A Continuous Prominence Score Based on Acoustic Features

Jean-Philippe Goldman¹, Mathieu Avanzi², Antoine Auchlin¹, Anne Catherine Simon³

¹Department of Linguistics, University of Geneva, Switzerland
²Chaire de linguistique française, Université de Neuchâtel, Switzerland
³Institut Langage & Communication, Centre VALIBEL, UCLouvain, Belgium

Abstract
Up to now, prominence detection has mainly been considered a binary matter, a syllable or a word being considered as prosodically prominent or not. This contribution aims at developing an automatic detection procedure of gradual prominence. Based on 4 prosodic parameters (relative duration, relative f0, f0 movement and pause duration), the system provides each syllable with a gradual score of prominence ranging from 0 (non-prominent syllable) to 4 (extra-prominent syllable). The automatic detection (ProsoProm) relies on a manually annotated corpus (18 minutes, or 3669 syllables, of speech annotated by three experts) and is cumulative (the relative weight of each parameter is taken into account in order to compute a global score for each syllable). The discussion of the results includes a qualitative analysis of misses and false detections. The agreement between automatic and (median) human annotation reaches a Kappa score of 0.8.

Index Terms: prosody, speech, prominence detection, cumulative prominence, automatic prosodic analysis, expert vs automatic prosodic annotation.

1. Introduction
Prominence studies are of great interest in the speech community as the recent wealth of research dealing with this topic attests it (see [1][2] among others, for a recent review). In this context, many algorithms dealing with prominence detection based on prosodic features only, or taking into account morpho-syntactic information, have been proposed, with a very satisfying score of correct identification when compared to a manual/human annotation. In the studies presenting the existing algorithms dealing with English [3], German [4], Italian [5], French [6][7] or with many languages at the same time [8], prominence is viewed as a binary object: a syllable or a word is considered as prominent or as non-prominent by virtue of a certain number of acoustical parameters. Nevertheless, as it has been shown by scholars dealing with perception of prominence by naïve or expert listeners [9][10], prominence is not binary in nature: a linguistic item can be perceived as more or less prominent in the speech flow [11]. To our knowledge, little work has been done about the modeling of gradual prominence perception and detection (see for works on pitch accent languages [4] or [12]).

Regarding French, the only existing algorithm was provided by [7]. Based on the hypothesis according to which, with a great number of acoustic parameters involved in the identification of prominence, the more the fixed thresholds are exceeded, the more strongly the prominence is perceived, the algorithm results from a computation of the values of silent pause, relative duration and pitch averages, which are considered as traditional parameters involved in the perception of prominence in French [13][14]. In this paper, we present the methodology we followed and the first results we obtained by comparing a corpus manually labeled for continuous prominence by three experts and automatically processed for gradual prominence detection.

2. Material

2.1. Data
The data we used for this pilot study are extracted from the PFC database [15]. 3 Parisian speakers (2 male, 1 female) were first asked to read a text including 398 words phrased in 22 sentences (mean duration: 2'24''). They were then recorded in a free interview task where they were asked to make conversation about familiar topics. A minimum of 200 seconds of monologue speech was extracted per speaker from these spontaneous recordings. In all, the corpus is 18'41'' long.

Table 1. Speech material description

<table>
<thead>
<tr>
<th>Style</th>
<th>Duration</th>
<th>File</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>7'15''</td>
<td>File-1a</td>
<td>2'44''</td>
</tr>
<tr>
<td></td>
<td></td>
<td>File-1b</td>
<td>2'24''</td>
</tr>
<tr>
<td></td>
<td></td>
<td>File-1c</td>
<td>2'05''</td>
</tr>
<tr>
<td>Conversation</td>
<td>11'28''</td>
<td>File-2a</td>
<td>3'17''</td>
</tr>
<tr>
<td></td>
<td></td>
<td>File-2b</td>
<td>4'51''</td>
</tr>
<tr>
<td></td>
<td></td>
<td>File-2c</td>
<td>3'20''</td>
</tr>
</tbody>
</table>

2.2. Manual Annotation
For each of the six audio files, three annotators (three authors of the paper) received the associated Praat TextGrid file [16], which provided a 3-layer segmentation structure with a phone, syllable and word tiers. These tiers were obtained automatically thanks to EasyAlign tool [17], 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 singled out using specific symbols. A fourth empty tier, duplicated from the syllabic tier but containing only pauses and disfluency markers, was filled in by the three annotators independently. The coding methodology is the same as the one proposed in [7], and 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 speech spans ranging from 2 to 5 seconds. Those spans are not given the status of “breath groups”, or “prosodic units”, but correspond roughly to complete units, often followed by a pause. Annotators had then to fill in the syllabic intervals of...
the empty tier by marking symbol “4” for strong prominences, and “3” for the syllables where they hesitated between strong and weak prominences. Then, they listened again to the speech span for a maximum of three times, labeling weak prominent syllables with number “2”, and the syllables where they hesitated as to whether they were weak prominences or non-prominent with symbol “1”. They left the other intervals empty (those correspond to non-prominent syllables), and started over the operation with the next span, and so on, until the whole file was processed. 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 auditory processing. The 3 manual annotation tiers were merged into a unique one, in which each interval contains the median of the 3 corresponding intervals of each annotator. In all, the data is composed of 4301 intervals of which:

- 524 were considered as pause or hesitation by at least one annotator, and thus put aside for this study;
- 3777 were annotated.

The following tables show the distribution of the three annotators and the resulting median annotation over the 3777 syllables.

Table 2. Annotation distribution over 5 classes (from 0 to 4) for the 3 human annotators and their median in number of syllables and percent

<table>
<thead>
<tr>
<th></th>
<th>AS</th>
<th>AA</th>
<th>AM</th>
<th>Median</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2495</td>
<td>2702</td>
<td>2507</td>
<td>2568</td>
<td>68%</td>
</tr>
<tr>
<td>1</td>
<td>176</td>
<td>125</td>
<td>195</td>
<td>150</td>
<td>4%</td>
</tr>
<tr>
<td>2</td>
<td>496</td>
<td>464</td>
<td>379</td>
<td>431</td>
<td>11.5%</td>
</tr>
<tr>
<td>3</td>
<td>516</td>
<td>347</td>
<td>213</td>
<td>366</td>
<td>9.5%</td>
</tr>
<tr>
<td>4</td>
<td>192</td>
<td>236</td>
<td>495</td>
<td>262</td>
<td>7%</td>
</tr>
</tbody>
</table>

Table 2 shows that each annotator has a privileged mode (AS=3, AA=2, AM=24+4). This may be due to:

1. Different interpretation of coding guidelines, e.g. which cases deserve maximal score. A previous human coding (see [6] and [20]) involved a “post-coding mediation” among annotators, aimed at leveling inter-annotators divergences, not just getting a median out. This step was not processed here.

2. Different windowing among annotators: no strict span was given to all three annotators, who might have chosen different hearing spans, in some cases; this was not controlled by the procedure.

Finally, the median distribution should help in the choice of the probability distribution function for the final mapping from the continuous prominence degree to the integer scale. The Spearman’s rank correlation coefficient or Spearman’s ρ is used to compare the annotators’ agreement, as in the table below. The coefficients in the Median column have higher values as each annotator is closer to the median than to the other annotators.

Table 3. Spearman’s coefficient comparing the three annotators, the median

<table>
<thead>
<tr>
<th></th>
<th>AS</th>
<th>AM</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>0.809</td>
<td>0.757</td>
<td>0.893</td>
</tr>
<tr>
<td>AS</td>
<td>0.813</td>
<td>0.934</td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>0.889</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Automatic Annotation

In previous work, the former method for prominence detection was based on thresholds described in [6] and [20]. Several prosodic parameters are computed for each syllable:

- the mean pitch (in semitones)
- the duration (in seconds)
- the pitch movement (in semitones)
- the duration of the following pause (in seconds)

A relativization of the first two parameters is processed within a local context, that is, by comparing the computed syllable with the two preceding syllables and the following one, if no pauses is interposed. If at least one of the parameters is above its optimized threshold, the syllable is considered as prominent. The principle of continuous, or gradual, prominence detection relies on the same four parameters but normalizes and combines them instead of considering each parameter on its own (see [7] for the origin of the procedure). This approach pursues two goals: first, it provides each syllable with a degree of prominence (ranging from 0 to 4), and hence allows to distinguish strong vs. weak prominences. Second, the addition of several prosodic parameters into one global score allows detecting some kind of prominences that would be missed by the previous threshold-based method. These weak prominences may be realized by dispatching their strength into two or three prosodic parameters, among which none would reach any threshold.

As a first step of signal processing, a pitch detection and stylization must be realized. We give credit to a two-pass pitch detection as suggested in [18], but nevertheless proceeded to a manual verification. The pitch stylization is done by the Prosogram ([19]), which proceeds in two steps: (i) the algorithm finds the vocalic nuclei with the help of the intensity and voicing parameters within every syllable; (ii) it stylizes the intonation curve on the nucleus into a static or dynamic tone, on the basis of a perceptual glissando approach. Only 3669 syllables (out of the 3777) were assigned a tone by Prosogram. The resulting distributions of these four parameters are of various forms:
The relative pitch adopts a normal distribution, but the relative duration is more a log-normal distribution, and thus must first be log-transformed before being combined with the others. The pitch movement has a bimodal distribution. One must note that only 14% of the syllables are dynamic (i.e. non static), hence are represented here. Finally, the distribution of the duration of the following pause (concerning 8% of the syllables) is rather bimodal (representing short pause < 250ms and long pauses) and also log-normal. The parameters are normalized (their means were negligible and only the standard deviations were kept) and combined as in the following formula:

\[
promCUM = w_{F0} \frac{relF0}{3} + w_{DUR} \frac{\ln(relDUR)}{0.6} + w_{MVT} \frac{MVT}{5} + w_{PAUSE} \frac{PAUSE}{1.5}
\]

Equation 1. Cumulative model for gradual prominence where:
- promCUM represent the cumulative degree of prominence
- relF0 and relDUR, the relative mean pitch and duration
- MVT the pitch movement
- PAUSE the pause after

This cumulative approach is justified by the fact that, in our view, the articulation effort depicting the strength of prominence may be dispatched in one or more acoustic parameters. Thus, adding the efforts of every parameter should present the global importance of the prominence. Each parameter is weighed (with a value of 1 for a start). The distribution of the unique cumulated prosodic parameter is represented in Figure 2. In order to transform this parameter into a usable integer scale, a generalized sigmoid function (or S-curve) is applied. The resulting prominence degree is rounded to the nearest integer in order to be in the 0-to-4 scale.

\[
f(x) = \frac{K}{1 + e^{-\lambda(x-x_0)}}
\]

Equation 2. S-curve function where:
- K, the upper asymptote, is fixed to 4 in order to compare to the manual annotation
- \(\lambda\) is the growth rate
- \(x_0\) is the time of maximum growth.

In this figure, the 5 rectangles represent how the cumulative normally-distributed prominence degree is transformed into an 0-to-4 scale, thanks to an optimized sigmoid curve. One must note that the intensity parameter was not considered as it is very dependent on the recording settings (input level, distance to microphone). Even relative intensity might be miscalculated due to variation of dynamic gain. Nevertheless, the intensity is somehow taken in account by ProsoGram to determine the vocalic nuclei.

4. Results

4.1. First result

An exhaustive optimization research of \(\lambda\) and \(x_0\) was performed in order to maximize the Spearman coefficient between the manual and automatic annotation of the 3669 considered syllables. The first result led to \(\rho=0.685\) (with \(\lambda=3\) and \(x_0=1.3\)).

4.2. Refinements

4.2.1. Pitch movement

Several attempts were realized considering the pitch movement:
- raw movement: considering that a rising tone gives more weight to a prominence, and that a falling tone, on the contrary, lighten prominence (as used in the previous section)
- absolute movement: the falling tones are considered positively
- positive movement only: i.e. only rising movements are considered, while the falling ones are considered as flat.

The best result is obtained with the positive movement option.

<table>
<thead>
<tr>
<th>Table 4. Several strategies for considering movement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Raw movement</td>
</tr>
<tr>
<td>Absolute movement</td>
</tr>
<tr>
<td>Positive movement</td>
</tr>
</tbody>
</table>

4.2.2. Relativization span adjustment

The usual span for relativizing the syllable mean pitch and duration was 2-1 (2 syllables before and 1 syllable after), as in [6]. The following attempts explore various spans from 1-1 to 5-5. A better \(\rho\) is obtained with a 3-3 span, rising to \(\rho=0.727\)

4.2.3. Weight adjustments

Other adjustments were made for the 4 parameters. Only the weight for pause \(w_{PAUSE}=7.5\) and the relative duration \(w_{DUR}=0.85\) gave better results. All in all, the best Spearman’s \(\rho\) is 0.776 (this is equivalent to a quadratic Kappa coefficient of 0.803, testifying to a very good agreement). Comparatively, the Analor system described in [7], relying on the same procedure, provided a Spearman’s \(\rho\) of 0.694 for the same corpus (or a \(k=0.73\)). The following contingency table shows the detailed confusion between automatic and manual annotation.

4.2.4. Conclusion

In this section, we have shown that the articulation effort depicting the strength of prominence can be dispatched in one or more acoustic parameters. Thus, adding the efforts of every parameter should present the global importance of the prominence. Each parameter is weighed (with a value of 1 for a start). The distribution of the unique cumulated prosodic parameter is represented in Figure 2. In order to transform this parameter into a usable integer scale, a generalized sigmoid function (or S-curve) is applied. The resulting prominence degree is rounded to the nearest integer in order to be in the 0-to-4 scale, thanks to an optimized sigmoid curve. One must note that the intensity parameter was not considered as it is very dependent on the recording settings (input level, distance to microphone). Even relative intensity might be miscalculated due to variation of dynamic gain. Nevertheless, the intensity is somehow taken in account by ProsoGram to determine the vocalic nuclei.
4.3. Results grouped in a 2-class annotation

Another point of comparison could be done with a two-class annotation, i.e. prominent vs. non-prominent. Classes 0 and 1 were considered as a non-prominent, whereas the classes 2, 3 and 4 were marked as prominent. A Kappa score of 0.61 was computed for the former threshold method [6], whereas the presented cumulative method yields a Kappa of 0.67 (and Analor a Kappa of 0.63).

Table 6. Confusion matrix between manual and automatic annotations for 2 classes

<table>
<thead>
<tr>
<th>Manu/Auto</th>
<th>Non-Prom</th>
<th>Prom</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Prom</td>
<td>2571</td>
<td>76</td>
<td>2647</td>
</tr>
<tr>
<td>Prom</td>
<td>365</td>
<td>657</td>
<td>1022</td>
</tr>
<tr>
<td>Total</td>
<td>2936</td>
<td>733</td>
<td>3669</td>
</tr>
</tbody>
</table>

5. Discussion

A qualitative diagnostics was performed on extreme mismatches between the human and the automatic annotation machine, especially on the cases where the disagreement was by 4 classes. For the 4 misses (the machine estimated the syllable of class 0, whereas the humans annotated it as 4), a final low prominence was not detected because of the very low pitch of the nucleus and the absence of the following pause. In certain cases, human annotators detected a strong prominence without any acoustic parameter involved: in those cases, syntactic finality plays a strong part in the perception. On the contrary, the 20 cases of over-detection (i.e. octave jumps (which should lead to a more robust technique of stylization), cases of acoustically prominent syllables in grammatical positions where prominence is not expected (in those cases, the human is “deaf” to prominence because the syllable is phonologically not stressable; see [21]).

6. Conclusion

We presented an original technique for automatic gradual prominence annotation based on a cumulative prominence score. This approach outperformed a former method based on the same acoustic parameters but on a decision strategy with thresholds, and a similar system. Two observations can be done on the methodology. First, one can argue that the evaluation was done in a closed test (i.e. using the same data for optimization and evaluation) instead of an open test or a n-fold cross validation. Second, instead of a sequential optimization of the parameters (span then weights), one should expect a more robust tuning methodology. This first attempt may be refined and leaves room for future work. Further evaluation should also be done on other speech material of various contents, in order to evaluate the robustness of the actual training. Eventually, some further investigations should be done to better detect final lowering syllables and optionally hesitation syllables.

7. Acknowledgements

This work is part of the Swiss-FNS project “Prosodic and linguistic characterization of speaking styles” (fond n°100012-134818) and of the Belgian FNRS project “Périphérie gauche des unités de discours” (convention n°2.4524.11 - 17468261).

8. References