UN SYSTÈME PRÉDICTIF DE LA STRUCTURATION SYNTAXICO-RYTHMIQUE D'UN ÉNONCÉ À L'AIDE D'INFORMATIONS PROSODIQUES.

Philippe LANGLAIS†, Jean-Luc COCHARD‡, Henri MÉLONI†

†Laboratoire d'Informatique – 339, Chemin des Meinajariès – BP 1228 – 84911 Avignon Cedex 9
Tél.: 90 84 35 25 – Fax: 90 84 35 01 – e-mail: {langlais,meloni}@univ-avignon.fr
‡IDIAP – 4 rue du Simplon – CH-1920 Martigny
Tél.: (+41) 26 22 76 64 – Fax: (+41) 26 22 78 18 – e-mail: cochard@idiap.ch

ABSTRACT

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.

1. INTRODUCTION

Plusieurs études ont montré par le passé les nombreuses possibilités d'utilisation de la prosodie dans les systèmes de reconnaissance automatique de la parole (RAP)[Lea 80, Vaissière 88]. Les systèmes existants [Carbonell 82, Bonin 90, Nasri 90] ont cependant tous contribués, malgré eux, à montrer que l'énoncé de règles prosodiques aussi complexes soient-elles, n'est pas d'une grande efficacité en reconnaissance automatique de la parole. Les principales raisons de cet échec sont premièrement analysées introduisant ainsi notre système de prédiction automatique de la structure syntaxico-rythmique d'un énoncé. Les résultats de prédiction mesurés sur deux bases de parole continue (phrases et nombres décimaux) sont ensuite présentés.

2. POSITION DU PROBLÈME

2.1. Difficultés

Parmi les nombreuses difficultés de l'intégration de la prosodie dans les systèmes de RAP, il convient de distinguer celles qui sont imputables aux systèmes (qui soulèvent entre autre le problème fondamental et non trivial de la fusion des connaissances) de celles qui sont inhérentes à l'analyse prosodique.

D'un point de vue purement prosodique, et abstraction faite des problèmes non réglés totalement de la mesure des paramètres et de leur correction microprosodique et perceptuelle, nous rappelons la principale cause de l'échec à l'intégration de la prosodie en reconnaissance : la prosodie est le fruit d'une interaction complexe entre différents niveaux de structuration du message (syntaxique, sémantique, pragmatique, rythmique, ...) qui peuvent être conflictuels [Vaissière 88].

2.2. Quelles solutions ?

Face à la multiplicité des facteurs intervenant sur les variations des paramètres prosodiques, il semble difficile à un expert de mener à bien son analyse sans y introduire un biais que VanSanten appelle la "piecemeal analysis" (analyse locale de quelques facteurs a priori importants) [Santen 94]. Cet auteur remarque alors très justement que le recours à l'outils statistique permet de s'affranchir de cette analyse locale qui n'est pas souhaitable. On assiste d'ailleurs depuis quelques années à une tendance de plus en plus marquée au remplacement des connaissances formalisées par des experts par des composantes statistiques [Veilleux 93, Pagel 95].

Nous pensons qu'une solution satisfaisante doit tenir compte des avantages de chacune des deux approches et nous préconisons l'usage de méthodes statistiques pour épauler l'expert dans son analyse. C'est en ce sens qu'a été conçu le système que nous allons présenter.

3. LE SYSTÈME

3.1. Étiquetage prosodique

L'étiquetage prosodique consiste à réduire l'ensemble des paramètres prosodiques à un nombre limité d'indices pertinents, avec le minimum de perte d'information. Ce problème délicat n'a actuellement pas de solution unanime (cf. le workshop de Stockholm'95 sur ce thème). En l'absence de consensus, nous avons retenu un jeu assez classique de 40 étiquettes (9 indices de durée, 9 indices d'intensité et 22 indices pour la f0) qui caractérisent chaque noyau vocalique : maximum et minimum de chaque paramètre sur l'énoncé, émergence de la valeur d'un paramètre sur un noyau (n) par rapport aux valeurs correspondantes sur les segments adjacents (n-2,n-1,n+1,n+2), différents codages en niveaux (1 à 4) de la valeur de f0 d'un noyau vocalique sur une échelle correspondant au découpage en 4 zones de la dynamique du paramètre sur l'ensemble de l'énoncé, etc.

Cet étiquetage simple et entièrement automatique (nécessitant pour seule entrée le signal de parole) ne fait intervenir aucune étape de correction microprosodique ou perceptive et ce en raison des expérimentations et remarques exposées dans [Langlais 95].

Cette étape de caractérisation paramétrique est par nature criticable, aussi nous contentons nous de remarquer que notre système est ouvert à l'ajout de nouvelles étiquettes (pour autant qu'elles soient automatiquement calculables) ou au contraire à la suppression d'autres.

3.2. Le principe

Nous émettons l'hypothèse que la distribution des configurations prosodiques sur la totalité d'un énoncé n'est pas du tout aléatoire mais est au contraire suffisamment régulière pour autoriser des prédictions structurelles à partir des seules informations prosodiques.

Pour vérifier cette hypothèse, nous avons élaboré le système ProStat qui propose deux fonctionalités :

1. l'aide à l'analyse prosodique de grand corpus par un expert. Le système dispose d'une interface graphique permettant de visualiser des contours paramétriques divers. Il renseigne (par l'usage d'un jeu restreint de requètes) son utilisateur sur les corrélations (mesurées sur un corpus donné) entre des indices prosodiques et différents niveaux d'organisation ou points particuliers du message.

2. La prédiction de la structure linguistique d'un énoncé à partir des seuls indices prosodiques calculés automatiquement. Le travail de l'expert consiste uniquement à rassembler un corpus de parole et d'en fournir la description linguistique (syntaxique et/ou sémantique, etc.).

Dans cette étude, nous avons décidé de réduire notre champ d'analyse aux seules interactions du rythme (que nous définissons ici par la distribution du nombre de voyelles des différents groupes d'un énoncé) et de la syntaxe sur la distribution des indices prosodiques.

Le principe de base que nous ne développerons pas, faute de place, est la création automatique d'un graphe orienté, à partir d'un corpus d'apprentissage [Langlais 95, pp. 141 à 144]. Chaque énoncé du corpus est décrit par sa décomposition grammaticale sous forme arborescente (fournie manuellement dans notre cas), par un alignement phonétique obtenu par des modèles markoviens développés à l'IDIAP et par son treillis prosodique automatiquement calculé à partir du signal de parole. Chaque arc de ce graphe est une contrainte de nature syntaxique et/ou rythmique dérivée du corpus d'apprentissage. Chaque nœud du graphe décrit une structure syntaxico-rythmique particulière et contient des informations comme le nombre de fois où il a été visité lors de l'apprentissage, le nombre d'occurrences de chaque étiquette prosodique apposée à l'initiale ou en finale d'un groupe quelconque de la structure décrite, etc. Plus on avance dans le graphe et plus la structure syntaxico-rythmique décrite est précise.

En dotant notre système d'une métrique simple, capable de fournir une distance entre le treillis prosodique d'un énoncé inconnu et les informations contenues dans un nœud donné du graphe, nous disposons d'un outil capable de réaliser des prédictions syntaxico-rythmiques.

Nous allons maintenant décrire deux expériences qui valident notre système et confirment l'hypothèse que nous avons formulée.

3.3. Résultats

Ces deux expériences ont commencé par une étape préalable d'apprentissage (que nous décrivons brièvement pour chaque tâche) qui a nécessité l'intervention d'un expert pour fournir les décompositions syntaxiques sous forme arborescente de chaque énoncé des corpus d'apprentissage.

Nous avons ensuite demandé à notre système de formuler des hypothèses syntaxico-rythmiques valuées complètes (arbre syntaxique complet ainsi que le nombre de voyelles de chacune de ses feuilles) à l'aide du graphe issu de l'apprentissage et des treillis prosodiques automatiquement calculés pour des énoncés issus de différents corpus de tests.

3.3.1 Les phrases isolées

Un corpus de 500 phrases isolées, répétitions en nombre inégal de 80 phrases différentes par 50 locuteurs à travers un canal téléphonique assez bruité a ici servi de corpus d'apprentissage. Ces phrases de structures syntaxiques simples (principalement des phrases composées dans un ordre variable d'un groupe sujet, d'un groupe verbal et d'un ou de plusieurs compléments circonstanciels) contenaient de 4 à 17 voyelles.

ProStat a montré des capacités très intéressantes qui permettent son utilisation pour la reconnaissance de phrases :

1) sur l'ensemble des énoncés utilisés pour l'apprentissage, plus de 90% des 500 phrases sont classées en première position (parmi un choix moyen de 15 possibilités) ;

2) sur un corpus de test de 300 phrases (la majorité étant des répétitions différentes des 80 phrases de base du corpus d'apprentissage, pas nécessairement prononcées par les mêmes locuteurs), la première hypothèse affecte correctement près de 60% des énoncés à la bonne structure. Trois à cinq hypothèses suffisent pour assurer l'association de l'énoncé à sa structure syntaxico-rythmique (voir figure 1).

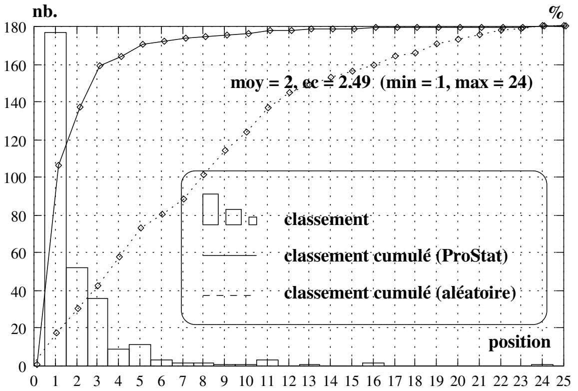

Figure 1. *Classement des hypothèses fournies par ProStat pour les phrases du corpus de test (abscisse = rang de la bonne proposition, ordonnée à gauche = nb. de propositions formulées, ordonnée à droite = pourcentage cumulé de propositions justes). La courbe en pointillé correspond à une proposition aléatoire.*

3.3.1 Les nombres décimaux

Un corpus de 500 nombres décimaux prononcés via le même canal téléphonique par une cinquantaine de locuteurs a servi de corpus d'apprentissage dans cette expérience (une grammaire classique des nombres a ici fourni les arbres grammaticaux de chaque nombre du corpus).

Là encore le système s'est avéré apte à associer les treillis prosodiques des nombres du corpus d'apprentissage à leur bonne structure syntaxico-rythmique dans plus de 80% des cas (avec en moyenne 15 structures possibles par nombre).

Sur un corpus de test de 298 nombres, le système a montré un taux de prédiction en tête d'un peu moins de 50%, ce qui tend à indiquer que l'information prosodique des nombres décimaux est moins riche que celle des phrases prononcées isolément.

En analysant les hypothèses non classées premières, il est apparu très clairement que le mot "virgule" était très souvent bien positionné (voir figure 2). Nous n'avons pas réussi à approcher ce score par la formulation de règles locales spécifiques, ce qui semble confirmer l'importance de l'information prosodique prise sur la globalité d'un énoncé.

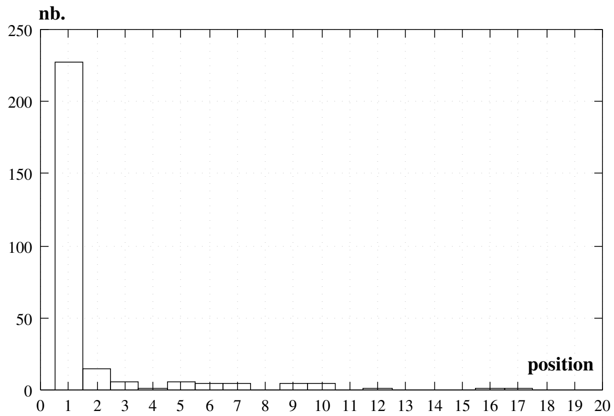

Figure 2. *Classement des hypothèses fournies par ProStat pour les nombres du corpus de test en considérant uniquement l'exactitude de la position du mot virgule dans la chaîne.*

4. CONCLUSIONS

De ces deux expériences, nous pouvons retenir que l'information prosodique — même extraite automatiquement — permet de formuler des hypothèses syntaxico-rythmiques avec un taux de réussite significatif. Le choix de notre approche sans connaissance *a priori* permet également de valider l'hypothèse d'une distribution non aléatoire des indices prosodiques sur la globalité d'un énoncé.

Si l'on peut à nouveau convenir que l'information prosodique doit être intégrée avec profit dans les systèmes de reconnaissance de la parole, nous n'avons cependant pas montré, pour chaque type d'utilisation, les efforts importants qui restent nécessaires pour aboutir à un emploi optimal de ces données et de ces connaissances. Des travaux supplémentaires doivent être menés afin de définir — pour une tâche donnée — un sous-ensemble d'indices prosodiques pertinents (actuellement, chaque indice participe de manière égale à la notation par le système des différentes hypothèses structurelles) ; les tests doivent de plus être étendus à des énoncés aux structures plus variées.

Bibliographie

[Bonin 90] J.J. Bonin et J.M. Pierrel. Fréquence fondamentale et durée pour la détection de frontières syntagmatiques en parole continue. Dans *XVIIIème JEP*, Montréal, 1990.

[Carbonell 82] N. Carbonell, J.P. Haton, F. Lonchamp, et J.M. 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.

[Langlais 95] P. Langlais. *Utilisation de la prosodie en reconnaissance automatique de la parole.* Thèse, Université d'Avignon, 1995.

[Lea 80] Wayne A. Lea. Prosodic aids to speech recognition. Dans *Trends in Speech Recognition*, W.A. Lea, éditeur. Prentice Hall, 1980.

[Nasri 90] M.K. Nasri, G. Caelen-Haumont, et J. Caelen. Utilisation de règles prosodiques en reconnaissance de la parole. Dans *XVIIIèmes JEP*, 1990.

[Pagel 95] V. Pagel, N. Carbonell, et J. Vaissière. Spotting prosodic boundaries in continuous speech in french. Dans *XIIIth International Congress of Phonetic Sciences*, volume 4, pages 308–311, Stockholm, 1995.

[Santen 94] J.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, 1994.

[Vaissière 88] J. Vaissière. The use of prosodic parameters in automatic speech recognition. Dans *Recent Advances in Speech Understanding and Dialog Systems*, volume F46. NATO ASI Series, 1988.

[Veilleux 93] 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, 1993.