Toward a Continuous Modeling of French Prosodic Structure: Using Acoustic Features to Predict Prominence Location and Prominence Degree

Mathieu Avanzi[1, 2], Nicolas Obin[2, 3], Anne Lacheret-Dujour[2], Bernard Victorri[4]

[1]Université de Neuchâtel; [2]MoDyCo, Université de Paris Ouest Nanterre; [3]Ircam, Paris; [4]Lattice/ENS, Paris

mathieu.avanzi@unine.ch, nicolas.obin@ircam.fr, anne@lacheret.com, bernard.victorri@lattice.fr

Abstract

The aim of this paper is to present a tool developed in order to generate French rhythmical structure semi-automatically, without taking grammatical cues into account. On the basis of a phonemic alignment, the software first locates prominent syllables by considering basic acoustic features such as F0, duration and silent pause. It then assigns a degree of prominence to each syllable identified. The estimation of this degree results from a computation of the values of silent pause, relative duration and height averages used for prominence detection in the first step. The second part of the article presents an experiment conducted in order to validate the algorithm's performances, by comparing the predictions of the software with a continuous manual coding carried out by four annotators on a 4-minute stretch of corpus (788 syllables) involving read aloud speech, map task and spontaneous dialogue. The performance of the algorithm is encouraging: a Fleiss' kappa calculation estimates the rate at 0.8, and a correlation agreement calculation at 91%, in the best cases.

Index Terms: prominence, automatic detection, degree of prominence, prosodic structure, French.

1. Introduction

There are two complementary ways to represent the prosodic structure of an utterance; common to both of them is the notion of **syllabic prominence** [1]. In the first method, the modeling consists in concentrating on accentual/rhythmical phenomena, in order to construct the **metrical grid** of an analyzed utterance [2]. The second procedure consists in making explicit the implementation of the tones associated with stressed and unstressed syllables, in order to generate the **tonal patterns** associated with accentual groups [3]. This article deals with the first aspect of the modeling.

Traditionally, the procedures used to construct the metrical grid of a given utterance are top-down procedures. In practice, for French, one can predict accentual prominences by identifying lexical words and their dependent clitics, and arrange the location of final and non-final stressed syllables with the help of rhythmical constraints specific to this language (eurhythmic principles such as clashes and lapses avoidance for example). Then, in order to estimate the prominence degree of the successive syllables, one can use the degree of embedding of the constituent in the syntactic structure of the whole utterance, and its informational status (see for references [4][5][6][7]).

Constructing the metrical grid of a given utterance automatically, however, constitutes a great challenge nowadays. The problem is that, from a descriptive perspective, the rules established by prosodic phonology are extremely complicated to apply. In fact, it is necessary to control a certain number of factors and to understand exactly how they interact. While interesting developments concerning the interaction of syntactic, semantic/pragmatic and rhythmic cues in French phonological phrasing were introduced a few years ago by optimality theory (see for French [6] and [7], among others), more recent work dealing with extra-sentential elements in spontaneous speech has shown that things are not as straightforward as was initially thought, *i.e.* that major syntactic boundaries do not always coincide with major prosodic boundaries ([8], and for similar conclusions on English, see the recent work of [9]). For these reasons and others that cannot be detailed here for lack of space, we followed an alternative procedure to build a system to generate metrical grids of French utterances. In our approach, we use a bottom-up methodology to take into account the acoustic correlates of accentual prominences without formulating any hypotheses about the functional constraints they are associated with, and without taking grammatical features into account. In this paper, we present the methodology followed to achieve this purpose.

2. Automatic Processing

This section presents the automatic processing. First, from a phonemic alignment, our system conducts a detection of prominent syllables (§2.1). Then, on the basis of this detection, it estimates the degree of prominence of each syllable considered as prominent (§2.2). The section concludes with an illustration of the detection.

2.1. Prosodic Prominence location

The prominent syllables detection procedure is implemented under Matlab in an interface called ANALOR. It is presented in detail in [10]. Briefly, the detection relies on the calculation of four prosodic parameters: (i) normalized duration average of the current syllable compared with the three preceding and the three following syllables; (ii) height average of the current syllable compared with the three preceding and the three following syllables (taking into account the F0 points of the vowels only, *i.e.* not all the F0 points of the syllables, as the ones carried by the consonants are considered as less important with respect to tonal perception [11]); (iii) estimation of the amplitude of the rising tone on the vowel of the current syllable, if any; (iv) presence of a silent pause after the current syllable if the syllable is not connected with a hesitation or a false start (silent pauses being considered as strong cues for the end of an accentual group in French [12]).

A syllable is considered as prominent if (a) one of these first three parameters reaches a certain threshold (on the basis of a corpus-based learning procedure involving a 70-minute long corpus annotated for prominence studies by two experts, containing various genres and French metropolitan, Swiss and

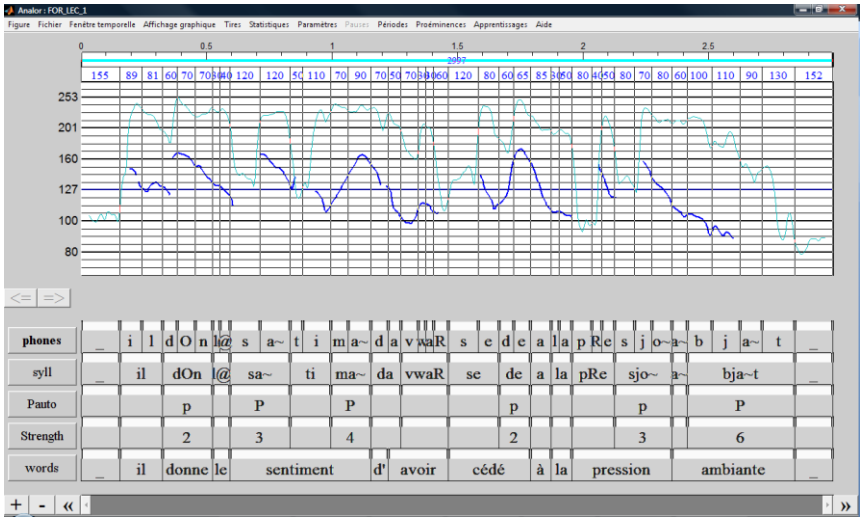

Figure 2: *Analor screen shot, illustrating the automatic identification of prominence. Analysis of the utterance: "il donne le sentiment d'avoir cédé à la pression ambiante" [read aloud speech]. In the top part, the evolution of F0 can be measured in hertz (values are on the left) or in semi-tones (the interval between two horizontal lines is one semi-tone). In the bottom part, the transcription tiers are, from top to bottom: phones, syllables (both in SAMPA), prominent syllables, strength of the prominence and words.*

Belgian speakers; we estimated that the optimal thresholds for height and duration averages varied between 1.48 and 1.76, 1.38 st. and 2.68 st. respectively, and that the threshold for rising tone varied between 1.71 st. and 3.67 st.); and/or (b) a silent pause (whatever its duration) follows the current syllable (the annotation of silent pauses was made during the semi-automatic phones alignment step; "false" pauses such as pre-occlusive silences were automatically excluded).

2.2. Prominence Degree Categorization

In order to estimate the degree of prominence of the syllables detected as prominent, we adopted the following hypothesis: the greater the number of acoustic parameters involved in the identification of prominence, the more the fixed thresholds are exceeded, the more the prominence is perceived as strong. In practice, here is how we proceeded.

First, for each of the first three criteria used to detect the location of stressed syllables in the preceding step (relative duration, relative height and rising tone), we attribute a score between 0 and 10. This score is determined according to the difference with the optimal threshold fixed during the corpus-based learning procedure. A value equal to the threshold gives a score of 5; a value of 0 (*i.e.* 100% lower than the threshold) gives a score close to 0, and a value of twice the threshold (*i.e.* 100% above the threshold) gives a score close to 10. The exact formula used is:

$$f(x) = 10 \cdot \left(\tfrac{1}{2} + \tfrac{1}{2} \cdot \tanh(2 \cdot \lambda \cdot \frac{x-t}{t}) \right) \quad (1)$$

where x is the value of the current syllable for the given criterion, t the threshold, and λ the slope of the function (changing this makes the slope more or less steep; by default its value is 1.5). Concerning the silent pause criterion, the score is 0 or 10 since it is a binary criterion.

Finally, the strength of the prominent syllable is obtained by computing the weighted average of the four scores:

$$strength = \frac{f_D(x_D).wght_D + f_H(x_H).wght_H + f_R(x_R).wght_R + f_P(x_P).wght_P}{wght_D + wght_H + wght_R + wght_P} \quad (2)$$

where D is the duration value, H the height value, R the rise value and P the silent pause value. The weight (*wght*) for the three continuous criteria is 1, while that for silent pause is 0.5.

2.3. Illustration

The result of the automatic identification of the location of prominences and of their strength is shown on figure 2. Syllables detected as prominent are marked "p" or "P" in a dedicated tier (named "Pauto"), and the score of prominence (rounded to the nearest unit) in the tier just below (named "Strength").

3. Manual Coding Procedure

We conducted an experiment to evaluate the robustness of the algorithm of prominence degree estimation. This section presents the data, the task (§3.1) and the results of the inter-annotator consistency (§3.2).

3.1. Material and Task

Four French native speakers (two of the authors and two PhD students in French linguistics) were asked to annotate the prominences and to indicate their degree of strength in a 4-minute long set of recordings. The recordings comprised four files: a map-task (50 sec., 171 syll.), an extract from a spontaneous interaction between two women (57 sec., 172 syll.), a read-aloud newspaper article (72 sec., 360 syll.) and a read dialogue sampled from a teaching manual for foreign learners of French (39 sec., 85 syll.), see [13] for the presentation of this material. For each of the four files, the annotators received the associated Praat textgrid file [14], which provided a 4-layer segmentation structure: segmentation within a phones string, syllabic string, words string and breath groups string, all displayed on four temporally-aligned tiers. These tiers were obtained automatically thanks to a script [15], and checked manually by the supervisor of the experiment. During the checking phase, disfluencies (syllables associated with a hesitation, a false start or an overlap) were marked specifically. A fifth empty tier, duplicated from the syllabic

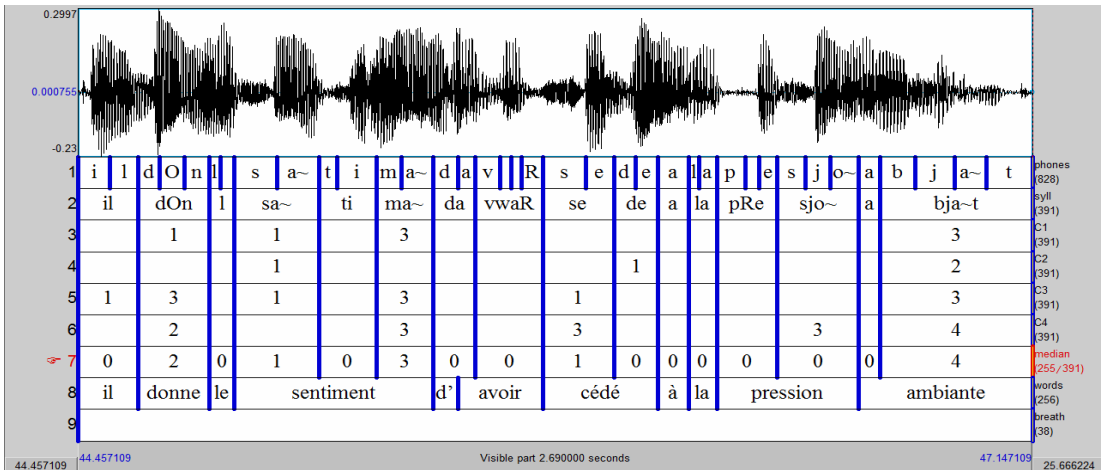

Figure 3: *Praat screen shot of the manual coding. Analysis of the utterance: "il donne le sentiment d'avoir cédé à la pression ambiante" [read aloud speech]. The transcription tiers are, from top to bottom: phones, syllables (both in SAMPA), manual coding of the four annotators (C1, C2, C3 and C4), reference coding (the median of all the annotators) and words.*

tier but containing pauses and disfluency markers, had to be filled in by the annotators. In all, the material corresponds to 739 syllabic intervals.

The coding methodology is structured in the following way: each annotator browses the file from left to right and organizes the work in two steps. First, annotators were asked to listen at most three times to each breath group, and to fill in the intervals of the empty syllable tier by marking number "4" for strong prominences, and number "3" for the syllables where they hesitated between strong and weak prominences. Then, within each prosodic group thus created, they listened to it again for a maximum of three times, marking weak prominence with number "2", and the syllables where they hesitated as to whether they were weak prominences or non-prominent with number "1". They left the other intervals empty, and recommenced the operation with the next breath group, and so on, until the whole file had been processed. The annotators learnt the task with the supervisor of the experiment on a minute-long stretch of corpus (63 sec., 335 syllables), consisting of a monologue of spontaneous speech (interview with a shopkeeper from southern France). Note that since the annotators do not have access to the acoustic parameters (melodic and intensity line, spectral information), the identification of prominences is based only on perceptual processing.

3.2. Inter-annotator Consistency

The **kappa statistic** has been widely used in the past decade to assess inter-annotator agreement in prosodic labeling tasks [16], [17], and in particular the reliability of inter-annotator agreement in the case of a categorical rating [18]. Among the many versions proposed in the literature, we selected the **Fleiss' kappa** [19], which provides an overall agreement measure over a fixed number of annotators in the case of categorical rating (unlike Cohen's Kappa [20] which only provides a measure of pairwise agreement).

We projected the continuous coding onto a categorical scale, considering 2, 3 and 4 as a single category of prominence, and excluding the syllables marked with the number 1 (this marker being considered as a categorical hesitation marker) in the inter-annotator consistency calculation. The result gave a Fleiss' kappa of 0.6713, which is a substantial agreement.

Then, in order to get a reference coding tier to compare with the automatic annotation, we computed the results by calculating the median of the annotations for each file. Absurd codings (*e.g.* when one annotator put 4 whereas none of the others marked the syllable as prominent) were excluded from the analysis. Measures with too great a dispersion (*e.g.* when one annotator put 1, another put 4 and none of the others filled the interval in) were marked 1 (which is the indecision marker). For all the other cases, the reference coding is the median. Figure 3 gives an illustration of the manual coding.

4. Comparing Automatic and Manual Coding of Prosodic Structure

The prominence degree algorithm was run on all the data, with the following thresholds: for the map task: D = 1.71; H = 2.6; R = 2.47); for dialogue; D = 1.54; H = 1.38; R = 2.48), and for read speech: D = 1.61; H = 1.43; R = 2.07.

At this point of the development, we have a corpus of 739 syllables for which we have two types of information: (i) a discrete classification in five ordinate classes which results from the manual annotation of a consortium of annotators (from 0 to 4), and (ii) an automatic classification on a scale of values between 0 and 10. In order to compare these two annotations, we needed to make the automatic annotation discrete by choosing four thresholds t_1, t_2, t_3 and t_4 so as to obtain an ordinate classification in five classes: class 0 if the strength is between 0 and t_1, class 1 if the strength is between t_1 and t_2, etc. up to class 4 if the strength is between t_4 and 10.

In order to determine the thresholds, we seek to minimize the gap between the two classifications. The gap for a given syllable is 0 if the two classifications are identical for this syllable; it is 1 if the two classifications differ by one rank (for example the manual classification is 0 and the automatic classification is 1, or the manual classification is 3 and the automatic classification is 2, and so on; the gap is 2 if they differ by two ranks, etc. The total gap between the two classifications is the sum of the gaps for all the syllables. The formula which gives the gap G is the following:

$$G = \sum_i |a_i - m_i| \qquad (3)$$

where m_i and a_i are the manual class and the automatic class of the i^{st} syllable.

In this way, it is easy to determine the threshold which minimizes the total gap between the two classifications (it is possible to demonstrate mathematically that we just have to minimize the number of mistakes relating to each threshold separately). For the 739 syllables of our corpus, we obtain the following thresholds: t_1 = 2.35; t_2 = 2.85; t_3 = 3.55 and t_4 = 6.5. The minimum gap is 257. The matrix of confusion (which details the mistakes) is the following:

Table 1. *Matrix of confusion between manual and automatic classes of prominence degrees.*

		Automatic classes					Total
		0	**1**	**2**	**3**	**4**	
Manual classes	**0**	492	18	12	9	0	**531**
	1	19	7	5	4	0	**35**
	2	23	5	3	11	0	**42**
	3	10	3	15	40	14	**82**
	4	2	1	1	16	29	**49**
Total		**546**	**34**	**36**	**80**	**43**	**739**

The correlation measure shows that 571 syllables are correctly classified (namely a rate of 77%), which is quite encouraging. More interestingly still, if we admit as acceptable the syllables for which the gap is one (classed 0 instead of 1, 3 instead of 2, etc.), in other words, if we add the two sub-diagonals in light grey, we obtain 674 syllables, namely a rate of 91% of acceptable answers.

We also calculated the kappa measure to compare the two classifications. The standard kappa [19] does not give good results: the value is around 0.5 (0.493). Weighted kappa, in contrast, which is used when all the non-concordances do not have the same importance [21], gives around 0.7 (0.696) with a linear weighting (a non-concordance is n times more serious for a gap of n than for a gap of 1). And if we take a quadratic weighting (a non-concordance is n^2 times more serious for a gap of n than for a gap of 1), the kappa approximates 0.8 (0.807).

5. Discussion & Conclusion

In this paper, we have presented an algorithm which generates the prosodic structure of a given utterance using acoustic features only. On the basis of the phonemic alignment, our system proceeds to the location of prominent syllables, and then estimates the prominence strength of the syllables detected as prominent. We have compared the performance of the software with a manual annotation carried out by four annotators on a 4-minute long corpus, involving different genres. From this comparison, it appeared that the performance was quite encouraging, with a correlation measure giving a rate of 91% of good classification, and a Fleiss' kappa approximating 0.8, in the best cases.

Admittedly, to be acceptable, the validation of the prominence degree algorithm must be conducted on a larger corpus, involving more annotators and more discourse genres. Nevertheless in its present state, this kind of tool may help to advance our knowledge of French prosodic structure, making it possible to validate or invalidate well-known rules on the prosody/syntax interface.

6. Acknowledgements

This research has been funded by two institutions: the Swiss National Science Foundation (under grant n°PBNEP1-127788, Neuchâtel University), and the Agence Nationale de la Recherche (ANR-07-CORP-030- 01, "Rhapsodie – Corpus prosodique de référence du français parlé").

7. References

[1] Ladd, D. R. "Notes on the phonology of prominence", Working Papers Lund University, 41:10-15, 1993.

[2] Liberman, M. and Prince, A. "On stress and linguistic rhythm", Linguistic Inquiry, 8:249-336, 1977.

[3] Pierrehumbert, J. The Phonology and Phonetics of English Intonation, PhD thesis, MIT, 1980.

[4] Dell, F. "L'accentuation dans les phrases en français", in Dell, F., Hirst, D.J. & J.-R. Vergnaud (eds), Forme sonore du langage : Structure des représentations en phonologie, Paris, Hermann, 65-122, 1984.

[5] Martin, Ph. "Prosodic and rhythmic structures in French", Linguistics, 25:925-949, 1987.

[6] Delais-Roussarie, E. "Phonological phrasing and accentuation in French", in Nespor, M. & Smith, N. (eds), Dam Phonology, La Haye, Holland Academic Graphics, 1-38, 1996.

[7] Delais-Roussarie, E. and Post, B. "Unités prosodiques et grammaire de l'intonation: vers une nouvelle approche", Actes des 27èmes Journées d'Etudes sur la Parole, 2008.

[8] Avanzi, M., Gendrot, C. and Lacheret-Dujour, A. "Is there a prosodic difference between left-dislocated and heavy subjects? Evidence from spontaneous French", Speech Prosody Proceedings, 2010.

[9] Dehé, N. "Clausal parentheticals, intonational phrasing, and prosodic theory", Journal of Linguistics, 45:3:569-615, 2009.

[10] Avanzi, M., Lacheret-Dujour, A. and Victorri, B. "A Corpus-Based Learning Method for Prominence Detection in Spontaneous Speech", Prosodic Prominence, Speech Prosody Satellite Workshop Proceedings, 2010.

[11] House, H. Tonal Perception in Speech, Lund, University Press, 1990.

[12] Lacheret-Dujour, A. and Beaugendre, F. La prosodie du français, Paris, CNRS, 1999.

[13] Avanzi, M. and Delais-Roussarie, E. Regards croisés sur la prosodie du français, Journal of French Language Studies, 21, 2011, http://www2.unine.ch/conscilaprosodie/page26423.html

[14] Boersma, P. and Weenink, D. Praat: doing phonetics by computer (version 5.1), www.praat.org, 2010.

[15] Goldman, J.-Ph. EasyAlign: a semi-automatic phonetic alignment tool under Praat, http://latlcui.unige.ch/phonetique, 2008.

[16] Syrdal, A. K. and McGory, J. Inter-transcribers Reliability of ToBI Prosodic Labelling. Proceedings of the International Conference on Spoken Language Processing, Beijing, China, 3:235-238, 2000.

[17] Smith, C. L. Naïve listeners' perceptions of French prosody compared to the predictions of theoretical models. Proceedings of the third Symposium Prosody/Discourse Interfaces, Paris, September 2009.

[18] Carletta, J. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, 22:2:249-254, 1996.

[19] Fleiss, J. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76:5:378–382, 1971.

[20] Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20:37-46, 1960.

[21] Fleiss, J. and Cohen, J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, Educational and Psychological Measurement, 33, 613-619, 1973.