Prominence detection without syllabic segmentation

Philippe Martin

UFR Linguistique, Université Paris Diderot, Paris, France
philippe.martin@linguist.jussieu.fr

Abstract

Detection of prominence, whether automatically or manually through perception tests, is pivotal in the interpretation of data in a prosodic theoretical framework. This is particularly true for French, where phonologically stressable syllables are not necessarily stressed. To assert a prominence character to syllables is mandatory to evaluate prosodic theories, especially those which predict the phonetic features of melodic contours (rise, fall, height, etc.) located on those syllables.

Some algorithms are already available to detect prominent syllables automatically, but most involve a precise segmentation of speech into syllables, vowels and consonants, a task which generally requires a reasonable good quality of recording, exempt from background noise and echo. In order to avoid the problematic segmentation into phonetic units, we propose here an algorithm for prominence detection operating differently and based on readily available phonetic properties of speech, at the exeption of spectral properties.

1. Introduction

Detection of prominence, whether automatically or manually through perception tests, is pivotal in the interpretation of data in a prosodic theoretical framework. This is particularly true for French, where phonologically stressable syllables are not necessarily stressed. To assert a prominence character to syllables is mandatory to evaluate prosodic theories, especially those which predict the phonetic features of melodic contours (rise, fall, height, etc.) located on those syllables.

It is well known that the characterization of prominence by human operators is particularly difficult in French [1], [8] to the point that it has been argued that using a speech processing algorithm constitutes a better solution as its use cannot be influenced by other factors external to the actual acoustic properties involved [2]. It would also relieve human operators from difficult decisions, particularly painful for large corpora. Recently, automatic approaches have been implemented in various speech analysis programs such as Praat [5], [3] or WinPitch [2].

2. Existing methods

To date, one of the most popular method [5] is based on the "prosogram" which proceeds by stylizing graphically melodic variations according to a glissando value. If the melodic contour, described by its duration and variation of Fo, is below a threshold, the contour is replaced by a straight line at 2/3 of the contour Fo variation. If it is above the threshold, it will appear as a linear variation linking the starting and ending point of the contour. The glissando threshold has been established first by Rossi [6] on pure tones and synthetic vowel.

This representation implicitly integrates a simplified model of perception, combining melodic contours prominence duration and change of fundamental frequency. From the prosogram curves, one can then apply some criteria to detect prominence among syllables, either by visual inspection or automatically.

Despite its wide use, this glissando representation of melodic curves presents some important drawbacks, e.g.:

1. The glissando threshold has an adjustable parameter, to take into account the possible variations occurring in the perception tests were this threshold was established;
2. As the representation is linear, there is no provision for non linear melodic contours, convex, concave, bell shaped, frequently observed in idiosyncratic or regional variations of French;
3. The glissando parameter is supposed to be adjusted for the entire analyzed recording, not considering possible variations during the speaker performance;
4. The whole process is very sensitive to the quality of the segmentation into syllables and vowels. Automatic segmentation must be carefully checked visually, a time consuming process for spontaneous speech recordings.

Some implementations, e.g. [7] take care of drawback No 3 by adjusting a prominence detection parameter, and by implementing some local criteria taking the difference of syllable prominence. Likewise, some realizations bypass the syllabic segmentation [9] to avoid their pitfalls.

3. Principle of operation

Duration

Segmentation into pseudo-syllables, i.e. an approximation of syllable obtained without explicit segmentation into consonants and vowels, is usually done either by building an histogram of the intensity bimodal distribution and retaining the segments with the highest intensity mode, or selecting intensity peaks displaying a sufficient fall (e.g. -6 dB) on both sides [10]. The latter approach was chosen here: the duration of the segment above this intensity threshold is retained as duration of the pseudo-syllable. The threshold can be computed as well from an histogram of regularly time sampled intensity values.

Fundamental frequency

The fundamental frequency movement corresponding to the segments retained above the intensity threshold is then defined as follows (Fig. 1): starting from the Fo at the level of the intensity peak, the Fo curve is expanded to the left and to the right until either a pseudo-syllable boundary is reached, or when Fo is null or invalid. This avoids the retaining of erroneous or missing values frequently found at the beginning or at the end of voiced segments, as the most reliable Fo

values are found for large intensity values. If no Fo is found at the intensity peak, the pseudo-syllable is discarded.

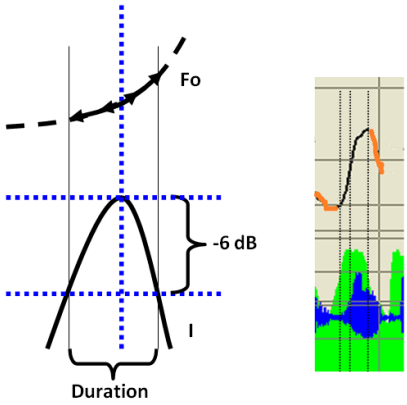

Fig. 1. Pseudo-syllable duration and Fo segment

To take into account non linear variations of Fo, the Fo curve is "straightened" by taking values every 50 ms and adding the absolute value of Fo variation, so that bell shape curves are not represented by the Fo values taken at their end points (Fig. 2). This would better approximate the perception effect of non linear variations, often ignored in other implementations.

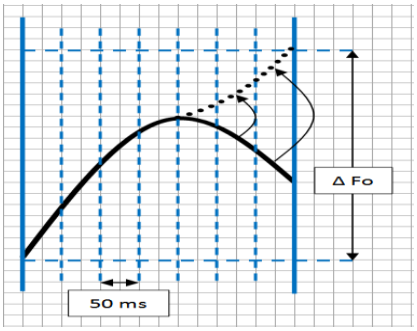

Fig. 2. "unwrapping" of a bell shaped Fo curve

Each pseudo-syllable is represented by 2 values: -6dB duration and Fo "corrected" fundamental frequency variation. The intensity is ignored at this point.

Prominence window

A relative prominence index is then evaluated for a shifting "prominence window" of n pseudo-syllables, with n = 7 as default value. This index is computed separately for duration and Fo variation, and could be easily extended to other parameter such as intensity differences between retained pseudo-syllables. For each prominence window, pseudo-syllables are ranked from n to 1 (see an example on Fig. 3, with n = 7 and duration ranking). The prominence window is then shifted by one pseudo-syllable, and the pseudo-syllables are ranked again. The process is repeated until the end of the sequence of pseudo-syllables (In order to be displayed at a larger size, Fig. 3 and 4 are sown at the end of this paper).

Final prominence ranking

Rankings of each pseudo-syllable are then added and normalized, to take into account the partial ranking done at the beginning and at the end of the sequence (see table on Fig. 3.). Final ranking of all pseudo-syllables gives the retained prominences (last row of table Fig. 3). Likewise, Fig. 4 shows a similar process this time done on Fo change values (see end of the paper).

Fig. 5 and 6 display the pseudo-syllables with relative prominence index above a 5.6 threshold for duration (in blue), and above 5.8 threshold for Fo changes (in red). These thresholds can be chosen arbitrary, or in such a way to return a predetermined number of prominent syllables.

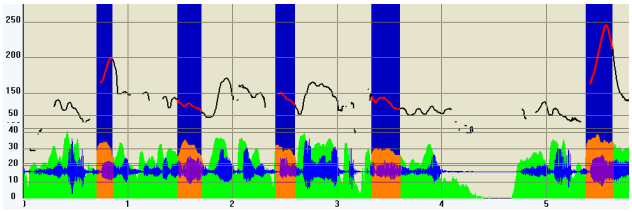

Fig. 5. Retained 5 most prominent pseudo-syllables by duration, threshold 5.6

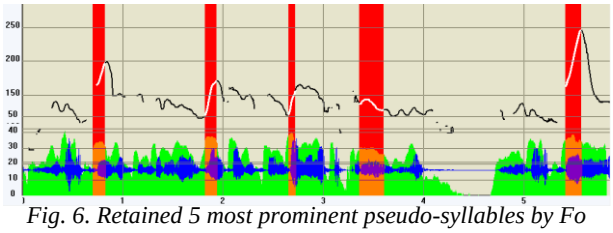

Fig. 6. Retained 5 most prominent pseudo-syllables by Fo change, threshold 5.8

Performance

If we compare the duration and Fo most prominent pseudo-syllable of our example (taken from [12]) to 1) the prominence predicted by phonological rules (thus excluding secondary or emphatic stress) from the orthographic transcription and 2) to the intuitively perceived prominence, we obtain results summarized on Table 1:

	PIED	***MOI***	***je***	***TION***	***dans***	***MEN***	***SANT***	***PIED***
lex	X	X		X		X	X	X
sec			X		X			
dur	**X**	**X**		**X**		**X**		**X**
Fo	**X**		**X**		**X**	**X**		**X**

Table 1: perceived prominent syllables, relative duration, relative Fo.

This table indicates with a X the phonological (lex row), secondary perceived stress (sec row), relative duration (dur row) and relative Fo change (Fo row).

The actual orthographic transcription of the example is:
j'y vais à pied je suis chez moi je m'conditionne dans mon appartement en me disant j'y vais à pied

It contains a total of 25 syllables, and detection based on peaks of the intensity curve detected 22 pseudo-syllables.

a) Five syllables perceived as most prominent

j'y vais à ***PIED*** *je suis chez* ***MOI*** ***je*** *m'condi****TION****ne* ***dans*** *mon apparte****MENT*** *en m'di****SANT*** *j'y vais à* ***PIED***

5 lexical stressed syllables (***PIEDS***, ***MOI***, *m'condi****TION****ne*, *di****SANT***, ***PIED***) ,2 syllables with secondary stress (***je***, ***dans***).

b) Five pseudo-syllables with largest duration

PIED, ***MOI***, *m'condi****TION****ne*, *apparte****MENT***, ***PIED***.

c) Five pseudo-syllables with largest change of Fo

PIED, ***je***, ***dans***, *apparte****MENT***, ***PIED***.

By union of duration and Fo detected pseudo-syllables, we only miss the manually perceived prominence on *m'di****SANT***.

4. Underlying principles

Prominence window

It is well known that prominence pertains to a relative value in a certain time frame in the sentence or in the discourse. The listener memory is not capable to retain and rank more than a limited number of occurrences of syllables duration, intensity and Fo characteristics, and operates in a limited time window. At a normal speech rate, the value of 7 syllables has been suggested [11], which justifies the default value chosen for the prominence window. Other values or a time span can easily be used instead. Shifting the prominence window ensures the averaging of the relative prominence on a sufficient time span. This corresponds roughly to the manual procedure of [4], where the manual operators allow themselves the possibility to go left and right around the tested syllable to better establish its prominence character.

Separate duration and Fo prominence

Perceived prominence results from a complex integration of acoustic parameters such as duration and Fo variation, as well as phonetic events such as the presence of vowels followed by a voiced consonant, and also linguistic events such as the detection of a stressable class such as a verb, noun, adverb or adjective. As the method presented here is not using phonetic and linguistic information, the prominence index for duration and Fo are used separately and not combined as in other methods. This enforces the

Relative prominence

Absolute values of duration and Fo changes were often wrongly interpreted from stylization of Fo contours [5] but are now seldom used, as relative values are now used instead. This principle is used here, retaining indexes above a certain threshold. If information about the linguistic content is available, such as an orthographic or phonetic transcription, the 7 syllables rule can be applied, giving the minimum number of stressed syllables and thus the minimum number of syllabic pseudo-syllables.

Prominence is not stress!

There are cases where automatic detection of prominence may be misleading. A typical example is given by major continuation contours in Italian, which are often realized with a somewhat flat or gently falling Fo contour on the stressed syllable and a sharply rising contour on the final post-stressed syllable. An example is given Fig. 7, with the sentence *alcuni di questi edifici prefabbricati si sono revelati pericolosi*.

In this example, the phonologically stressed syllable *ca* in *prefabbri****ca****ti* may be judged as less prominent than the final non stressed syllable *ti* on the base of Fo variation, whereas it could appear as more prominent than the post-stressed syllable on the base of its duration.

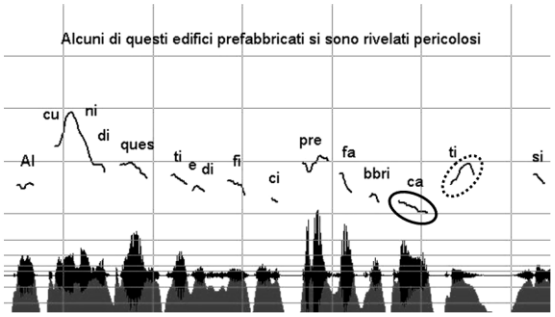

Fig. 7. A phonologically possibly misleading detection of syllabic prominence:

5. Conclusions

As discussed in detail in [1], expert detection of prominent syllables results from a complex interaction of phonetic, phonological, lexical, syntactic and semantic knowledge. To obtain similar performances from an automatic process would require access to the same multi levels information, which is certainly not the case for algorithms using only acoustic information. It follows that automatic detection of prominent syllables, or pseudo-syllables, is essentially a tool helping the linguist to access and process pertinent prosodic data more rapidly and efficiently.

The process presented here does not require a pre-existing segmentation into syllables, consonants and vowels, as it operates only from the intensity and fundamental frequency curves. Pseudo-syllables are obtained from selected intensity peaks, and, using a shifting "prominence window" of 7 pseudo-syllables, corresponding roughly to the listener short term memory, a prominence index is evaluated separately for duration and Fo change values of each pseudo-syllable.

Preliminary results are very encouraging, as they open the path to computer assisted detection of prominence, where the operator keeps the final decision. The method can easily incorporate other parameters such as pause duration after pseudo-syllables, or intensity differences between consecutive pseudo-syllables, etc.

For the reasons exposed above, we do not believe that any algorithm could identify syllabic prominence without some false alarms or missing detections. Such tools must be considered as an aid to decision for linguistically and phonetically trained operators, allowing faster and possibly more reliable analysis of large speech corpora.

6. References

[1] Martin, Philippe (2005) La transcription des proéminences accentuelles : mission impossible ? *Revue PFC*, septembre 2005.

[2] Avanzi, Mathieu et Philippe Martin (2007) Un outil pour la détection automatique des proéminences accentuelles dans les corpus oraux, *XXV CILPR 2007*, Innsbruck, 3-8 septembre 2007.

[3] Goldman, Jean-Philippe (2007) *EasyAligner: a semi-automatic phonetic alignment tool under Praat*. http://latcui.unige.ch/phonetique.

[4] Avanzi, Mathieu, Jean-Philippe Goldman, Anne Lacheret-Dujour, Anne-Catherine Simon & Antoine Auchlin (2007) Méthodologie et algorithmes pour la détection automatique des syllabes proéminentes dans les corpus de français parlé, *Cahiers of French Language Studies*, 13/2, 2-30.

[5] Mertens, Piet (2004) Un outil pour la transcription de la prosodie dans les corpus oraux *Traitement Automatique des langues* 45 (2) 109-130.
[6] Rossi, Mario (1978) Interaction of intensity glides and frequency glissandos, Language and Speech (21) 4, 384-396.
[7] Obin, Nicolas, Goldman, J.-P., Avanzi, M. & Lacheret-Dujour, A. (2008) Comparaison de trois outils de détection automatique des proéminences en français parlé *Actes des 27èmes journées d'étude sur la parole (JEP 08)*, Avignon, 8-13 juin 2008. non paginé
[8] Goldman, Jean-Philippe & Avanzi, M. (2007) Vers un algorithme de détection (semi-)automatique des proéminences en français parlé *Actes des 7èmes Rencontres des Jeunes Chercheurs sur la Parole (RJCP07)*, Paris, 05-06 juillet 2007, 84-87.
[9] Mertens, Piet & Christophe d'Alessandro (1995) *Pitch contour stylization using a tonal perception model*, Proc. 13th International Congress of Phonetic Sciences, Vol. 4, 228-231.
[10] Martin, Philippe (1979) Automatic Location of Stressed Syllables in French *Proceedings of the International Congress of Phonetics*, Miami, 1091-1094.
[11] Wioland, François (1985) *Les structures rythmiques du français*, Slatkine-Champion, Paris.
[12] Branca-Rosoff, Sonia (2009) Corpus de Français Parlé Parisien (CFPP2000) http://ed268.univ-paris3.fr/syled/ressources/Corpus-Parole-Paris-PIII/

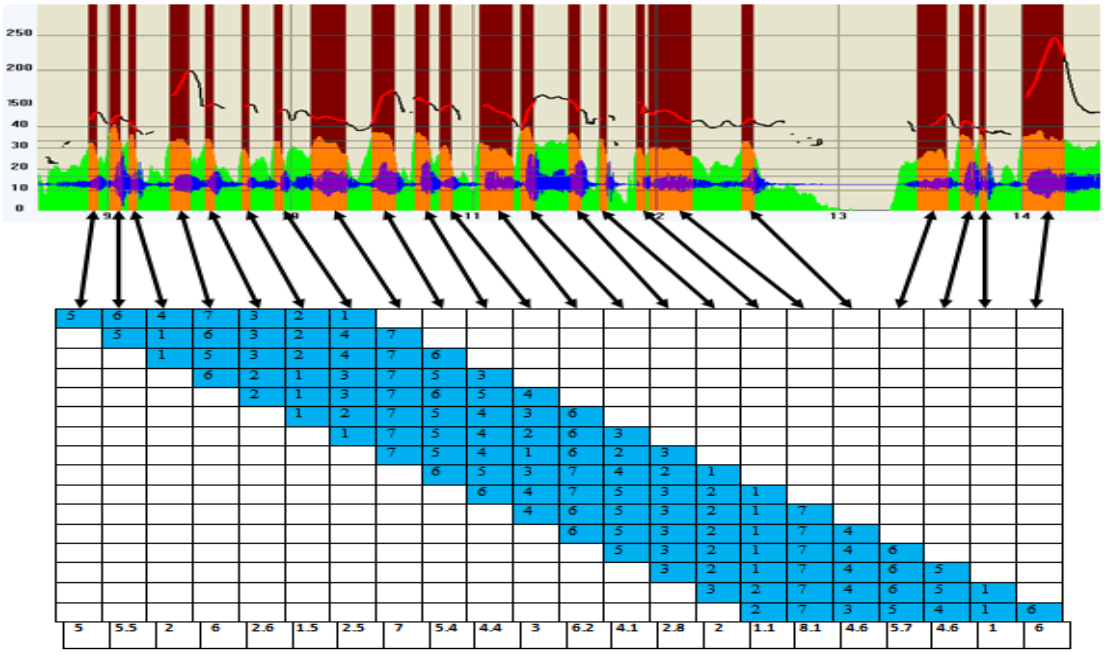

Fig. 3. Relative duration prominence computed on a shifting window of 7 pseudo-syllables. Adding the relative values gives the overall relative prominence on the whole recording.

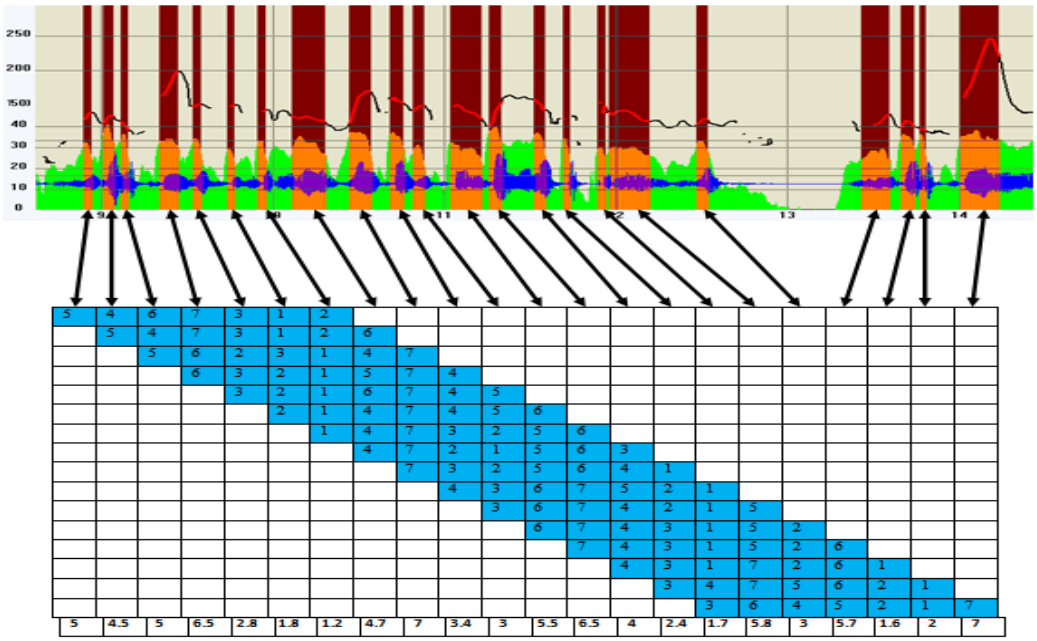

Fig. 4. Relative Fo prominence computed on a shifting window of 7 pseudo-syllables. Adding the relative values gives the overall relative prominence on the whole recording.