Pondération d'indices prosodiques dans un système automatique de prédiction syntaxico-rythmique d'un énoncé

†*Philippe Langlais and* ‡*Jean-Luc Cochard*

†Centrum för talteknologi (CTT) – KTH
Box 70014 — S-10044 Stockholm
Tél.: ++46 8 790 75 65
Fax: ++46 8 790 78 64
http://www.speech.kth.se/

‡CT-SPI, Swisscom
Ostermundingenstr. 93 — CH-3000 Berne
Tél.:++41 31 342 87 34
Fax: ++41 31 342 29 53
Jean-Luc.Cochard@Swisscom.com

ABSTRACT

This paper concerns the study of information derived from the melodic, temporal and intensity characteristics of the material to be recognized in a speech recognition system, in French.

More precisely, it describes experiments we achieved at the suprasegmental levels with a system that outperform automatic correlation between prosodic labels and linguistic organization of a message to decode. An overview of the system is described along with the results of experiments carried out to determine which prosodic indexes are best-suited for syntactic and rhythmycal prediction.

1. INTRODUCTION

Il est bien connu que les structures prosodiques et syntaxiques mises en œuvre dans le processus de communication ne sont ni identiques ni non corrélées. Savoir quand et comment ces deux structures sont mises en relation est d'un intérêt évident dans des tâches telles que la désambiguïsation d'hypothèses concurrentes dans un système de dialogue. Ce type d'application explique en partie le regain d'intérêt récent de l'intégration nécessaire de la prosodie en reconnaissance automatique de la parole [2, 6, 7, 4].

Cet article présente des expériences visant à répondre à ces deux questions :

- Est-ce que l'information prosodique est, lorsque qu'extraite automatiquement, suffisante à la prédiction de la structure syntaxique et/ou rythmique d'un énoncé inconnu ?
- Dans l'affirmative, existe t-il des indices prosodiques, parmi l'ensemble de ceux utilisés, qui jouent un rôle prépondérant dans le processus de prédiction ?

2. BASE DE DONNÉES

Les deux expériences décrites ici utilisent un sous-ensemble d'une base de parole téléphonique réalisée à l'IDIAP en collaboration avec Swisscom [1].

2.1. *Le corpus de phrases*

Un corpus d'apprentissage de 500 phrases a été sélectionné. Prononcées par 50 appelants (via différentes lignes téléphoniques), ces phrases sont des répétitions en nombre inégal de 80 phrases différentes, de structure syntaxique simple contenant de 4 à 17 voyelles (voir figure 1).

⋆ Le tout apparaît dans un bilan clair.
⋆ Cette révélation tardive souligne la gravité du malaise.

Figure 1: Exemples de phrases utilisées.

À des fins de tests, un corpus de 300 phrases a aussi été sélectionné. La plupart de ces phrases étant des répétitions (généralement par d'autres locuteurs), des phrases du corpus d'apprentissage.

2.2. *Le corpus de nombres décimaux*

Un corpus d'apprentissage de 500 nombres décimaux (voir figure 2) enregistré dans les même conditions (lignes téléphoniques, 50 locuteurs, etc.) a été reuni.

⋆ Treize mille deux cent quarante virgule dix.
⋆ Quatre mille virgule huit.

Figure 2: Exemples de nombres décimaux utilisés.

Enfin, un corpus de test de 280 nombres décimaux a été collecté. 148 de ces nombres ont une structure "syntaxico-rhytmique" apparaissant au moins une fois dans le corpus d'apprentissage.

3. ÉTIQUETAGE PROSODIQUE

En dépit de systèmes d'étiquetage émergant comme ToBI [8], l'étiquetage prosodique n'a toujours pas trouvé de solution satisfaisante. Un ensemble d'étiquettes a été retenu dans cette étude (9 indices de durées, 9 d'intensité et 22 indices décrivant la fréquence fondamentale). Chaque indice, de nature binaire (présent/absent), décrit chaque noyau vocalique (émergence locale d'un paramètre au regard des voyelles adjacentes, émergence globales, extrema, etc.). Ce processus simple et complètement automatique [5] calcule pour un signal de parole donné une matrice prosodique qui constitue la seule entrée du processus de reconnaissance que nous décrivons maintenant.

4. LE SYSTÈME DE PRÉDICTION

4.1. *Hypothèse de travail*

Le but des expériences présentées est de valider ou d'infirmer l'hypothèse de travail suivante : la distribution des configurations des paramètres prosodiques sur l'ensemble d'une phrase n'est pas aléatoire, mais répond bien au contraire à des contraintes structurelles (syntaxiques, rythmiques, sémantiques) suffisamment régulières pour un traitement automatique.

4.2. *Description générale*

Le système décrit ne fait aucune hypothèse quant à l'organisation des différentes contraintes régissant les paramètres prosodiques pas plus qu'il ne présuppose l'existence d'entités prosodiques particulières comme par exemple *l'accent*, notion pour laquelle nous n'avons pas trouvé de caractérisation acoustique "robuste". L'approche fondatrice du système est une étude menée par les données des corrélations entre différents niveaux linguistiques et les indices prosodiques mesurés automatiquement. Le système est composé de deux étapes décrites ci-après.

4.3. *Phase d'apprentissage*

Les étiquettes prosodiques calculées depuis un signal de parole, ainsi qu'un arbre syntaxique associé (dont les feuilles sont alignées au signal) constituent les entrées du processus d'apprentissage. Le formalisme descriptif des arbres syntaxiques respecte les règles suivantes où α et β représentent des symboles non terminaux (définis par l'utilisateur) :

$\alpha(\beta,n)$	i.d. α est un groupe de n voyelles décrit par β,
$\alpha(\epsilon,n)$	i.d. α est une feuille de n voyelles,
$\alpha(\beta,\varepsilon)$	i.d. α est un groupe décrit par β sans spécification du nombre de voyelles,
$\alpha.\beta$	i.d. α et β,

Un exemple d'arbre est reporté en figure 3a. La phase d'apprentissage consiste à créer un graphe décrit de la manière suivante :

- chaque nœud (p) contient la caractérisation prosodique d'une structure syntaxico-rythmique ($\mathcal{S}_p$),

- chaque arc (a_{ij}) représente une contrainte (syntaxique ou rythmique) affinant la description de la structure décrite par le nœud i.

Les arcs sont déterminés à partir des arbres syntaxiques disponibles dans le corpus d'apprentissage en appliquant, niveau par niveau, les règles de généralisation suivantes (voir figure 3) :

a- un groupe α de n voyelles décrit par β est un groupe α de n voyelles ;

b- un groupe α de n voyelles décrit par β est un groupe de n voyelles ;

c- un groupe α de n voyelles est un groupe α ;

d- un groupe α de n voyelles est un groupe.

$$P = \left[\begin{array}{lll} a- & \alpha(\beta,n) & \rightarrow & \alpha(\epsilon,n) \\ b- & \alpha(\beta,n) & \rightarrow & \mathcal{G}(\epsilon,n) \\ c- & \alpha(\beta,n) & \rightarrow & \alpha(\epsilon,\varepsilon) \\ d- & \alpha(\beta,n) & \rightarrow & \mathcal{G}(\epsilon,\varepsilon) \end{array} \right.$$

où $\mathcal{G}$ est un symbole désignant un groupe quelconque. Une observation (i.e. un arbre syntaxique) de profondeur p peut générer jusqu'à $\frac{p\times(p+3)}{2}$ nœuds dans le graphe d'apprentissage. En pratique, le nombre de recouvrements (dépendant de l' application visée) limite la croissance du graphe.

Chaque nœud de ce graphe tient le compte des indices prosodiques localisés au début ou à la fin de chaque feuille de la structure décrite par le nœud. Chaque structure syntaxico-rythmique est ainsi décrite par une matrice prosodique de n lignes (n étant le nombre d'indices considérés : 40 ici) et de k colonnes ; k étant borné par le double du nombre de feuilles de la structure décrite.

À la fin du processus d'apprentissage (complètement décrit dans [5, pp. 134–138]) un utilisateur peut interroger le système afin d'obtenir des informations comme des contours paramétriques moyens associés à une structure donnée de phrase, ou encore des statistiques sur les points de rendez-vous entre la syntaxe, le rythme et la prosodie.

a)	$PH(\ SS(GN(Art(\epsilon,1).NC(\epsilon,3),4),4).$ $SV(VB(\epsilon,1).GN(ART(\epsilon,1).$ $NC(\epsilon,1).ADJ(\epsilon,1),3),5),9)$
b)	1- $PH(SS(\epsilon,4).SV(\epsilon,5),9)$, 2- $PH(\mathcal{G}(\epsilon,4).\mathcal{G}(\epsilon,5),9)$, 3- $PH(SS(\epsilon,\varepsilon).SV(\epsilon,\varepsilon),9)$, 4- $PH(\mathcal{G}(\epsilon,\varepsilon).\mathcal{G}(\epsilon,\varepsilon),9)$

Figure 3: a) Structure syntaxico-rythmique de la phrase : "Une pique-niqueuse mange une pomme verte" avec les symboles utilisateur suivants : PH phrase, SS groupe sujet, GN groupe nominal, ART article, NC nom commun, SV groupe verbal, VB verbe, ADJ adjectif. **b)** nœuds crées par l'application des règles de production au niveau 2 de la structure décrite en a).

4.4. *Prédiction structurelle*

Ce stage consiste à émettre des hypothèses sur la structure syntaxico-rythmique d'un énoncé, étant donné la matrice prosodique le caractérisant (M_o qui est une matrice $n \times l_o$, l_o étant le nombre présumé de voyelles dans le signal) et un graphe d'apprentissage (G). Ce processus met en jeu un parcours de graphe dont les points importants sont rappelés ici (voir [5, pp. 141–144] pour plus de détails).

Soit p un nœud appartenant à G, décrivant une structure ($\mathcal{S}_p$) de f feuilles ; une matrice prosodique $n \times l_p$ (avec $f < l_p < 2 \times f$) lui est associée. Les deux étapes du parcours de graphe sont les suivantes :

réduction de M_o **par** $\mathcal{S}_p$ **:** étape qui consiste à trouver un ensemble de f couples (d_i, f_i) de colonnes de M_o, correspondant à un découpage possible du signal de parole en f constituants (ceux décrits par $\mathcal{S}_p$). Les conditions suivantes sont également vérifées : pour chaque $i \in [1, f]$, $d_i = f_i + 1$, $f_i \geq d_i$, $d_1 = 1$, $f_{l_p} = l_o$. Les matrices résultantes sont notées M_{o_p}.

mesure de similarité : entre deux matrices M_{o_p} et M_p (de même dimension) par le calcul de $d(M_{o_p}, M_p)$:

$$\left\{ \begin{array}{ll} d(M_{o_p}, M_p) & = \frac{\sum_{i=1}^{n} \left(\alpha_i \sum_{j=1}^{l_p} \delta_{ij} \right)}{\sum_{i=1}^{n} \alpha_i} \\ \text{et } \delta_{ij} & = \left\{ \begin{array}{l} M_p(i,j) \;\; si \;\; M_{o_p}(i,j) = 1 \\ 0 \;\; sinon \end{array} \right. \\ \text{et } \alpha_i & \text{poids du } i^{eme} \text{ indice} \end{array} \right.$$

La valuation d'une hypothèse (potentiellement tout nœud de G) est finalisée en retenant le meilleur chemin de la racine de G au nœud considéré. Le score associé à un chemin particulier est la moyenne des scores de similarité entre les matrices M_{o_q} et M_q, pour chaque nœud q du chemin.

5. Expériences

5.1. *Protocole de test*

Dans ces deux expériences, nous avons imposé à notre système de réaliser une prédiction complète de la structure syntaxico-rythmique d'un énoncé inconnu ; c'est à dire de déterminer un arbre syntaxique complet ainsi que la spécification du nombre de voyelles de chacune de ses feuilles. Seule la matrice prosodique calculée à partir du signal de parole est utilisée.

Le nombre moyen de réponses possibles était de 15 aussi bien pour les phrases que pour les nombres décimaux.

5.2. *Expérience A*

Dans cette expérience, nous avons fait l'hypothèse que chaque indice prosodique jouait un rôle identique dans le processus de reconnaissance (i.e. $\alpha_i = 1$ pour tout indice i).

Phrases Le système montre des aptitudes intéressantes : sur le corpus d'apprentissage plus de 90% des hypothèses formulées en première position sont les bonnes (figure 4a). Sur le corpus de test, ce taux avoisine 60% (4b). Ces résultats peuvent être positivement comparés à ceux obtenus par un score aléatoire (figure 4c).

Nombres décimaux Là encore, le système présente un bon taux de reconnaissance (plus de 80%, voir la figure 4d). Les performances sur le corpus de test réduit aux nombres dont la structure a déjà été vue lors de l'apprentissage sont moins bonnes (proches de 50%, voir la figure 4e). L'analyse précise des hypothèses a cependant montré que même en cas d'erreur, le système était fiable quand à la prédiction de l'emplacement du mot "virgule" (figure 4f).

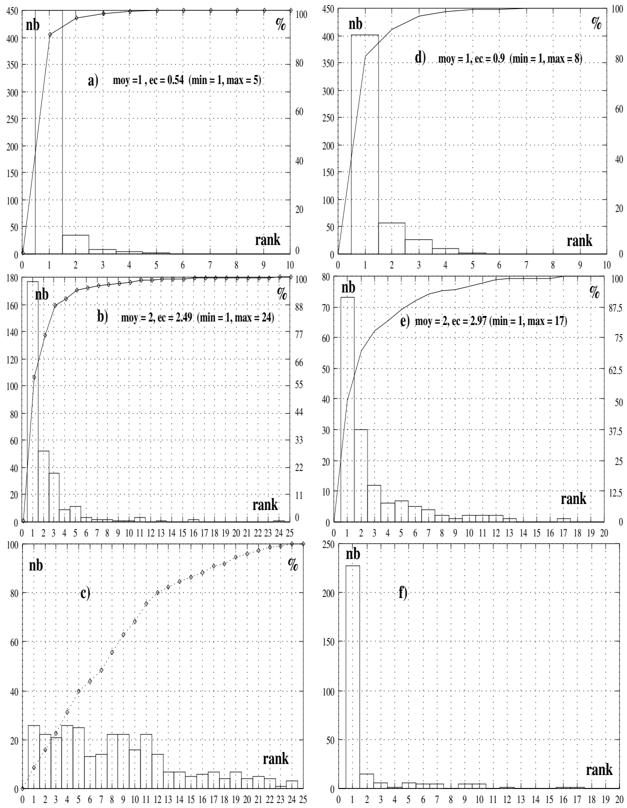

Figure 4: L'axe des abscisses indique le rang dans la cohorte d'hypothèse de la bonne solution. L'axe des ordonnées à gauche indique le nombre d'hypothèses réalisées et à droite, le pourcentage cumulé de succès à un rang donné. **a)** prédictions sur le corpus d'apprentissage. **b)** prédictions sur le corpus de test (300 sentences). **c)** prédictions sur le corpus de test à l'aide d'une mesure de similarité aléatoire. **d)** prédictions sur le corpus d'apprentissage des nombres. **e)** prédictions sur le corpus de test des nombres. **f)** prédiction de l'emplacement du mot "virgule" dans les nombres du corpus de test.

Nous avons essayé par un ensemble de règles mises au point sur le corpus d'apprentissage de décrire prosodiquement le mot "virgule" et son contexte droite et gauche immédiat. Cet ensemble de règle appliqué au corpus de test n'a jamais permis d'approcher les performances de notre système. Ceci tend à confirmer le fait que l'étude de l'information prosodique sur l'ensemble d'un énoncé (phrase, nombre, mot, etc.) semble plus pertinente que l'analyse locale des phénomènes prosodiques [9]. Ceci illustre une des limitations importantes de l'intégration efficace de la prosodie dans les systèmes de reconnaissance actuels.

5.3. *Expérience B*

Deux expériences supplémentaires ont été réalisées afin de déterminer les indices prosodiques les plus pertinents dans notre système. Dans la première (B1), nous avons évalué séparément la performance de chaque indice pour la tâche de prédiction. Dans la seconde expérience (B2), nous avons tenté de définir la combinaison optimale (au regard des performances sur nos corpus d'apprentissage) des poids (α_i) associé à chaque indice prosodique i.

Dans les deux expériences nous avons utilisé une fonction d'erreur E où r_i est le rang de la solution dans la cohorte d'hypothèses formulées et N le nombre de tests.

$$E = \frac{\sum_{i=1}^{i=N}(r_i - 1)}{N}$$

Expérience B1 Nous reportons quelques observations faites à partir des résultats obtenus sur la tâche de prédiction des structures complètes des phrases du corpus de test.

a) Prendre en considération tous les indices avec un poids égal à l'unité donne de meilleurs résultats qu'en ne considérant qu'un sous-ensemble des indices.

b) En regardant les taux d'erreur mesurés lorsque les prédictions sont réalisées en ne considérant les indices que d'un seul paramètre, on constate que les indices de durée sont les plus pertinents, suivis par les indices d'intensité. Les étiquettes prosodiques caractérisant le minimum et le maximum de chaque paramètre sont les indices prosodiques les plus bruités.

c) Plus précisemment, l'allongement de la durée de la voyelle (plus de 20ms au dessus de la durée moyenne des voyelles sur la phrase considérée) est l'indice le plus pertinent. Les indices de pente caractérisant le paramètre de fréquence fondamentale (montant, descendant, plat) sont ensuite les plus utiles.

Expérience B2 Dans cette expérience, nous avons supposé que E était une fonction dépendant seulement des n variables $\alpha_1, \alpha_2, ..., \alpha_n$ (i.e. les poids prosodiques). Nous avons alors utilisé la méthode d'optimisation non dérivative du *Simplex* [3] afin de trouver automatiquement l'ensemble de coefficients minimisant la fonction d'erreur E.

Les poids ainsi obtenus confirment les observations faites dans l'expérience B1. Le taux de performance mesuré sur la tâche de prédiction des structures des phrases du corpus de test a ainsi été augmenté de 0.8% (gain proche d'un rang en moyenne).

6. CONCLUSION

Le système présenté ici permet de prédire la structure syntaxico-rythmique d'un signal de parole à partir des seules informations prosodiques calculées automatiquement. Les premières expériences (tous les indices prosodiques jouant un rôle identique dans la fonction de notation) indiquent pour les phrases des performances de 90% sur le corpus d'apprentissage, de 60% sur le corpus de test et de moins de 10% avec un score aléatoire. Ceci permet donc de valider notre hypothèse de travail et de démontrer que notre système d'étiquetage prosodique est pertinent.

Dans les expériences visant à déterminer l'information la plus utile, nous notons que l'indice d'allongement de la durée des voyelles est l'indice le plus pertinent. Nous rejoignons en cela les observations faites dans des études antérieures [10]. D'autres expériences sur des corpus de taille plus importante sont cependant nécessaire.

REMERCIEMENTS

Nous tenons à remercier l'institut IDIAP pour la base de données utilisée dans cette étude ainsi que Pierre Jourlin pour ses précieux conseils.

BIBLIOGRAPHIE

[1] G. Chollet, J-L. Cochard, A. Constentinescu, C. Jaboulet, and P. Langlais. *Swiss-French Polyphone and PolyVar : Telephone Speech Databases to develop Model Inter- and Intra-speaker Variability*, chapter 8, pages 117–135. CSLI, 1997.

[2] A. Hunt. A generalised model for utilising prosodic information in continuous speech recognition. In *ICASSP*, volume II, pages 169–172, Adelaïde (Australia), 1994.

[3] Nelder J.A. and Mead R. A simplex method for function minimization. *Comput. J.*, pages 308–313, Janv. 1965.

[4] Hirose K. *Disambiguating Recognition Results by Prosodic Features*, volume 19, chapter 21, pages 327–342. Springer-Verlag, 1997.

[5] P. Langlais. *Traitement de la prosodie en reconnaissance automatique de la parole*. PhD thesis, Université d'Avignon, 1995.

[6] P. Langlais and J-L. Cochard. The use of prosodic agents in a cooperative automatic speech recognition system. In *International Congress of Phonetic Sciences*, volume 4, pages 292–295, Stockholm, Sweden, August 13–19 1995.

[7] Ostendorf M. and Ross k. *A Multi-level Model for Recognition of Intonation Labels*, chapter 19, pages 291–308. Springer-Verlag, 1997.

[8] K. Silverman, M. Beckman, J. Pitrelli, C. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. Tobi : a standard for labeling english prosody. In *ICSLP*, volume 2, pages 867–870, Banff, Alberta, Canada, 1992.

[9] Aubergé V. *Developing a structured lexicon for synthesis of prosody.*, chapter in Talking Machines: Theories, Models, and Designs, pages 307–321. Elsevier Science, 1992.

[10] J. Vaissière. The use of prosodic parameters in automatic speech recognition. In H. Niemann et al., editor, *Recent Advances in Speech Understanding and Dialog Systems*, volume F46. NATO ASI Series, Springer-Verlag Berlin Heidelberg, 1988.