ESTIMATING PROSODIC WEIGHTS IN A SYNTACTIC-RHYTHMICAL PREDICTION SYSTEM

Philippe Langlais
LIA, 339 chemin des Meinajariès, BP 1228, 84911 Avignon Cedex 9, France
langlais@univ-avignon.fr

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. Firstly 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

It is well known that prosodic structure and syntactic structure are not identical; neither are they unrelated. Knowing when and how the two correspond could help in the disambiguation of competing syntactic hypotheses in a speech understanding system. This practical reason explains the renewed interest in the use of prosody in ASR [8, 2, 4, 6].

This paper reports our experiments trying to answer these two general questions :

- Is prosody reliable enough to allow efficient syntactic and rhythmical prediction structure of an unknown message ?.
- If yes, is there any prosodic information that play a larger role in the performance of the structural prediction ?

2 DATABASE

For the two experiments described in this paper we used part of a telephone speech database recorded at IDIAP laboratory in collaboration with the Swiss-Telecom [1].

2.1 Isolated sentences corpora

A corpus of 500 isolated sentences (newspaper excerpts), made of 80 different sentences of simple syntactic structure from 4 to 17 vowels uttered by 50 speakers (via a noisy telephone line), was used for learning purposes (see figure 1). This set of sentences was also used to optimize the prosodic weights in the second experiment.

⋆ Le tout apparaît dans un bilan clair.
(Everything appears in a clear conclusion.)

⋆ Cette révélation tardive souligne la gravité du malaise.
(This late revelation highlights the seriousness of the situation.)

Figure 1: *Example of learning sentences.*

For testing purposes, a corpus of 300 sentences was selected (most of them being repetitions of the original 80 sentences of the learning corpus, not necessarily by the same speakers).

2.2 Decimal numbers corpora

A corpus of 500 decimal numbers (see examples in figure 2) uttered by 50 speakers via the same telephone line has been used for learning purposes.

⋆ Treize mille deux cent quarante virgule dix.
(Thirteen thousand, two hundred forty point ten.)
⋆ Quatre mille virgule huit.
(For thousand point eight.)

Figure 2: *Example of decimal numbers from the learning corpus.*

A corpus of 280 decimal numbers was also selected for testing purposes (among them, 148 numbers have their syntactic-rhythmical structure encountered at least once during the learning process).

3 PROSODIC LABELLING

Despite emergence of ToBI [7], prosodic labelling has not found a unanimous solution yet. Therefore, a fairly classical set of 40 labels is retained in this study (9 duration labels, 9 intensity labels and 22 F0 labels) characterizing each vocalic nucleus. This simple and fully automatic labelling process (see [5] for a full description) computes from an input speech signal, a

prosodic matrix, which is the input of the recognition process we are going to discuss.

4 THE PREDICTION SYSTEM

4.1 Our work hypothesis

We formulate the hypothesis that the distribution of prosodic configurations on the whole sentence is not random, but on the contrary regular enough to permit structural (syntactic, rhythmical, semantic, ...) predictions in an automatic way. This is a point we wanted to make with our system.

4.2 Basic description

The main characteristic of the system we propose (see [6] for a general description) is that it does not have any a priori either on the hypothetical hierarchy of the different constraints governing the prosodic parameters, or on particular prosodic entities such as stress for which we have found no robust acoustic characterization (at least for French). The basic principle of our approach is a data-driven study of correlations between linguistic levels and prosodic labels automatically computed.

The system involves two stages that are described below.

4.3 Learning phase

Prosodic labels and a syntactic tree (whose leaves are aligned on the speech signal) are provided for each sentence of a learning corpus. The formalism used for the tree description respects the following rules (α and β belonging to the user's set of symbols :) :

$\alpha(\beta, n)$	means α is a group of n vowels described by β,
$\alpha(\epsilon, n)$	means α is a leaf of n vowels,
$\alpha(\beta, \varepsilon)$	means α is a group described by β with no mention of its number of vowels,
$\alpha.\beta$	means α and β,

An example of such a structure is given in figure 3. The learning stage consists in plotting a graph which can be described as follows :

- each node (p) contains the prosodic characterization of a syntactic-rhythmical structure ($\mathcal{S}_p$),
- each arc (a_{ij}) stand for a syntactic or rhythmical constraint refining the structure described by node i .

Arcs are defined from the learning sentence structures by applying, level by level, one of the following P-rules (see figure 3 for an example) with the following meaning :

a- an α group of n vowels described by β is also an α group with n vowels ;

b- an α group of n vowels described by β is also a group of n vowels ;

c- an α group of n vowels is also an α group ;

d- an α group of n vowels is also a group.

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

where $\mathcal{G}$ is a symbol for any type of group.

Each node keeps count of the labels located at the beginning or at the end of the associated structure leaves, in such a way that each syntactic-rhythmical structure is prosodically described by means of a (normalized) matrix with n lines (n is the number of the different labels – 40 in this study) and k columns (k is maximized by the number of leaves of the described structure $\times$ 2).

After the learning stage — which is fully described in [5, pp. 134 to 138] — a user can ask the system to provide him with meeting points between syntax, rhythm and prosody, and can also ask for a parametric contour matching a given linguistic organization.

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)** *Ex : syntactic-rhythmical structure given for the utterance : "Une pique-niqueuse mange une pomme verte" (A picknicker is eating a green apple) with the set of user's symbols : PH sentence, SS subject group, GN noun group, ART article, NC common noun, SV verb group, VB verb, ADJ adjective.* **b)** *nodes created by the application of the production rules to level 2 of the structure described in a).*

4.4 Structural prediction phase

This stage consists in putting forward syntactic-rhythmical hypotheses from the learning graph (G) and the prosodic labelling of the speech signal to decode (considered as an $n \times l_o$ matrix M_o, l_o being the assumed number of vowels). This process involves the parser briefly described below (see [5, pp. 141 to 144] for further details).

Given the following notations : p is a node belonging to G, describing a structure ($\mathcal{S}_p$) of f leaves ; it

is prosodically described by an $n \times l_p$ matrix (with $f < l_p < 2 \times f$). The parser requires two basic steps :

M_o **reduction from** $\mathcal{S}_p$ **:** consisting in finding a set of f couples of M_o columns (d_i, f_i) corresponding to the possible limits (in terms of vowels) of the f constituents. This process verifies the constraints on the number of vowels of each $\mathcal{S}_p$ leaf as well as the following conditions : for each $i \in [1, f]$, $d_i = f_i + 1$, $f_i \geq d_i$, $d_1 = 1$, $f_{l_p} = l_o$. The resulting matrices are noted M_{o_p}.

similarity measurement : between two matrices M_{o_p} and M_p (which have the same dimension) by computing $d(M_{o_p}, M_p)$:

$$\begin{cases} d(M_{o_p}, M_p) & = \dfrac{\sum_{i=1}^{n} \left(\alpha_i \sum_{j=1}^{l_p} \delta_{ij} \right)}{\sum_{i=1}^{n} \alpha_i} \\ \text{et } \delta_{ij} & = \begin{cases} M_p(i,j) \;\; if \;\; M_{o_p}(i,j) = 1 \\ 0 \;\; else \end{cases} \\ \text{and } \alpha_i & \text{weight of the } i^{th} \text{ label} \end{cases}$$

The valuation of an hypothesis (any node of the learning graph) is achieved by keeping the best score path from the root of G to the considered node. The scoring of a particular path is the average of the similarity measurement of M_{o_q} and M_q matrices for each node q of the path.

5 EXPERIMENTS

5.1 Test Protocol

For both experiments, we ask our system to make full syntactic-rhythmical hypotheses of unknown sentences (that is, the prediction of the full syntactic tree, as well as the number of vowels for each leaf) using only the prosodic matrix.

The average number of possible answers for a given sentence is 15.

5.2 Experiment A

In this experiment, we made the assumption that each prosodic label plays an equal role in this scoring function (that is, α_i was equal to unity for each label i.).

5.2.1 *Isolated sentences*

The system has shown interesting capacities which allow its use for sentence recognition. On the learning corpus, more than 90% of the hypotheses have been classified in first position (see figure 4a). In the test corpus the first hypothesis assigns almost 60% of sentences to the right structure (see figure 4b). This results can be positively compared as those obtained with a random scoring (see figure 4c)

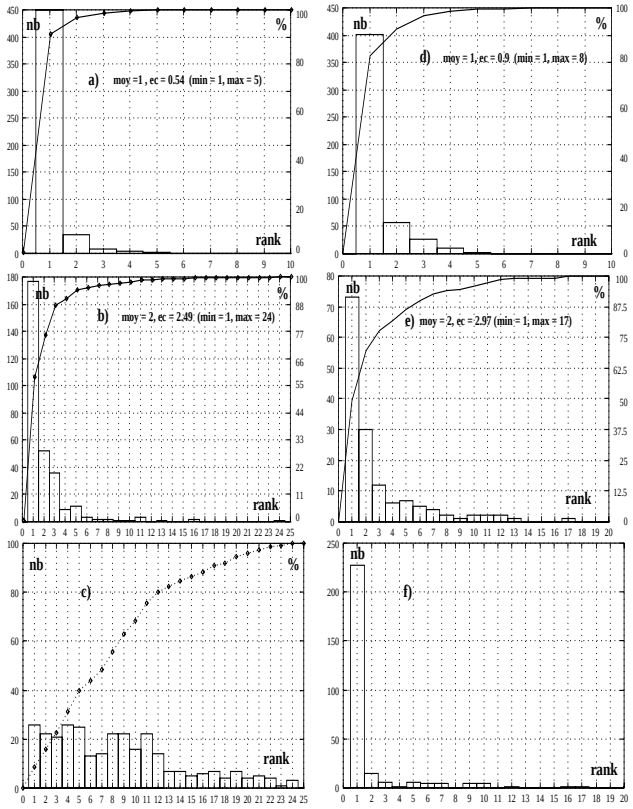

Figure 4: the x-axis represents the rank of the hypothesis correctly formulated by the system; the y-axis indicates, on the left, the number of hypotheses and, on the right, the percentage of correct hypotheses formulated at a given rank. **a)** predictions on the learning corpus (500 sentences). **b)** predictions on the testing corpus (300 sentences). **c)** predictions on the testing corpus by means of random scoring. **d)** predictions on the number-learning corpus (500 numbers). **e)** predictions on the number-testing corpus. **f)** prediction of the location of word "virgule" (point).

5.2.2 *Decimal numbers*

Here again, the system has been able to match the prosodic lattices of the learning numbers to the correct syntactic-rhythmical structures in more than 80% of the cases, with an average of 15 possible answers (see figure 4d). On a test corpus of 148 numbers, the system has shown a prediction rate of slightly less than 50% (see figure 4e). The analysis of the non-first hypotheses has shown that the word "virgule" (point) was very often well located (see figure 4f). We were not able to reach this score by means of local specific rules, which confirms the importance of the prosodic information taken from the whole sentence.

5.3 Experiment B

Two experiments were carried out in order to look forward to the most powerful prosodic labels. For

the former, we proposed to evaluate separately the performance of each prosodic parameter for the prediction task. For the latter, we estimated prosodic weights (α_i) that outperforms the best results on a chosen corpus.

For both experiments, we defined an error-rate function E where r_i is the rank of the correct prediction hypothesis proposed by the system and N the number of tests :

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

5.3.1 *Experiment B1*

We briefly report the observations we made on a prediction task consisting of plotting full-syntactic-rhythmical structure both of the learning-sentences and the training-sentences.

a) Taking into account each label ($\alpha_i = 1 \forall i$) gives better results than taking into account only part of the label.

b) Looking at the error-rate when the prediction is achieved by considering only labels of one parameter at time, we observed that duration labels outperform the best, followed by intensity labels. The minimum and maximum of each parameter for the sentence are the noisiest ones.

c) More precisely, the lengthening of the duration-vowel (more than 20 ms above the sentence-average duration-vowels) seems to be the most promising label. The pitch-slope labels (rising, lowering, flat) are the next best suited.

5.3.2 *Experiment B2*

In this experiment, we assumed that E was dependent only on the n variables $\alpha_1, \alpha_2, ..., \alpha_n$ (that is the prosodic weights). We estimated the set of coefficients that minimize the above error-rate function by applying the nonderivative *simplex* method [3] to the training-sentence prediction task.

The weights obtained confirmed the observations we made in the previous experiment. The ranking improved by 0.8% (for the test-sentence prediction task) with the obtained combination. More data is however necessary to reinforce this first result.

6 CONCLUSION

We achieved a system that allows linguistic structural prediction of a speech signal by means only of prosodic information automatically computed. The first experiments we report (with all prosodic labels playing an equal role in the scoring function) are fairly promising (more than 90% of hypotheses ranked in first position on the learning corpus, 60% on the testing corpus and less than 10% with a random scoring function.).

We are carrying out the weighting of these labels in order to find the most pertinent ones. Our first experiments tend to indicate that each label is not of equal importance in our prediction task. Lengthening of the duration tends to be the most powerful label. Further experiments on larger databases needs to be carried out to insure these observations.

Acknowledgments

We thank particularly IDIAP Institute for the database we used in this study and Pierre Jourlin for his advice on this paper.

References

[1] Gérard Chollet, Jean-Luc Cochard, Philippe Langlais, and Robert Van Kommer. Swiss-french polyphone : a telephone speech database to develop interactive voice servers. In *Linguistic Databases*, Gröningen, 1995.

[2] Andrew 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] Philippe Langlais. Microprosodic study of isolated french word corpora. In *4th European Conference on Speech Communication and Technology*, Madrid, Spain, September 18–21 1995.

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

[6] Philippe Langlais and Jean-Luc 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] 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.

[8] N.M. Veilleux and M. Ostendorf. Probabilistic parse scoring with prosodic information. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 51–54, Minneapolis, Minnesota, 27-30 April 1993.