Towards Automatic Annotation of Temporal Features in Discourse: the Case of Syllabic Duration in Spontaneous French.

Cyril Auran & Albert Di Cristo
Laboratoire Parole & Langage, Université de Provence
E-mail: cauran@wanadoo.fr, albert.dicristo@lpl.univ-aix.fr

ABSTRACT

Numerous discourse functions systematically resort to the prosodic resources found in the speakers' management of temporal features. These resources concern pauses, syllabic duration modifications and speech rate. However, the description and modelling of temporal phenomena constitutes a particularly delicate endeavour, mainly because, contrary to intonation and intensity, they cannot be based upon the direct interpretation of analytic representations. The problem is even more important as the aim is to develop tools for the automatic annotation of prosodic phenomena in discourse. This paper presents the methodology and the first results of such a tool, focusing on syllabic duration. The analysis shows that the order of influence of the three parameters retained to account for the perception of subjective duration is: normalised duration > silence > velocity. This paper details their respective influences and discusses the interest of our approach in automatic annotation of syllabic duration in spontaneous speech.

1. INTRODUCTION

This paper focuses on the issue of the automatic coding of the variations of syllabic duration in French connected speech. A dramatic trend has appeared in the last few years concerning the importance of prosody in spontaneous spoken discourse ([2], [4]) A closer look clearly shows that it is intonation which remains the main centre of interest in these studies ([9]). Prosody in a given language, however, constitutes a complex system the core of which is represented by two sub-systems respectively concerned with the tonal and temporal organisation of utterances. Without questioning the importance of tonal organisation, which constitutes the very basis for the intonation systems of languages, we consider that temporal organisation also assumes a set of linguistic and paralinguistic functions (such as the segmentation and hierarchical organisation of spoken discourse units or pragmatic contextualisation functions) which play a major role in discourse structure and processes. Temporal organisation can be viewed as a homogenous set concerning variations of syllable and segment durations, speech rate and pauses.

The description and modelling of temporal elements turns out to be quite delicate, mainly because it can not rely on the direct interpretation and the stylisation of an analytical representation (duration curve?), as is the case for intonation (F0 curve) and energy (intensity curve). The interpretation of variations in temporal phenomena requires the definition of units of measure (phoneme, syllable or other), which are submitted to interfering production and phonotactic constraints. The issue grows even more problematic when one wishes to go beyond laboratory speech and read utterances and deal with spontaneous speech for which numerous factors constitute supplementary sources of variation

In our approach to the relations of prosody to discourse ([6]), we argue in favour of the use of a multi-linear representation grid concerned with the tonal and temporal organisations of discourse. One of the most important principles in the setting up of the grid is that all its elements must be represented as categorical entities. The symbols selected for the discrete coding of tonal organisation thus specify the value of tonal segments (**M**id, **T**op, **B**ottom, **H**igher, **L**ower …), register levels (**N**ormal, **Rai**sed, and **Lowered**) and span (**N**ormal, **Ex**panded and **Re**duced). This paper focuses more specifically on the issue of the categorical coding of temporal organisation, mainly through the analysis of syllabic duration. In this perspective, we compare a subjective estimation of this parameter with an automatic statistical procedure based on objective data in order to evaluate the accuracy of its predictions.

2. METHODOLOGY

Our methodology involves three main phases, based on the analysis of a corpus consisting of a 3-minute radio interview between two female French speakers.

Phase 1 consisted in two procedures: the subjective annotation of phenomena and automatic extraction of parameters in the corpus.

The subjective annotation procedure consisted in asking seven French native experts from the Laboratoire Parole & Langage to code such perceptive phenomena as accent, emphasis, speech rate, pause, breath and syllabic durations. The coding was performed on an auditory basis only, without resorting to any signal processing system. The coding of syllabic duration was carried out using the symbols [L], [XL] and [R] ("long", "very long" and "reduced" respectively).

The second procedure related to the automatic extraction of the following speech signal parameters, the choice of which

ISBN 1-876346-48-5

is justified in section 4:

• silences were detected automatically using a Praat script based on voicing and intensity level data, with a classical 200 ms threshold;

• intrasyllabic F0 velocity (or slope) was automatically calculated on a ST/sec scale using another Praat script, based on the amdf detection of F0 variations within the stable part of syllabic nuclei;

• objective segmental duration, eventually, was also retrieved from the speech signal, using yet another Praat script, and were normalised (Perl scripts) in two successive steps, first neutralizing "co-intrinsic" effects using the Di Cristo algorithm ([5]), before neutralizing "intrinsic" effects with the application of z-score transforms, a method inspired by Campbell's 1992 algorithm ([3]). Segmental durations were then categorized (last Praat script for this phase) using the symbols mentioned for the subjective procedure, on the basis of a priori thresholds on the z-score values;

• speech rate, finally, was automatically categorized using a basic algorithm which labelled sequences of more than two [L]/[XL] as "slow" and sequences of more than two [R] syllables as "fast".

Phase 2 consisted in calculating correlations between the data from the subjective and objective procedures of the first phase. Two complementary methods were used to treat accent and syllabic duration perception: on the one hand, correlation coefficients were resorted to in order to give a first approximation of the links between pairs of given perceptive phenomena and objective parameters; on the other hand, multiple regression algorithms were used in order to isolate and classify the objective parameters related to the perception of a given subjective phenomenon. This paper focuses on the second method, implemented through the use of Wei-Yin Loh's GUIDE (Generalized, Unbiased, Interaction Detection and Estimation) software ([7]), the results of which are detailed in the following section. More particularly, subjective lengthening (which groups syllables coded as [L] and [XL]) for each of the 996 syllables in the corpus was analyzed in relation with such objective parameters as normalised duration, velocity (raw and absolute values), speech rate and silence before and after each item.

Phase 3 consisted in implementing the output of the second phase, developing a prototype algorithm for the automatic prediction of subjective syllabic duration on the basis of parameters from the speech signal. In this perspective, normalised duration and post-syllabic silence (the first two parameters given by phase 2) were taken into account and used as input to the Praat script devoted to this automatic annotation procedure. The algorithm relies on three major elements:

• the "L threshold", used to determine the category ("Lengthened" vs "Normal or Short") of each syllable in the corpus;

• the presence of a silence following each syllable, which consisted in categorizing such syllables as [L] instead of [N] and [XL] instead of [L] respectively;

• the effect of the preliminary neutralization of "co-intrinsic" effects via the phase 1 algorithm mentioned above.

The results of the multiple regression analyses mentioned earlier, together with the evaluation of the output of this automatic annotation procedure, are detailed in the following section.

3. RESULTS

The results of our work will be presented in two parts, corresponding respectively to the output of phases 2 and 3.

The multiple regression algorithm used during phase 2, as far as subjective syllabic duration is concerned, allowed us to confirm and qualify the intimate links suggested by the correlation coefficients method (Table 1).

Corr. Coef.	Norm. Length	Silence after	Raw Velocity	Silence before	Absolute Velocity
Subjective Length	**0,589**	**0,380**	**0,156**	**-0,144**	**-0,079**

Table 1: Correlation coefficients

Indeed, these results suggest an ordered influence of at least the first two parameters (normalised length and following silence), the correlation coefficients of which exceed the classical 0.35 significance threshold. The influence (and order) of these two parameters is confirmed by the multiple regression method (of which Figure 1 is a hierarchical graphic representation).

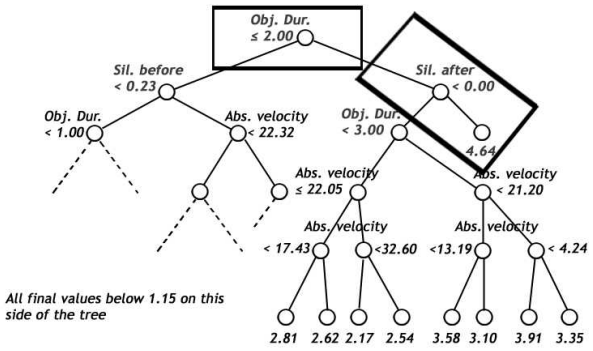

Figure 1: Regression tree on subjective lengthening

Objective duration is isolated as the hierarchically most salient parameter influencing subjective syllable length (horizontal rectangle) while silence following a syllable (slanting rectangle) is the second most important parameter, with a noteworthy unambiguous categorization of the syllable as "lengthened".

These results, together with the relative ambiguousness of

the role played by absolute F0 velocity on the syllabic nucleus (see bottom right part of figure 1), constitute arguments in favour of the use of normalised length and silence in the implementation of our phase 3 algorithm.

Following van Rijsbergen ([8]), we will resort to three standard measures in order to give an accurate account of the quality of the predictions: *precision*, *recall* and *F-measure*. In our case, *precision* corresponds to the proportion of subjectively lengthened syllables accurately predicted as such by the algorithm (N. accurate / N. relevant); *recall* is the proportion of possible categorisations accurately predicted (either "lengthened" or "not lengthened") by the algorithm (N. accurate / N possible); *F-measure*, eventually, is the harmonic mean of the two preceding measures.

The following figures (2 and 3) represent the precision of the predictions as a function of "L threshold", using both normalisation algorithms (figure 2) or the "Campbell algorithm" alone (figure 3) and either normalised length alone ("NL" condition) or normalised length and silence ("NL + Silence" condition) as parameters.

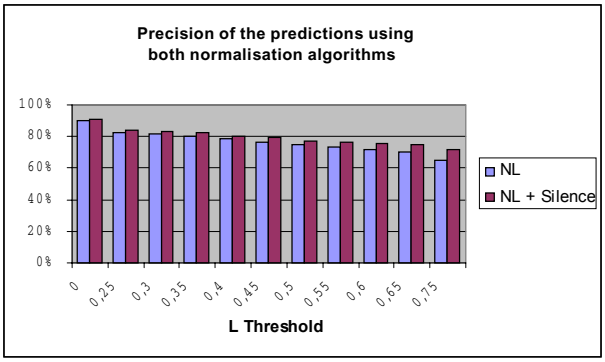

Figure 2: Precision using both normalisation algorithms

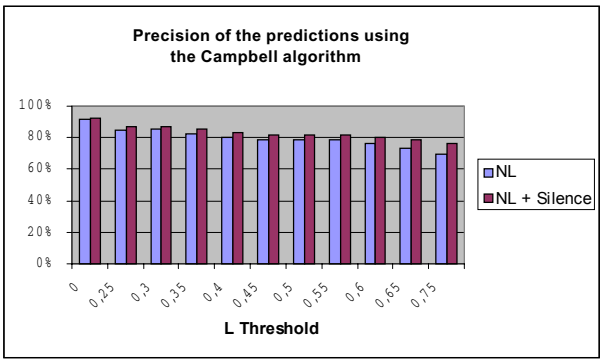

Figure 3: Precision using the "Campbell algorithm" alone

These results suggest three remarks:

• the use of both parameters leads to an increase in precision (an average of 3.20% when both algorithms are used, and 2.94% otherwise);

• the simultaneous use of both normalisation algorithms leads to a decrease in precision (an average of 3.21% with one parameter and 3.47% with both parameters). This will be discussed in the following section of this paper;

• precision decreases as L Threshold increases. This is due to the fact that a growing number of syllables are predicted as lengthened, which obviously increases the number of syllables accurately predicted as lengthened, but also leads to a growing number of erroneous predictions.

This justifies the resort to recall for a better evaluation of the predictions. Figure 4 represents recall as a function of "L threshold", using the "Campbell algorithm" alone (first 2 series, "1 alg." condition) or both normalisation algorithms (last 2 series, "2 alg." condition) and either normalised length alone ("NL" condition) or normalised length and silence ("NL + Silence" condition).

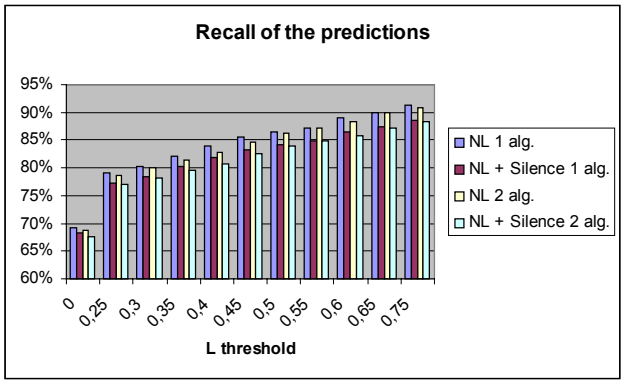

Figure 4: Recall of the predictions

Figure 4 illustrates the overall better quality of the predictions of the method using the "Campbell algorithm" alone (by an average of 0.50%). The progressive increase of recall as L threshold increases is related to the fact that more syllables are predicted as "not lengthened", a category which constitutes 89.76% of the total number of subjectively encoded syllables and logically influences the measure in a dramatic way.

As is typical in information retrieval, we are thus confronted with a situation where precision and recall are inversely proportional to each other. On figure 5, precision is plotted against recall: the preferred condition will then correspond to the uppermost curve ("Campbell algorithm" alone, both parameters being taken into account; series "NL + Silence 1 alg." on the chart).

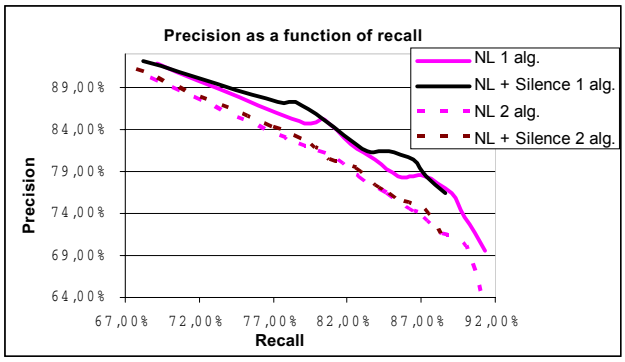

Figure 5: Precision as a function of recall

ISBN 1-876346-48-5

This result is confirmed by the F-measure (figure 6):

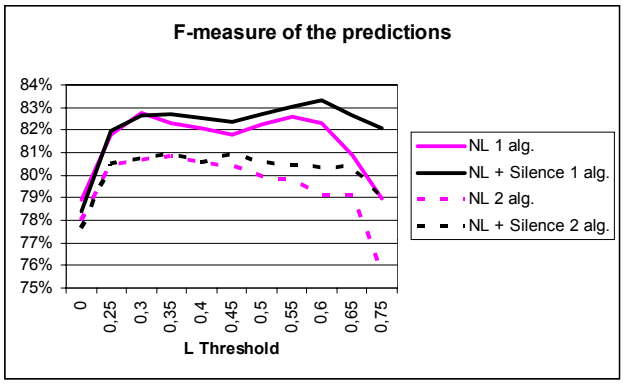

Figure 6: F-measure for the predictions

The data allow us to conclude that our prediction algorithm is optimal for an L Threshold of 0.6 (F = 83.32%) when the "Campbell algorithm" is used alone and both parameters (normalised length and following silence) are taken into account: precision is then 80.39%, and recall 86.47%.

4. DISCUSSION

These results raise several questions concerning the central issue of the prediction of subjective syllabic length on the basis of objective data extracted from the signal.

First, one may question the number and nature of parameters required for an optimal prediction of subjective lengthening. Intonation (through F0 variations) has often been shown to play a role in this respect, but neither correlation coefficients nor multiple regressions were able to evaluate its importance. A non-final version of our prediction algorithm integrating this parameter was implemented, but results showed no gain in precision and even a slight decrease in recall. This situation may be related to a poor quality of the F0 velocity detection method, or to more complex interactions between relevant parameters; new scripts are being considered, together with the possible implementation of d'Alessandro & Mertens' algorithm ([1]), itself devoted to the stylization of perceived pitch. We will note, however, that this latter possibility may result in introducing elements of heterogeneity in the parameters taken into account, "perceived" pitch (though predicted by the algorithm) constituting (predicted) subjective data, quite different in kind from the objective parameters used so far.

The second interesting point raised by our results questions the validity of the algorithm used to neutralise "co-intrinsic" effects, which systematically resulted in poorer predictions. More particularly, this issue is related to the relevance of mapping experimental results obtained on laboratory speech onto spontaneous speech: situated discourse, induces constraints which may well influence "co-intrinsic" effects in a substantially different manner, yet to be explored. This justifies the scheduled setting up of a series of new studies about "co-intrinsic" effects in spontaneous French and Russian at the Laboratoire Parole & Langage.

We may eventually mention the issue of the role played by speech rate in the subjective assessment of syllabic duration. Speech rate wasn't incorporated in the analyses because the algorithm chosen to account for this variable was dependent on the automatic a priori prediction of syllable length (phase 2); other algorithms are being considered which will characterize speech rate as independent from predicted syllable length, possibly using the distribution of production/perception units as a function of time.

5. CONCLUSION

This paper argues in favour of a methodology taking into account both subjective and objective procedures in the categorization of prosodic sub-systems other than intonation in spontaneous discourse. Subjective syllabic duration, which plays a major role in discourse segmentation and hierarchical organisation, has been shown to be reliably predictable on the basis of automatic procedures, thus allowing future integration in automatic content-analysis systems for spontaneous French.

REFERENCES

[1] Alessandro, C. d' & Mertens, P., "Automatic pitch contour stylization using a model of tonal perception.", *Computer Speech and Language 9(3)*, pp. 257-288, 1995.

[2] E. Blaauw, *On the Perceptual Classification of Spontaneous and Read Speech*, OTS, Utrecht University, 1995.

[3] N. Campbell, *Multi-level timing in speech*, PhD thesis, University of Sussex, 1992.

[4] E. Couper-Kuhlen and M. Selting, *Prosody in Conversation*, Cambridge MA: Cambridge University Press, 1996.

[5] A. Di Cristo, *De la Microprosodie à l'Intonosyntaxe*, State Thesis, Université de Provence, 1978.

[6] A. Di Cristo *et al.*, "An integrative approach to the relations of prosody to discourse: towards a multilinear representation of an interface network", *Prosodic Interfaces* International AAI Workshop, Nantes, March 27-29, 2003.

[7] W.-Y. Loh, "Regression trees with unbiased variable selection and interaction detection", *Statistica Sinica*, *vol. 12*, pp. 361-386, 2002.

[8] C.J. van Rijsbergen, "Retrieval Effectiveness", in *Progress in Communication Sciences*, *vol.1*, M.J. Voigt, Ed., pp. 91-118. London: Butterworth, 1979.

[9] A. Wichmann, *Intonation in Text and Discourse*, London: Longman, 2000.

ISBN 1-876346-48-5