A Corpus-based Learning Method for Prominence Detection in Spontaneous Speech

Mathieu Avanzi[1, 2], *Anne Lacheret-Dujour*[2] *& Bernard Victorri*[3]

[1] Chaire de linguistique française, Université de Neuchâtel, Neuchâtel, Switzerland
[2] MoDyCo, Université Paris Ouest Nanterre, Paris, France; [3] Lattice, CNRS, Paris, France
mathieu.avanzi@unine.ch, anne@lacheret.com, bernard.victorri@ens.fr

Abstract

The aim of this paper is to present a software tool called ANALOR, which allows semi-automatic prominence detection in spontaneous French. On the basis of a manual annotation performed by two experts on a 70-minute long corpus including different regional varieties of French (Belgian, Swiss and metropolitan French) and various discourse genres (from read speech to spontaneous conversations), our system conducts a learning-method in order to determine the best thresholds for prominence prediction. This procedure appreciably improves detection, with consistency between automatic identification and the human labeling rising from 75.3 without training to 79.1 of f-measure after corpus-based learning.

Index Terms: prominence, discourse genre, corpus-based learning method, automatic detection.

1. Introduction

Nowadays, the automatic detection of prominence is considered by experts as an international challenge for the processing and linguistic analysis of spoken corpora, whatever the linguistic topic is (intono-syntactical rules, discourse/prosody interface, pragmatic effects of accentuation, marking of expressivity and emotion, etc.). Traditionally, automatic prominence detection has been based on (i) manual annotation, which is used as a reference for the automatic learning step and the development of a prosodic model of discourse; (ii) in-depth knowledge of the acoustic correlates of prominence perception. Many studies have addressed the problem of prominence detection over the last decade, and algorithms are still emerging [1][2], particularly for French [3][4][5][6]. In this paper, we present one of these algorithms (ANALOR), focusing on the learning methods and the theoretical prerequisites underpinning its constitution.

2. Comparing three systems for French

The earliest studies on the perceptual and automatic identification of prominence in French were conducted within the PFC Project ([7][8]), and were continued by an informal consortium of linguists in a certain number of publications (see [9] for more details). They gave birth to three systems: ANALOR (see [3] and §3 below for the most recent description of the tool), PROSOPROM [4] and IRCAMPROM [5]. A study comparing the performances of these algorithms on the basis of a 50-minute annotated corpus of spontaneous speech was published in [6].

2.1. Constitution principles

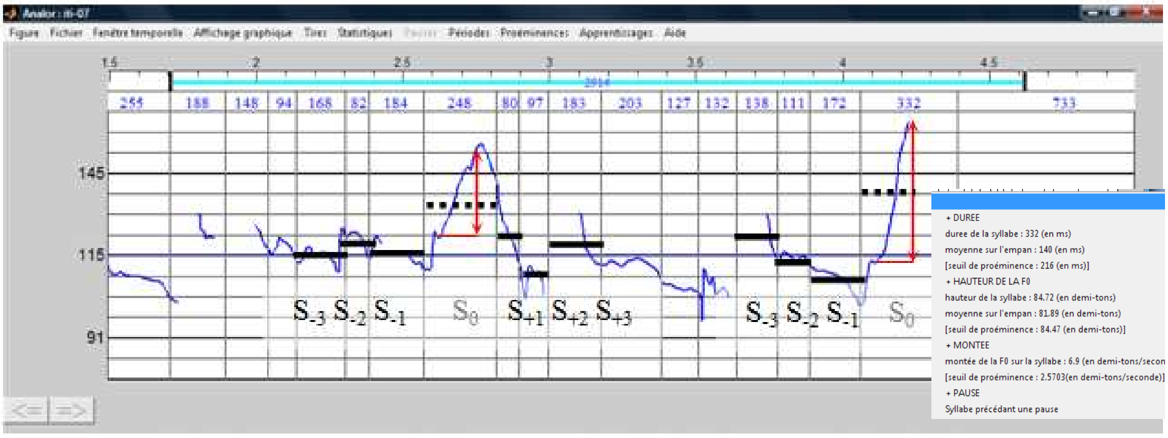
The three systems share at least three principles: (i) prominence is syllabic; (ii) as prominence is a local phenomenon [10], the context-window for identification of prosodic variations must be a constrained one; (iii) the acoustic parameters involved in prominence perception are numerous, but f0 and duration are the most important ones concerning French. Beyond these three principles, they follow different options.

Thus, among the numerous acoustic parameters involved in prominence perception, the three software programs do not focus one the same prosodic features. From this point of view, ANALOR is the least sophisticated of the three. It considers the presence of a subsequent silent pause (a silent pause being considered as a strong clue for the identification of the end of a prosodic group in French [11]), and it calculates significant variations in relative height and relative duration averages to detect the syllables which stand out from their environment like a figure on a ground. PROSOPROM does the same, one difference however is that it also considers the presence of a rising tone on the current syllable (if the amplitude of the contour reaches a certain value, the syllable will be considered as prominent). IRCAMPROM is the most complex of the three tools, as the prominence detection it conducts consists in the manipulation of ten acoustic parameters, comprising duration (syllable duration, local speech rate and nucleus duration); pitch (f0), and spectral (specific loudness) features.

Concerning the context-window for relative calculations of significant prosodic variations, ANALOR employs the “prosodic period” (a unit defined by the presence of a silent pause following a contour of a certain amplitude, and associated with a melodic reset, see [3] for further details), i.e. it uses a more or less large dynamic unit (like any other discourse unit, the size of a “prosodic period” varies greatly from one speaker to another). PROSOPROM considers a static and constrained domain for prominence detection, namely the immediate syllabic context (the two preceding syllables and the following one), while IRCAMPROM mixes the two strategies by taking into account the immediate syllabic context (one syllable before, one after) and the inter-pausal group.

2.2. Performances

A subpart of the C-PROM database (see [9] and see §3.4. below) has been used to train, evaluate and compare the three algorithms. The set of recordings (50-minutes long) was composed of a total of 12851 syllables (semi-automatically parsed with the EasyAlign [12] Praat [13] script), and annotated for prominence analysis by two experts. During the manual annotation by these two experts, 973 syllables were excluded (because they were associated with an elongation connected with a hesitance or because they presented specific



+ DUREE
duree de la syllabe : 332 (en ms)
moyenne sur l'empan : 140 (en ms)
[seuil de proéminence : 216 (en ms)]
+ HAUTEUR DE LA F0
hauteur de la syllabe : 84.72 (en demi-tons)
moyenne sur l'empan : 81.89 (en demi-tons)
[seuil de proéminence : 84.47 (en demi-tons)]
+ MONTEE
montée de la F0 sur la syllabe : 6.9 (en demi-tons/secon
[seuil de proéminence : 2.5703(en demi-tons/seconde)]
+ PAUSE
Syllabe précédant une pause

Figure 1: *ANALOR screen shot of the utterance: "euh jusqu'à l'église Notre Dame vous prenez la première à gauche" [mp-7]. On the abscissa, temporal values are given in milliseconds; on the ordinate, the values of F0 in a logarithmic scale can be seen. Duration labels are given in milliseconds. Annotation tiers are, from top to bottom: phones, syllables (both in SAMPA), manual annotation ("prommanu", indicating prominence syllables (P), excluded syllables (z), silent pause (_) and breath (*)) and graphemic words.*

prosodic properties, such as being in a post-focus position or containing a schwa, see [9] for the details of the procedure) and 3244 syllables were annotated as prominent. Out of the remaining syllables, 8634 units are non-prominent and non-excluded syllables.

Two of the software programs were automatically trained on the basis of the manual annotation. While PROSOPROM conducted a discriminant analysis, IRCAMPROM conducted both a discriminant analysis and a context-dependency to determine the best threshold for automatic prominence labeling. The results showed that the tool which had not been trained was the less robust. Indeed, ANALOR presents an f-measure of 69.7%, with a tendency to under-detection (recall: 63.6% and precision: 77.2). PROSOPROM performs slightly better (f-measure: 71.7%), but tends to over-detection (recall: 78.9%, and precision: 68.2%) while the performance of IRCAMPROM is the highest (f-measure 75.4%) and the most well-balanced (recall: 76.4%, and precision: 74.5%).

2.3. Summary and consequences for future work

This study is interesting with regard to three points. (i) It revealed that too large a window for the relative acoustic calculation (such as the prosodic period) engenders under-detection; too small a window (such as the one PROSOPROM uses), however, is not more adequate, because it tends to over-detection. A compromise between the two seems to be the optimum solution. (ii) The comparison of the three algorithms reveals that increasing the acoustic parameters did not lead to significant improvements in performance (when the four acoustic PROSOPROM parameters were compared with the ten IRCAMPROM features, a MacNemar test (p = 0.005) showed that there was no significant difference between the performances of these two tools). In linguistics, as elsewhere, the simpler the system, the better. We therefore decided to manipulate as few criteria as possible. (iii) The experiment was also instructive because it showed the limits in the generalization ability of a rule based system and the need for a corpus-based training for prominence detection.

3. Prominence detection with ANALOR

These observations have guided our improvements to the ANALOR software. In the next section, we present the changes we have made to the original algorithm as it was described in [3].

3.1. Acoustic parameters calculation

The automatic algorithm still relies on basic relative acoustic parameters such as f0, duration and pause. Three changes have however been made.

The first concerns the context-window for relative calculations of duration and height averages. While in the original algorithm we considered the intonational period to estimate the significant melodic and duration variations, we now use a fixed environment, determined by the three preceding syllables and the three following ones of the current syllable. This strategy was chosen in order to determine a processing window close to the accentual phrase (this unit, which is hard to define on grammatical criteria, is generally considered as composed at most of seven syllables, see [14] and [15]).

One other change we made concerns relative syllabic duration. Ideally, a syllable duration model based on intrinsic syllable properties and normalized local speech rate for spontaneous speech should be used (for the first experiment on read aloud sentences see [16]). However, such a model still has to be reliably developed and tested. We therefore took only the number of phonemes of the syllable into account, thus avoiding the bias of syllable weight: for example a syllable composed of five phones is by nature longer than a mono-phonemic one, as has been demonstrated by work on syllabic quantities (see for example [17]).

The last modification made concerns rising tones. Like PROSOPROM, ANALOR now detects a syllable as prominent if it bears a rising tone reaching a certain amplitude. One difference however between the two tools is that in ANALOR, the rise amplitude is measured on the vocalic part of the syllable, not on the whole part of it. This restriction to nucleus is based on the fact that the melodic variations on consonants are less relevant perceptually than those borne by vowels [18]).

Figure 1 illustrates how ANALOR calculates the different acoustic parameters used for prominence detection. The algorithm calculates, for the current syllable (S_0), the following features: its relative height and duration average compared with the f0 and averages of the three preceding syllables (S_{-3}; S_{-2} and S_{-1}) and the three following ones S_{+1}; S_{+2}

and S_{+3}), the presence of a rise if there is a positive movement of f0 on the syllabic nucleus, and the presence of an adjacent silent pause (the latter label is based on the pre-manual syllable segmentation of the corpus). F0 measures are given in semi-tones, while duration measures are calculated without any unit. Note that the contextual relativization is blocked if there is a syllable marked as excluded in the labeling tier (based on the pre-manual annotation of the corpus) or a silent pause in the immediate context of the current syllable. In the example, the last syllable of the utterance is followed by a pause. Duration and f0 measures are thus calculated only with reference to the three preceding syllabic intervals. Clicking on the current syllable makes a small window appear in which one can consult the different measures calculated.

3.2. Method for prominence threshold optimization

As mentioned in the previous section, prominence detection is performed on the basis of a multi-criteria analysis which relies on five parameters. The silent pause parameter does not need to be trained (the duration of such a prosodic object is not important in a prominence identification task, as we consider that the presence of a pause is sufficient to activate prominence, see [3] for the details of the argumentation), but the other four have to be. They are:

- The relative syllabic duration threshold, S_D
- The weight given to the number of phones in the calculation of the syllable duration W_{Ph}
- The relative syllabic height average threshold, S_H
- The intra-vocalic amplitude rise threshold, S_R

The method we decided to follow in order to obtain the best parameters for automatic prominence identification consisted in carrying out a supervised corpus-based learning. The aim was to hone the f-measure performance by comparing systematically the results with the human annotations.

The algorithm used for automatic learning is based on a decreasing step-size random search in the parameter space from an initial relevant value. More precisely, if *V* is a vector of the space (a 4-D space, the vector *V* having as components S_D, W_{Ph}, S_H and S_R), the algorithm can be described as follows:

Let δ_i be the step size, V_i the value of the parameter set, and F_i the F-measure at step *i* of the procedure. We perform a random search to find a new value of *V* which improves the F-measure by searching in the neighborhood of V_i defined by step size δ_i. That is to say we try the *V* values of the form:

$$V = V_i + \delta_i \cdot \Phi \cdot V_i \qquad (1)$$

where Φ is a uniformly distributed random vector in the hypercube unit.

As long as we find a better value for V, we continue by replacing V_i by this value. If N_{max} attempts are made without finding a better value, we proceed to step *i*+1 of the procedure with a step size $\delta_{i+1} = \delta_i$ /2. The procedure stops when the step size becomes smaller than a given value δ_{min}. The results given below were obtained with $N_{max} = 250$, $\delta_1 = 0.4$ and $\delta_{min} = 0.01$.

To conclude this description of the corpus-based learning method, it should be pointed out that this algorithm is efficient if and only if the initial values of the parameters are sufficiently close to the optimal values. In other words, the initial values were fixed on the basis of a linguistic analysis, calling on specific linguistic knowledge. For this study, we considered the following initial values: $S_D = 2$; $W_{Ph} = 3.3$; $S_H = 2$ and $S_R = 3$. For a justification of the value of these thresholds fixed *a priori*, the reader is referred to [11], [19], [20] and [21].

3.3. Material

The C-PROM corpus was used to train the algorithm and compare its performance with a manual annotation. As the corpus is fully presented in [9], only a brief summary is given here. The corpus is 70 minutes long, comprises 7 genres, with, from the more to the less formal: Read Speech (RS), Political Speeches (PS), Conferences (CF), News Broadcasts (NB), Radio Interviews (RI), Map Tasks (MT) and Life Stories (LS); in all, 29 native speakers of French (13 females, 16 males) from Belgium, Switzerland and France are involved. On the basis of pre-manual syllable segmentation, two expert transcribers annotated the prominent syllables of the corpus, and labelled elongations associated with a hesitance, post-tonic schwas and post-focus syllables. One of the authors of this study re-annotated post-tonic schwas and post-focus syllables as prominent or non-prominent syllables. He also excluded the syllables preceding a pause connected with a syntactic interruption, in order to filter the silent pauses. During this coding phase, he also corrected certain annotations (removed, deleted or added some syllables boundaries).

Table 1. *Details of the corpus study, with, from left to right: discourse genre, duration, number of syllables, number of prominences (P), non-prominent/non-excluded syllables (NP) excluded syllables (Z), and valid syllables.*

Disc. genre	**Duration (sec.)**	**Nb. Syll.**	**P**	**NP**	**Z**	**Valid_ syll.**
RS	401	1830	470	1357	0	470
PD	635	2174	632	1539	1	633
NB	621	3165	825	2279	58	883
CF	687	3133	818	2202	108	926
RI	627	2591	684	1806	90	774
MT	590	2222	562	1490	162	724
LS	622	2663	685	1763	199	884
TOTAL	**4183**	**17730**	**4676**	**12436**	**618**	**17112**

Table 1 shows the detail of the corpus used for this study. It comprises 17730 syllables, among which 4676 were marked as prominent (P), 618 excluded via the manual annotation tier (Z), and 12436 which were neither associated with a hesitance or a syntactic interruption, nor were prominent (NP). The algorithm uses the 17112 valid syllables (P + NP syllables) to train itself.

3.4. Evaluation

Following the method described in §3.2., we trained, for each discourse genre, the intuitive thresholds initially fixed. The measure selected to assess agreement between the manual annotation and the automatic identification is the f-measure, that is to say the harmonic average between precision and recall [22]. Table 2 shows the performance of our tool for each discourse genre.

Table 2. *% of F-measure for each discourse genre, before and after training. Average for all the discourse genres is given in the grey columns. The column “gain” indicates the jump before and after learning.*

Genre	**initial performance**	**trained performance**	**Gain**

	Prec.	Rec.	F-ms	Prec.	Rec.	F-ms	
RS	79.86	71.7	75.56	76.41	77.87	77.13	1.57
PS	75.07	83.39	79.01	82.35	81.86	82.16	3.15
NB	74.57	73.58	74.07	75.7	82.3	78.86	4.79
CF	76.11	73.23	74.64	79.18	79.95	79.56	4.92
RI	71.88	82.6	76.87	79.3	80.89	80	3.13
MT	75.31	76.51	75.9	79.86	79	79.43	3.53
LS	83.27	61.75	70.91	73.44	80.73	76.91	6.00
TOT.	**76.58**	**74.68**	**75.28**	**78.03**	**80.37**	**79.15**	**3.87**

As we can see, the corpus-based learning improved the results by about 3.87% of f-measure: the performance before training is 75.3%, against 79.15% after training. The best progression is for the LS discourse-genre (6%) and the worst for RS (1.57%). Concerning the agreement rate between manual annotation and automatic detection, it can be seen that the best score is for PS, while the worst is for LS recordings. Globally, the performance reached by our tool (79.15%) is close to the inter-annotator consistency found by [9] (where it was estimated at 82.8% of F-measure), which is quite encouraging.

While it may be a little adventurous to compare two experiments which were not carried out with exactly with the same material, we can conclude that the modifications made considerably enhanced the detection precision of our tool. The improvement introduced also made it possible to adjust the precision and the recall, and to achieve a more well-balanced detection. When precision and recall between the two performances are compared, results show that, apart from PS and LS, the detection is sufficiently well-balanced. Moreover, in comparison with the first experiment, the final results no longer tend to over- or under-detection (recall = 78.03 and precision = 80.37).

4. Discussion & Conclusion

The aim of this paper was to present a software tool for semi-automatic prominence detection in spoken French. From a pre-aligned and annotated transcription, the ANALOR algorithm calculates the value of a certain number of prosodic contextual variations, involving f0, duration and pause features. On the basis of a manual prominence annotation, it then estimates the best thresholds associated with the activation of syllabic salience. The performances obtained on the corpus studied gave encouraging results, as they reveal that the variation between human and automate was nearly the same as the variation between two humans. Our following investigations will focus on the detection of elongation connected with a hesitance, and integrate syntactic tagging, in order to conduct a fully automatic prominence detection in spontaneous speech. ANALOR can be downloaded from: http://www.lattice.cnrs.fr/Analor.html. Sources are in free access.

5. Acknowledgments

Funding for this research was provided by two institutions: the Swiss National Science Foundation (under grants n°PBNEP1-127788 and n°100012-113726/1, Neuchâtel University), and the Agence Nationale de la Recherche (ANR-07-CORP-030-01, “Rhapsodie – Corpus prosodique de référence du français parlé”). Thanks also to B. Victorri (Lattice/ENS, Paris) for friendly help in the different stages of the ANALOR software design.

6. References

[1] Tamburini, F. & C. Caini, “An Automatic System for Detecting Prosodic Prominence in American English Continuous Speech”, International Journal of Speech Technology, 8, 33-44, 2005.

[2] Rosenberg, A & J. Hirschberg, “Detecting pitch accent using pitch corrected energy-based predictors,” Interspeech’07, 2777–2780, 2007.

[3] Avanzi, M. Lacheret-Dujour, A. & Victorri, B. “ANALOR. A Tool for Semi-Automatic Annotation of French Prosodic Structure”, Proceedings of Speech Prosody’08, 119-122, 2008.

[4] Goldman, J.-P.; Avanzi, M.; Lacheret-Dujour, A.; Simon, A.C.; Auchlin, A., “A Methodology for the Automatic Detection of Perceived Prominent Syllables in Spoken French”, Interspeech, Antwerp, Belgium, 2007, 98-101.

[5] Obin, N. Rodet, X. & Lacheret-Dujour, A. “Prominence model: a probabilistic framework,” ICASSP, Las Vegas, NV, USA, 2008, 3993–3996.

[6] Obin, N., Goldman, J.-P., Avanzi, M. & Lacheret-Dujour, A. « Comparaison de trois outils de détection semi-automatique des proéminences dans les corpus de français parlé », Actes des 22èmes JEP, Avignon, 2008.

[7] Durand, J., Laks, B. & C. Lyche. “La phonologie du français contemporain: usages, variétés et structure”, in Pusch, C. & W. Raible (eds.). Romance Corpus Linguistics - Corpora and Spoken Language, Tübigen, Gunter Narr Verlag, 93-106, 2002.

[8] Poiré, P., « La perception des proéminences et le codage prosodique », Bulletin PFC, 6, 69-79.

[9] Avanzi, M., Simon, A.C. Goldman, J.-P. & A. Auchlin, “C-PROM. An Annotated Corpus for French Prominence Studies”, Prosodic Prominence: Perceptual and Automatic Identification (Speech Prosody 2010 workshop), Chicago, USA, 2010.

[10] Terken, J. “Fundamental Frequency and Perceived Prominence of accented syllables”, Journal of the Acoustical Society of America, 89, 1768-1776, 1991.

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

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

[13] Boersma, P. & D. Weenink, Praat: doing phonetics by computer (Version 5.1). www.praat.org, 2009.

[14] Wiolland, F., Les structures rythmiques du français, Paris: Slatkine-Champion, 1985.

[15] Martin, Ph., “Prosodic and rhythmic structures in French”, Linguistics, 5/5, 1987, 925-949.

[16] Obin, N, Rodet, X., & A. Lacheret-Dujour, “A Multi-Level Context-Dependent Prosodic Model Applied to Durational Modeling”, Interspeech, Brighton, UK, 2009.

[17] Dell, F. Generative Phonology and French Phonology, Cambridge, Cambridge University Press, 1980.

[18] House, D., Tonal Perception in Speech, Lund, University Press, 1990.

[19] Rossi, M. "Le seuil différentiel de durée", in Papers in Linguistics and Phonetics to the Memory of Pierre Delattre, Vol. 54, A. Valdman (ed.), Collection Janua Linguarium, Mouton, The Hague, Indiana University, 1972.

[20] D'Alessandro, C. & P. Mertens, “Automatic pitch contour stylization using a model of tonal perception”, Computer Speech and Language 9/3, 257-288, 1995.

[21] Garnier-Rizet, Elaboration d'un module de règles phonético-acoustiques pour un système de synthèse à partir du texte en français, Phd, LIMSI/CNRS, Paris, 1994.

[22] van Rijsbergen, C.J. Information Retrieval, Butterworths, London, 1979.