	0	0	23	0	0	8	35	6	26	0	0
FO	74	0	0	20	27	0	0	51	0	0	10	20	7	14	0	0
EFO	43	0	0	16	37	0	0	28	0	0	7	25	3	11	0	0
RO	79	0	0	21	27	0	0	53	0	0	16	30	2	4	0	0
ERO	2	0	0	1	50	0	0	3	0	0	1	33	0	0	0	0
MINFO	12	3	25	0	0	5	42	10	3	30	2	20	0	0	2	20
MAXFO	12	1	8	5	42	1	8	10	0	0	4	40	2	20	1	10
MINMS	12	1	8	4	33	0	0	10	4	40	0	0	0	0	0	0
MAXMS	12	0	0	1	8	10	83	10	0	0	5	50	0	0	4	40
MAXERO	12	1	8	7	58	0	0	10	2	20	8	80	0	0	0	0
MINERO	12	2	17	0	0	2	17	10	3	30	0	0	2	20	4	40

	G	PH(8)								H	PH(8)							
		SS(2)				SV(6)					SS(4)				SV(4)			
		D		F		D		F			D		F		D		F	
étiquette	total	nb	%	nb	%	nb	%	nb	%	total	nb	%	nb	%	nb	%	nb	%
AL	50	1	2	6	12	0	0	21	42	43	2	5	15	35	8	19	11	26
ED	39	0	0	7	18	5	13	0	0	29	0	0	15	52	6	21	0	0
FO	93	0	0	19	20	15	16	0	0	68	0	0	26	38	6	9	0	0
EFO	47	0	0	14	30	5	11	0	0	33	0	0	16	49	0	0	0	0
RO	89	0	0	19	21	11	12	0	0	66	0	0	27	41	11	17	0	0
ERO	3	0	0	1	33	1	33	0	0	1	0	0	0	0	0	0	0	0
MINFO	14	0	0	0	0	2	14	6	43	10	0	0	1	10	0	0	2	20
MAXFO	14	1	7	5	36	1	7	2	14	10	0	0	5	50	0	0	0	0
MINMS	14	0	0	2	14	3	21	0	0	10	2	20	0	0	1	10	0	0
MAXMS	14	1	7	0	0	0	0	7	50	10	1	10	5	50	0	0	3	30
MAXERO	14	1	7	8	57	2	14	0	0	10	0	0	3	30	0	0	0	0
MINERO	14	4	29	0	0	1	7	2	14	10	6	60	0	0	0	0	0	0

Table 5.10: Étude de l'influence du nombre de voyelles sur le comportement prosodique des phrases du corpus PolyPhrase composées de deux syntagmes de surface (un groupe sujet suivi d'un groupe verbal) — *suite*.

		SS(3)						SV(3)						CIRC(7)									
		ART(1)		NC(2)				VB(1)		NC(2)				ART(1)		NC(2)				ADJ(4)			
		B		D		F		B		D		F		B		D		F		D		M		F	
étiquette	total	nb	%	nb	%	nb	%	nb	%	nb	%	nb	%	nb	%	nb	%	nb	%	nb	%	nb	%	nb	%
AL	79	2	3	2	3	22	28	2	3	2	3	9	11	4	5	0	0	4	5	9	11	1	1	22	28
ED	82	0	0	0	0	24	29	3	4	1	1	16	20	7	9	1	1	8	10	21	26	1	1	0	0
FO	142	0	0	6	4	27	19	10	7	4	3	29	20	8	6	2	1	13	9	19	13	24	17	0	0
EFO	49	0	0	5	10	13	27	2	4	1	2	8	16	2	4	1	2	2	4	10	20	5	10	0	0
RO	143	0	0	9	6	1	1	27	19	27	19	13	9	0	0	16	11	20	14	13	9	17	12	0	0
ERO	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	100	0	0	0	0	0	0
MINFO	12	3	25	0	0	0	0	0	0	0	0	0	0	1	8	0	0	0	0	2	17	0	0	6	50
MAXFO	12	1	8	2	17	6	50	0	0	0	0	0	0	0	0	0	0	0	0	1	8	1	8	1	8
MINMS	12	3	25	1	8	0	0	0	0	0	0	0	0	1	8	4	33	1	8	0	0	2	17	0	0
MAXMS	12	0	0	0	0	6	50	0	0	0	0	1	8	0	0	0	0	0	0	1	8	0	0	4	33
MAXERO	12	1	8	2	17	0	0	6	50	1	8	1	8	0	0	0	0	0	0	1	8	0	0	0	0
MINERO	12	1	8	0	0	1	8	0	0	0	0	0	0	7	58	0	0	0	0	3	25	0	0	0	0
PAUSE	4	0	0	0	0	1	25	0	0	0	0	3	75	0	0	0	0	0	0	0	0	0	0	0	0

Table 5.11: Décompte et localisation des étiquettes prosodiques pour les 12 observations du corpus PolyPhrase vérifiant le critère syntaxico-rythmique : PH(SS (GN (ART(1). NC(2),3),3). SV(VB(1). CO(GN(ART(0).NC(2),2),2),3). CIRC(PREP(0).GN(ART(1). NC(2). ADJ(4),7),7),13).

Nous avons tenu à présenter également les indices prosodiques mesurés pour une structure syntaxico-rythmique complète présente au moins dix fois dans notre corpus PolyPhrase. La table 5.11 reporte le décompte et la localisation des différents étiquettes prosodiques apposées aux phrases vérifiant la structure :

$$PH\left[\begin{array}{ll} SS & \left[GN\left[\begin{array}{ll} ART & (1) \\ NC & (2) \end{array}\right]\right] \\ SV & \left[\begin{array}{l} VB(1) \\ CO\left[GN\left[\begin{array}{ll} ART & (0) \\ NC & (2) \end{array}\right]\right] \end{array}\right] \\ CIRC & \left[\begin{array}{l} PREP(0) \\ GN\left[\begin{array}{ll} ART & (1) \\ NC & (2) \\ ADJ & (4) \end{array}\right] \end{array}\right] \end{array}\right]$$

Le peu de données considérées ici ne nous autorise que quelques commentaires succints. Il semble que les articles soient peu soumis à un marquage prosodique particulier si ce n'est par les étiquettes de minima globaux des paramètres prosodiques. La frontière syntaxique entre le groupe sujet et le groupe verbal semble quant à elle bien marquée, ainsi que la frontière syntaxique entre le groupe verbal et le complément circonstanciel. On note également un marquage prosodique (plus léger cependant) à l'intérieur du complément circonstanciel le découpant équitablement en deux parties, soit à la fin du syntagme nominal, soit au début de l'adjectif. On pourra également noter que les pauses, peu nombreuses, sont toutes localisées aux frontières syntaxiques de surface.
Nous pourrions multiplier les observations du type de celles que nous venons de présenter et comblerions ainsi certainement une partie des lacunes de cette analyse trop sommaire. On sent néanmoins sur ces quelques exemples l'ampleur du travail et l'imposante

couverture syntaxico-rythmique des données d'observations que réclamerait une étude méticuleuse.
Il n'est peut-être pas inintéressant de rappeler ici quelques résultats que nous avions obtenus [79] lors d'une étude corrélative sur 44 phrases phonétiquement équilibrées du corpus BDSON (celles pour lesquelles nous disposions de l'alignement phonétique). Dans cette étude, nous souhaitions montrer que l'étiquetage prosodique (qui est celui qui a été présenté dans le chapitre 2) obtenu de manière automatique était suffisamment proche — pour une étude corrélative — du même étiquetage obtenu cette fois-ci à partir d'une segmentation manuelle des phrases en phonèmes. Notre conclusion était alors que l'étiquetage automatique proposé pouvait être utilisé sans dégradation notable des résultats. Nous reportons la table 5.12 et la figure 5.23 qui présentaient alors les pourcentages de différentes étiquettes et de leurs combinaisons localisées à l'initiale et en finale de mots et de syntagmes. Il ressortait alors, que des combinaisons d'indices (comme l'allongement de la voyelle qui possède la *f0* la plus élevée) s'avéraient un bon prédicteur de fin de syntagme et/ou de mots. Le faible nombre de phrases observées ne nous permettait cependant pas de fonder le réel potentiel de ces indices. Notons que les phrases phonétiquement équilibrées du corpus BDSON possèdent une très large majorité de mot disyllabiques et monosyllabiques ce qui laisse peu d'occasions à un prédicteur de fin mots de se tromper !
Nous renvoyons le lecteur à [79] pour plus de détails sur cette étude, et terminons cette analyse en présentant quelques exemples de courbes de fréquence fondamentale mesurées pour plusieurs prononciations des mêmes phrases par différents locuteurs, et ce dans le but d'illustrer les régularités mais également les variations entre les diverses réalisations. Les figures 5.24, 5.25 et 5.26 présentent donc des courbes du fondamental pour trois phrases de notre corpus. Ces courbes sont présentées telles qu'elles nous sont fournies par notre détecteur de fréquence fondamentale, aussi peut-on observer — entre autres choses — des petits dysfonctionnements inévitables (surtout sur de la parole téléphonique relativement bruitée) qui contribuent à la difficulté d'une investigation prosodique.

		AL	ED	MAX FO	MIN FO	MAX MS	F0	EF0	AL + ED	AL + MAX_FO	AL ED FO	*marque*
		111	102	44	44	44	123	105	62	22	45	204
fin de phrase	44	35			38	16						40
fin de sujet	44	20	23	18	1	16	29	27	17	11	16	32
fin de gpe verbal	44	29	15	3	19	11	18	17	10	2	9	40
fin de complément	33	24	9	8	17	11	10	9	9	5	6	25
début de phrase	44	5		1	2	1						5
début de sujet	44	2			1	1						3
début de gpe verbal	44	7	7	1	1	3	5	2	5	1	2	12
début de complément	33	5	3	2	1	1			1			7
fin de mots	339	97	86	38	43	42	104	92	55	21	45	201
fin mots grammaticaux	172	9	13	4	3	5	14	8	5	2	3	31
fin mots lexicaux	167	88	73	34	40	37	90	84	50	19	42	170
fin de syntagme	111	73	47	29	37	38	57	53	36	18	31	97
début de syntagme	111	14	10	3	3	5	5	2	6	1	2	22

Table 5.12: Corrélations entre des positions intéressantes dans la phrase et des configurations d'étiquettes prosodiques apposées sur 44 phrases phonétiquement équilibrées du corpus BDSON. La première ligne de valeurs représente le nombre total d'étiquettes positionnées sur le corpus ; la première colonne précise le nombre d'occurrences de chaque position. La colonne *marque* indique un allongement de durée ou une émergence quelconque (durée ou *f0*).

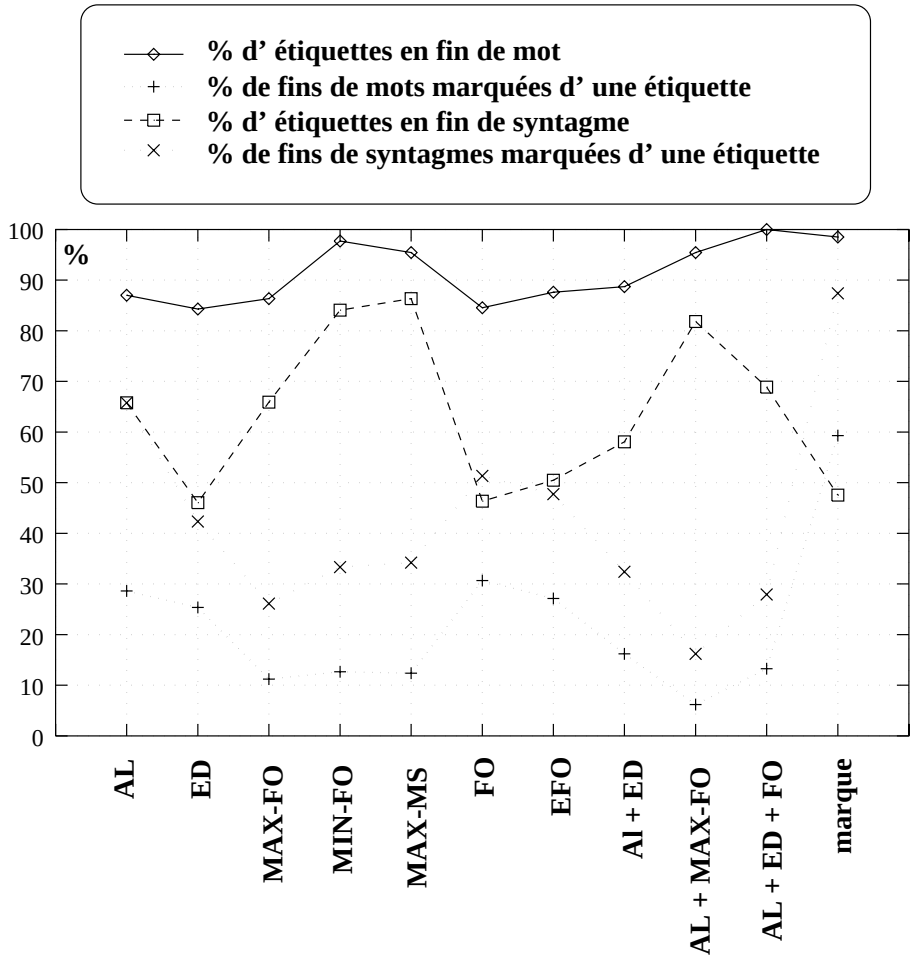
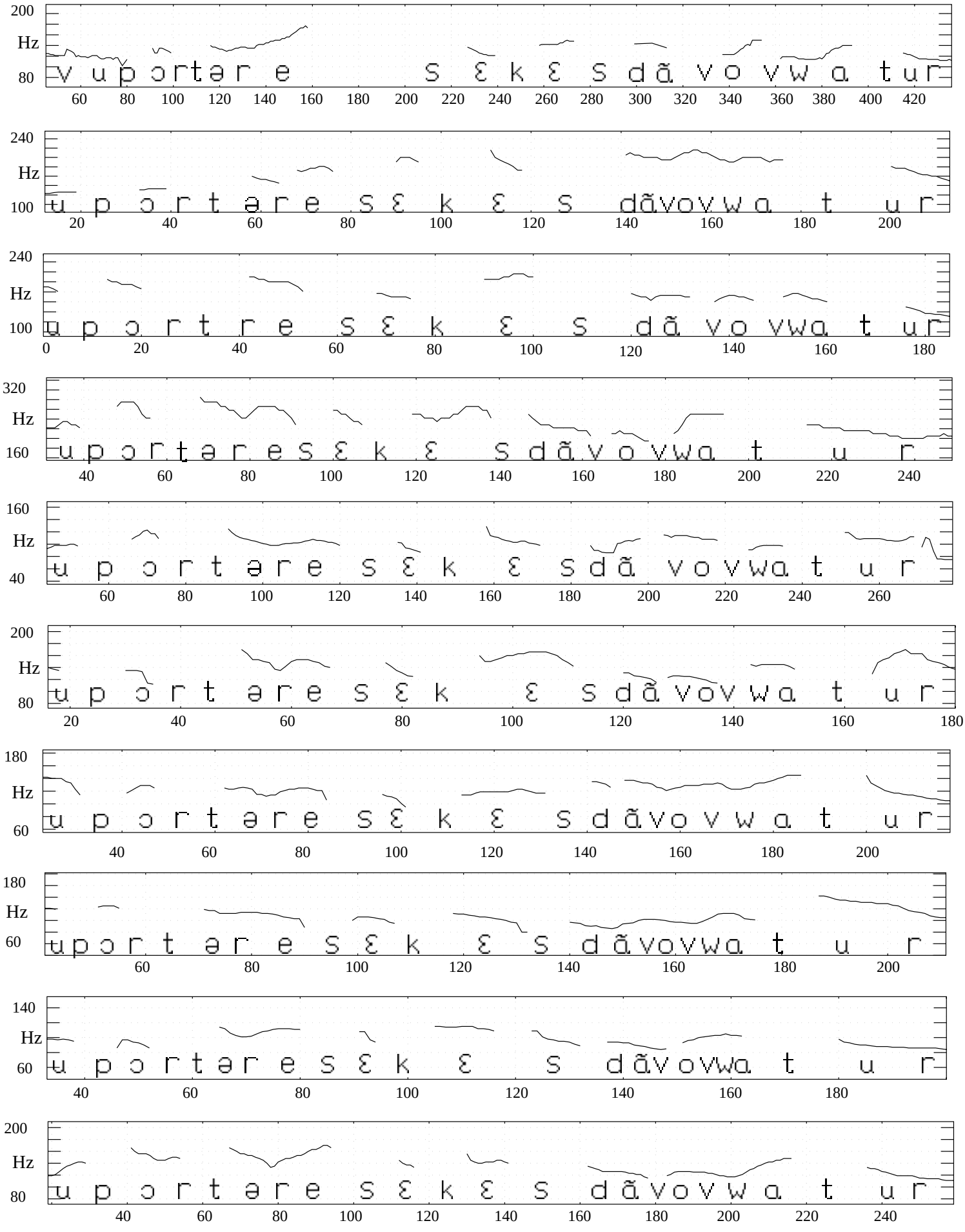


Figure 5.23: Pourcentages de corrélation entre des configurations d'indices prosodiques et les positions finales de mots et de syntagmes.


200
Hz
80
v u p ɔ r t ə r e s ɛ k ɛ s dã v o v w a t u r
60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400 420
240
Hz
100
u p ɔ r t ə r e s ɛ k ɛ s dãvovwa t u r
20 40 60 80 100 120 140 160 180 200
240
Hz
100
u p ɔ r t r e s ɛ k ɛ s dã v o vwa t u r
0 20 40 60 80 100 120 140 160 180
320
Hz
160
u p ɔ r t ə r e s ɛ k ɛ s d ã v o vwa t u r
40 60 80 100 120 140 160 180 200 220 240
160
Hz
40
u p ɔ r t ə r e s ɛ k ɛ s d ã v o v wa t u r
60 80 100 120 140 160 180 200 220 240 260
200
Hz
80
u p ɔ r t ə r e s ɛ k ɛ s d ã v o v w a t u r
20 40 60 80 100 120 140 160 180
180
Hz
60
u p ɔ r t ə r e s ɛ k ɛ s d ã v o v w a t u r
40 60 80 100 120 140 160 180 200
180
Hz
60
u p ɔ r t ə r e s ɛ k ɛ s d ã v o v w a t u r
60 80 100 120 140 160 180 200
140
Hz
60
u p ɔ r t ə r e s ɛ k ɛ s dãvovwa t u r
40 60 80 100 120 140 160 180
200
Hz
80
u p ɔ r t ə r e s ɛ k ɛ s d ã v o v w a t u r
40 60 80 100 120 140 160 180 200 220 240

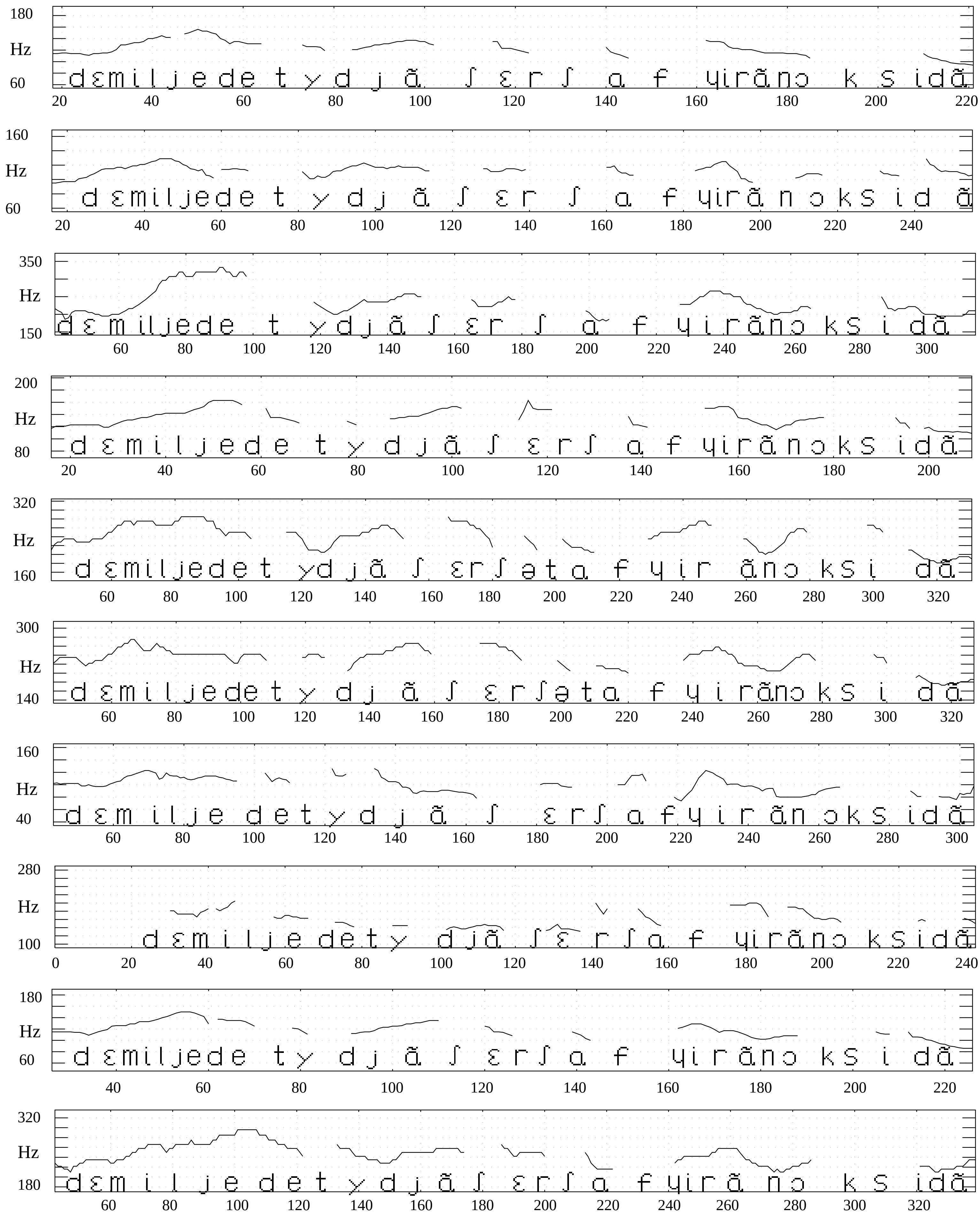
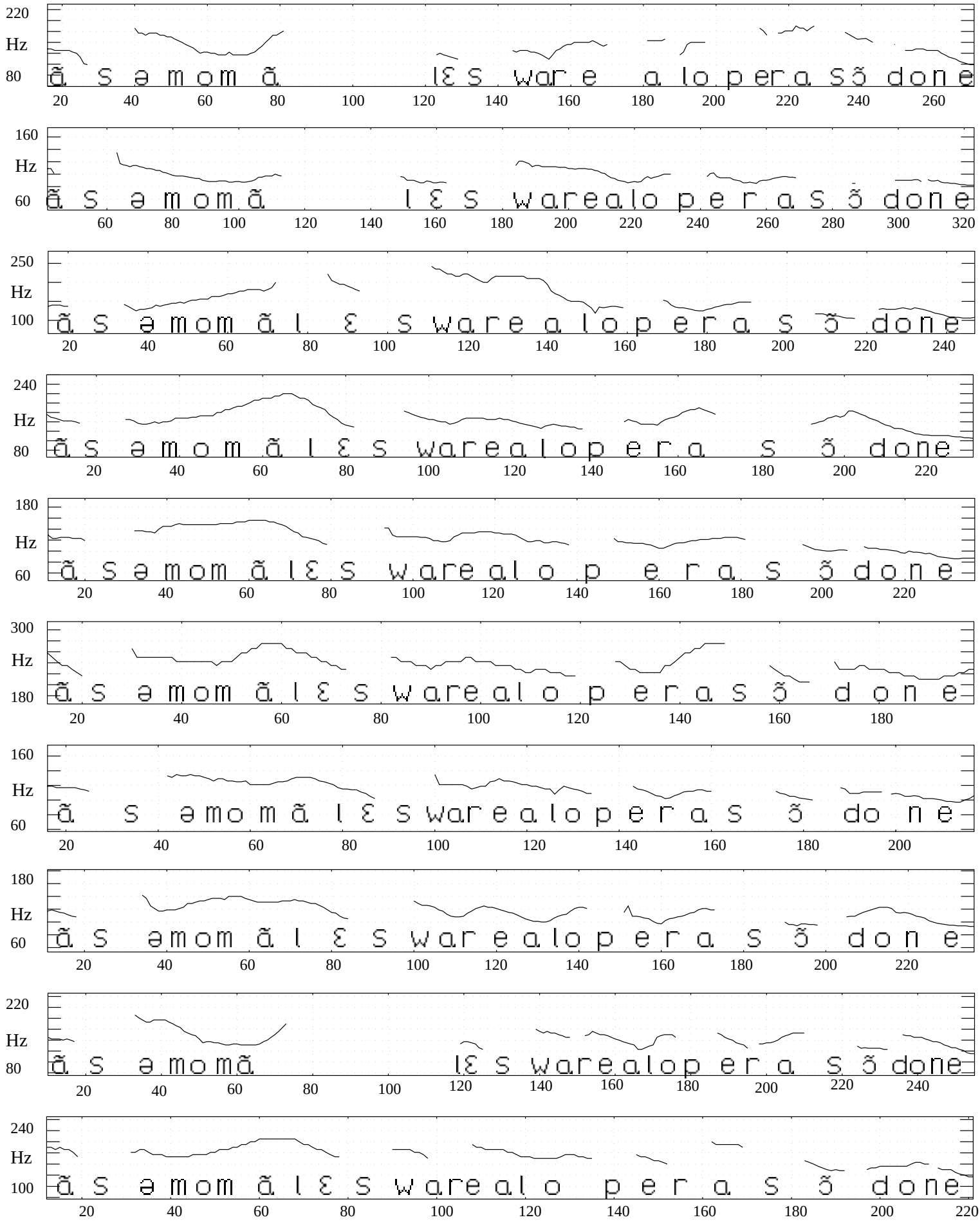


Figure 5.35: Courbes de fréquence fondamentale de différentes réalisations de la ph


220
Hz
80
ã s ə m o m ã lɛ s wa r e a l o p e r a s õ d o n e
20 40 60 80 100 120 140 160 180 200 220 240 260
160
Hz
60
ã s ə m o m ã l ɛ s w a r e a l o p e r a s õ d o n e
60 80 100 120 140 160 180 200 220 240 260 280 300 320
250
Hz
100
ã s ə m o m ã l ɛ s w a r e a l o p e r a s õ d o n e
20 40 60 80 100 120 140 160 180 200 220 240
240
Hz
80
ã s ə m o m ã l ɛ s w a r e a l o p e r a s õ d o n e
20 40 60 80 100 120 140 160 180 200 220
180
Hz
60
ã s ə m o m ã l ɛ s w a r e a l o p e r a s õ d o n e
20 40 60 80 100 120 140 160 180 200 220
300
Hz
180
ã s ə m o m ã l ɛ s w a r e a l o p e r a s õ d o n e
20 40 60 80 100 120 140 160 180
160
Hz
60
ã s ə m o m ã l ɛ s w a r e a l o p e r a s õ d o n e
20 40 60 80 100 120 140 160 180 200
180
Hz
60
ã s ə m o m ã l ɛ s w a r e a l o p e r a s õ d o n e
20 40 60 80 100 120 140 160 180 200 220
220
Hz
80
ã s ə m o m ã lɛ s w a r e a l o p e r a s õ d o n e
20 40 60 80 100 120 140 160 180 200 220 240
240
Hz
100
ã s ə m o m ã l ɛ s w a r e a l o p e r a s õ d o n e
20 40 60 80 100 120 140 160 180 200 220

Les résultats

À l'instar des nombres, nous allons maintenant vérifier l'aptitude du système ProStat à classer et prédire des hypothèses syntaxiques et/ou rythmiques à l'aide du graphe d'apprentissage dont nous avons précédemment décrit les caractéristiques principales. Notre intuition est que l'information prosodique contenue dans les phrases doit être plus importante que celle contenue dans les nombres et que notre système devrait pour cela obtenir des résultats meilleurs que ceux recueillis pour les nombres. Nous allons vérifier cela dans une première expérience où il a été demandé à notre système de proposer un classement structurel (rythmique et syntaxique) des données d'apprentissage et ce afin de vérifier que l'information prosodique captée par notre système est au moins capable de classer les données sur lesquelles les informations prosodiques ont été mesurées. Les résultats de ce test sont consignés sur la figure 5.27. Il apparaît très clairement que le système est apte à classer en première position plus de 90% des 500 observations du corpus PolyPhrase (le nombre moyen de classes pour ce corpus est proche de 15). Notons qu'une analyse sommaire des observations non classées en tête nous a permis de constater que les arbres syntaxiques obtenus manuellement ne sont pas toujours homogènes dans le choix des symboles. Particulièrement, la distinction entretenue par les symboles $LIEU$ — complément circonstanciel de lieu — et $CIRC$ — complément circonstanciel de manière ou de temps — n'est pas maintenue par l'organisation prosodique et contribue à la dispersion des résultats. Notre objectif premier étant de valider notre approche, il nous a semblé cependant inutile d'optimiser les entrées du système ProStat pour améliorer les résultats qui nous conviennent pleinement en l'état.

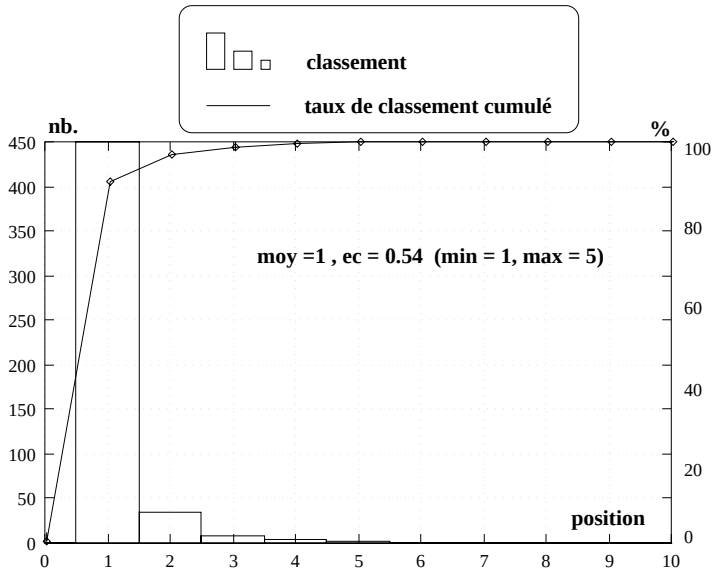


Figure 5.27: Classement des hypothèses fournies par le système ProStat pour les 500 phrases du corpus PolyPhrase.

La second test présente cette fois-ci les résultats obtenus pour le corpus de test PolyPhraseTest

en demandant toujours au système ProStat de se prononcer sur des structures syntaxico-rythmiques complètes (*i.e.* parmi les feuilles du graphe). On y observe qu'une fois de plus les observations classées en tête sont largement majoritaires (plus de 55 %) bien qu'en proportion moindre. On notera que sur les 348 phrases du corpus de test, 301 ont leur structure syntaxico-rythmique modélisée dans le graphe d'apprentissage et que le nombre moyen de classes possibles pour une observation est toujours voisin de 15. Afin de pouvoir apprécier ces résultats, nous reportons sur la figure 5.29 le classement obtenu par notre système en attribuant à chaque P-nœud une note aléatoire.

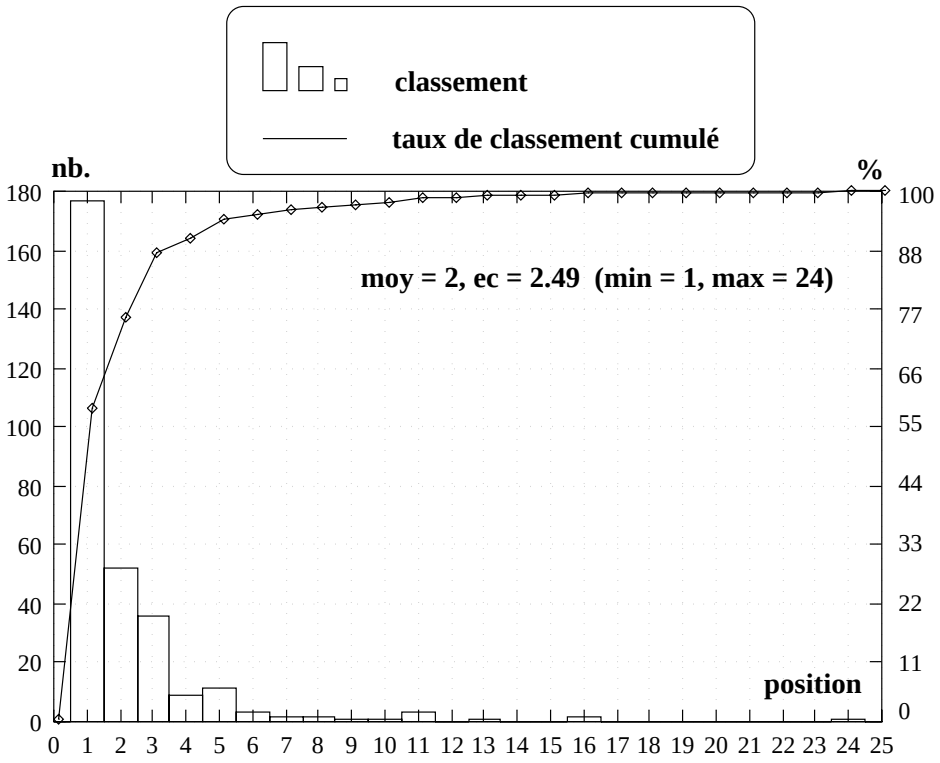


Figure 5.28: Classement des hypothèses fournies par le système ProStat pour les 301 phrases du corpus PolyPhraseTest dont la structure syntaxico-rythmique est présente dans le graphe d'apprentissage.

Dans une dernière expérience (voir la figure 5.30), nous avons demandé à notre système de classer les observations du corpus PolyPhraseTest sur la simple information contenue en début et fin de syntagme de surface sans tenir compte de la nature exacte du syntagme c'est-à-dire en ne retenant comme critère que le nombre de voyelles de chaque groupe de surface comme dans l'exemple suivant :

$$PH \begin{bmatrix} SS & (5) \\ SV & (4) \\ CIRC & (3) \end{bmatrix} \equiv PH \begin{bmatrix} SS & (5) \\ CIRC & (4) \\ SV & (3) \end{bmatrix}$$

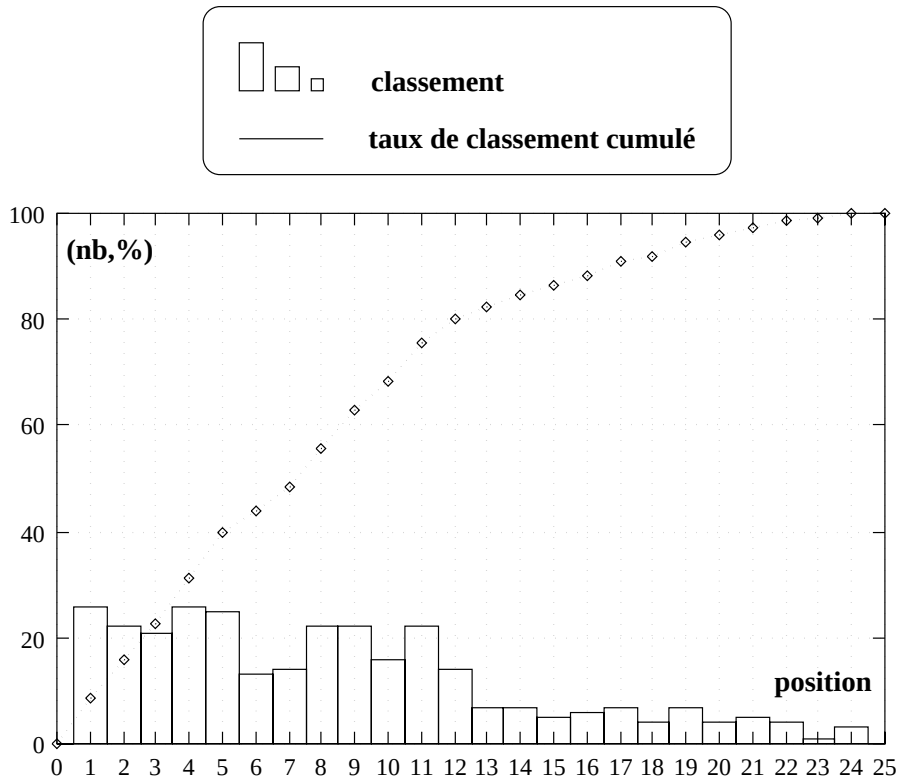


Figure 5.29: Classement des hypothèses fournies par le système ProStat avec une notation aléatoire pour les 301 phrases du corpus PolyPhraseTest dont la structure syntaxico-rythmique est présente dans le graphe d'apprentissage.

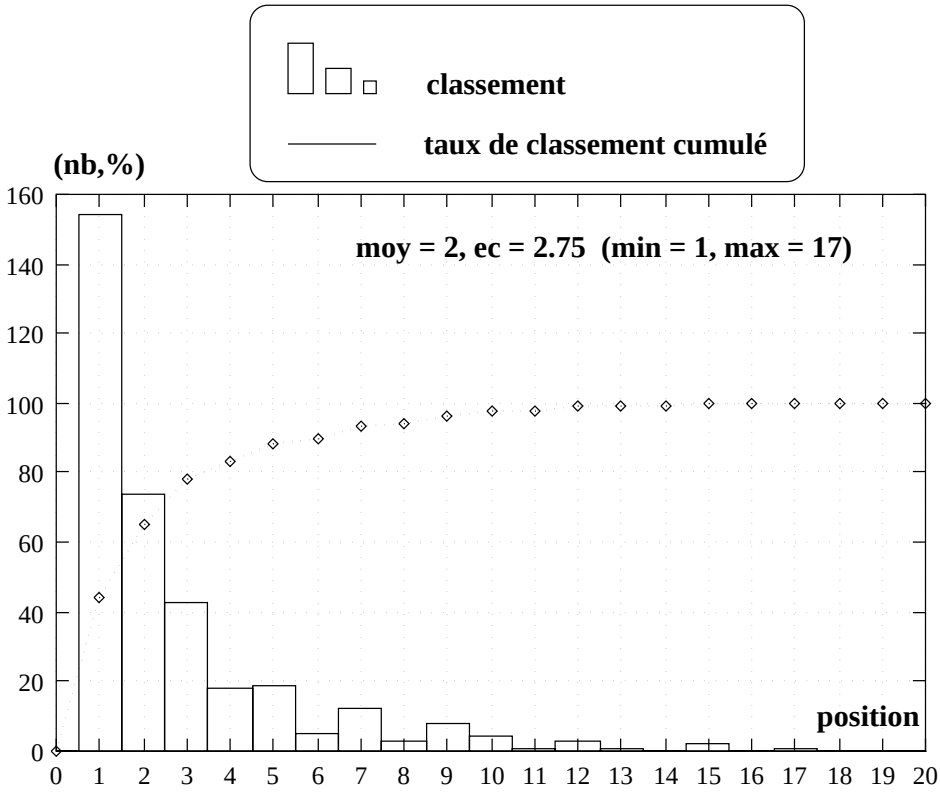


Figure 5.30: Classement des hypothèses fournies par le système ProStat pour les observations du corpus PolyPhraseTest en ne considérant que l'information localisée à l'initiale et en finale de syntagme de surface sans tenir compte de la nature exacte des syntagmes.

L'analyse des valeurs reportées sur la figure 5.30 permet de constater que les résultats sont sensiblement moins bons que ceux présentés auparavant puisque moins de 45% des observations sont classées en tête contre près de 60% dans l'expérience précédente. Ceci s'explique normalement par le faible nombre de voyelles prises en compte pour l'attribution de la note d'une observation (*ex.* : dans le cas d'une observation contenant deux syntagmes de surface, au plus quatre voyelles seront prises en compte au moment de la notation). Ceci semble indiquer qu'une notation prosodique globale de la phrase est préférable à une notation localisée en des points particuliers d'une observation.

5.6 Des améliorations possibles

Comme nous l'avons dit au cours des différentes expériences que nous venons de présenter, les résultats obtenus n'ont été l'objet d'aucune investigation particulière visant à les améliorer et ceci principalement en raison de nos objectifs premiers. Il faut cependant bien avouer qu'une intégration de notre système dans une application particulière gagnerait certainement en robustesse si nous prenions soin de concentrer nos efforts sur les

quelques points suivants :

- Une première amélioration que nous pourrions apporter à notre système serait de nous interroger sur la pertinence des indices prosodiques que nous mesurons depuis le signal de parole. Une procédure automatique itérative pourrait pour cela ajuster les coefficients α_i — coefficients de pondération des indices prosodiques qui pour le moment sont tous égaux à l'unité — en optimisant les résultats que l'on obtiendrait sur une base de tests dédiée.

- Dans un deuxième temps, le calcul d'autres indices pourrait alors être envisagé comme par exemple le passage à la ligne de déclinaison présenté récemment par Vaissière [158]. Notons à ce propos qu'il serait intéressant d'observer l'incidence d'un indice particulier sur les résultats obtenus par notre système ProStat.

- Une autre amélioration pourrait consister à prendre en considération non plus simplement les seules informations localisées en début et en fin de groupe mais également celles localisées en milieu de groupe.

- Un apprentissage sur un plus grand nombre de données ne pourrait qu'être bénéfique aux performances de notre système.

Enfin et pour conclure sur les améliorations à apporter à ProStat, il est sage de remarquer que tout système possède des défauts et que la liste que nous venons de dresser pourrait fort bien s'étendre davantage encore.... Ainsi la prise en compte d'informations sémantiques complètement écartées — faute de compétences en la matière — de ce travail ou encore l'étude d'un algorithme de syllabation automatique qui nous permettrait de ne plus baser notre étude sur de simples voyelles seraient certainement des points intéressants à traiter. Précisons cependant qu'il ne s'agit pas là d'une simple recherche d'optimisation de l'existant comme le sont les points de la précédente liste.

Chapitre 6

Bilan du travail présenté et perspectives

Il est maintenant temps de dresser un bilan — si possible critique — de ce premier contact avec la prosodie que nous venons de présenter.

Nous avons débuté notre étude par la présentation des outils développés pour l'étiquetage prosodique automatique. Le choix des indices prosodiques retenus est certainement critiquable, mais le caractère ouvert de notre système ne s'oppose aucunement à l'intégration de nouveaux indices pourvu qu'ils soient calculables automatiquement. Il n'en reste pas moins que la restriction du champ d'investigation du système aux seules voyelles ne nous satisfait pas au vu des nombreux travaux qui présentent la syllabe comme l'unité privilégiée de l'étude prosodique.

Nous nous sommes ensuite attachés à étudier les phénomènes microprosodiques en réalisant l'étude de plusieurs corpus de mots isolés, certains d'entre eux ayant été spécifiquement enregistrés pour les besoins de cette analyse. Le bilan de ce travail montre qu'une très petite partie des phénomènes couramment décrits dans les études concernées sont observables de manière significative par les techniques que nous avons utilisées. Les informations les plus pertinentes (majoritairement le voisement obtenu lors du calcul de la courbe de fréquence fondamentale) ont été intégrées à un module d'accès lexical dans le cadre d'une application à la reconnaissance de mots prononcés isolément, améliorant ainsi sensiblement ses performances. À la suite de cette étude, notre position sera cependant d'appeler à la plus grande prudence quant à l'intégration d'informations prosodiques segmentales dans une composante de reconnaissance automatique de la parole. En particulier, l'opération courante qui consiste à éliminer des paramètres prosodiques bruts, les variations à caractère microprosodique à l'aide de coefficients de pondération intrinsèques et co-intrinsèques ne nous semble pas pertinente dans une approche automatique.

Nous avons également vérifié que, pour le français, et *a contrario* de la langue anglaise, la prise en compte d'informations prosodiques suprasegmentales n'apportait que peu ou pas d'amélioration pour la réduction des cohortes de mots toujours dans le cadre restreint de la reconnaissance de mots prononcés isolément.

Dans la dernière partie de ce mémoire, nous avons présenté un système simple d'appren-

tissage à partir d'exemples capable de réaliser automatiquement une étude corrélative entre un ensemble d'indices prosodiques et différents niveaux d'organisation du message. Nous avons démontré l'efficacité de ce système à appréhender des contraintes structurelles multiples (ici syntaxiques et rythmiques) souvent en conflit et dont il est difficile de proposer une hiérarchisation sur la base de simples règles. Une validation du système a été proposée par l'étude de deux corpus multi-locuteurs enregistrés à travers le réseau téléphonique : le premier composé de nombres décimaux et le second constitué de phrases isolées aux structures syntaxiques simples. Nous avons en particulier montré l'aptitude de notre système à prendre en compte des informations prosodiques non formalisées en comparant les taux de classement des hypothèses syntaxico-rythmiques proposées par le système sur des corpus de tests spécifiques, avec un classement obtenu de manière aléatoire. Nous avons ensuite présenté quelques améliorations possibles de ce système en vue d'une intégration robuste dans de futurs systèmes de reconnaissance de la parole.

Quittons maintenant la première personne du pluriel usitée tout au long de ce mémoire — et qui traduit bien le fait qu'il n'est pas le fruit d'un travail isolé mais au contraire d'une collaboration étroite entre un étudiant et deux laboratoires : le LIUAPV et l'IDIAP — pour retrouver la première personne du singulier employée au tout début de l'introduction ; ceci afin de m'éloigner légèrement des travaux présentés et dresser un bilan plus personnel de ce mémoire.

Durant une période non négligeable de mes travaux, j'ai considéré la prosodie comme une source de connaissances trop peu fiable pour être utilisée en reconnaissance de la parole autrement que dans des applications à caractère très limité. Ceci explique par exemple le soin que j'ai pris dans cette étude à montrer la faiblesse des informations microprosodiques — lorsqu'elles sont appréhendées par des techniques automatiques — me permettant ainsi de renouer avec les nombreuses observations des divers corpus de phrases dans lesquelles je ne retrouvais que peu des phénomènes intrinsèques et co-intrinsèques répertoriés. J'ai ensuite commencé à mettre en place des systèmes de reconnaissance aux prétentions réduites à l'aide d'outils disponibles au LIUAPV et à l'IDIAP et tenté d'y ajouter des informations prosodiques formalisées par mes soins. Les résultats que j'obtenais alors — s'ils étaient souvent meilleurs avec la prise en compte de la prosodie — relevaient davantage d'ajustements spécifiques aux corpus étudiés et ne constituaient en aucune façon une preuve de la validité des informations prosodiques en reconnaissance de la parole. Il faut avouer que cette difficulté à formaliser des règles prosodiques à caractère général m'a conforté dans mon impression initiale. Dès lors, j'ai compris qu'il était vain de vouloir décrire les phénomènes prosodiques "possibles" par un système de règles cohérent et qu'une solution au problème pouvait être la réalisation d'un système ne possédant aucun *a priori* et se contentant simplement d'étudier statistiquement les corrélations entre des indices prosodiques mesurés depuis le signal de parole et divers niveaux d'organisation. Étant persuadé — expérience à l'appui — qu'aucun des indices prosodiques mesurés depuis le signal de parole — par les techniques que j'ai présentées — ni aucune combinaison de ces indices ne permettait d'émettre d'hypothèse structurelle fiable à l'endroit où ils étaient observés, j'ai pensé qu'une prise en compte de l'information prosodique sur toute la phrase répondant à des contraintes non observables directement pouvait peut-être réconcilier

à mes yeux la prosodie et la reconnaissance automatique de la parole. J'ai développé pour cela le système ProStat en le rendant le plus général possible et en le dotant d'une interface de visualisation et d'interrogation afin d'en faire un outil d'investigation simple à utiliser. C'est en tentant d'extraire des connaissances de ce système sur des bases de parole particulières, que l'idée m'est venue d'utiliser ce système de manière complètement automatique. J'ai donc doté ProStat d'une métrique simple l'autorisant à formuler des hypothèses structurelles à partir d'un treillis d'indices prosodiques. Les résultats que j'ai présentés ont eu raison de mon scepticisme puisqu'ils étaient le fruit de connaissances que je n'avais pas formalisées. Mon opinion quant à la prosodie étant maintenant plus claire et disposant d'une plate-forme d'étude me satisfaisant, il reste de nombreux points que j'aimerais traiter ultérieurement et dont je décris brièvement les plus importants.

- J'aimerais dans un premier temps utiliser le système ProStat à des fins très locales comme la désambiguïsation ou la prédiction de la modalité d'une phrase, ceci afin de comparer les résultats avec ceux de la littérature consacrée [167, 111, 74] et tenter d'extraire de l'information pertinente à partir des situations d'échec ;

- Une deuxième direction que j'aimerais suivre serait d'intégrer ProStat dans un système de reconnaissance complet afin de pouvoir évaluer précisément sa contribution. Un travail est actuellement entrepris dans ce sens avec le système $\mathsf{ETC}_{\mathsf{vérif}}$ [29] qui, par son architecture multi-agents, se prête bien à l'ajout de composantes de natures diverses.

- J'aimerais également étendre les compétences du système ProStat à d'autres niveaux d'organisation linguistique et essentiellement la sémantique ; cela ne pourra se faire que dans le cadre d'une application aux prétentions bien délimitées.

- Enfin, la prosodie intervient dans la résolution de nombreux problèmes (qui sortent cependant du cadre limité des travaux présentés) comme le traitement automatique de la parole spontanée, les situations de dialogue, l'identification du locuteur, la détection de mots clés dans la parole continue, *etc.*

J'espère que le travail présenté, s'il n'apporte pas de réponse définitive, permettra cependant — par son regard "applicatif" — de montrer les limites et les perspectives d'utilisation de la prosodie en reconnaissance automatique de la parole.

Remerciements

Parce qu'une thèse n'est pas seulement le fruit de quelques années studieuses[1], je souhaite exprimer mes remerciements les plus sincères à ceux qui m'ont accompagné pendant toute cette période.

⋈ ⋈ ⋈

Je sais pertinemment que ce ne sont pas quelques lignes rapidement écrites à la fin d'un mémoire qui effaceront les absences dues à mon sens familial peu aiguisé, mais mes premiers remerciements vont tout naturellement à mes parents et à mon frère qui connaissent mon attachement bien au-delà des mots. Je vais pouvoir annoncer à mon père, qui s'inquiétait de voir son fils étudier à cet âge, que j'ai enfin fini !

⋈ ⋈ ⋈

J'aimerais ensuite exprimer toute ma gratitude à un petit groupe d'amis qui sont les premiers artisans de ce mémoire.

À Henri et Claudie Méloni qui sont venus jusqu'en Suisse, sacrifiant pour la coutume locale foie et estomac pour un peu de fendant et davantage de chocolat, j'exprime ici toute ma reconnaissance pour avoir su me motiver à un moment où j'étais proche de baisser les bras. Ce soutien concrétisé à chacun de mes retours dans la cité papale m'a été une aide appréciable.

Je tiens à remercier tout particulièrement Gérard Chollet, sans la présence de qui je n'aurais pas pu passer ces deux magnifiques années dans ce havre de paix, peuplé de vaches mauves, qu'est le Valais. À ses côtés, j'ai pu me familiariser avec des techniques nouvelles pour moi telles que le surf et les chaînes markoviennes et j'ai réussi à prendre le recul nécessaire à la rédaction de ce mémoire. Je ne saurais trop le remercier de l'impressionnante littérature qu'il m'a transmise durant toute cette période et qui a contribué grandement à élargir mes horizons prosodiques, ainsi que de la constante préoccupation qu'il a eue à me voir terminer mon mémoire.

À Jean-Luc Cochard qui n'a pas eu à me préciser sa ressemblance avec Kevin Costner tellement elle est frappante, je souhaite exprimer ma profonde gratitude pour la confiance

[1] Le nombre d'années pouvant varier sensiblement d'un individu à l'autre !

qu'il m'a toujours témoignée, qui, si je ne lui trouve toujours pas d'explication, n'en est pas moins touchante ! Je me dois de le remercier également chaleureusement pour l'intérêt qu'il a manifesté à mes travaux ainsi que pour son assistance à l'utilisation de LaTeX qui m'a été précieuse jusqu'au dernier moment de la rédaction.

⋈ ⋈ ⋈

Je souhaite également remercier Marc El-Bèze pour la simplicité, la disponibilité et la spontanéité qu'il a toujours à mon égard à chacune de ses explications.

Ma visite à l'institut de phonétique de Paris dirigé par Jacqueline Vaissière reste un souvenir agréable ; j'y ai rencontré une personne qui n'a pas hésité à réserver une journée pour m'expliquer en toute sympathie son approche de la prosodie. Si je n'ai pas su entretenir cette relation privilégiée, je peux tout de même la remercier chaleureusement de son aide et de son dynamisme que je retrouve dans ses articles.

Je souhaite aussi remercier les membres de l'institut de phonétique d'Aix et plus particulièrement Pascale Nicolas, Albert Di Cristo et Daniel Hirst ; nos rencontres ont toujours été très agréables et je ne peux que regretter qu'elles n'aient pas été plus fréquentes.

Enfin, je remercie également Gérard Bailly et Piet Mertens pour leurs conseils et leurs encouragements spontanés lors des $19^{èmes}$ Journées d'Étude sur la Parole de Bruxelles.

⋈ ⋈ ⋈

Partir c'est s'offrir le plaisir égoïste de fuir son quotidien le temps de reprendre en d'autres lieux, avec d'autres personnes, les mêmes habitudes. Je souhaite exprimer ici mes remerciements à la toujours grandissante (et hélas toujours majoritairement masculine) équipe de l'IDIAP...

À Georg, enfant de la Forêt noire, système damager et professeur de Go émérite,

à Gilbert le montagnard qui m'a éclairé de ses conseils mathématiques,

à Olivier, mon compère "caviste",

à Martine qui a préféré l'air pur de la banlieue parisienne à l'air pollué du Valais,

à Cathy, ex-système damager mais néanmoins valaisanne,

à Magali, partenaire officielle de Hagen-Dazs,

à Cédric et Anne ...sans commentaire,

à Stéphane, récidiviste remarqué, adepte de la légèreté à temps partiel,

à Dominique, éleveur de kiwis valaisans,

ainsi qu'à tous les intermittents du spectacle : JP, Hassan et Colette, Patrick, Christophe, Éric, Indu, Perry le cycliste, Miguel (tu do bem), Guig, Philippe (I et II), Gilles, Dilia,

Sandrine, Hubert et ses patins à roulettes, Andrei, Andrea, Robert, Murielle, David, Bert et Jerœn,
de cœur un grand merci.

Des liens que je tiendrai secrets, associent également l'achèvement de ce mémoire à la présence de l'*ISO team*, en les personnes de Patricia, Imelda et Brigitte.

Je remercie également Anne-Marie au nom de ma sympathie pour l'ensemble de sa famille grandissante.

Ne pas remercier Migros et son self-service serait pour moi nier avec une mauvaise foi évidente mes attaches suisses, les 300 demi-poulets, et autres plats ingurgités tout au long de ces deux années, les sourires échangés ...

Je n'oublie pas non plus les joyeux compagnons du CREM qui ont contribué à mon intégration réussie au pays de la viande séchée, du fendant, des croûtes au fromage, des fondues et autres spécialités légères.

Enfin, je remercie également Fabrice, ex-douanier des portes de la Forclaz, pour son amitié débordante.

⋈ ⋈ ⋈

Partir c'est également s'apercevoir avec émotion, qu'à chacun de vos retours au pays, un groupe d'amis est là qui vous témoigne son affection.

À Fred, acteur principal du film "*l'aventure c'est l'aventure*" réalisé pour le bien de tous et financé par le fonds de réserve liracois,
à Philou, dans le double rôle du barbu et du savant[2],
à Thynou, dans le double rôle de Dom-Juan des côtes du Rhône et de membre actif de l'association "*Accueil aux sans abris helvetico-avignonnais*",
à Zeff, dans le rôle de l'homme des Zairs, qui a assuré seul toutes les cascades aériennes,
à Jean-Jacques, dans le rôle du gentleman-farmer ardéchois, étalon du prêt-à-porter,
à Yves K., dans le rôle du skipper du Bénodet team,
à Corinne, dans le rôle de victime volontaire de la SEPR,
à Marie-Luce, la Florence Arthaud du Pont-des-deux-Zeaux,
à Antoun, Ferrouze et leurs trois enfants, dans le rôle de la famille *Trèjantille*,
à Abbas, dans le rôle de l'homme ténébreux,
à Pierrot (dit *Pierrot l'embrouille*), dans le rôle de l'homme de frappe du LIUAPV,
à Stéphane, dans les rôles d'objecteur sans conscience et d'apprenti dragueur,

[2]De ses propres paroles, il préfère être surnommé le Quid que la rousse ...

à Laurence, jeune talent, **très** remarquée dans la scène torride où elle conduit sa Ducat,

à Mariette, dans le rôle de la femme fatale,

à Vava et Dédou, dans le nouveau rôle de parents,

à Paquita, dans le rôle de la danseuse latino-américaine mais juste un petit peu ...,

à Loule, dans le rôle du patron du Koala,

à Madeleine, dans le rôle de la découvreuse de talents ...,

à Patoun, dans le rôle de Alain Gabé, motard à l'estomac tendre et au cœur gros comme ça (à moins que ce ne soit le contraire ...),

à Éric, dans le rôle du plongeur et à Thérèse sa femme,

ainsi qu'à tous les figurants, mais non moins appréciés, du *Philou's club* (Nana et son pédalo, Domi et son footballeur Philippe, Édith et Serge, Joëlle et Vincent, Pascal ...), du *Lirac's club* (Jean-Claude animateur qu'on ne présente plus, Renaud, Gilles, François, Alex, Schwa, Gérard l'éthéré, Warda, Soraya, Jean-Christophe, le frêle Manu et sa compagne Odile) et les autres (Bernadette, Corinne, Magali, Cathy et Isaac),

à tous un très grand merci pour ce film aux rebondissements multiples.

J'allais oublier la musique : elle a été écrite et jouée par Jean-Marc Boï ("*Mes années Rock-en-Roll*" distribué chez *JMB production*).

Je remercie enfin Fantuz Moto qui a assuré — sans compter la dépense — la mécanique capricieuse de feu mon bolide dont j'ai fini par avoir raison.

⋈ ⋈ ⋈

Mes amitiés au nancéien Vincent, apprenti-prosodicien, spécialiste de la crème jaune, guitariste à ses heures bleues et dont la sympathie pressentie par e-mail m'a été largement confirmée lors du congrès de phonétique de Stockholm.

⋈ ⋈ ⋈

Et puis, si dans cet élan spontané, je commettais l'indélicatesse d'en oublier certains, il ne me resterait plus qu'à m'incliner, ne pouvant faire mieux que de déplorer qu'une thèse ne soit pas 200 pages de remerciements et quelques pages de résumé.

Bibliographie

[1] J. Allen. Synthesis of speech from unrestricted text. *IEEE*, 64:422–433, 1976.

[2] C. Astesano, A. Di Cristo, et D.J. Hirst. Discourse-based empirical evidence for a multi-class accent system in french. Dans *XIIIth International Congress of Phonetic Sciences*, volume 4, pages 630–633, Stockholm, 13–19 August 1995.

[3] B.S. Atal. Efficient coding of lpc parameters by temporal decomposition. Dans *IEEE-ICASSP*, pages 81–84, 1983.

[4] Véronique Aubergé. *La synthèse de la parole : "des règles au lexique"*. Thèse, ICP / CRISS, Grenoble, 1992.

[5] Gérard Bailly. Un analyseur interactif adapté à la génération de la prosodie : limites et perspectives. Séminaire Prosodie et Reconnaissance, Octobre 1982.

[6] Gérard Bailly, Thierry Barbe, et Haïdong Wang. Automatic labelling of large prosodic databases : tools, methodology, and links with a text-to-speech system. Dans *Talking Machines: Theories, Models, and Designs*, G. Bailly, C. Benoît, et T.R. Sawallis, éditeurs, pages 323–333. Elsevier Science Publishers, 1992.

[7] J. Baker. Dragon dictate-30k: Natural language speech recognition with 30,000 words. Dans *Eurospeech*, volume 2, pages 161–163, September 1989.

[8] Marie-Hélène Banel et Nicole Bacri. On metrical patterns and lexical parsing in french. Dans *Speech Communication*, volume 15, pages 115–126. Elsevier Science B.V., 1994.

[9] Thierry Barbe et Marité Janot-Giorgetti. évaluation d'un détecteur de f0 sur des voix normales et sur des voix pathologiques. Séminaire d'intelligence artificielle de Luminy, 1989.

[10] Plínio Barbosa et Gérard Bailly. Characterization of rythmic patterns for text-to-speech synthesis. Dans *Speech Communication*, volume 15, pages 127–137, 1994.

[11] Frédéric Béchet. *Système de traitement de connaissances phonétiques et lexicales : Application à la reconnaissance de mots isolés sur de grands vocabulaires et à la recherche de mots cibles dans un discours continu.* Thèse, Université d'Avignon et des Pays de Vaucluse, janvier 1994.

[12] Frédéric Bimbot, Gérard Chollet, Paul Deléglise, et Claude Montacié. Temporal decomposition and acoustic-phonetic decoding of speech. Dans *IEEE-ICASSP*, pages 445–448, New York, April 1988.

[13] Eleonora Blaauw. The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. Dans *Speech Communication*, Elsevier Science B.V., éditeur, volume 14, pages 359–375, 1994.

[14] B.A. Blesser. *Perception of spectrally rotated speech.* Thèse, Massachusetts Institute of Technology, 1969.

[15] Louis-Jean Boë. *étude de l'interaction source laryngienne - conduit vocal dans la détermination des caractéristiques intrinsèques des consonnes du français.* Bulletin de l'Institut de Phonétique de Grenoble, 1973.

[16] Jean-Jacques Bonin et Jean-Marie Pierrel. Fréquence fondamentale et durée pour la détection de frontières syntagmatiques en parole continue. Dans *XVIIIème Journées d'étude sur la Parole*, pages 21–25, Montréal, 28-31 mai 1990.

[17] J. Breckenridge et M. Liberman. The declination effect in perception. Bell Laboratories Publ. Mimeographed, 1977.

[18] L. Breiman, H.J. Friedman, R.A. Olshen, et C.J. Stone. *Classification and regression trees.* Wadsworth & Brooks, Pacific Grove, CA, 1984.

[19] Rémi Bulot. *Techniques d'intelligence artificielle pour la reconnaissance de la parole, application au décodage acoustico phonétique.* Thèse, Université Aix-Marseille II, 1987.

[20] G. Caelen. Une représentation syntaxique adaptée à la prosodie. Dans *Journal d'acoustique*, volume 2, pages 137–146, 1989.

[21] Geneviève Caelen-Haumont. Stratégies des locuteurs en réponse à des consignes de lecture d'un texte : analyse des interactions entre modèles syntaxiques, sémantiques, pragmatique et paramètres prosodiques. Thèse de doctorat d'état - spécialité ès lettres, Institut de la Communication Parlée, Grenoble, 29 novembre 1991. Soutenue à Aix-en-Provence.

[22] W.N. Campbell. Prosodic encoding of english speech. Dans *International Conference on Spoken Language Processing*, pages 663–666, Banff, Canada, 1992.

[23] N. Carbonell et J.J. Bonin. Détection de frontières syntagmatiques en parole continue : utilisation de la fréquence fondamentale. Dans *17ème JEP*, pages 163–167. SFA, 1988.

[24] Noëlle Carbonell, Jean-Paul Haton, François Lonchamp, et Jean-Marie Pierrel. élaboration expérimentale d'indices prosodiques pour la reconnaissance ; application à l'analyse syntaxico-sémantique dans le système Myrtille II. Séminaire Prosodie et Reconnaissance d'Aix-en-Provence, Octobre 1982.

[25] René Carré. Acoustic characteristics of vowel nasalization. Quart. Progres. Rep. Res. Lab. of Electr. MIT, 1975. 270-274.

[26] M. Chen. Vowel length variation as a function of the voicing of the consonant environment. *Phonetica*, 22:129–159, 1970.

[27] M.J. Cheng, L.R. Rabiner, A.E. Rosenberg, et C.A. McGonegal. Some comparisons among several pitch detection algorithms. Dans *je n'ai pas l'entête du papier ...*, pages 332–335, 1900.

[28] Gérard Chollet, Yves Grenier, et S.M. Marcus. Temporal decomposition and non-stationary modeling of speech. Dans *Eurasip*, pages 365–368, 1986.

[29] Jean-Luc Cochard et Philippe Froixdevaux. Environnement multi-agents de reconnaissance automatique de la parole en continu. Dans *Actes des 3èmes Journées Francophones sur l'Intelligence Artificielle Distribuée et les Systèmes Multi-agents*, pages 101–110, mars 1995.

[30] Jean-François Bonastre. *Stratégie analytique orientée connaissances pour la caractérisation et l'identification du locuteur.* Thèse, Université d'Avignon et des Pays de Vaucluse, janvier 1994.

[31] François Dell. L'accentuation dans les phrases en français. Dans *Formes sonores du langage. Structure des représentations en phonologie*, Hermann, éditeur, pages 65–122. F. Dell and D. Hirst and J.R. Vergnaud, 1984.

[32] François Grosjean et Jean-Yves Dommergues. Les structures de performance en psycholinguistique. *L'Année Psychologique*, pages 513–536, 1983.

[33] François Grosjean et James Paul Gee. Prosodic structure and spoken word recognition. Dans *Cognition special issues*, A Bradford book, éditeur, pages 135–155. Elsevier Science, first MIT press edition édition, 1987. reprinted from Cognition :IJCS, Volume 25.

[34] Françoise Emerard. Les diphones et le traitement de la prosodie dans la synthèse de la parole. *Bulletin de l'institut de Phonétique de Grenoble*, VI:103–147, 1977.

[35] R. Collier. On the communicative function of prosody : some experiments. Annual Progress Report 28, IPO, 1993.

[36] W.E. Cooper et J. Paccia-Cooper. *Syntax and Speech.* Harvard Univ. Press, Cambridge, MA, 1980.

[37] Albert Di Cristo. *Soixante et dix Ans de Recherches en Prosodie. Bibliographie alphabétique, thématique et chronologique.* Université de Provence, 1975.

[38] Albert Di Cristo. Méthodes et modèles d'analyse dans les recherches sur l'intonation. Dans *Revue d'Acoustique*, 1978.

[39] Albert Di Cristo. Aspects phonétiques et phonologiques des éléments prosodiques. *Modèles linguistiques*, Tome III(Fascicule 2):24–83, 1981. Presse universitaire de Lille.

[40] Albert Di Cristo. *De la microprosodie à l'intonosyntaxe.* Université de Provence, 1985. Texte de la thèse d'état soutenue par l'auteur en décembre 1978.

[41] Albert Di Cristo et Daniel Hirst. *à paraître.*

[42] Albert Di Cristo et Daniel Hirst. Modelling french micromelody : analysis and synthesis. *Phonetica*, 43:11–30, 1986.

[43] Christophe d'Alessandro, Piet Mertens, et Frédéric Beaugendre. Automatic stylisation of intonation : application to speech synthesis. Dans *Second ESCA/IEEE workshop on Speech Synthesis*, pages 155–158, New York, September 12-15 1994.

[44] Stéphanie de Tournemire. Recherche d'une stylisation extrème des contours de f0 en vue de leur apprentissage automatique. Dans *XXèmes Journées d'étude sur la Parole*, pages 75–80, Trégastel, 1–3 juin 1994.

[45] Elisabeth Delais. Prédiction de la variabilité dans la distribution des accents et les découpages prosodiques en français. Dans *XXèmes Journées d'étude sur la Parole*, pages 379–384, Trégastel, 1–3 Juin 1994.

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

[47] Pierre Delattre. *Comparing the phonetic features of english, french, german and spanich.* Heidelberg, New York, Philadelphia, julius groos verlag édition, 1965.

[48] Denise Deshaies, Isabelle Guaïtella, et Claude Paradis. La perception de l'accent en français du Québec et en français de France. Dans *XXèmes Journées d'étude sur la Parole*, pages 81–86, Trégastel, 1–3 Juin 1994.

[49] V. Digalakis, M. Ostendorf, et J.R. Rohkicek. Fast search algorithms for connected phone recognition using the stochastic segment model. Dans *Speech and Natural Language Workshop*. DARPA, 1990.

[50] P. Dumouchel. Suprasegmental features and continuous speech recognition. Dans *ICAASP*, volume II, pages 177–180, 1994.

[51] Marc El-Bèze. *Choix d'unités appropriées et introduction de connaissances dans des modèles probabilistes pour la reconnaissance automatique de la parole.* Thèse, Paris X, IBM France, Juillet 1990.

[52] Marc El-Bèze. Les modèles de langage probabilistes : quelques domaines d'application. LIPN : Université Paris-Nord, 18 Janvier 1993. Document d'habilitation à diriger des recherches.

[53] F. Emerard et C. Benoît. Base de données prosodiques pour la synthèse de la parole. *J. Acoustique*, 1:303–307, Décembre 1988.

[54] F. Emerard, L. Mortamet, et A. Cozannet. Prosodic processing in a text-to-speech synthesis system using a database and learning procedures. Dans *Talking machines : Theories, Models and Designs*, G. Bailly, C. Benoît, et T.R Sawalis, éditeurs, pages 225–254. Elsevier Science, 1992.

[55] Zsuzsanna Fagyal. Leçon de 'déclinaison' de marguerite duras et de marguerite yourcenar. Dans *XXèmes Journées d'étude sur la Parole*, pages 511–516, Trégastel, 1–3 juin 1994.

[56] G. Faure. La description phonologique des systèmes prosodiques. *Zeitschfift für Phonetik*, 24(5):349–359, 1971.

[57] Hiroya Fujisaki. From intonation to information : analysis and interpretation of prominence in spoken japanese. Dans *International symposium on Prosody*, pages 7–18, Yokohama, Japan, September 18 1994.

[58] Hiroya Fujisaki. Modeling the generation process of f0 contours as manifestation of linguistic and paralinguistic information. Dans *XIIth International Congress of Phonetic Sciences*, Aix-en-Provence, 19–24 août 92. hors volume.

[59] Philippe Gilles. *Décodage phonétique de la parole et adaptation au locuteur.* Thèse, Université d'Avignon et des Pays de Vaucluse, 20 janvier 1993.

[60] P. Gross. Sur la place de l'intonation dans une grammaire transformationnelle. Dans *5èmes Journées d'étude sur la Parole*, volume II, pages 2–9, 1975.

[61] Isabelle Guaïtella. Détermination des dimensions rythmique et symbolique de la courbe de fréquence fondamentale en parole spontanée. Prépublication des actes du séminaire prosodie de la Baume-lès-Aix, 20 et 21 octobre 1992.

[62] Wolfgang Hess. *Pitch determination of speech signals.* Springer-Verlag, 1983.

[63] Julia Hirschberg et Pilar Prieto. Training intonational phrasing rules automatically for english and spanish text-to-speech. Dans *Second ESCA/IEEE workshop on Speech Synthesis*, pages 159–162, New York, September 12-15 1994.

[64] Daniel Hirst. Prosodie et structures de données en phonologie. Dans *Formes sonores du langage. Structure des représentations en phonologie*, Hermann, éditeur, pages 43–63. F. Dell and D. Hirst and J.R. Vergnaud, 1984.

[65] Daniel Hirst, Nancy Ide, et Jean Véronis. Coding fundamental frequency patterns for multi-lingual synthesis with INTSINT in the MULTEXT project. Dans *Second ESCA/IEEE workshop on Speech Synthesis*, pages 77–80, New York, September 12-15 1994.

[66] Daniel Hirst, Pascale Nicolas, et Robert Espesser. Coding the f0 of a continuous text in french : an experimental approach. Dans *XIIème Congrès International des Sciences Phonétiques*, volume 5, pages 234–237, 19–24 août 1991.

[67] D.J. Hirst. Prediction of prosody : an overview. Dans *Talking Machines : Theories, Models and Applications*, G. Bailly et C. Benoît (eds), éditeurs. Elsevier Science Publishers, 1992.

[68] Andrew Hunt. A generalised model for utilising prosodic information in continuous speech recognition. Dans *ICASSP*, volume II, pages 169–172, 1994.

[69] U. Jensen, R.K. Moore, P. Dalsgaard, et B. Lindberg. Modelling intonation contours at the phrase level using continuous density hidden markov models. Dans *Computer Speech and Language*, volume 8, pages 247–260, 1994.

[70] Denis Jouvet. *Reconnaissance de mots connectés indépendamment du locuteur par des méthodes statistiques.* Thèse, école Nationale Supérieure des Télécommunications, 15 Juin 1998.

[71] D.H. Klatt. Review of text-to-speech conversion for english. *Journal of the Acoustical Society of America*, 54(4):1102–1104, 1987.

[72] D.H. Klatt et W.E. Cooper. *Perception of segment duration in sentence contexts*, pages 69–86. Springer Verlag, Berlin, communication & cybernetics édition, 1975.

[73] R. Kompe, A. Batliner, A. Kießling, U. Kilian, H. Niemann, E. Nöth, et P. Regel-Brietzmann. Automatic classification of prosodically marked phrase boundaries in german. Dans *ICASSP*, volume II, pages 173–176, 1994.

[74] R. Kompe, E. Nöth, A. Kießling, T. Kuhn, M. Mast, H. Niemann, K. Ott, et A. Batliner. Prosody takes over: Towards a prosodically guided dialog system. Dans *Speech Communication*, volume 15, pages 157–167. Elsevier Science Publisher, 1994.

[75] Roland Kuhn, Ariane Lazarides, Yves Normandin, Julie Brousseau, et Elmar Nöth. Applications of decision tree methodology in speech recognition and understanding. Dans *CRIM/FORWISS Workshop*, pages 220–232, September 1994.

[76] L. Lamel et J.L. Gauvain. Experiments on speaker-independent phone recognition using BREF. Dans *ICASSP*, 1992.

[77] Philippe Langlais. état d'avancement du projet polyphon. Rapport interne, 1994.

[78] Philippe Langlais et Henri Méloni. Integration of a prosodic component in an automatic speech recognition system. Dans *3rd European Conference on Speech Communication and Technology*, volume 3, pages 2007–2010, 21–23 September 1993.

[79] Philippe Langlais, Henri Méloni, et Jacqueline Vaissière. étiquetage prosodique ascendant. Dans *19èmes Journées d'étude sur la Parole*, Bruxelles, 1992.

[80] Wayne Lea. Segmental and suprasegmental influences on fundamental frequency contours. Dans *Consonant type and tone (Proceedings of the First Annual Southern California Round Table in Linguistics)*, L. Hyman, éditeur, volume 1, Los Angeles : University of Southern California Press, 1973.

[81] Wayne A. Lea. Prosodic aids to speech recognition. Dans *Trends in Speech Recognition*, W.A. Lea, éditeur, chapter 8, pages 166–205. Prentice Hall, New Jersey, 1980.

[82] C.H. Lee, R. Rabiner, R. Pieraccini, et J. Wilpon. Acoustic modeling for large vocabulary speech recognition. *Computer Speech Language*, 4:127–165, 1990.

[83] Ilse Lehiste. *Contemporary issues in experimental phonetics*, chapter 7: Suprasegmental features of Speech, pages 225–239. Academic Press, n. lass édition, 1976.

[84] P. Léon. Où sont les études sur l'intonation ? Dans *VIIth Congress of Phonetic Sciences*, pages 113–156, Montréal, 1972.

[85] W. Levelt. Monitoring and self-repair in speech. *Cognition*, 14:41–104, 1983.

[86] P. Lieberman. Intonation, perception and language. M.I.T. Press, 1967.

[87] Vincent Lucci. La variabilité intonative dans quelques types de français oral (lecture, conférence, interview). recherches sur la prosodie du français, 1979. pub. de l'Université des langues et lettres de Grenoble.

[88] B. Malmberg. Analyse prosodique et analyse grammaticale. Word, 23, 1967.

[89] Alain Marchal. *Les sons et la parole*, volume 2. Guy Connolly, guérin édition, 1980. Collection langue et société.

[90] Alain Marchal. L'intonation : quel modèle ? Séminaire Prosodie et Reconnaissance, Octobre 1982.

[91] Joseph-Jean Mariani. Esope : un système de compréhension de la parole continue. Master's thesis, Université Pierre et Marie Curie, Paris 6, 9 juillet 1982.

[92] W.D. Marslen-Wilson et A. Welsh. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychol.*, 10:29–63, 1978.

[93] Philippe Martin. Une théorie syntaxique de l'accentuation en français. Dans *L'accent en français contemporain*, Studia Phonetica, éditeur, volume 15, pages 1–12. Fnagy I. and Léon P., 1979.

[94] Philippe Martin. Phonetic realisation of prosodic contours in french. Dans *Speech Communication*, volume 1, pages 282–294. North Holland Publishing Company, 1982.

[95] J.J. McDowald. The reliability of ratings by linguistically untrained subjects in response to stress in speech. *J. Psycholing. Res.*, 3:247–259, 1974.

[96] H. Méloni et P. Gilles. Décodage acoustico-phonétique ascendant. *Traitement du signal*, 8(2):107–114, 1991.

[97] Henri Méloni, Frédéric Béchet, et Philippe Gilles. Reconnaissance analytique de mots isolés d'un grand lexique. Dans *XIXèmes Journées d'étude sur la Parole*, pages 195–199, 1992.

[98] Henri Méloni et Jacques Guizol. Utilisation de paramètres prosodiques dans un système de reconnaissance automatique de la parole continue. Séminaire Prosodie et Reconnaissance d'Aix-en-Provence, Octobre 1982.

[99] P. Mermelstein. Automatic segmentation of speech into syllabic units. *The Journal of the Acoustical Society of America*, 58((4)):880–883, october 1975.

[100] Piet Mertens. Automatic segmentation of speech into syllables. Dans *European Conference on Speech Technology*, M.A. Laver, J. & Jack, éditeur, volume 2, pages 9–12, Edinburgh, 2–4 september 1987.

[101] Piet Mertens. Automatic recognition of intonation in French and Dutch. Dans *European Conference on Speech Communication and Technology*, Tubach J.P. et Mariani J.J., éditeurs, volume 1, pages 46–50, Paris, September 1989.

[102] Piet Mertens. *Le français parlé. études grammaticales*, chapter IV, pages 159–176. Sciences du langage, CNRS édition, 1990.

[103] Piet Mertens. Local prominence of acoustic and psychoacoustic functions and perceived stress in French. Dans *12th International Congress of Phonetic Sciences*, volume 3, pages 218–221, Aix-en-Provence, 19–24 août 1991.

[104] A.I.C. Monaghan. Intonation accent placement in a concept-to-dialogue system. Dans *Second ESCA/IEEE workshop on Speech Synthesis*, New York, September 12-15 1994.

[105] A.I.C. Monaghan et D.R. Ladd. Manipulating synthetic intonation for speaker characterisation. Dans *ICASSP*, volume 7, pages 453–456, 1991.

[106] Alex Monaghan. Generating synthetic prosody : means & ends. Prépublication des actes du séminaire prosodie de la Baume-lès-Aix, 20 et 21 octobre 1992.

[107] W.A. Munson. The growth of auditory sensation. *J.A.S.A.*, 19:584–591, 1947.

[108] Christine H. Nakatani et Julia Hirschberg. A corpus-based study of repair cues in spontaneous speech. *Acoustical Society of America*, 95(3):1603–1616, March 1994.

[109] L.H. Nakatani et J.A. Schaffer. Hearing 'words' without words: prosodic cues for word perception. *The Journal of the Acoustical Society of America*, 63((1)):234–245, january 1978.

[110] M.K. Nasri, G. Caelen-Haumont, et J. Caelen. Utilisation de règles prosodiques en reconnaissance de la parole. Dans *XVIIIèmes Journées d'étude sur la Parole*, Montréal, 28-31 mai 1990.

[111] H. Niemann, E. Nöth, E.G. Schukat-Talamazzini, A. Kießling, R. Kompe, T. Kuhn, K. Ott, et S. Rieck. Statistical modeling of segmental and suprasegmental information. Dans *NATO-ASI*, pages 237–260, Bubion, 1993.

[112] Pascal Nocéra. *Utilisation conjointe de réseaux neuronaux et de connaissances explicites pour le décodage acoustico-phonétique.* Thèse, Université d'Avignon et des Pays de Vaucluse, 1992.

[113] M. Ostendorf, P.J. Price, J. Bear, et C.W. Wightman. The use of relative duration in syntactic disambiguation. Dans *DARPA workshop on Speech and Natural Language*, pages 26–31, Hidden Valley, Juin 1990.

[114] Vincent Pagel, Noëlle Carbonell, et Jacqueline Vaissière. Spotting prosodic boundaries in continuous speech in french. Dans *XIIIth International Congress of Phonetic Sciences*, volume 4, pages 308–311, Stockholm, 13–19 August 1995.

[115] Valérie Pasdeloup. *Modèle de règles rythmiques du Français appliqué à la synthèse de la parole.* Thèse, Université de Provence (Aix-en-Provence), 1990.

[116] Valérie Pasdeloup. Organisation de l'énoncé en phases temporelles : analyse d'un corpus de phrases réitérées. Dans *XVIIIèmes Journées d'étude sur la Parole*, pages 254–257, Montréal (Quebec), Canada, 28–31 mai 1990.

[117] G. Pérennou, M. de Calmès, I. Ferrané, et J.M. Pécatte. Le projet BDLEX de bases de données lexicales du français écrit et parlé. Dans *Actes du séminaire lexique*, pages 41–56, Toulouse, 1992. IRIT-UPS.

[118] Guy Perennou et Geneviève Caelen. Utilisation de la prosodie pour la reconnaissance de la parole continue dictée. Séminaire Prosodie et Reconnaissance, Octobre 1982.

[119] Guy Pérennou et Nadine Vigouroux. L'indication de l'accent dans les transcriptions peut-elle faciliter l'alignement automatique ? Dans *Prépublication des actes du Séminaire Prosodie*, pages 39–49, La Baume-lès-Aix, 21-22 mars 1991. GRECO PRC-CHM.

[120] J. Perrot. Fonctions syntaxiques, énonciation, information. *Bull. Société de Linguistique de Paris*, LXXIII:85–101, 1978.

[121] Niels Reinholt Petersen. Perceptual compensation for segmentally conditioned fundamental frequency perturbation. *Phonetica*, 43:31–42, 1986.

[122] J. Pierrehumbert. The perception of fundamental frequency declinaison. *J.A.S.A.*, 66:363–369, 1979.

[123] Patti Price, Mari Ostendorf, Stefanie Shattuck-Hufnagel, et Cynthia Fong. The use of prosody in syntactic disambiguation. Dans *DARPA workshop on Speech and Natural Language*, pages 372–377, Pacific Grove, Février 1991.

[124] L.R. Rabiner et B.H. Juang. An introduction to Hidden Markov Models. *IEEE ASSP magazine*, Janvier 1986.

[125] JaeYeol Rheem, MyungJin Bae, et SouGuil Ann. A spectral AMDF method for pitch extraction of noise-corrupted speech. Dans *3rd European Conference on Speech Communication and Technology*, volume 3, pages 2021–2024, Berlin, 21–23 septembre 1993.

[126] M. Rossi, A. Di Cristo, P. Martin, et Y. Nishinuma. *L'intonation de l'acoustique à la sémantique*. Klincksieck, Paris, 1981.

[127] Mario Rossi. L'intensité spécifique des voyelles. Dans *Seventh International Congress of Phonetic Sciences*, André Rigault et René Charbonneau, éditeurs, pages 574–588. Mouton, 22-28 august 1971.

[128] Mario Rossi. Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole. *Phonetica*, III:229–237, 1972.

[129] Mario Rossi. L'intonation prédicative dans les phrases transformées par permutation. *Linguistics*, 103:64–94, 1er mai 1973.

[130] Mario Rossi. La perception des glissandos descendants dans les contours prosodiques. *Phonetica*, 35:11–40, 1975.

[131] Mario Rossi. Interaction des glissements d'intensité et des glissements de fréquence. Dans *XIVth international Conference on Acoustics*, High Tatra, 1976.

[132] Mario Rossi. The perception of non repetitive intensity glides on vowels. *Journal of Phonetics*, 6:9–18, 1978.

[133] Mario Rossi. L'intonation et l'organisation de l'énoncé. *Phonetica*, 42:135–153, 1985.

[134] Mario Rossi. A model for predicting the prosody of spontaneous speech (ppss model). Dans *Speech Communication*, volume 13, pages 87–107, North-Holland, 1993. Elsevier Science Publishers.

[135] Mario Rossi. A principle-based model for predicting the prosody of speech. Dans *Levels in Speech Communication : relations and Interactions*, C. Sorin et al., éditeur, pages 159–170. Elsevier Science B.V., 1995.

[136] Mario Rossi et Albert Di Cristo. Un modèle de détection automatique des frontières intonatives et syntaxiques. Dans *XIèmes Journées d'étude sur la Parole*, pages 141–164, Strasbourg, 1980.

[137] Mario Rossi et Albert Di Cristo. En quête des indices prosodiques de segmentation de l'énoncé. Dans *Prosodie et reconnaissance automatique de la parole*, pages 141–164. G.A.L.F., 1992.

[138] Jean-Jacques Schneider. *En cours de rédaction.* Thèse, Université d'Avignon et des Pays de Vaucluse, 1995.

[139] Chilin Shih et Benjamin Ao. Duration study for the at&t mandarin text-to-speech system. Dans *Second ESCA/IEEE workshop on Speech Synthesis*, pages 29–44, New York, September 12-15 1994.

[140] Hiroshi Shimodaira et Mitsuru Nakai. Prosodic phrase segmentation by pitch pattern clustering. Dans *ICASSP*, volume II, pages 185–188, 1994.

[141] E. Shriberg, J. Bear, et J. Dowding. Automatic detection and correction of repairs in human-computer dialog. Dans *Speech and Natural Language Workshop*, Harriman, éditeur, pages 419–424, 1992.

[142] Robert Sokol, Pierre-Yves Glorennec, Guy Mercier, et Krzyztof Wolinski. Utilisation d'un réseau neuro-flou pour la distinction du trait voisé/non-voisé. Dans *XXèmes Journées d'étude sur la Parole*, pages 153–158, Trégastel, 1–3 juin 1994.

[143] Thierry Spriet. *Traitements formels de connaissances linguistiques dans un système de reconnaissance automatique de la parole continue : SYRAPAC.* Thèse, Université d'Avignon et des Pays de Vaucluse, 20 janvier 1993.

[144] H. Strik. *Physiological control and behaviour of the voice source in the production of prosody.* Thèse, University of Nijmegen, 1994.

[145] J. 't Hart. Pitch contour stylisation on a high-quality analysis-synthesis system. I.P.O. Annual Progress report, 1977.

[146] J. 't Hart et R. Collier. Integrating different levels of intonation analysis. *Journal of Phonetics*, 3:235–255, 1975.

[147] J. 't Hart, R. Collier, et A. Cohen. *A perceptual study of intonation : an experimental-phonetic approach to speech melody*. Cambridge University Press, 1990.

[148] David Talkin et Colin W. Wightman. The aligner : Text to speech aligment using markov models and a pronunciation dictionary. Dans *Second ESCA/IEEE workshop on Speech Synthesis*, pages 89–92, New York, September 12-15 1994.

[149] Louis F.M. ten Bosch. Automatic classification of pitch movements via MLP-based estimation of class probabilities. Dans *ICASSP*, pages 608–611. IEEE, 1995.

[150] Christof Traber. F0 generation with a database of natural f0 patterns and with a neural network. Dans *ESCA workshop on Speech Synthesis*, pages 141–144, Autrans, Septembre 1990.

[151] D. Tuffelli. A pitch detection algorithm with hypothesis and test strategy by means of fast surface AMDF. Dans *IEEE*, volume 18 of *B*, 1984.

[152] Jacqueline Vaissière. *Contribution à la synthèse par règles du français*. Thèse, Université des langues et lettres de Grenoble, Novembre 1971.

[153] Jacqueline Vaissière. *Automatic procedure for segmenting continuous speech into prosodic words, in French*, volume 2, pages 193–208. Recherches Acoustiques, Centre National d'études des Télécommunications édition, 1976.

[154] Jacqueline Vaissière. Premiers essais d'utilisation de la durée pour la reconnaissance en mots dans un système de reconnaissance. Dans *8èmes Journées d'étude sur la Parole*, pages 345–352, Aix en Provence, 1977.

[155] Jacqueline Vaissière. *A suprasegmental component in a french speech recognition system : Reducing the number of lexical hypotheses and detecting the main boundary*, volume VII, pages 109–125. Recherches Acoustiques, Centre National d'études des Télécommunications édition, 1982.

[156] Jacqueline Vaissière. Language-independent prosodic features. Dans *Prosody : models and measurements*, A. Cutler et D.R. Ladd, éditeurs, chapter 5, pages 53–67. Springer-Verlag, 1983.

[157] Jacqueline Vaissière. The use of prosodic parameters in automatic speech recognition. Dans *Recent Advances in Speech Understanding and Dialog Systems*, H. Niemann et al., éditeur, volume F46. NATO ASI Series, Springer-Verlag Berlin Heidelberg, 1988.

[158] Jacqueline Vaissière. Caractérisation des variations individuelles du contour de fréquence du fondamental observées dans des phrases lues en anglais. Dans *XXèmes Journées d'étude sur la Parole*, pages 87–92, Trégastel, 1–3 juin 1994.

[159] Jan P.H. van Santen. Using statistics in text-to-speech system construction. Dans *Second ESCA/IEEE workshop on Speech Synthesis*, pages 240–243, New York, September 12-15 1994.

[160] Jan P.H. van Santen et Joseph P. Olive. The analyse of contextual effects on segmental duration. Dans *Computer Speech and Language*, volume 4, pages 359–390, 1990.

[161] N.M. Veilleux et M. Ostendorf. Probabilistic parse scoring with prosodic information. Dans *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 51–54, Minneapolis, Minnesota, 27-30 April 1993.

[162] Alexander Waibel. *Prosody and Speech Recognition*. Thèse, Carnegie-Mellon University, Pittsburgh, Pennsylvania, october 27 1986.

[163] W.S.Y Wang. The many uses of f0. Dans *Papers in linguistics and phonetics to the memory of Pierre Delattre*, volume 4, pages 487–503. Valdman, mouton, the hague édition, 1972.

[164] Karen Ward et David G. Novick. Prosodic cues to word usage. Dans *ICASSP*, pages 620–623. IEEE, 1995.

[165] W. Ward. Understanding spontaneous speech : the phoenix system. Dans *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 365–367, New York, 1991.

[166] Colin W. Wightman et Mari Ostendorf. Automatic labeling of prosodic patterns. Dans *IEEE transactions on speech and audio processing*, volume 2, pages 469–481, 1994.

[167] C.W. Wightman, N.M. Veilleux, et M. Ostendorf. Use of prosody in syntactic disambiguation : an analysis-by-synthesis approach. Dans *DARPA workshop on Speech and Natural Language*, pages 384–389, Pacific Grove, février 1991.

[168] S.J. Young et P. Woodland. *HTK Manual version 1.4*. Cambridge University Engineering Department, 1990.

Annexes

Conventions phonétiques

Certaines sorties de nos programmes proposent une notation phonétique qui diffère (pour des raisons de simplicité de manipulation informatique) de la notation phonétique standard IPA. Nous indiquons ici la correspondance de ces symboles avec la notation standardisée.

notation IPA	ici	notation IPA	ici	notation IPA	ici
	eu	∂	ee	œ	oe
o	au		oo	a	aa
i	ii	u	ou	e	ei
	ai	$\tilde{\text{œ}}$	un	$\tilde{}$	in
$\tilde{}$	an	$\tilde{}$	on	y	uu
j	yy	μ	uy		ww
ρ	pp	t	tt	κ	kk
b	bb	d	dd	g	gg
f	ff	s	ss		ch
	vv	z	zz		jj
m	mm	n	nn		gn
ℓ	ll		rr	$\sharp$	##

Matrice de confusion pour l'évaluation de nos modèles de phonèmes

```
----------------------- Overall Results -------------------------
PHRASE: %Correct=23.44 [H=432, S=1411, N=1843]
PHONE:  %Corr=73.03, Acc=62.52 [H=15486, D=2254, S=3464, I=2230, N=21204]
-----------------------------------------------------------------
Confusion Matrix
       s  e  o  u  a  o  a  o  i  i  u  o  a  e  a  e  w  u  y  m  n  g  p  t  k  b  d  g  f  s  c  v  z  j  l  r
       i  u  e  n  u  o  n  n  n  i  u  u  i  i  a  e  w  y  y  m  n  n  p  t  k  b  d  g  f  s  h  v  z  j  l  r
       l                                                                                                        Del [ %c / %e ]
 sil 3690  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  2  0  0  0  0  0  0  0  0  4  0  0  1  0  2  0  0  1  0  0  0   19  [99.7/ 0.1]
  eu    4 85  4  0  0  0  7  0  0  4  4  0  6  0 12 20  0  0  2  0  0  0  0  4  0  0  0  0  0  0  1  1  0  2  1   33  [54.1/ 0.3]
  oe    1  3 38  0  0  1  0  0  0  0  0  0  6  0  4  0  1  0  0  2  1  0  1  0  0  0  0  0  0  2  0  1  0  0  1  1   22  [60.3/ 0.1]
  un    0  0  0 18  0  1  2  1  0  0  0  0  0  0  1  0  4  0  0  0  5  0  0  1  1  0  0  0  0  0  0  0  0  0  0  0    4  [52.9/ 0.1]
  au    0  5  0  0202  6  7  8  1  0  1  1  0  0  1  2  1  0  0  0  0  0  0  2  0  0  0  0  1  0  0  0  0  0  0  3   16  [83.8/ 0.2]
  oo    0  5  1  0  6280  3  3  0  1  1  3  1  0 18  3  9  0  0  0  3  0  1  4  0  0  1  1  0  1  1  1  0  0  1  0   41  [80.5/ 0.3]
  an    3  1  0  0 27  4476 22 12  1  0  5  2  4  3 18  4  0  0  1  6  1  2  3  1  0  1  0  1  6  0  0  2  0  1 12   54  [76.9/ 0.7]
  on    0  0  0  0  0  1  1 14244  0  2  0  0  0  2  4  0  0  0  0  5  1  0  0  2  0  0  0  0  0  0  0  1  0  0  0  1    9  [87.8/ 0.2]
  in    5  1  3  4  0  2  9  3104  2  0  1  8  2 28  2  1  0  0  1  2  1  0  6  2  0  1  0  0  0  0  0  0  0  0  0  1   23  [55.0/ 0.4]
  ii    7  2  1  0  6  1  2  4  1975 24  1  2 32  5  7  1  0 11  0  3  1  6 10  2  0  2  1  2  1  0  1  6  3  4  3   91  [86.5/ 0.7]
  uu    1  1  0  0  0  0  0  1  0 23242  0  2  5  1  3  1  0  1  0  3  0  0  3  4  0  0  3  2  0  0  0  0  0  3  1   20  [80.7/ 0.3]
  ou    3  2  0  0  1  7  8  0  0  2  0114  1  0  6 10  1  0  0  1  1  0  7  2  2  1  5  0  0  2  1  1  0  0  0  1   32  [63.7/ 0.3]
  ai    0  0  0  0  0  0  0  0  1  0  3  3  0300  2 17 16  1  1  2  0  4  0 11  9  7  0  2  3  0  1  0  0  1  0  2  0   74  [77.7/ 0.4]
  ei    0  0  0  0  0  0  2  0  4  0  9  1  0  4322  0  3  0  0  0  0  2  0  3  4  4  0  1  0  0  0  0  0  0  0  0  3   20  [89.0/ 0.2]
  aa    4  6  3  7  8 40 33 30 18  9  1  3 60  71357 84  3  1  0  3  1  0  9 11 12  0  4  0  1  6  2  4  1  1  3 27  171  [77.1/ 1.9]
  ee   20 16  0  0  0  2  3  6  0 18 11  5  6  1 26249  0  0  1  0  2  2  2  4  0  0  0  1  0  1  0  0  4  0  1  1  183  [65.2/ 0.6]
  ww    0  1  1  0  1  7  1  0  0  0  1  6  0  0  0 0340  1  1  4  2  0  5  7 19  0  3  2  2  1  0  2  0  0  0  5   95  [82.5/ 0.3]
  uy    5  1  0  0  0  0  0  0  0  0  2  1  0  2  1  1 0127  0  3  1  9  2  6  1  0  2  2  1  0  0  2  7  1  0  2   31  [70.9/ 0.2]
  yy    1  0  0  0  0  1  0  0  0  7  0  0  1  1  0  0  1 1189  1  0  0  0 10  2  0  4  7  0  0  0  0 14  0  0  2   20  [78.1/ 0.2]
  mm    0  1  3  1  1  0  2  3  0  0  0  1  2  3  1  1  6  1 0500 75  9  0  6  8  2  0  3  2  2  0  3  2  0  1  4   87  [77.8/ 0.7]
  nn    0  2  0  0  3  0  1  4  1  3  3  0  0  3  5  2  2  0  0 27496  7  3  2  6  1 10  4  1  2  0  5  0  0  6  3   79  [82.4/ 0.5]
  gn    0  0  0  0  0  0  0  1  0  2  5  0  0  0  0  0  3  0  0  0  0  3 20  0  0  0  0  0  0  0  1  1  0  0  0  0  0    7  [55.6/ 0.1]
  pp    2  0  0  0  0  0  0  5  0  1  1  2  0  0  0  1  5  0  0  0  0  0  0189 13  5  0  4  0  0  4  0  1  0  1  0  1   22  [80.4/ 0.2]
  tt    1  2  0  0  0  5  5  3  5  2  0  0  6  4  0  5  9  5  7  2  0  6  0 15696 34  0 14  1  1 11  0  1  2  5  4  8  166  [81.0/ 0.8]
  kk    5  0  0  1  1  1  1  0  2  1  4  1  2  3  3  8  9  5  8  1  2  0 35 28643  4  7 16  0  9  0  2  1  0  8 11   91  [78.2/ 0.8]
  bb    0  0  0  0  0  1  3  0  0  0  0  0  3  0  1  0  1  1  0  1  0  3  4  0  5  5  0 71  5  4  0  0  0  0  0  0  1   24  [65.1/ 0.2]
  dd    1  0  0  0  0  0  1  1  6  0  1  0  4  2 16 12  3  0  0  0  0 14  1  3  5  4  7515 23  1  0  0  0  5  2  2  3   95  [81.5/ 0.6]
  gg    0  2  0  0  0  3  0  4  0  1  1  0  0 11  0  2  1  1  9  0  3  0  1  1 13  2  5185  0  1  0  2  1  1  1  1   32  [73.4/ 0.3]
  ff    0  0  0  0  0  1  1  2  3  0  1  0  0  0  2  6  1  8  3  1  0  1  0  4  3  3  1  0  2198 34  0  2  0  1  2  2   33  [70.2/ 0.4]
  ss    3  0  0  1  2  1  4  2  1  7  2  2  2  7  5 12  1  2  0  1  4  0 18 12 10  0  2  2 34903  7  3 10  4  2  9  129  [84.0/ 0.8]
  ch    0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  3  0  0  0  0  0  0  0  0  0  0  3 50  0  0  0  0  0    3  [87.7/ 0.0]
  vv    0  1  0  1  0  0  0  0  0  3  1  0  0  1  0  3 12  0  0  7  5  0  6  0  0  7  4  1  1 18  0165  7  2  4  8   60  [64.2/ 0.4]
  zz    1  0  0  0  0  6  1  0  2  0  0  1  2  3  1  1  0  0  1  2  1  0  0  0  2  0  8  0  0  6  1 10205  5  1  3   35  [77.9/ 0.3]
  jj    0  0  0  0  0  0  0  0  2  1  2  0  0  0  2  1  2  0  0  1  0  1  0  0  5  4  0  0  0  0  6  4  0  0 98  1  2   20  [74.2/ 0.2]
  ll    2  2  0  0 10  2  3 10  3  0  5  1  0 20 12  4 13  4  0  6 29  0  2 25  3  3 11  4  0  8  0  5  3 3313 19  208  [59.6/ 1.0]
  rr    3  3  0  3 13  8 17 43  4  1  7  2  3  3 26 13 15  6  7  5 25  0 13  4 21  4  8  1  4 18  1 10 19  0 14887  205  [73.2/ 1.5]
 Ins   90 24 13  7 39 33 48109 33 58 30 30 55 99128104 82 26 45 32 71  7 44231166  4 59 40 27 99  3 18 67 35 51214
```

Exemple de feuille d'appel PolyVar

SVP APPELEZ LE (026) 21 22 25

Un ordinateur vous répondra et vous posera quelques questions.

NUMÉRO D'IDENTIFICATION : 124

IDIAP vous remercie de votre collaboration. Votre voix sera enregistrée et utilisée pour la recherche et le développement en traitement automatique de la parole.

La session commence par les demandes suivantes:

Donnez votre numéro d'identification	*(votre réponse)*
Déclinez votre identité et prononcez votre adresse	*(votre réponse)*
Etes-vous de sexe masculin ?	*(votre réponse)*
Donnez un numéro de téléphone familier	*(votre réponse)*
Quelle est votre langue maternelle ?	*(votre réponse)*
Quelle est votre date de naissance ?	*(votre réponse)*
Dans quelle ville avez-vous commencé votre scolarité ?	*(votre réponse)*

À l'annonce des numéros, veuillez fournir la réponse correspondante proposée dans la colonne de droite.

1. *annulation*
2. *0380 302832 40*
3. *taxes*
4. *Lors de son ouverture, elle emploiera une centaine de personnes.*
5. *Casino*

---

6. *700 000 dcl.*
7. *cinéma*
8. *concert*
9. *L'inauguration est programmée pour l'ouverture de la prochaine saison.*
10. *Zinal* $\Longrightarrow$ *Z i n a l*

---

11. *Corso*
12. *exposition*
13. *54 037,59 $US*
14. *Il se garantira du froid avec un bon capuchon.*
15. *trafic*

---

16. *Galerie du Manoir*
17. *L'an prochain se déroulera à Paris le Congrès forestier mondial, du 17 au 26 septembre.*
18. *Il est conseillé de réserver sa place au 22 23 47 ou au 22 92 72.*
19. *Gianadda*
20. *guide*

---

21. *manifestation*
22. *six heures moins dix du matin*
23. *télégramme*
24. *Au cri du sergent, les hommes du poste étaient sortis pêlemêle.*
25. *renseignement*

---

26. *message*
27. *581 119 FS*
28. *mardi 16 octobre 1984*
29. *Bon point pour la gestion de la santé.*
30. *mode d'emploi*

---

31. *Louis Moret*
32. *80 020,118*
33. *dérangement*
34. *Ovronnaz* $\Longrightarrow$ *O v r o n n a z*
35. *musée*

---

36. *Les récentes catastrophes naturelles de 1987 et de 1990 ont laissé des traces qui en disent long sur ces questions.*
37. *précédent*
38. *quitter*
39. *international*
40. *Brig* $\Longrightarrow$ *B r i g*

---

41. *suivant*
42. *0 2 4 0 2 5*
43. *Les sondages archéologiques terminés, une société d'économie mixte sera créée fin janvier.*
44. *On se criait de loin : Bonjour !*

La session se termine alors par quelques questions :

Formulez une requête au 111. *(votre réponse)*
Quel est votre niveau scolaire :
1- apprentissage, 2 = moyen, 3 = universitaire, 4 = autre) ? *(votre réponse)*
Formulez maintenant vos commentaires sur cette session. *(votre réponse)*

La session est maintenant terminée.

N.B. Si vous désirez obtenir des renseignements supplémentaires ou formuler des remarques, téléphonez au : (026) 22 76 64.

Nous vous remercions de votre appel.

Liste des mots de FeLex

1) abattais	2) abattis	3) abattons	4) abattue	5) abêties
6) abondais	7) abonda	8) abonder	9) aboutis	10) absentant
11) absentaient	12) abusât	13) acajou	14) accablaient	15) accentuent
16) acclamât	17) accessits	18) acclamé	19) accomplit	20) acculas
21) accumule	22) accusas	23) acquittés	24) activât	25) adaptez
26) adaptas	27) affadie	28) affabules	29) affama	30) affamais
31) affichés	32) affichons	33) affilies	34) affina	35) affixale
36) affûter	37) affûtée	38) agaçait	39) agença	40) agavé
41) agendas	42) agitant	43) aimanta	44) aimantons	45) alanguit
46) alchimie	47) alcalin	48) alignons	49) alignant	50) alignas
51) alitons	52) allongeons	53) almanachs	54) alphabets	55) alunies
56) amadoue	57) amassées	58) amatis	59) amendé	60) amendât
61) amincie	62) ananas	63) animons	64) apiquais	65) apiquât
66) aplanit	67) aplatis	68) appendaient	69) appendus	70) appontât
71) appontez	72) assagie	73) assemblées	74) assemblées	75) assemblons
76) assembla	77) assimiles	78) associe	79) astiquant	80) attablé
81) attaquant	82) attaquais	83) attaquât	84) attaquer	85) attendit
86) attendons	87) attendu	88) attentais	89) attifant	90) attifa
91) autopsie	92) avachis	93) avalais	94) avanie	95) avant-goûts
96) avilies	97) bâtissais	98) badigeons	99) baladés	100) balançât
101) baldaquins	102) balbutient	103) baluchon	104) bambochez	105) bambochas
106) bambocher	107) bannissez	108) baptisez	109) baptismaux	110) bécota
111) bécotons	112) bégaiement	113) bélougas	114) benjamins	115) bicéphale
116) bichonnant	117) blondissons	118) bombasin	119) bougonnons	120) boulonnât
121) boulonnant	122) boulonné	123) butinât	124) butinons	125) cagibis
126) calfatait	127) calicots	128) calvitie	129) campagnol	130) cancanaient
131) cantonnât	132) cantonnons	133) capitaux	134) capitales	135) capitulent
136) captivas	137) capuchons	138) cascada	139) cascadais	140) cascader
141) cascadant	142) catalpa	143) catissant	144) cavalas	145) cavaler
146) cavalée	147) centennales	148) chicanant	149) chicanés	150) chicanons
151) chiffonnons	152) chipotas	153) chuchotas	154) chuchotant	155) cimentait
156) cisalpin	157) clandestin	158) classifie	159) clavicule	160) clignotas
161) cokéfient	162) colmatant	163) combattues	164) combattais	165) commandés
166) comestible	167) commença	168) compassai	169) compassas	170) compatit
171) compassons	172) complotas	173) composât	174) condamnez	175) condamnas
176) condamner	177) condensez	178) condensât	179) condensons	180) confessât
181) confettis	182) confondons	183) confondez	184) congédient	185) consentie
186) consentait	187) constatez	188) constatons	189) contactaient	190) contemplant
191) contemplais	192) contentés	193) contentât	194) contestons	195) convaincus
196) convainquez	197) convainquit	198) convoquas	199) convola	200) convoquait
201) culbuté	202) culbutant	203) culminas	204) culottant	205) culbutons
206) culminant	207) cultivée	208) cumulas	209) cumulons	210) damassât
211) damassais	212) damassons	213) débâtis	214) débandai	215) débattus
216) déboulez	217) décaissant	218) décaissas	219) décanté	220) décampé
221) décampons	222) décapée	223) décapons	224) décapsule	225) décatît
226) décéda	227) décennale	228) déchantas	229) décessons	230) déchantez
231) déchaussons	232) décochas	233) décommit	234) déconfis	235) décongèlent
236) découcha	237) décousu	238) dédaignés	239) dédaignas	240) dédoublât
241) défaussât	242) défendue	243) défolie	244) défonçons	245) défonçant
246) dégaina	247) déganta	248) dégauchît	249) déglaçât	250) déjauni
251) délaçons	252) délaçant	253) délassa	254) délégua	255) démâtons
256) démangea	257) démangeaient	258) démêlât	259) démenties	260) démodons
261) démoli	262) démoda	263) démoulât	264) démoulons	265) dénantît
266) déneigea	267) dépassât	268) dépêchât	269) dépeignis	270) dépendons
271) dépendez	272) déplaça	273) déplantât	274) déplombez	275) dépoli
276) dépotas	277) désaxât	278) désempli	279) désolons	280) détapons
281) détaxons	282) détendus	283) détendons	284) dévêtit	285) descendues
286) descendez	287) dessanglée	288) dessanglât	289) destitues	290) diffamât
291) diffamais	292) difficile	293) diffusé	294) digital	295) dilatais
296) dilatons	297) diphtonguai	298) diphtonguant	299) diminue	300) diphtonguer
301) disculpai	302) discuter	303) discutées	304) dispensons	305) disposât
306) disposant	307) disputez	308) disputons	309) dissemblable	310) dissimule

311) dissipée 312) dissipas 313) dissipons 314) distanças 315) distancie
316) distendis 317) distendus 318) distillât 319) distinguos 320) distinguons
321) divisât 322) divisez 323) divulguées 324) divulguons 325) dyslexies
326) dynamo 327) ébaubît 328) ébaudis 329) éboulant 330) éboula
331) éboulons 332) écatît 333) échampi 334) échangeait 335) échangeât
336) échanson 337) éclatant 338) écobuent 339) écouta 340) éjacule
341) élégies 342) élégants 343) élément 344) élongées 345) élongeons
346) émascule 347) émonda 348) émondons 349) émondait 350) émoulant
351) émoulu 352) épaissis 353) épanchais 354) épandait 355) épandu
356) épinça 357) épongeât 358) épousât 359) établi 360) étanchons
361) étêtas 362) étendais 363) étendis 364) évacues 365) évangile
366) évolue 367) effaça 368) effaçons 369) effectuent 370) emballées
371) emballât 372) embauchais 373) embattons 374) embêtés 375) embêta
376) embauchons 377) embellît 378) embêtons 379) emblavât 380) embossaient
381) embolies 382) emboutis 383) embusquant 384) embusquas 385) emmanchez
386) emmancha 387) emmancher 388) emmêlons 389) empâtaient 390) empala
391) empâter 392) empalons 393) empalez 394) empauma 395) empêchas
396) empestas 397) encageons 398) encageas 399) encaqua 400) encaquées
401) encensez 402) enchâsser 403) enchaîna 404) enchantez 405) enchaussas
406) enchaussant 407) enclencher 408) encollons 409) endettés 410) endettât
411) endentaient 412) endiguas 413) enfantais 414) enfanter 415) enfanta
416) enflammais 417) enflammât 418) enfonçais 419) enflammons 420) enfoncer
421) engainant 422) engainas 423) engageât 424) englacez 425) engloutie
426) engonça 427) engonçais 428) engonçant 429) enjambés 430) enjôlons
431) enjamber 432) enlaçât 433) enlacé 434) enlaidît 435) enlaçons
436) enlisât 437) ennoblies 438) enneigeas 439) ensablez 440) ensabla
441) ensachée 442) ensablant 443) entablant 444) entacher 445) entachas
446) entachai 447) entassais 448) entendez 449) entêtant 450) entendant
451) entendus 452) entendit 453) envasa 454) envidas 455) envolant
456) envoûtons 457) escomptons 458) espingole 459) esquintons 460) essouffla
461) estampas 462) estampons 463) estompons 464) exceptât 465) exempta
466) exhaussa 467) expédies 468) fabuleux 469) falbala 470) falsifient
471) fandangos 472) fantassin 473) fatigable 474) fatiguas 475) faux-semblant
476) fécondant 477) fécondait 478) féconda 479) financées 480) finançât
481) flatulent 482) fulminai 483) fulminât 484) fumigeas 485) fumigeais
486) fustigeaient 487) fustigeant 488) fustigeons 489) galvanos 490) génépis
491) gigotant 492) glapissaient 493) glatissons 494) glapissons 495) glycocolle
496) glycémies 497) gonfanon 498) gymkhanas 499) habitaient 500) habitats
501) habitacles 502) habiter 503) habitue 504) hallali 505) hidalgos
506) homélie 507) homuncule 508) humecta 509) humilient 510) humectons
511) hypostyle 512) ici-bas 513) ichneumon 514) illisibles 515) imbattable
516) imbéciles 517) imago 518) immanent 519) immangeable 520) immisça
521) immobiles 522) impalpables 523) impassibles 524) implacable 525) implantée
526) implanta 527) impossibles 528) incassable 529) incivil 530) incombaient
531) incombas 532) incombons 533) inconnu 534) inconstant 535) indécis
536) indicible 537) indiquant 538) indocile 539) indompté 540) indistincts
541) ineffables 542) ineptie 543) infatues 544) infectât 545) infléchi
546) ingénies 547) ingénu 548) inhabile 549) inhalait 550) inhiba
551) inhumons 552) inhumas 553) injecté 554) injectons 555) inocule
556) inondais 557) insolvables 558) insoumis 559) insondable 560) installas
561) installant 562) intendants 563) intentez 564) intestins 565) inusable
566) inutiles 567) inventons 568) inventas 569) inventai 570) investis
571) invivable 572) invincible 573) invisible 574) isabelle 575) jacassai
576) jacassons 577) jugulât 578) justifie 579) kidnappaient 580) kidnapper
581) kidnappons 582) laitonnas 583) langenthal 584) lapidant 585) lapidons
586) lavabos 587) licencient 588) licitées 589) licitons 590) ligament
591) ligotât 592) ligotons 593) limita 594) liminal 595) liquéfie
596) liniments 597) liquidées 598) liquidons 599) lumignons 600) machinales
601) maculons 602) magasin 603) magnifie 604) malaxaient 605) malaxas
606) malhabile 607) malfaçons 608) malfamée 609) mandater 610) maniements
611) manipules 612) mastiquais 613) mastaba 614) matinale 615) mausolée
616) mécomptât 617) méconnues 618) mécomptons 619) mélangeant 620) mélangea
621) mélangeais 622) mévendons 623) mikados 624) mijota 625) minuscules
626) mitigeant 627) mitigeait 628) mitigeons 629) mitonnons 630) monacale

631) monoskis 632) mouloudji 633) multiplies 634) muscadet 635) musulmans
636) mutilas 637) mutité 638) mystifies 639) naviguât 640) naviguons
641) nicolas 642) noctambule 643) nonchalant 644) notifies 645) nouveau-nés
646) obsédons 647) offensât 648) offensons 649) okapi 650) ostensibles
651) oulémas 652) pâtissé 653) pâtissons 654) pacifies 655) pactisé
656) paginât 657) palatal 658) palisson 659) panachez 660) panachas
661) panachons 662) passagers 663) pavanons 664) pédoncule 665) photopile
666) picotons 667) pigmentait 668) pigmentant 669) pignochons 670) pimentées
671) pissenlit 672) pitonnons 673) pistonnons 674) plaisantant 675) plaisantons
676) planifie 677) plastiquez 678) plastiquât 679) platini 680) plus-value
681) pneumonie 682) polenta 683) pomponnât 684) postdatai 685) pubescent
686) pugilat 687) pyjama 688) qualifie 689) quémandé 690) quémandons
691) saccadait 692) saccager 693) saccageait 694) saccagea 695) salamis
696) salissons 697) salivons 698) sanglotas 699) sanglotons 700) sanglotai
701) satané 702) satisfaits 703) segmentons 704) sibilants 705) signalés
706) signifie 707) similis 708) simulas 709) simulant 710) siphonnât
711) sodomie 712) somnolant 713) soubattue 714) soubattant 715) soubattis
716) sous-jacent 717) sous-tendu 718) spadassins 719) spécifies 720) spontanée
721) statu_quo 722) statufie 723) stimulant 724) stipulas 725) stipulais
726) stimuli 727) stipendies 728) stimulons 729) stipulons 730) stupéfie
731) subalpins 732) subaigu 733) subséquent 734) substitue 735) subsistons
736) succédas 737) succombons 738) succombez 739) suffixa 740) suffixons
741) suffixaient 742) suffoquant 743) supplantant 744) supplantée 745) supplanter
746) supplantons 747) supplicie 748) suppléments 749) supputez 750) suscitais
751) suscitons 752) susceptible 753) suspendue 754) sustentait 755) symphonie
756) syndicats 757) syndiquons 758) tapissait 759) tapissant 760) taquinas
761) téléski 762) tentacules 763) thégonie 764) titubas 765) toboggans
766) toccatas 767) tomaisons 768) tombola 769) tuméfie 770) tumescents
771) tympanon 772) ulula 773) ululé 774) ululons 775) unifient
776) usager 777) ustensiles 778) utopie 779) vaccina 780) vagissez
781) vagissant 782) validons 783) vendanger 784) vendangeant 785) vendangeât
786) vendangeons 787) vicinal 788) vicomté 789) vidangés 790) vis-à-vis
791) viticole 792) vitellins 793) vivifient 794) vivotât 795) vivotons
796) volatil 797) volonté 798) wagons-lits 799) xanthophylles 800) zigzaguons

Liste des mots d'AviLex1

1) à 2) accès 3) adjectif 4) aérienne 5) aïeul
6) allée 7) ambitieux 8) analyse 9) anniversaire 10) appétit
11) argument 12) aspiratrice 13) attribut 14) autonomie 15) aviateur
16) bail 17) barbare 18) bébé 19) bicyclette 20) blindé
21) bonté 22) bourgeoise 23) brillant 24) brutalité 25) cage
26) cane 27) carburant 28) casserole 29) certain 30) chance
31) chasseuse 32) cheval 33) chrétien 34) ciseaux 35) climat
36) colle 37) comité 38) compagnon 39) conception 40) confus
41) consonne 42) contrebande 43) correct 44) courte 45) cri
46) cultivateur 47) débrouillarde 48) définition 49) démocratique 50) désaccord
51) dessus 52) dictée 53) discret 54) document 55) drap
56) éclatant 57) égale 58) éléphant 59) employé 60) enterrement
61) épidémie 62) espionne 63) éternel 64) évidemment 65) exécution
66) extrait 67) familial 68) féminin 69) fleur 70) format
71) fraîche 72) gamme 73) générale 74) glaciale 75) grâce
76) grave 77) guère 78) hausse 79) hier 80) hôtesse
81) identique 82) immoral 83) imposant 84) inconscient 85) indispensable
86) infinitif 87) inondation 88) intelligent 89) intime 90) israélite
91) jeûne 92) jumeau 93) latins 94) lentement 95) lieu
96) locution 97) loyer 98) magistrat 99) maladresse 100) manche
101) marins 102) matériel 103) mécanisme 104) ménager 105) message
106) microscope 107) minéral 108) mobile 109) mondiale 110) mouche
111) musicien 112) nation 113) négligence 114) noël 115) nu
116) oblong 117) odorat 118) opinion 119) ordonnance 120) originalité
121) pacte 122) papier 123) paresseuse 124) partielle 125) patience
126) pêche 127) périlleux 128) peuple 129) pierre 130) plafond
131) plombier 132) poétique 133) pomme 134) pose 135) pouce
136) précieuse 137) préparatif 138) prévision 139) privilégié 140) profitable
141) propos 142) qualificative 143) quoi 144) raisonnement 145) ravi
146) réclamation 147) réfléchi 148) rein 149) renvoi 150) réserve
151) réveil 152) risque 153) rouage 154) ruse 155) salive
156) sauf 157) seconde 158) sens 159) serrure 160) sincérité
161) soin 162) sorte 163) souscription 164) spectateur 165) stupéfait
166) suite 167) sursis 168) synthétique 169) tapisserie 170) témoignage
171) terre 172) tigresse 173) torture 174) tragédie 175) travaux
176) tronc 177) une 178) vaste 179) venue 180) veston
181) vigilante 182) violet 183) voici 184) vrai 185) abandonnerai
186) aborder 187) abrégerons 188) abstiens 189) accéderons 190) accompagnerons
191) accourons 192) accueilli 193) achetant 194) adapterai 195) administrant
196) adoucirai 197) afficherai 198) affronterai 199) agissez 200) agréant
201) aligner 202) allongerai 203) aménagerai 204) anéantissez 205) apaisons
206) aplatissez 207) appelle 208) apportant 209) approuverai 210) arranger
211) arroser 212) assassinerons 213) assiégerons 214) assurant 215) atteignez
216) atterrissez 217) attristons 218) avancerai 219) avoue 220) baise
221) baptise 222) bâtissez 223) bénissons 224) blanchissant 225) bombarder
226) bornant 227) bouleverse 228) braient 229) brillons 230) brouillerai
231) brunissent 232) calculerai 233) capitulerons 234) causant 235) chargerai
236) chercher 237) cirer 238) clorai 239) coïnciderons 240) combattrons
241) commettre 242) compléterai 243) comprise 244) concernerons 245) condenser
246) conformer 247) conquièrent 248) conserver 249) consomme 250) construisant
251) contesterai 252) contredite 253) convenez 254) correspondu 255) coulerons
256) couronnant 257) craignent 258) crèverons 259) croissons 260) cuisent
261) danserai 262) débarrasserons 263) débouche 264) décède 265) déchausserons
266) décollerons 267) découragerons 268) dédirai 269) défendez 270) déguiser
271) délibérer 272) déménage 273) démissionnerons 274) démoralisons 275) dépassons
276) déplacer 277) déposerai 278) déroberai 279) déshonorer 280) désobéis
281) détournant 282) devenons 283) dévoue 284) diminuerai 285) discute
286) disposerons 287) dissoudrai 288) diviser 289) dormir 290) dressant
291) éblouissons 292) échapper 293) éclairerons 294) écoulant 295) écrouerai
296) effectue 297) effrayerons 298) élèverons 299) emballe 300) embrasse
301) émigrerons 302) empêche 303) empresserai 304) enclouerai 305) enduire
306) enflerai 307) engloutissons 308) enlèverons 309) enrichirai 310) entendez

311) entreprendre 312) envahir 313) envolant 314) épellerons 315) épuiserons
316) essayerai 317) établissent 318) étendant 319) étranglant 320) évaluons
321) éviter 322) examiner 323) exclure 324) exiger 325) expliquons
326) exposons 327) fabrique 328) fauche 329) fendrai 330) figurons
331) flambe 332) foncerai 333) forger 334) fouillons 335) franchis
336) frémirai 337) frottons 338) gâchons 339) garerai 340) geins
341) gênerons 342) goûter 343) gratter 344) grince 345) grossissant
346) habiller 347) heurte 348) ignorons 349) imiterai 350) imposerai
351) indignerai 352) infliger 353) inscris 354) inspecterons 355) insulte
356) interpellant 357) intervenu 358) invoquons 359) jaunir 360) jouerai
361) justifierons 362) lancer 363) léguer 364) lierai 365) lirons
366) luis 367) maigrirons 368) manierons 369) maquillant 370) maudire
371) mélangeons 372) mériterons 373) modère 374) montons 375) mouchons
376) mouvons 377) murissent 378) nationalisons 379) nettoie 380) notons
381) nuisons 382) obligerons 383) obtenons 384) offrir 385) opérant
386) organisant 387) oublier 388) palis 389) parcourent 390) parlerons
391) parviennent 392) paierai 393) penche 394) percer 395) permettons
396) peser 397) pincerons 398) plaignant 399) pleurant 400) pondre
401) poster 402) poursuivrons 403) pratiquerai 404) précisant 405) préméditons
406) prescrire 407) présiderons 408) prêtons 409) prive 410) produisez
411) promener 412) propageons 413) prouvons 414) punir 415) quitter
416) raccourcissent 417) raffolons 418) rajeunirons 419) ramenons 420) ranimerons
421) rapportant 422) rassurant 423) ravissons 424) réalisant 425) recevrons
426) récolter 427) réconcilierai 428) reconnaîtrons 429) recrute 430) récupérer
431) redoublerai 432) refaire 433) reflétons 434) refusons 435) réglerai
436) rejetterons 437) relâchant 438) remettons 439) remplissent 440) renferme
441) renseigne 442) répandez 443) répartissez 444) repentent 445) répondons
446) repoussant 447) reprocherons 448) résignons 449) respectant 450) resserrons
451) résultant 452) retarderons 453) retourne 454) retrouvant 455) révèle
456) reverdirons 457) rincerons 458) ronger 459) rougirons 460) ruinerai
461) salissons 462) sauver 463) séchons 464) séduite 465) sentant
466) siégeons 467) signifierons 468) sollicitant 469) soufflons 470) souillant
471) soumettrai 472) soutenant 473) subissent 474) sucrerai 475) suivons
476) supprime 477) surprendrai 478) survenir 479) suspends 480) tu
481) tardant 482) télégraphier 483) tenons 484) timbrerons 485) tombant
486) torpillerons 487) traduisez 488) trairons 489) transformerai 490) transportant
491) tricher 492) trompe 493) tuerons 494) utiliserons 495) vantons
496) vengeons 497) vexerai 498) visons 499) verrons 500) voterai

Liste des mots d'AviLex2

1) à jeun
2) ébranler
3) écacher
4) échantillonner
5) échappatoire
6) échauffourée
7) échidné
8) échouer
9) éclabousser
10) écoinçon
11) écourgeon
12) écusson
13) égratigner
14) éléphant
15) éléphanteau
16) électrocardiogramme
17) électrochoc
18) électrocuter
19) électroménager
20) élongation
21) élucubration
22) émonctoire
23) émoustiller
24) épagneul
25) épargner
26) éparvin
27) épicondyle
28) épiloguer
29) épiphénomène
30) épistémologie
31) éplucher
32) époumoner
33) épouser
34) épouvantable
35) épouvantail
36) épouvanter
37) éradication
38) érugineux
39) étalon
40) état-major
41) étincellement
42) étouffe-chrétien
43) étouffer
44) étoupille
45) étuve
46) étuver
47) évangéliser
48) évanouir
49) éventualité
50) évincer
51) évolutionnisme
52) Peau-Rouge
53) aérodynamique
54) abaisser
55) abandon
56) abandonner
57) abaque
58) abat
59) abolitionnisme
60) abracadabrant
61) aide
62) aide-comptable
63) aide-mémoire
64) aider
65) aiglefin
66) aigrette
67) ailloli
68) ainsi
69) ambages
70) ambassade
71) ambassadeur
72) ambiance
73) ambidextre
74) amble
75) ambre
76) amphibologie
77) anarcho-syndicaliste
78) anfractuosité
79) antédiluvien
80) antigouvernemental
81) au demeurant
82) au-dedans
83) au-dehors
84) aubépine
85) aubade
86) aubaine
87) aube
88) aubergine
89) auburn
90) aucun
91) audio-visuel
92) auditionner
93) auge
94) aulne
95) auquel
96) autobiographie
97) autosuggestion
98) bégueule
99) béguin
100) bélouga
101) béluga
102) bénignité
103) baby-foot
104) bain
105) bakchich
106) be-bop
107) bec-croisé
108) bec-de-corbin
109) bec-de-perroquet
110) belle-de-jour
111) benjamin
112) benzène
113) beugler
114) beurre
115) beurrer
116) beuverie
117) bibelot
118) biberon
119) bible
120) bibliographie
121) bibliophile
122) bibliothèque
123) bibliothécaire
124) biceps
125) biche
126) bien
127) bien-aimé
128) bienheureux
129) bienveillamment
130) bleu
131) bleuâtre
132) bloc
133) bloc-moteur
134) bloc-notes
135) blond
136) blondir
137) blouse
138) blouser
139) blue-jean
140) bluff
141) bluffer
142) bock
143) bolchevisme
144) bombage
145) bombardement
146) bombarder
147) bombe
148) bonne-maman
149) boogie-woogie
150) bookmaker
151) borne-fontaine
152) bougainvillier
153) bouleversement
154) brûle-gueule
155) brûle-parfum
156) brain-trust
157) breakfast
158) breitschwanz
159) bringuebaler
160) bru
161) brucellose
162) brume
163) brun
164) brutaliser
165) bungalow
166) cérébro-spinal
167) cash-flow
168) chérubin
169) chacun
170) chauve-souris
171) check-up
172) chimpanzé
173) chinchilla
174) chouannerie
175) chromolithographie
176) chuchoter
177) chuter
178) circonlocution
179) coeur-de-pigeon
180) cognition
181) cold-cream
182) colin-maillard
183) collaborationniste
184) combativité
185) combinaison
186) commun
187) contre-manifestant
188) contre-proposition
189) cornichon
190) corymbe
191) cotignac
192) coupe-ongles
193) courtisanerie
194) cover-girl
195) crève-la-faim
196) crayon-feutre
197) creuser
198) crochu
199) cromlech
200) culpabilité
201) cumulo-nimbus
202) cure-ongles
203) curriculum vitae
204) cut-back
205) débander
206) déboucler
207) déboulonner
208) déboutonner
209) débucher
210) décacheter
211) décemvir
212) décomposer
213) décomprimer
214) décongestionner
215) déconsidérer
216) décontenancer
217) déconvenue
218) décousu
219) décrochez-moi-ça
220) décubitus
221) décuscuteuse
222) décuver
223) dédain
224) dédoubler
225) défaveur
226) défavoriser
227) défense
228) défloraison
229) défoncer
230) défourrer
231) défunt
232) dégénérescence
233) dégazonner
234) dégingandé
235) dégouliner
236) dégoupiller
237) dégringoler
238) dégrouper
239) dégueuler
240) déguiser
241) déjeuner
242) délai-congé
243) déliquescence
244) démantibuler
245) démeubler
246) démilitariser
247) déminéraliser
248) démobiliser
249) démonstratif
250) démoustication
251) dénaturaliser
252) dénicher
253) dénivellation
254) dénouement
255) dépeigner
256) dépeupler
257) déplomber
258) dépopulation
259) dépoudrer
260) déséchouer
261) désaccoutumer
262) désagrégation
263) désapprobation
264) désapprovisionner
265) désassembler
266) désavantager
267) désaveu
268) désavouer
269) désembourgeoiser
270) désemparer
271) désenchaîner
272) désenchanter
273) désencombrer
274) désenfler
275) désenfumer
276) désenivrer
277) désenlaidir
278) désennuyer
279) désenrhumer
280) désensibiliser
281) désenvaser
282) désenvelopper
283) désenvenimer
284) déshumaniser
285) désinfecter
286) désintoxication
287) désintoxiquer
288) désinvestir
289) désobligeamment
290) désodoriser
291) désoeuvré
292) désolidariser
293) déstructuration
294) détacher
295) dévergonder
296) dévoué
297) dévouement
298) de profundis
299) delirium tremens
300) derechef

301) dessouder
302) destabiliser
303) disc-jockey
304) disjonction
305) distributionnalisme
306) drop-goal
307) drugstore
308) duffle-coat
309) dum-dum
310) eczémateux
311) effaroucher
312) effondrer
313) elzévir
314) emberlificoter
315) emmagasiner
316) emprunté
317) en un tournemain
318) encoignure
319) enrégimenter
320) ersatz
321) escampette
322) escarboucle
323) escarmouche
324) escourgeon
325) espadon
326) espagnolette
327) espingole
328) esprit-de-vin
329) esseulé
330) essoucher
331) essouffler
332) essuie-meubles
333) estafilade
334) estragon
335) esturgeon
336) et cetera
337) ethnique
338) eucalyptus
339) eucharistie
340) eugénique
341) eugénisme
342) euphémique
343) euphémisme
344) evzone
345) ex-voto
346) exactitude
347) exclusivité
348) excommunier
349) exhalaison
350) exhibitionnisme
351) exiguïté
352) expansionnisme
353) exsanguino-transfusion
354) exterritorialité
355) extrême-onction
356) extra-utérin
357) extrajudiciaire
358) extravagance
359) extravagant
360) feignant
361) feld-maréchal
362) feutre
363) folichon
364) fraîcheur
365) générosité
366) génuflexion
367) géosynclinal
368) gérontocratie
369) gérontologie
370) gentleman's agreement
371) gentleman-farmer
372) gingembre
373) gingival
374) gneiss
375) gorge-de-pigeon
376) gorgonzola
377) gréco-latin
378) grisou
379) grisouteux
380) guéridon
381) guérillero
382) gueule-de-loup
383) gueuse
384) hécatombe
385) héliogravure
386) hémistiche
387) hérisson
388) hétérogénéité
389) hafnium
390) hard-top
391) haute-fidélité
392) hebdomadaire
393) herd-book
394) hors-jeu
395) horse-guard
396) humble
397) hypersensibilité
398) imbécillité
399) immettable
400) impétuosité
401) impartialité
402) inébranlable
403) in partibus
404) in vivo
405) inadvertance
406) income-tax
407) inconsidérément
408) inconvénient
409) inconvertible
410) individualiser
411) infinitésimal
412) inlandsis
413) inlassable
414) insignifiant
415) institutionnaliser
416) intransigeance
417) intransitif
418) israélite
419) jazz-band
420) jazzman
421) jodhpurs
422) juke-box
423) jumbo-jet
424) jungle
425) junior
426) junte
427) ketchup
428) konzern
429) kumquat
430) kvas
431) kymrique
432) législatif
433) lépidodendron
434) l'un
435) laideur
436) laissé-pour-compte
437) leishmania
438) leitmotiv
439) les uns
440) lumbago
441) lundi
442) lyncher
443) mécanographie
444) mécanothérapie
445) médico-social
446) méditerranéen
447) mélancolie
448) mélancolique
449) méninge
450) méphistophélique
451) mésaventure
452) métabolique
453) métapsychique
454) métempsycose
455) maître-à-danser
456) mainlevée
457) mal-logé
458) mea culpa
459) medicine-ball
460) mercantile
461) meugler
462) mezzo-soprano
463) midship
464) mnémotechnique
465) moribond
466) municipalité
467) néanmoins
468) nébuleuse
469) négligemment
470) négro-africain
471) nénuphar
472) néo-impressionnisme
473) néoformation
474) névralgie
475) nauséabond
476) nec plus ultra
477) negro-spiritual
478) nerprun
479) non-réussite
480) nonchalamment
481) nonchalance
482) nunchaku
483) obnubiler
484) obtempérer
485) obvenir
486) oecuménique
487) oedémateux
488) oeil-de-boeuf
489) oenologie
490) oestrogène
491) offshore
492) oignonade
493) onzième
494) ophtalmologie
495) opportun
496) oto-rhino-laryngologie
497) ovni
498) pécheur
499) pédoncule
500) péjoratif
501) péninsule
502) péripatéticien
503) pétrochimie
504) pétun
505) parfum
506) pechblende
507) perlimpinpin
508) perméabiliser
509) permanganate
510) perpendiculaire
511) perquisitionner
512) persona grata
513) pfennig
514) piédouche
515) pied-de-mouton
516) pince-sans-rire
517) plénipotentiaire
518) plain-chant
519) plus d'un
520) pneumatique
521) pneumocoque
522) pognon
523) polochon
524) pont-l'évêque
525) pont-levis
526) pop-corn
527) porte-savon
528) postdater
529) postface
530) postprandial
531) potron-minet
532) prééminence
533) prébende
534) précipitamment
535) prédécesseur
536) prédominance
537) préfigurer
538) prégnance
539) prégnant
540) préoccupation
541) prépondérance
542) prérogative
543) présidentiel
544) presbytéral
545) prestidigitateur
546) probant
547) prochain
548) promulgation
549) puddler
550) pugnace
551) push-pull
552) quelqu'un
553) quelques-uns
554) quetzal
555) queue-d'aronde
556) queue-de-cheval
557) queue-de-morue
558) quinzaine
559) quinzième
560) réabsorber
561) réaccoutumer
562) réajuster
563) réapprovisionner
564) réarranger
565) récompenser
566) réconfort
567) réconforter
568) récupérer
569) réduplication
570) réemployer
571) réfrangible
572) réfrigérateur
573) réfringent
574) réfugier
575) régionalisme
576) régisseur
577) régulariser
578) réimprimer
579) réinfecter
580) réjouir
581) rémouleur
582) républicanisme
583) répugner
584) réquisitionner
585) réquisitoire
586) résidentiel
587) résoudre
588) résurrection
589) rétroprojecteur
590) rétroviseur
591) réveil-matin
592) révolutionnaire
593) révulsé
594) révulsion
595) ragtime
596) rectangulaire
597) rectiligne
598) responsabilité
599) revolver
600) rognon
601) rognonnade
602) roman-feuilleton
603) romsteck
604) ronchonner
605) ronronnement
606) ronronner
607) rougeoyer
608) rugbyman
609) rush
610) sèche-cheveux
611) sèche-linge
612) séjourner
613) sénatus-consulte
614) séquentiel
615) sérigraphie
616) saint-marcellin
617) saint-nectaire
618) samizdat
619) savon
620) sceau-de-salomon

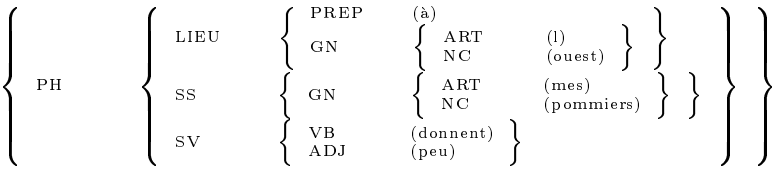
621) schlitte
622) sclérenchyme
623) scottish-terrier
624) script-girl
625) sebkha
626) seigneur
627) seigneurial
628) self-control
629) self-government
630) self-inductance
631) self-induction
632) self-made man
633) self-service
634) sergent-major
635) serviette-éponge
636) sex-shop
637) shogoun
638) show-business
639) shrapnel
640) shunt
641) smaragdite
642) snow-boot
643) soubresauter
644) sous-alimentation
645) sous-jacent
646) spermatozoïde
647) sténodactylo
648) stéréophonie
649) stop-and-go
650) submersion
651) subsistance
652) suffocation
653) surabonder
654) surréalisme
655) télé-enseignement
656) télécommande
657) télécommander
658) télécommunication
659) télédiffuser
660) télégraphier
661) téléguider
662) téléobjectif
663) témoignage
664) tétrarchie
665) tête-de-loup
666) technico-commercial
667) tempo
668) terminaison
669) terre-neuvas
670) terre-neuve
671) thérapeute
672) thérapeutique
673) thermodynamique
674) thermonucléaire
675) trébuchet
676) tréfondre
677) traîtreusement
678) trade-union
679) transsubstantiation
680) trench-coat
681) tribun
682) trognon
683) troubadour
684) tungstène
685) underground
686) untel
687) vérification
688) verbeux
689) verbosité
690) vergogne
691) vermifuge
692) vernaculaire
693) vertugadin
694) vesse-de-loup
695) vingt-trois
696) vlan
697) vraisemblance
698) yacht-club
699) zeugma
700) zloty

Liste des mots de PVM

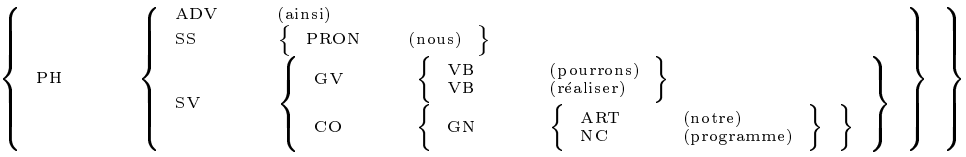
1) économie
2) état des routes
3) abonnement
4) adresse
5) adulte
6) agenda
7) aide
8) allemand
9) anglais
10) annulation
11) annuler
12) avalanches
13) avis régionaux
14) beaucoup
15) berthe
16) billet
17) cécile
18) carte de crédit
19) casino
20) chef téléopératrice
21) choisir
22) cinéma
23) concert
24) continuer
25) corriger
26) corso
27) cours des devises
28) crédit
29) début
30) dérangements
31) daniel
32) enfant
33) enneigement
34) espagnol
35) exemple
36) explications
37) exposition
38) faits divers
39) fin
40) français
41) françois
42) fumeur
43) galerie
44) galerie du manoir
45) gianadda
46) guide
47) henri
48) horaire
49) horoscope
50) ida
51) informations consommateurs
52) italien
53) jamais
54) jeanne
55) kilo
56) l heure
57) la bourse
58) le temps
59) les nouvelles
60) lire
61) louis moret
62) louise
63) météo
64) manifestation
65) marie
66) message
67) mode d emploi
68) moyennement
69) musée
70) nicolas
71) olga
72) oui
73) pas du tout
74) paul
75) petites annonces
76) place assise
77) politique
78) précédent
79) produit
80) quittance
81) quitter
82) réception
83) réduit
84) répéter
85) réservation
86) résumé
87) rarement
88) relevé de banque
89) renseignements internationaux
90) renseignements nationaux
91) robert
92) romanche
93) sans opinion
94) service des télécommunications
95) service international
96) services ptt
97) ski
98) sondage
99) souvent
100) sport
101) suivant
102) suzanne
103) télégramme
104) télévision
105) tarif
106) taxes
107) théâtre
108) thérèse
109) tous les jours
110) transfert
111) ulysse
112) validation
113) vents
114) victor
115) william
116) xavier

Arbres syntaxiques du corpus PolyPhrase

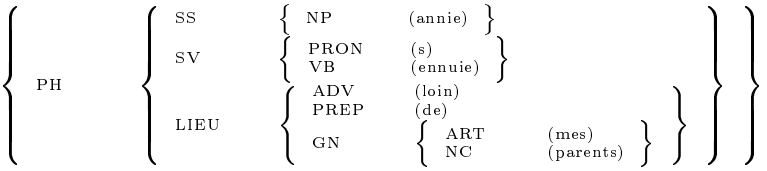
À l'ouest mes pommiers donnent peu.



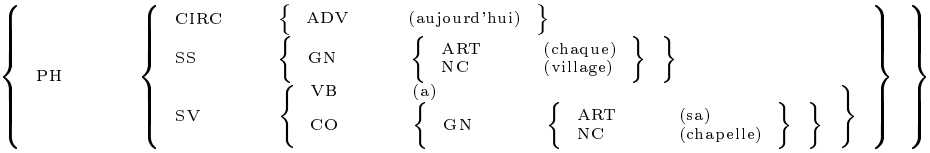
Ainsi, nous pourrons réaliser notre programme.



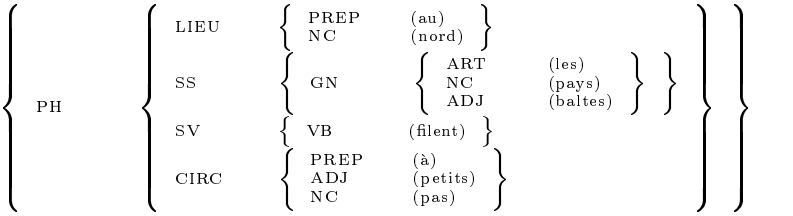
Annie s'ennuie loin de mes parents.



Aujourd'hui, chaque village a sa chapelle.

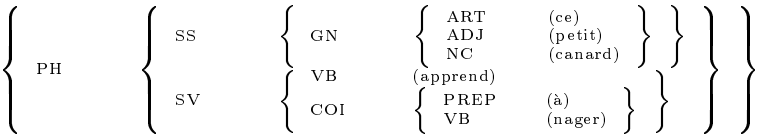


Au nord, les pays baltes filent à petits pas.

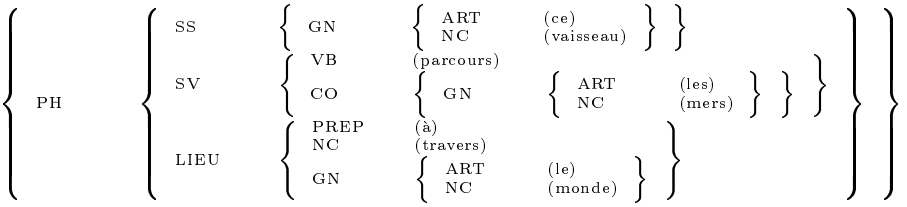


Ce bonbon contenait trop de sucre.

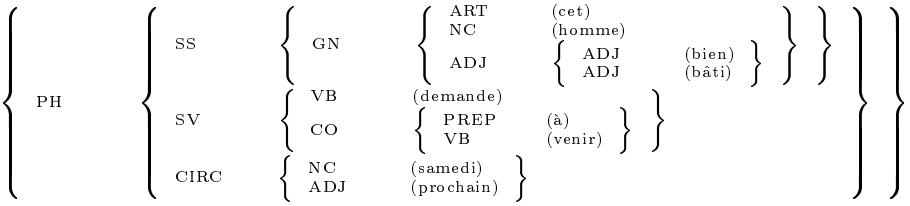
Ce petit canard apprend à nager.



Ce vaisseau parcourt les mers à travers le monde.

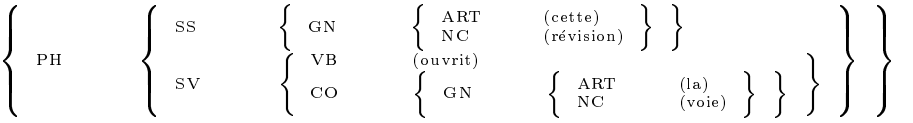


Cet homme bien bâti demande à venir samedi prochain.

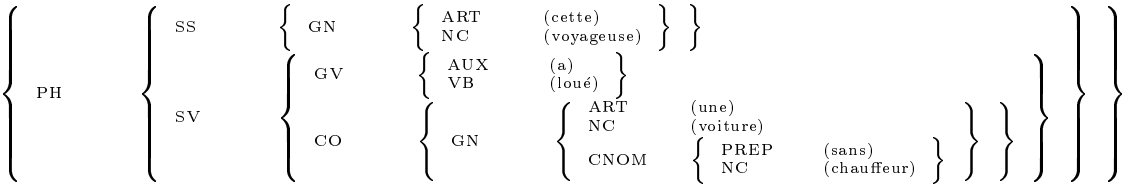


Cette cage contient mon oiseau.

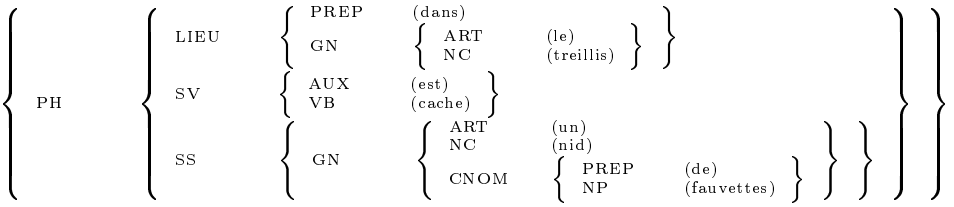
Cette révision ouvrit la voie



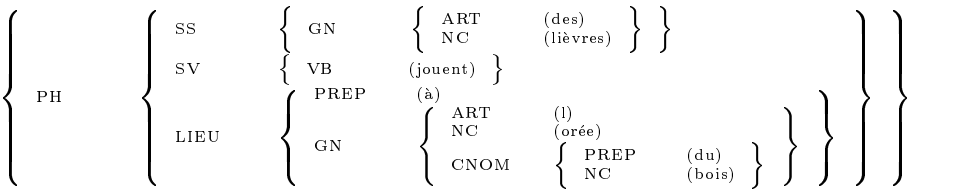
Cette voyageuse a loué une voiture sans chauffeur.



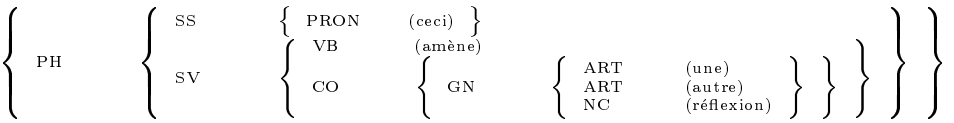
Dans le treillis est caché un nid de fauvettes.



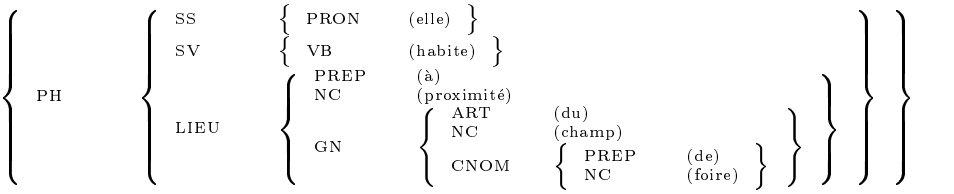
Des lièvres jouent à l'orée du bois.



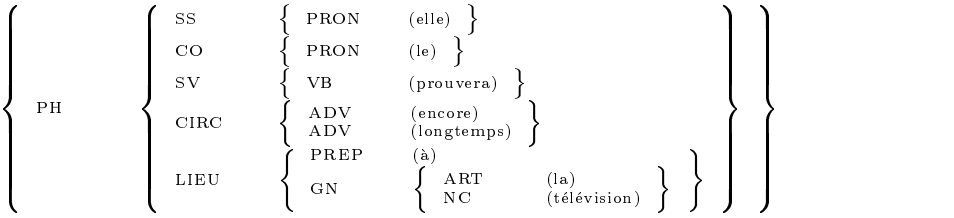
Ceci amène une autre réflexion.



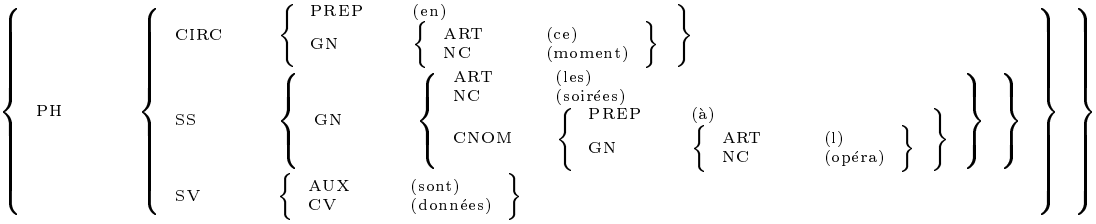
Elle habite à proximité du champ de foire.



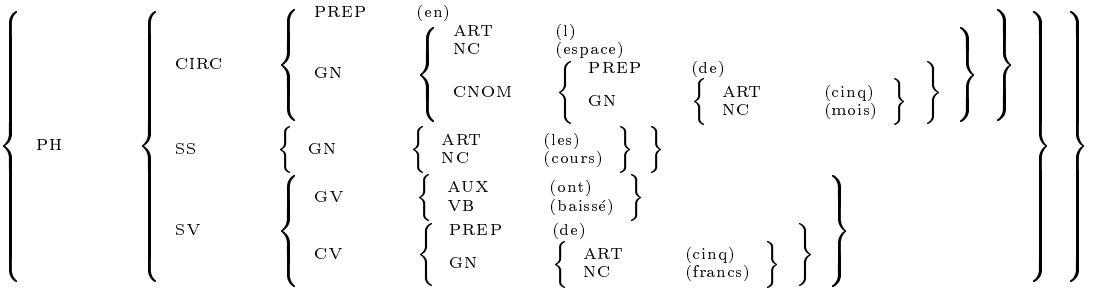
Elle le prouvera encore longtemps, à la télévision.



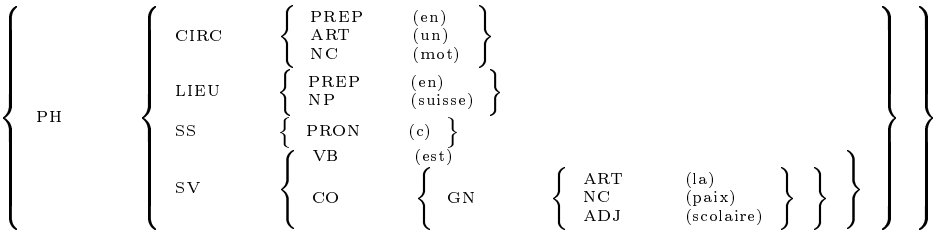
En ce moment, les soirées à l'opéra sont données.



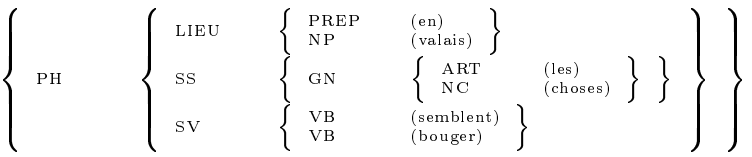
En l'espace de cinq mois, les cours ont baissé de cinq francs.



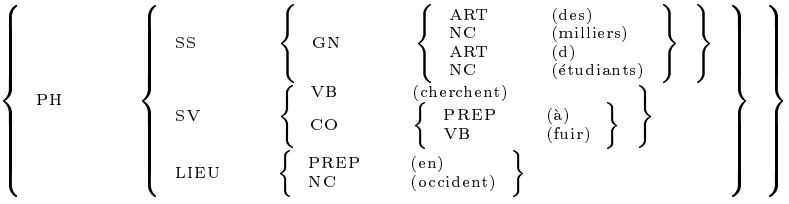
En un mot, en Suisse, c'est la paix scolaire.



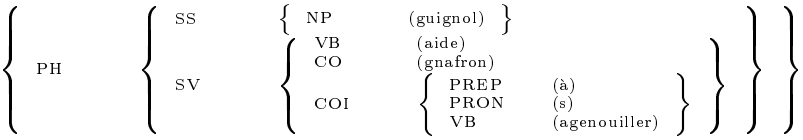
En Valais, les choses semblent bouger.



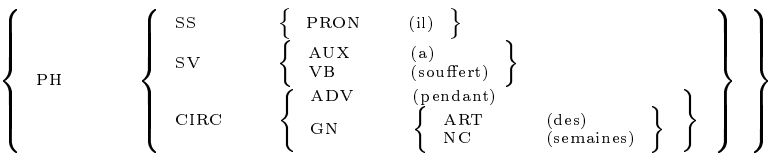
Des milliers d'étudiants cherchent à fuir en occident.



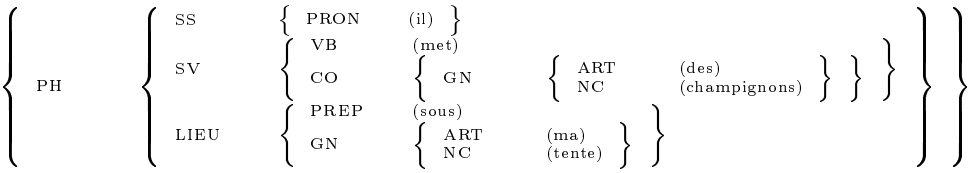
Guignol aide Gnafron à s'agenouiller.



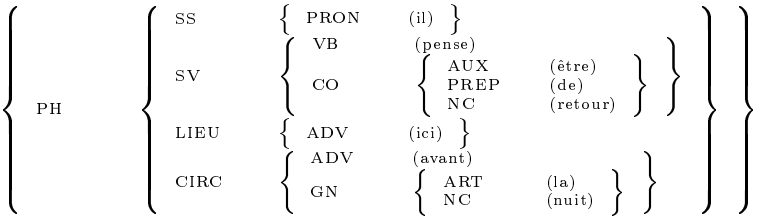
Il a souffert pendant des semaines.



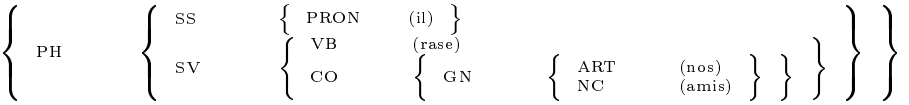
Il met les champignons sous ma tente.



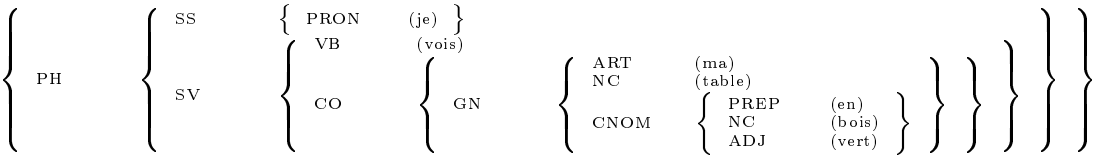
Il pense être de retour, ici avant la nuit.



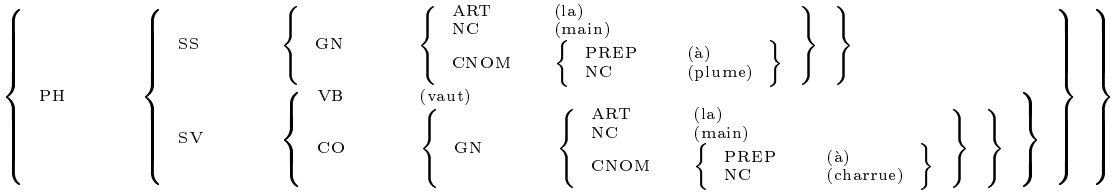
Il rase nos amis.



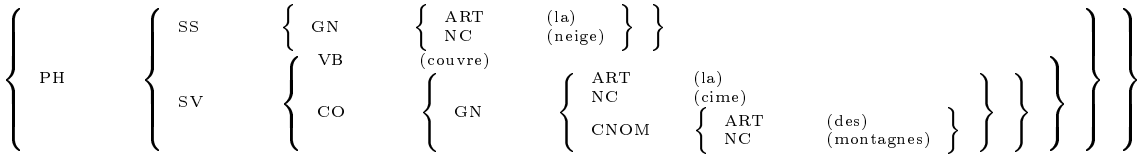
Je vois ma table en bois vert.



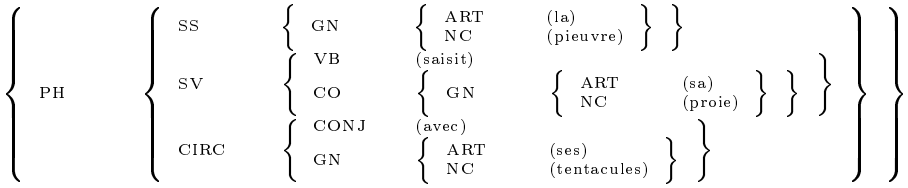
La main à plume vaut la main à charrue.



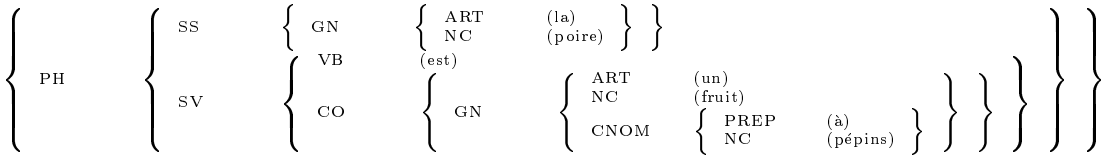
La neige couvre la cime des montagnes.



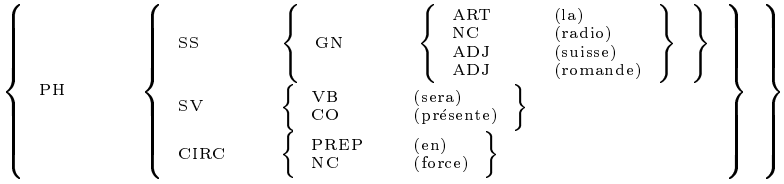
La pieuvre saisit sa proie avec ses tentacules.



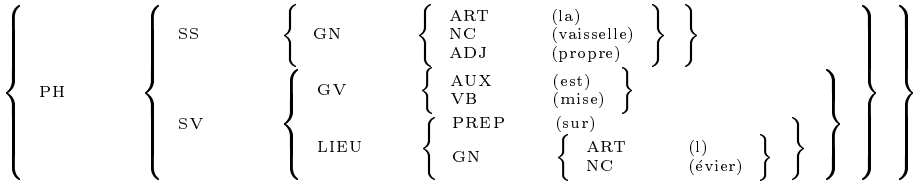
La poire est un fruit à pépins.



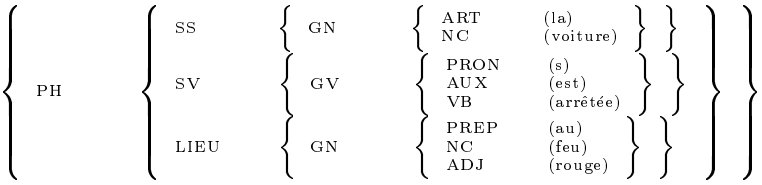
La radio suisse romande sera présente en force.



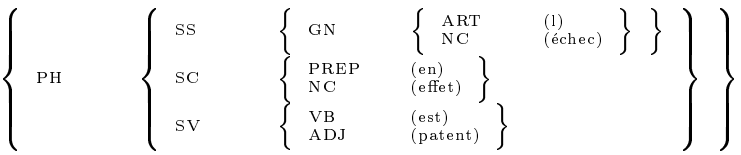
La vaisselle propre est mise sur l'évier.



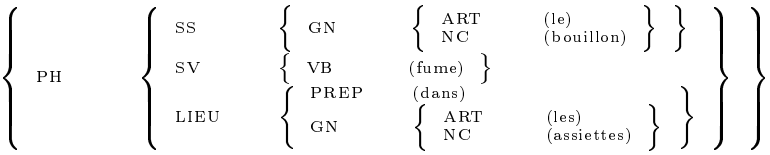
La voiture s'est arrêtée au feu rouge.



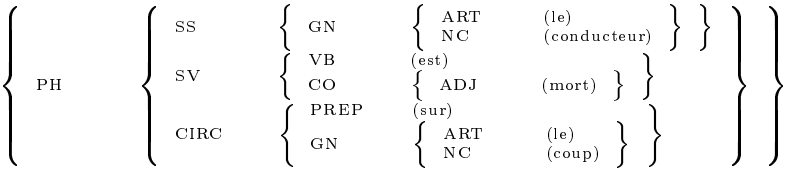
L'échec, en effet, est patent.



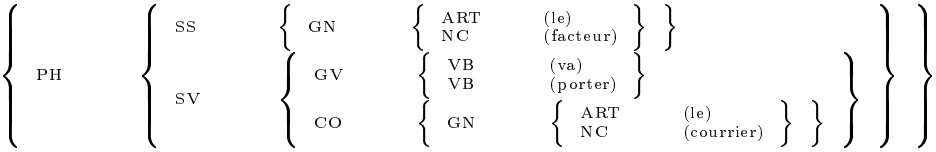
Le bouillon fume dans les assiettes.



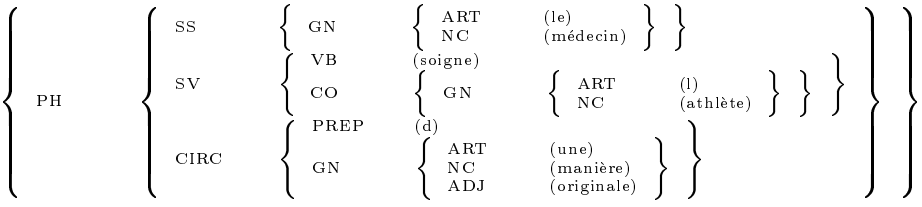
Le conducteur est mort sur le coup



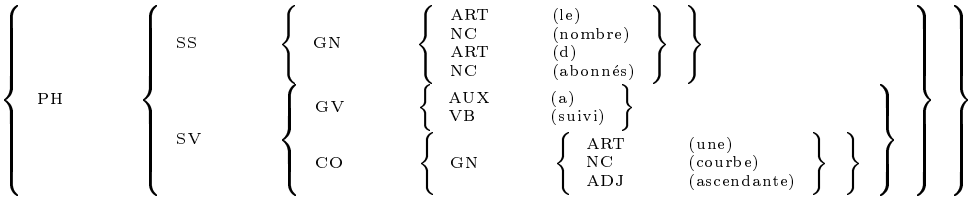
Le facteur va porter le courrier.



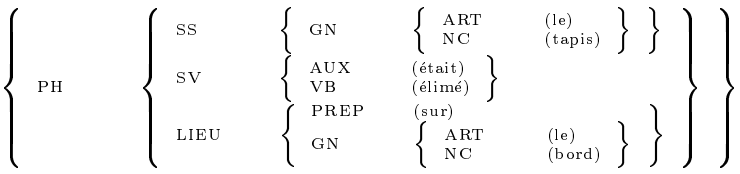
Le médecin soigne l'athlète d'une manière originale.



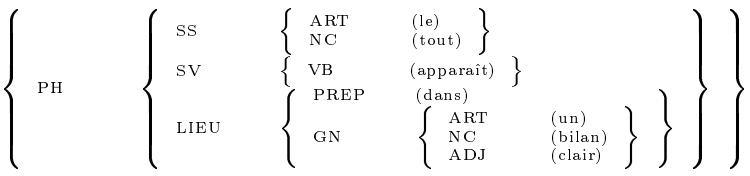
Le nombre d'abonnés a suivi une courbe ascendante.



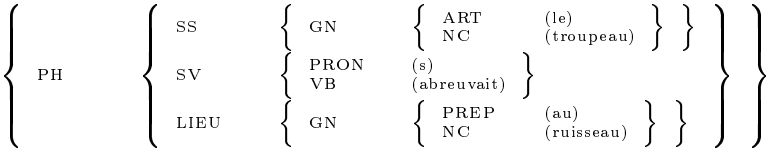
Le tapis était élimé sur le bord.



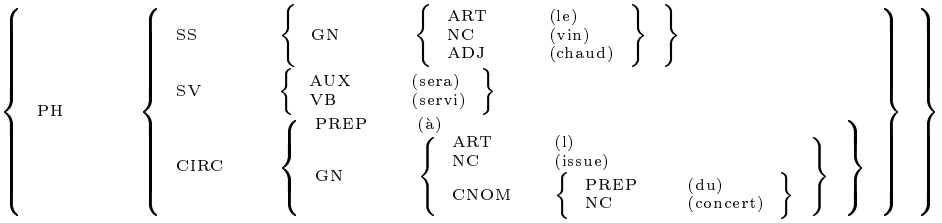
Le tout apparaît dans un bilan clair.



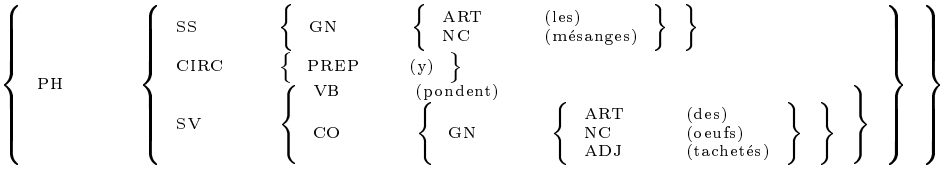
Le troupeau s'abreuvait au ruisseau.



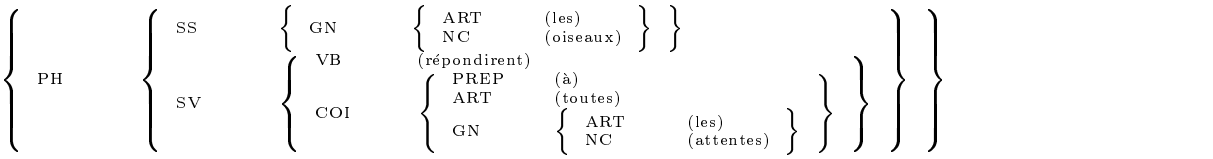
Le vin chaud sera servi à l'issue du concert.



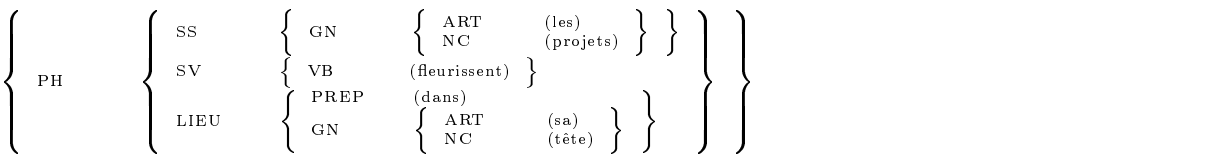
Les mésanges y pondaient des oeufs tachetés.



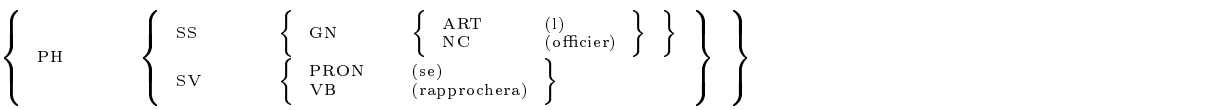
Les oiseaux répondirent à toutes les attentes.



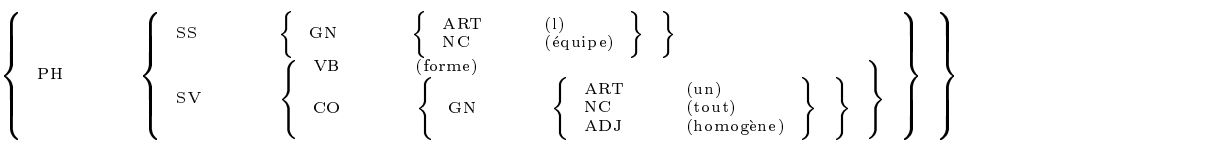
Les projets fleurissent dans sa tête.



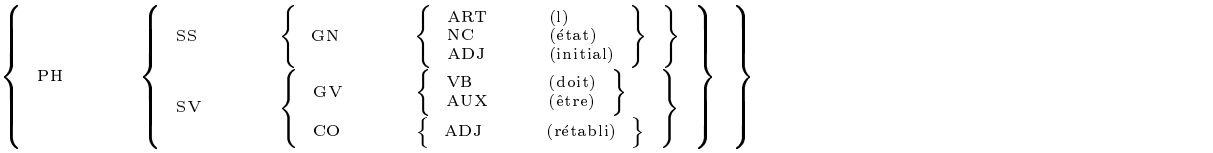
L'officier se rapprochera.



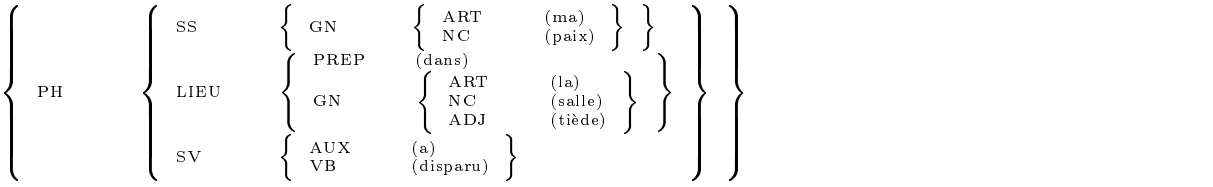
L'équipe forme un tout homogène.



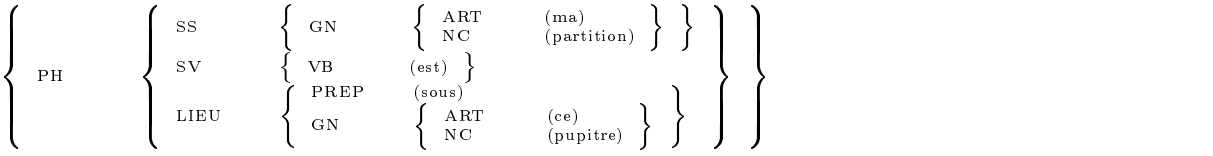
L'état initial doit être rétabli.



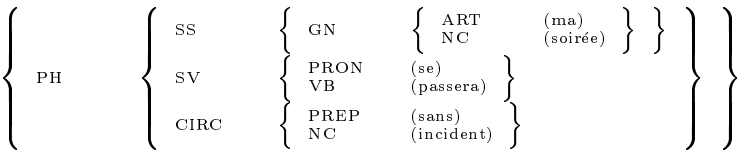
Ma paix dans la salle tiède a disparu.



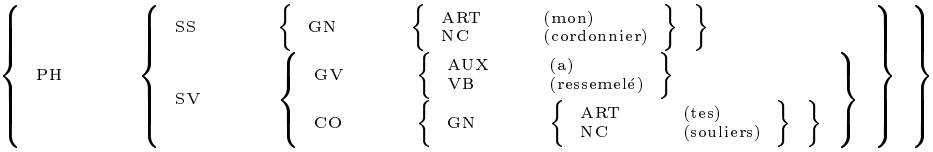
Ma partition est sous ce pupitre.



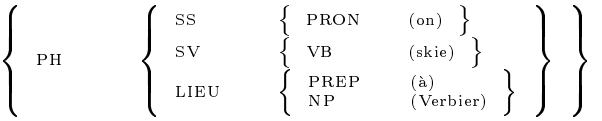
Ma soirée se passera sans incident.



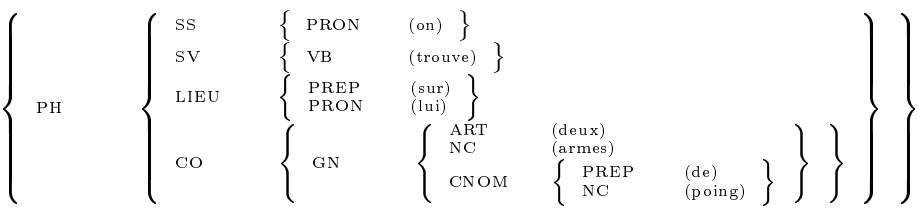
Mon cordonnier a ressemelé tes souliers.



On skie à Verbier.

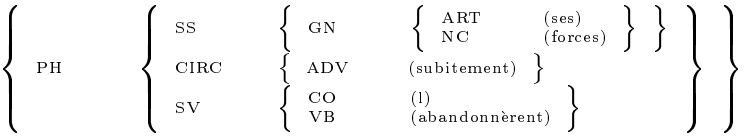


On trouve sur lui deux armes de poing.

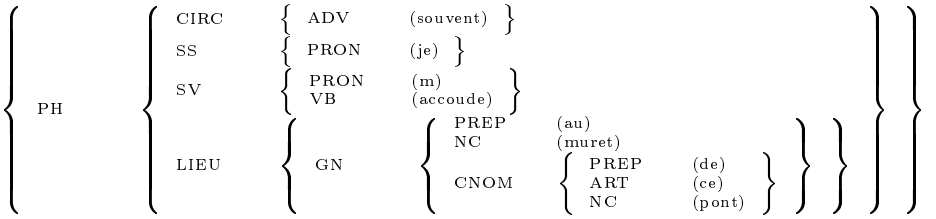


Sa voisine est inimitable.

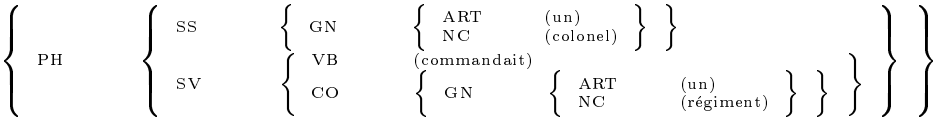
Ses forces subitement l'abandonnèrent.



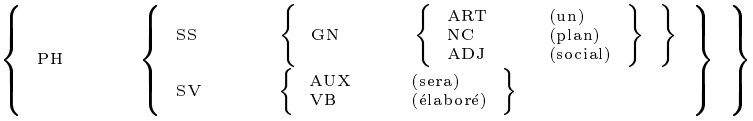
Souvent, je m'accoude au muret de ce pont.



Un colonel commandait le régiment.

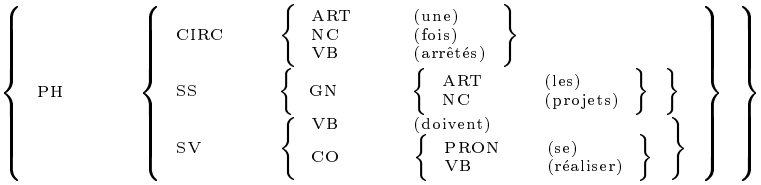


Un plan social sera élaboré.

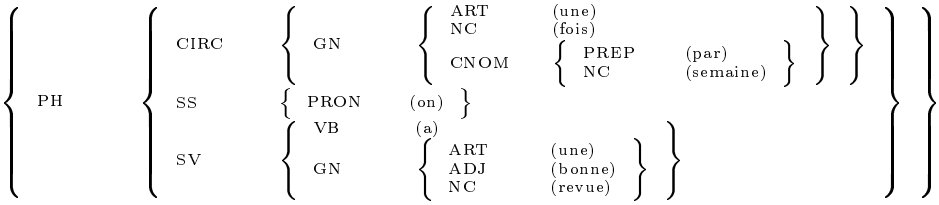


Une atmosphère obscure enveloppe la ville.

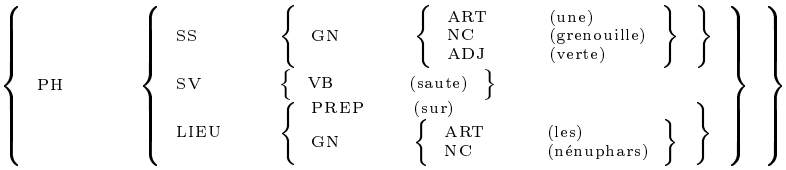
Une fois arrêtés, les projets doivent se réaliser.



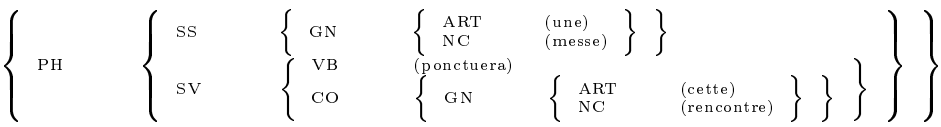
Une fois par semaine, on a une bonne revue.



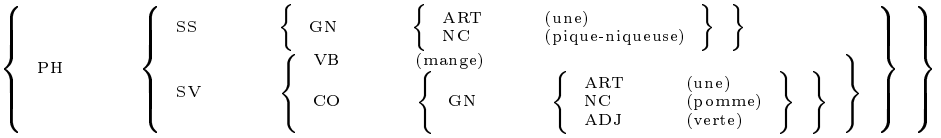
Une grenouille verte saute sur les nénuphars.



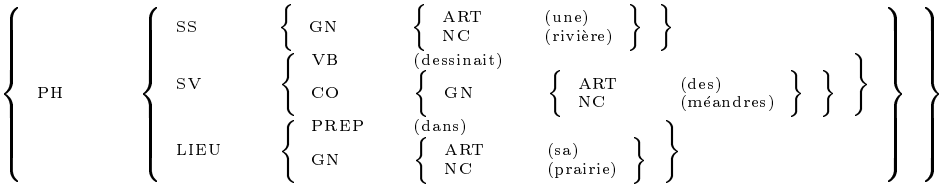
Une messe ponctuera cette rencontre.



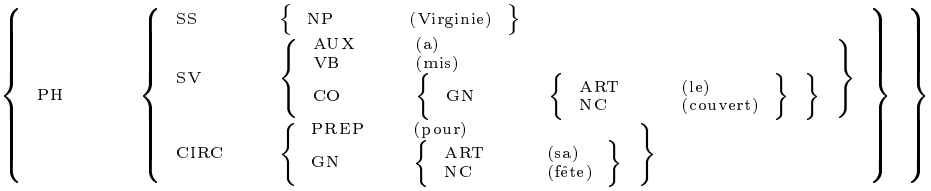
Une pique-niqueuse mange une pomme verte.



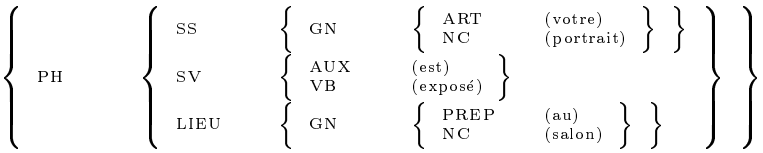
Une rivière dessinait des méandres dans sa prairie.



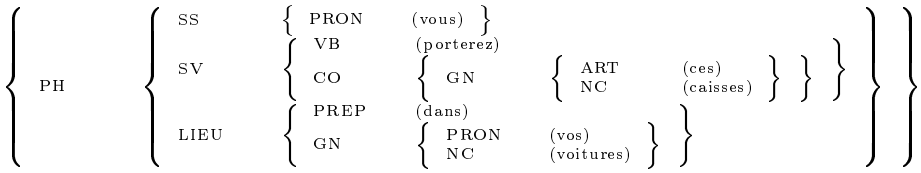
Virginie a mis le couvert pour sa fête.



Votre portrait est exposé au salon.

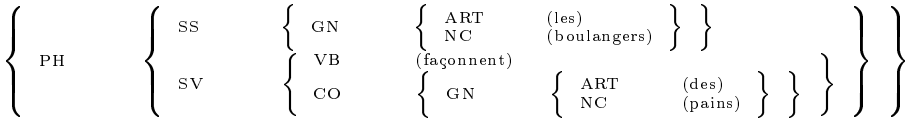


Vous porterez ces caisses dans vos voitures.

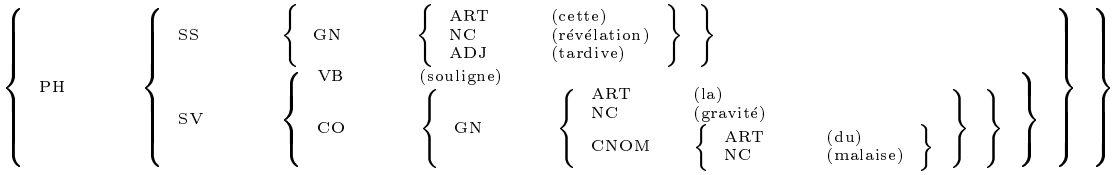


Vous rentrerez avec la vedette suivante.

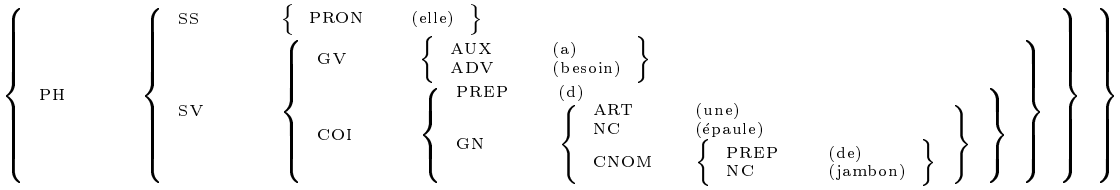
Les boulangers façonnent des pains.



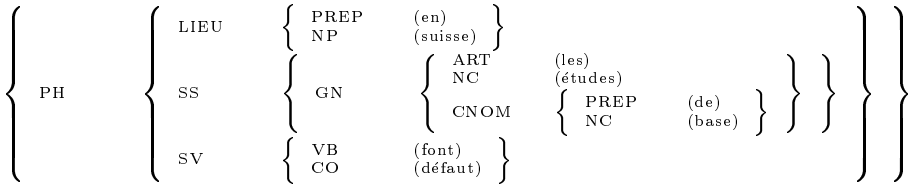
Cette révélation tardive souligne la gravité du malaise.



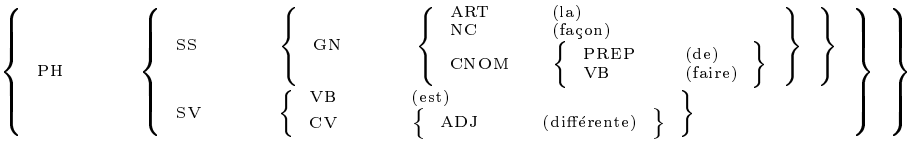
Elle a besoin d'une épaule de jambon



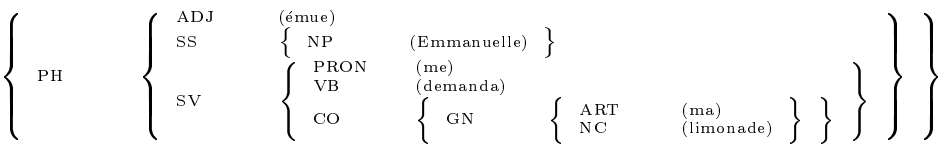
En Suisse, les études de base font défaut.



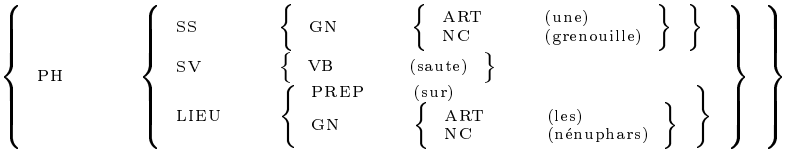
La façon de faire est différente.



Émue, Emmanuelle me demanda ma limonade.



Une grenouille saute sur les nénuphars



TRAITEMENT DE LA PROSODIE EN RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

RÉSUMÉ

Les travaux présentés concernent le traitement de la prosodie dans les systèmes de reconnaissance de la parole. Les principales étapes d'une approche prosodique classique (mesure des paramètres, corrections microprosodiques et perceptives, application de règles suprasegmentales) font l'objet de discussions qui introduisent les choix faits pour chacune d'elles.

Dans la première partie de ce mémoire, sont abordées en détail les variations microprosodiques (ou segmentales) des paramètres prosodiques. Un inventaire des principaux phénomènes abondamment étudiés par le passé est tout d'abord proposé. Chacun d'eux est alors étudié sur des corpus de mots prononcés isolément afin de déterminer d'une part, si l'emploi de techniques d'extraction automatique des paramètres autorise l'usage de ces variations en tant qu'indice pertinent lors d'une phase de décodage acoustico-phonétique, et d'autre part, de vérifier la robustesse d'un processus de correction des paramètres prosodiques objectifs à l'aide de coefficients microprosodiques dans le cadre d'un traitement automatique. L'étude montre que, dans le cadre restreint de mots isolés, peu de phénomènes microprosodiques sont observables de manière significative par les techniques retenues, rendant pour le moins incertaine toute tentative de correction des valeurs objectives des paramètres de durée, de fréquence fondamentale et d'intensité. Les indices pertinents ont été intégrés avec succès à un module d'accès lexical.

La seconde partie du mémoire présente les difficultés majeures liées à l'analyse prosodique suprasegmentale par un expert et tente d'expliquer le recours de plus en plus fréquent à l'outil statistique pour y parvenir. Un système d'étude corrélative automatique a été développé qui revendique d'une part, une assistance à l'analyse prosodique par un expert (en offrant des outils de visualisation et d'interrogation), et d'autre part, une fonction prédictive de la structure linguistique d'un message à décoder. Deux applications de ce système sont alors proposées, l'une de reconnaissance de nombres décimaux (notre système s'est par exemple montré apte à localiser précisement le mot *virgule* dans une chaîne inconnue à partir des informations prosodiques seules), l'autre de reconnaissance de phrases de type *lues* avec des résultats qui valident pleinement notre approche globale de résolution.

MOTS CLÉS

Prosodie, microprosodie, reconnaissance automatique de la parole, étude corrélative