ANALOR. A Tool for Semi-Automatic Annotation of French Prosodic Structure

Mathieu Avanzi, Anne Lacheret-Dujour, Bernard Victorri

ANALOR

A Tool for Semi-Automatic Annotation of French Prosodic Structure

Mathieu Avanzi[1]; Anne Lacheret-Dujour[2]; Bernard Victorri[3]

[1]U. Neuchâtel, Switzerland & U. Paris X Nanterre, France; [2]U. Paris X Nanterre & IUF, France; [3]Lattice, ENS, Paris, France

mathieu.avanzi@unine.ch; anne@lacheret.com; bernard.victorri@ens.fr

Abstract

In the area of large speech corpora, there is a definite need for common prosodic notation system based on efficient (semi)-automating tools of prosodic segmentation and labelling. In this context, we present the software program ANALOR, developed in order to process semi-automatically prosodic data. From a text-sound alignment, this computer tool detects major prosodic units, on the basis of global and local melodic variations. That leads to the segmentation of an utterance in **prosodic periods**. Inside those prosodic periods, **prominent syllables** are then automatically detected.

1. Introduction

Linguistics and speech technology have dealt with prosody from various points of view, which make a precise definition of the scope of research on prosody difficult. Nevertheless, a complete analysis is very useful as part of a linguistic analysis in order to determine the optimum number of functional prosodic units and to determine their nature according to precise acoustic correlates. In this context, most of the existing transcription systems, whether they engender a phonologic interpretation, like ToBI system [2], or not (see for example IViE [18] or IVTS [9] systems), necessarily share the point of view that prominence processing represents the cornerstone of the prosodic annotation. Actually, most of prosodic annotation systems are based on manual processing. This situation remains problematic for at least tow reasons.

First, it is a well-known fact that manual prosodic annotation varies greatly from an expert to another. See for example [16]'s experiment on the SEC. The authors report that the disagreement between the two experts who annotated the same sub-part of the whole corpus (nearly 4190 syllables) was of about 27 %. Regarding the assignment of tonetic stress marks, [5]'s calculations reveal that consensus was of about 55% at best. See also, for an example on spontaneous French, [17]. The author asked seven prosodic experts to annotate prominent syllables in a small stretch corpus (165 syllables). The proportion of syllables marked as prominent varied from 19% to 49%.

Others studies, like [4], showed that better results could be obtained if the annotators followed a strict protocol (set of symbols reduced, common training, etc.). Thus, they respectively got a very satisfactory inter-transcribers agreement on a 45-minute long spoken Dutch corpus. However, such a manual procedure is extremely time-consuming. The authors concluded that a non-expert annotator would need about 40 times the duration of the corpus to annotate minimal prosodic phenomena as strong and weak breaks, segmental lengthening and prominent syllables.

As a consequence, automating the procedure of prosodic annotation in spoken corpora is of great importance. In this paper, we present what the two different steps that punctuate the procedure of prosodic annotation proposed consist of. The division in major prosodic units (called **prosodic periods**) and the methodology which leads to the development of this first algorithm is described in section 2. Then, we present the algorithm used for the detection of **prominent syllables** (section 3). We finally conclude with the specificity of our software, compared with other quite similar tools (section 4).

2. Segmentation in major prosodic units (prosodic periods)

The concept of **prosodic period** stems from an inductive approach, which rests on the comings and goings between a manual observation of the data and a computer modelling. This method, introduced by [11] and [12], is structured around three fundamental steps.

2.1. Methodology

We first conducted a manual analysis on a small corpus of French radio talk (about two hours, with male and female speakers, see [12] for details). The goal was to isolate a set of phonetic cues (silent pauses of a certain minimal duration, major contour of specific amplitude) associated to what is commonly perceived of as a **strong prosodic break**.

The computer modeling, which comes after this first stage, rests on the processing of local and global pitch variations in a given time interval. The implementation of an automatic segmentation mechanism was thought up in order to systematically test the principles issued from manual processing.

Comparing the manual analysis to the automatic data processing permits the highlighting of differences between the two treatments. In fact, two tracks were then studied to report these differences: (i) questioned by the manual decision-making and/or (ii) redefining the criteria used for the automatic processing. In the first case, the computer modelling intervenes as a control mechanism, *i.e.* it allows for adjusting the intuition of the experimenter by pointing to incoherencies in the analysis, thus rationalizing the initial intuitions. In the second, the observation leads to refine the criteria used for the automatic data processing. It is after all these comings and goings between manual observation and automatic data processing that we developed a stable algorithm of automatic segmentation of the statements that we decided to call **prosodic periods**.

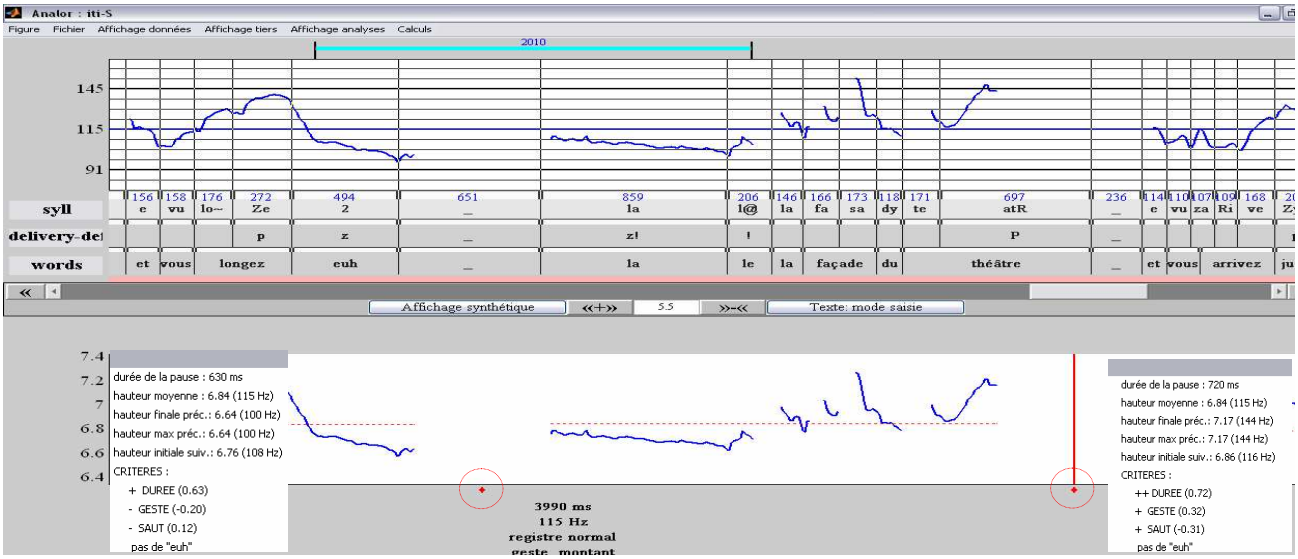

Figure 1: *Screenshot of ANALOR. Transcription of the sequence:*
« et vous longez euh – la le la façade du théâtre – et vous arrivez juste (…) » [iti-S]

In the abscissa, temporal values are given in milliseconds; in the ordinate, the values of F0 in a logarithmic scale can be seen. To run, the software needs alignment files in xml format. These can be obtained with a Praat script [3]. The user can make as many tier-alignments as he wishes to appear. Those tiers will appear in the **top window** of the screen, which are, from top to bottom: **syllable tier** (in SAMPA alphabet), obtained automatically thanks to the Praat script EasyAlign [7]; **manual categorization of syllables tier** (strong and weak prominences, respectively labelled: 'P' et 'p'; segmental lengthening: 'z'; syntactic interruptions: '!'; see [8] for the origin of this manual annotation) and the **graphic word tier**. In the **bottom window**, results obtained from automatic analysis can be visualized.

2.2. Algorithm

The algorithm that emerged from this bottom-up approach gave birth to the ANALOR software program (figure 1), implemented in Matlab. It relies on the characterization of terminal words' boundary contours, objects which may be associated to strong prosodic breaks in the speech flow. Those prosodic breaks depend on the combination of three acoustic markers: **silent pause**, **amplitude of the terminal contour** and **subsequent melodic resetting**. In practice, segmentation of a corpus into periods occurs if and only if the following four conditions are fulfilled:

- Occurrence of a pause of at least 300 ms;
- Detection of an F0 pitch movement reaching a certain amplitude, defined as the difference in height between the last F0 extremum and the mean F0 over the entire portion of the signal preceding the pause ;
- Detection of a "jump", defined as the difference in height between the last F0 extremum preceding the pause and the first F0 value following the pause ;
- Absence of « um » in the immediate vicinity of the pause.

It must be emphasized that the decision to recognize a periodic break does not depend on the exact values of the thresholds but on their size. In other words, when one parameter is very slightly to the chosen threshold, segmentation can occur only if the other parameters have values distinctly above the threshold. The values of the parameters activated during the decision of segmentation in period can be consulted in small boxes situated in the lower window (figure 1). From top to bottom, are noted: (1) the duration of the silent pause in milliseconds; (2) the height averages of the supposed period; (3) the final height of the supposed period; (4) the maximal height towards the end of the supposed period; (5) the initial height of the next period. Then, for the three calculated parameters, DURATION, (1), GESTURE, *i.e.* the distance between (2) and (4) or between (3) and (4), and the JUMP, *i.e.* the difference between (3) and (5), criteria '++' means that this parameter is widely above the threshold, (score = 2), '+', means that it is above the threshold (score = 1), '=' that it is of the order of the threshold (score = 0) and '-' below the threshold (score = 1). Consequently there is automatic cut, the total of the scores has to be superior to or equal to 2, without any negative score and without presence of a hesitance mark.

2.3. Illustration

In the bottom window (figure 1), a vertical red bar indicates the actual periodic cut. The values of the implied criteria can be posted by clicking the small red rhombuses situated under the plan of F0 (circled in red). So, in the analyzed sequence, "et vous longez euh – la le la façade du théâtre – et vous arrivez juste (…)", in spite of a break of duration sharply superior to the chosen threshold (630 ms, that is to say more than double), we do not observe a periodic break after "suivez", because of the weak resetting and of the weak amplitude of the terminal contour, as well as the presence of a "um" of hesitance before this break. On the other hand, the software detects a periodic break after "théâtre". This detection can be explained by the value of the duration pause, coupled with that of the jump and the melodic resetting.

3. Detection of accentual prominences

The second part of the implementation is still in development (see [1] for a first experimentation). It is relative to the modeling of the internal periodic structure. In order to appreciate it, we invoke the notion of **prominence**. In practice, accentual labelling does not rely on a structural feature of the word or word group such as lexical stress, but on a neutral phonetic definition of prominence, as a perceptual salience within background speech ([13], [14], [18], [19] and [20]). The main advantage of this approach is to be independent of any theoretical framework, and of any morphologic or syntactic considerations.

3.1. Algorithm

The automatic algorithm relies on basic relative acoustic parameters (see also [19] for a review of automatic

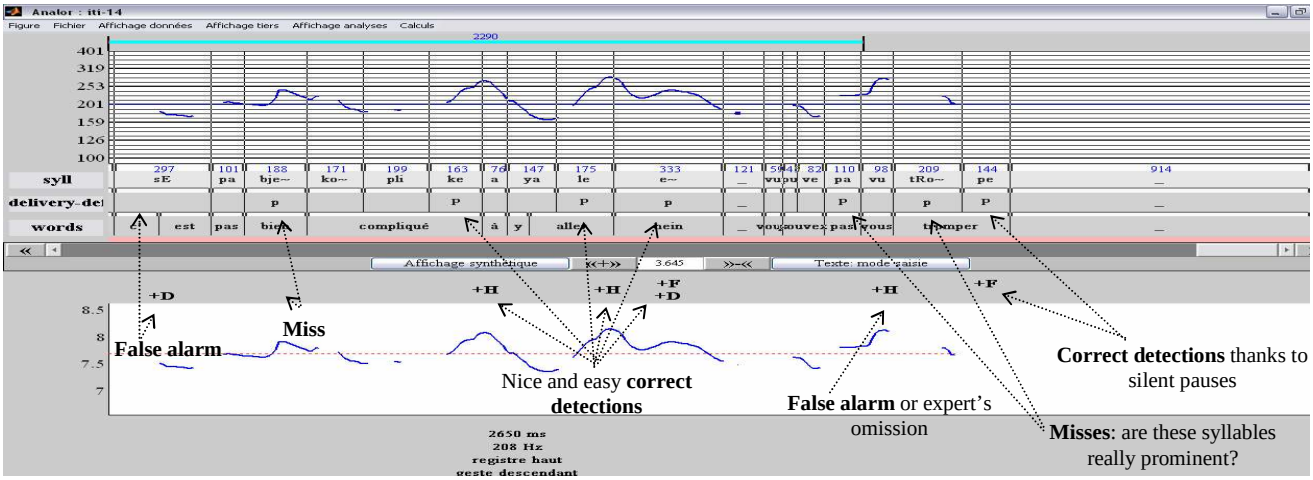

Figure 2: *Screenshot of ANALOR. Transcription of the period:*
« c'est pas bien compliqué à y aller hein vous pouvez pas vous tromper » [iti-14]
Duration and height prominent syllables are respectively labelled +D (for 'durée') and +H (for 'hauteur') in the bottom window. Syllables which are prominent because they are followed by a silent pause are labelled "+F" (for 'fin'). The user can compare directly the results of the automatic analysis with those of the manual analysis (median tier in the top window), and adjust afterwards the annotation

prominence detection). Today, we exploit the **pitch parameters** (pitch range, melodic movement) and **duration** of the syllabic segments composing a given period. Those parameters are calculated relatively, that is to say a syllable will be identified as prominent by the software if it stands out from its environment according to a certain threshold.

For the moment, the formula implemented is this one:

Let $M_h(\alpha)$ be the average and $E_h(\alpha)$ the distance-type of fundamental frequency on a given period P. A syllable is said prominent for height (and it is labelled "H", *cf.* figure 1) if it contains a local maximum pitch, marked h(s), verifying the condition

$$\mathbf{h(s) > M_h(\alpha) + K_h * E_h(\alpha)}$$

where K_h is an adjustable parameter (called "threshold height distance": its value is 1.5, by default).

In others words, the algorithm is based on a Gaussian distribution of F0 centred on the median axis (average of all the points of F0 for the given period), from which we calculate a standard deviation which allows the researchers to quantify the distribution of points around the average of F0, and beyond which we can detect a salient acoustic event. With a variable threshold, the interest of this kind of arithmetic is that it is strong concerning the inter-variability of different speakers. As a consequence, it does not matter much if the speakers modulate a lot or a little: the software will detect an event whatever happens to the significant variations of F0.

The formula is the same for duration.

Because we thought that these two parameters were not sufficient to track down all prominent syllables (like boundary tones which does not manifest a significant pitch or duration variation, see for example the last syllable of the period in figure 2), we decided to use **silent pauses** to hone the detection. Consequently, each syllable followed by a silent pause, whatever the pause duration, will be considered as prominent.

3.2. Evaluation phase

In order to validate this algorithm, we compared the results of the automatic detection with the consensual manual annotation made by two phonologist experts. The test corpus is composed of 18 minutes of spontaneous speech, segmented and aligned in syllables. It includes map tasks and radio interviews (respectively 8 and 10 minutes long). Male and female speakers are natives of Belgium and France. It is presented in detail in [8].

Among the 4432 syllables that the whole corpus contains, 1090 units were annotated as prominent by the experts of [8]'s study. 461 units were categorized as elongations connected to a hesitance (symbol 'z'); they were consequently excluded from the expertise to avoid disrupting the detection of the duration prominences. On the number of remaining syllables, 2881 units are non-prominent syllables.

The results ANALOR software gives are rather encouraging: the rate between the automatic approach and the manual annotation is of **83.8%**. Among this 83.8%, 19.4% and 64.4% were syllables respectively recognized non prominent both by ANALOR and the manual annotation; 7.8% were "misses" (segments labelled 'p' or 'P' by the experts but not recognized as such by the software), while 8.5% were "false alarms" (syllables identified as prominent by ANALOR but not by the experts). Taking account of silent pauses to refine the algorithm appeared to be of great importance: we obtained definitively better results (the score of correct detection was about 78.6% with only duration and pitch parameters).

To conclude, let us note that this score of 83.8% is quite similar to [8]'s results, which was about 84.1%. It is also quite similar to the best scores that are generally mentioned in the literature of others languages (see again [19] for a review).

3.3. Visualization

The user can consult the results of the automatic detection of the prominences in the bottom window (cf. figure 2), and thus compare those results with the manual coding (labelled prominent syllables 'P' or 'p' in the median tier, top window). If necessary (for example if the results are clashing), it is possible to correct the manual note or change the thresholds of automatic detection. The software thus allows making permanent comings and goings between the model and the phonetic analysis, driven by the empirical data: like this we can, at the same time, check the coherence of the analyses and improve the adequacy of the model to the analyzed data.

All the analyses and the results obtained with ANALOR can then be repatriated in TextGrid files (Praat format).

4. Discussion and conclusion

In this article, we presented a software program for prosodic analysis which constitutes a very useful tool assistant to facilitate the prosodic annotation of spontaneous spoken French corpora. A first formula leads to an automatic division of the prosodic continuum in major prosodic units, or prosodic periods, by basing itself on the interaction of melodic and temporal parameters. Within the identified segments, a second operation proceeds in a prominent syllables detection, according to a certain threshold of pitch and duration change.

The specificity of the elaboration of our tool is summed up in two points: (i) it is about an emergent approach of a **bottom-up** type, (ii) even though the approach is driven by strong hypotheses, it is on no account forced by a predetermined theoretical frame (it does not rest on a phonological categorization of the detected accents, as ToBI system [2] for example). Besides, contrary to other tools of the same type, ANALOR does not establish its measures on a stylised signal (like MOMEL [10] for example). Moreover, it is not reserved for the treatment of read sentences, stemming from laboratories in which they were recorded, contrary to the plug-in *WaveSurfer* developed by [19]. Finally, we would like to emphasize that ANALOR offers dynamic results, not static ones such as the plots allowed by ProsoProm [8], derived from [15]'s Prosogram.

In future works, we will develop a procedure for automatic detection of segmental lengthening resulting from a hesitance. For the moment, it must be pre-identified manually. The robustness of the tool on corpora of more varied genres also remains to be checked, and its performances must be compared to the competing tools mentioned *supra*.

ANALOR can be downloaded from: http://www.lattice.cnrs.fr/Analor.html.

Sources are in free access.

5. Acknowledgments

This publication was supported by the FNS, subsidy n°100012-113726/1, "La structure interne des périodes", hosted in Neuchâtel University, Switzerland.

6. References

[1] Avanzi, M.; Lacheret-Dujour, A.; Victorri, B., 2008, to appear. Analor: un outil d'aide pour la modélisation de l'interface prosodie – grammaire. Cahiers du Cerlico, 20.

[2] Beckman, M., Hirschberg, J.; Shattuck-Hufnagel, S., 2006. The Original ToBI System and the Evolution of the ToBI Framework. In Prosodic models and transcription: Towards prosodic typology, J. Sun-Ah (ed.). Oxford: University Press. Chap. II, 9-54.

[3] Boersma, P; Weenink, D., 2007. Praat: Doing Phonetics by Computer (Version 4.6). www.praat.org

[4] Buhmann, J.; Caspers, J.; van Heuven, V; Hoekstra, H.; Martens, J.-P.; Swerts, M., 2002. Annotation of Prominent Words, Prosodic Boundaries and Segmental Lengthening by Non Expert Transcribers in the Spoken Dutch Corpus. In Proceedings of the 2nd International Conference on Language Resources and evaluation (LREC 2002), In Rodriguez, M.C. & S. Araujo (eds). Paris: ELRA. 779-785.

[5] Campione, E., 2001. Étiquetage prosodique semi-automatique de corpus oraux: algorithmes et méthodologie. Thèse de doctorat. Aix-en-Provence: Université de Provence.

[6] Delais-Roussarie, E.; Post, B.; Portes, C., 2006. Annotation prosodique et typologie. Travaux Interdisciplinaires du Laboratoire Parole et Langage, 25. 61-95.

[7] Goldman, J.-Ph. 2007. EasyAlign: a Semi-Automatic Phonetic Alignment Tool under Praat. Available at http://latcui.unige.ch/phonetique

[8] Goldman, J.-P.; Avanzi, M.; Lacheret-Dujour, A.; Simon, A.-C.; Auchlin, A., 2007. A Methodology for the Automatic Detection of Perceived Prominent Syllables in Spoken French. In Proceedings of Interspeech'07, Antwerp, Belgium, August 27-31. 98-101.

[9] Grabe, E; Post, B.; Nolan, F., 2001. Modelling Intonative Variation in English: The IViE System. In Prosody 2000, S. Puppel & G. Demenko (eds). Poznan: Adam Mickiewicz University. 51-58.

[10] Hirst, D.; Espesser, R., 1993. Automatic Modelling of Fundamental Frequency Using a Quadratic Spline Function. Travaux de l'Institut de Phonétique d'Aix-en-Provence, 15. 75-85

[11] Lacheret-Dujour, A.; Victorri, B., 2002. La période intonative comme unité d'analyse pour l'étude du français parlé : modélisation prosodique et enjeux linguistiques ». In Verbum, 24/1-2. 55-73.

[12] Lacheret-Dujour, A., 2003. La prosodie des circonstants en français parlé. Leuven/Paris: Peeters.

[13] Martin, Ph., 2006. La transcription des proéminences accentuelles : mission impossible ?. *Bulletin PFC*, 6. 81-87.

[14] Mertens, P., 1991. Local Prominence of Acoustic and Psychoacoustic Functions and Perceived Stress in French. Proc. 12th ICPhS, vol. 3. 218-221.

[15] Mertens, P., 2004. The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. In Proceedings of Speech Prosody 2004, B. Bel & I. Marlien (eds). Nara (Japan), 23-26 March.

[16] Pickering B.; Williams B.; Knowles G., 1996. Analysis of Transcribers Differences in the SEC. In Knowles, G. Wichmann, A. & Alderson, P. (eds), Working with Speech: perspectives and research into the Lancaster/IBM Spoken English Corpus. London/New-York: Longman. 59-105.

[17] Poiré, P., 2006. La perception des proéminences et le codage prosodique. Bulletin PFC, 6. 69-79.

[18] Post, B.; Delais-Roussarie, E.; Simon, A-.C., 2006. IVTS, un système de transcription pour la variation prosodique. Bulletin PFC, 6. 51-68.

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

[20] Terken, J., 1991, Fundamental Frequency and Perceived Prominence. *Journal of the Acoustical Society of America*, 89. 1768-1776.