A Methodology For The Automatic Detection Of Perceived Prominent Syllables In Spoken French

J.-Ph. Goldman[1], M. Avanzi[2], A.-C. Simon[3], A. Lacheret[4], A. Auchlin[1]

1 U. of Geneva, Switzerland, 2 U. of Neuchâtel, Switzerland, 3 U. of Louvain-la-Neuve, Belgium, 4 U. Paris-X, France

goldman@lettres.unige.ch, mathieu.avanzi@unine.ch, anne-catherine.simon@uclouvain.be, anne@lacheret.com, auchlin@lettres.unige.ch

Abstract

Prosodic transcription of spoken corpora relies mainly on the identification of perceived prominence. However, the manual annotation of prominent phenomena is extremely time-consuming, and varies greatly from one expert to another. Automating this procedure would be of great importance. In this study, we present the first results of a methodology aiming at an automatic detection of prominence syllables. It is based on 1. a spontaneous French corpus that has been manually annotated according to a strict methodology and 2. some acoustic prosodic parameters, shown to be corpus-independent, that are used to detect prominent syllables. Some automatic tools, used to handle large corpora, are also described.

Index Terms: prosody, prominence, automatic detection

1. Introduction

In the area of large speech corpora, there is a definite need for a commonly accepted prosodic annotation for easy data exchange and comparison. The key to this annotation is prominence detection, as prosodic structure is built up and interpreted around the perceived saliencies. This difficult task is still a problem for human annotators as well as automatic approaches. In [1]'s experiment, based on perceptive identification of prominences, seven prosodic experts were asked to judge each syllable of a 3-minute spontaneous speech recording of a male speaker in French. Among the 165 uttered syllables, the proportion of syllables marked as prominent varied from 19% to 49%. This large difference shows that humans, even if expert, do not share the same notion of prominence.

In [2], a phonetic analysis was conducted on [1]'s material to estimate the correlation between prosodic acoustic parameters, namely pitch, intensity and duration for the perceived prominent objects. The results showed the complexity of these acoustic correlates. As a conclusion, the authors suggested further investigations to establish a reasoned parameter inventory and the fact that prominence annotation should always use some automatic approach. These automatic tools should take into account not only F0 (best perceived parameter) but also duration, as a local rate, see [3], and intensity. This latter parameter is often left apart whereas perception tests with speech synthesis showed that it should not be neglected. Finally, it is rather reasonable to study them simultaneously, and in particular, their correlations.

On the basis of these observations as well as on experimental phonetics hypotheses about perceptual threshold of accent (as in [4]), we suggest here an automatic method for prominence detection. In practice, accentual labelling does not rely on a structural feature of the word or word group such as lexical stress [5] [6] [7] but on a neutral phonetic definition of prominence, as a perceptual salience within a background speech. The main advantage of this approach is to be independent of any theoretical framework. The automatic algorithm relies on basic relative acoustic parameters (see also [8] for some review of automatic prominence detection). To tune this algorithm and evaluate it, a corpus was manually annotated with prominence according to a robust methodology as in [9], described in section 3, in which two annotators independently label the whole corpus, and then solve difference together. See also [10]'s methodology on Spoken Dutch Corpus prominence annotation. Our 18-minute corpus of spontaneous speech is described in section 2. The manual annotation was done according to a strict methodology. The final part explains how the whole process automatically detects the prominences and evaluates this approach.

2. Material

The data set used in the experiments consists of 6 recordings, mainly monologues. Two discourse situations are represented:

- **"Itineraries"** 4 recordings (2 male, 2 female speakers) in which people on the street were asked directions in a town. The duration varies from 46 seconds to 203 seconds. The total duration is 8 minutes
- **"Radio interviews**" two radio interviews (1 male, 1 female) which are each 5 minutes long.

The whole duration is about 18 minutes (1103 seconds). From the recordings and the corresponding orthographic transcription, a semi-automatic alignment was done with EasyAligner [11]. This tool produces a multi-tier segmentation (in Praat format [12]) with phones, syllables and orthographic words. The performances of this HTK-based [13] segmentation tool compare to human segmentation, as an evaluation showed that the proportion of phone boundary differences between human and machine that are less than 20ms, is over 80 %, which is comparable to humam-human agreement. This semi-automatic process requires only 3 to 5 times the duration of the corpus. Data were also *a posteriori* checked by hand.

3. Auditory detection of prominences

The first step of the study requires two expert phoneticians identifying prominent syllables in the corpus. In order to avoid the lack of reliability noted in similar identification task experiences [1], an explicit coding procedure was built up for the annotators with a robust methodology with two goals: 1. ensure a consensus between the annotators for the prominences and 2. exclude unwanted parts of the recording.

Annotation of spontaneous corpora run into the problems

that studies on read aloud speech would otherwise not encounter. Usual phenomena within spontaneous speech recording have to be specifically annotated to be excluded from the analysis or processed apart. If not, they would interfere with prominence annotation, and, in our study, weaken the significance of the results. Some of them are contextual like simultaneous speech, non-speech and noises. Recording never being perfect, some parts should be simply excluded from analysis if they are degraded speech. Some other are linked to speech production like hesitations, false starts, or interruptions.

One typical example of the latter type is very long syllables due to **hesitation**. Some human annotators, whose prominence judgment is based on acoustics, i.e. a syllable is prominent if it stands out from the surrounding ones, mark them as prominent. Some others, whose notion of prominence is linked to accentuation, consider that hesitation cannot be accented in French and thus cannot be prominent. For that reason, and because hesitation is marked acoustically (lengthened syllables), it should not be used as training data for our purposes. Another example is **post-tonic syllables**, i.e. syllables occurring after a final accent but within the same accentual group. Their realization may depend on the accent itself. In other words, they do not have accentuation autonomy and may be interpreted variously by the annotators. Finally, some occurrences of **disrupted syntactic structure** may not follow a canonical accentuation because of the disruption. For that reason, annotator's judgment may vary.

All of these events cannot be currently automatically detected with reliability and thus have to be marked manually. This is the **delivery tier**. It is meant to handle all unexploitable or uncertain parts of speech in order to exclude them from the analysis. Annotators were asked to supply this minimal (para-)linguistic information on syllables carrying typical phenomena of speech delivery, in addition to prominent/non-prominent status of each syllable.

Table 1 shows the various possible symbols for the double annotation: 1.prominence and 2.speech delivery.

Table 1 Annotation symbols

1. Prominence labelling	
P	strongly prominent syllable
p	weakly prominent syllable
o	non prominent syllable
2. Speech delivery labelling	
z	hesitation (lengthening)
@	post-tonic syllabic schwa (as in "c'est dinguE" as is [sEde~g@])
$	unaccented post-tonic syllables
!	interruption
*	breath
%	junk or inaudible or unexploitable part of the corpus (noise, simult., laugh, cough…)
_	silence (from the auto. align.)

For this manual annotation, a **strict methodology** has been established. Each annotator listens to small portions of a sound file (3.5 second fragments). Each fragment is replayed three times at most and the syllables perceived as prominent are coded accordingly with a "p" or a "P" symbol (these 2 possible degrees of prominence were merged as one in this study but will be used later). Annotation of prominences relies on auditory perception of salience, and not on the visual analysis of pitch or intensity (F0 movements, for example). In practice, each annotator starts from a text-empty tier, duplicated from the syllabic tier, and fills each interval with a prominence and/or a delivery symbol.

If a prominence in the penultimate syllable of a word is followed by a final schwa, the final syllable is coded with the symbol "@". This is for identifying prominence in final and non final word position, the final schwa having a special status in French.

The resulting tiers of the two annotators are then automatically compared and each disagreement among the annotators is marked off. Both annotators discuss them in order to agree on a final coding. A very little number of syllables were marked as belonging to a disrupted structure but could also be perceived as prominent and thus are marked with two symbols (like p! or P!).

In Table 2, a quantified distribution of prominences, valid non-prominences and non-valid prominences for the 6 sub-corpora is shown in number of syllables and in percentage. The first four are itineraries and the last two are radio interviews. One can see that the excluded part of the corpus represents about 10% of the syllables.

The last sub-corpora has a high percentage of non-valid syllables; this is due to a large quantity of simultaneous speech in the last part of the recording. Within the valid syllables, i.e. the first two columns, the proportion of prominent syllables is 27% on average.

The experiment showed that this task still requires a lot of time. Each annotator needs about 4 to 5 times the duration of the corpus for his own annotation.

Then the "agreement" session gathering both annotations may also take 5 times the real time. For comparison purposes, the inter-human agreement before discussion was as high as 93% among the valid syllables (i.e. not marked by a "delivery tier" symbol). And much of the disagreement occurred on the syllables marked by a special character of the delivery tier.

This rigorous double manual annotation of prominences and speech delivery allows us to study the former acoustically while excluding the latter. On the basis of this annotation, some studies on the specific objects, like hesitations and disruptions in discourse, can be considered.

Table 2 Syllable vital stats

	Prom	Non-prom	Non valid	Total
Iti-12	45 *30.40*	82 *55.00*	21 *14.18*	148 *100%*
Iti-14	103 *25.95*	261 *60.69*	66 *15.34*	430 *100%*
Iti-22	191 *23.29*	558 *68.04*	71 *8.65*	820 *100%*
Iti-D	105 *24.08*	293 *67.2*	38 *8.71*	436 *100%*
irtLF1r	333 *23.71*	946 *67.37*	125 *8.90*	1404 *100%*
irtWL1r	313 *26.21*	741 *62.06*	140 *11.72*	1194 *100%*
Total (n)	1090 24.59	2881 65.00	461 10.40	4432 100%

4. Automatic detection of prominences

The automatic detection of prominence is based on some

typical acoustic parameters of prosody, such as pitch, intensity and duration. This task is done in 2 successive steps: 1.nuclei segmentation and pitch stylisation and 2.computation of acoustic parameters of prosody for prominence detection.

The first step is based on an **adapted version of Prosogram** [14]. This tool has been developed for semi-automatic prosodic transcription. It computes and draws a stylized pitch contour based on a tonal perception model, and, optionally, on a phonetic segmentation. More precisely, within each syllable, the vocalic nucleus is found as the voiced portion, that "has sufficient intensity (using difference thresholds relative to the local peak)". Then, for each nucleus, the pitch is stylized in one or more segments. These segments can be stylized as flat or with a melodic slope with adjustable glissando and perceptual thresholds. The former semi-automatic phonetic and syllabic alignment will also be used.

Unfortunately, a non-negligible number of nuclei could not be found for various reasons such as the following: 1. the original Prosogram was developed for non-segmented speech, 2. the phones boundaries guessed by the automatic segmentation may not be exact, 3. the intensity parameter used for nuclei segmentation is not fully reliable as the local peak may occur in a non-voiced part of the vowel. Thus, Prosogram was slightly modified in order to ensure a maximum of stylized nuclei by allowing the nucleus to spread to the onset and the coda as well as implementing some back-off routines to force the Prosogram to find the nuclei. Some parameters like the intensity thresholds for nuclei segmentation have also been made adjustable by the user. This new flexibility of the tool opens to a future large exploration study. The pitch stylisation within Prosogram has been left as it is. As a result, the stylized nuclei increased from about 85% to more than 95% of the possible nuclei (i.e. syllables).

In the second step, two basic **acoustic prosodic parameters** are computed for each nucleus, i.e.:

1. syllable duration (in milliseconds) is preferred to nuclei duration which is too dependant on the voiced feature of onset and coda consonants
2. maximum pitch of the nuclei (in semitones)

These parameters are computed relatively to the adjacent syllables (a weigthed average of the preceding two and the following one with weights) to obtain a local rate (with no units) and a local pitch (in semitones) respectively. The pauses and the delivery tier are used to constrain the "relativisation" of the acoustic parameters, i.e. preceding and/or following syllables will not be taken into account if the current nuclei are next to an excluding symbol or a large pause. In this case, only the valid contiguous syllables will be used for comparison.

5. Results

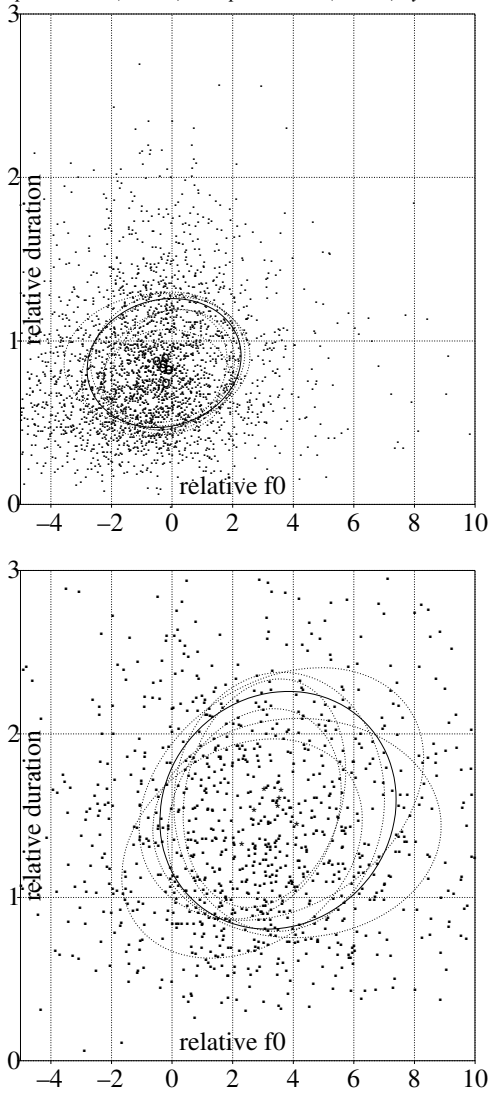
On the basis on these parameters, a discriminant analysis was conducted on the valid syllables for prominence/no-prominence status (i.e. 3971 syllables out of 4432) and showed rather promising results:

- 84.1% of overall correct classification (best sub-corpus had 87.8% and the poorest had 79.7%, both were itineraries)
- to validate the above closed-test result, a 10-th rotating leave-one-out ("jack-knife") test on the shuffled data gave also 84.1% of correct classification
- a cross-type test, i.e. the estimator is trained on radio speech and tested on itinerary interviews and vice versa, gave 83.9% on average
- a cross-gender test, i.e. trained on male speakers and tested on female speakers and vice versa, gave 84.1% .

These overall results show that the chosen acoustic parameters seem independent from the corpus, i.e. speaker gender, speech type and also speech rate (as for the fastest speaker, the rate is 4.74 syllables per second while the rate of the slowest is only 3.51 syllables per second)

In Figure 1, the 3971 syllables are represented in a plot showing the relative nucleus maximum pitch against the relative syllable duration. The manual labelling allowed us to draw the distribution for each type of syllable: the non-prominent ones above / the prominent ones below. One-sigma confidence ellipses on both types for each sub-corpus (in dotted line) and for the whole corpus (solid line) are also represented. One can notice that the two categories are rather well discriminated and that the two acoustic parameters are correlated.

Fig. 1: *Data distribution in relative pitch vs. relation syllable duration plane for manually labeled non-prominent (above) and prominent (below) syllables*



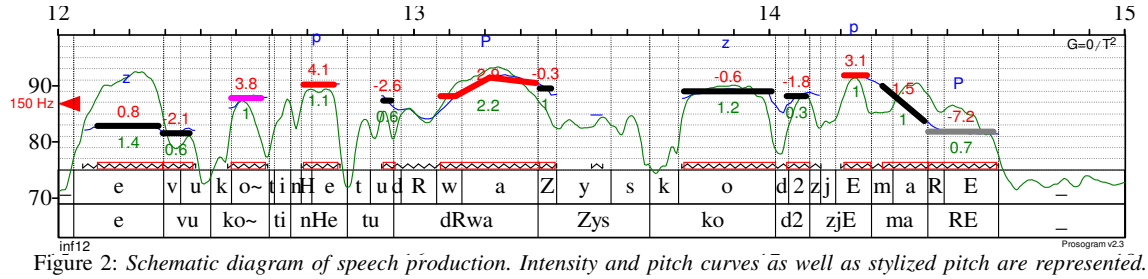


Figure 2: *Schematic diagram of speech production. Intensity and pitch curves as well as stylized pitch are represented. Acoustic parameter values and optional manual labeling appear below and above the stylized nuclei (from bottom to top: relative syllable duration, relative pitch height, pitch movement, manual labeling. Automatic segmentation also appears below. Utterance is:"et vous continuez tout droit jusqu'au prochain arrêt de tram"*

For information, a closed-test discrimination analysis on all the data including hesitations, post-tonic syllables and interruptions but leaving the junk ones out (marked with %), gave a lower score of 73.2% (compared to 84.1%). This shows that it is worth excluding the non-valid data, even if real in speech, from the statistics in order to build a robust estimator.

The original Prosogram allows nice plots of the stylised pitch together with the intensity parameter and the phonetic and syllabic segmentation, as we can see it in Figure 2. We also enhanced the plotting for research purposes by adding several informations:

- the value of the acoustic prosodic parameters for each nucleus (below is the relative duration of the syllable and above is the relative pitch of the nuclei and the cumulated absolute movement if not null)
- the delivery manual notation, i.e. prominent (with a P) or non-prominent (no sign) or others like hesitations, disrupted sequences… (i.e. symbols of Table 1)

The nuclei are coloured in red if the syllable was marked as prominent and the nucleus was also considered prominent by the automatic detection (these are the correct detections). The nuclei are coloured in grey if marked as prominent but has a low acoustic profile (a "miss") and magenta if marked as non-prominent but with high acoustic parameters (i.e. a "false alarm").

This approach allows us to work on an *ad hoc* printed version of a corpus on which speech is reduced to its significant information towards prominence, according to our hypotheses, i.e. intensity and pitch curves, pitch stylisation, phonetic and syllabic segmentation and values of the selected acoustic parameters. Every incorrect detection can *a posteriori* be diagnosed (wrong segmentation, wrong or uncertain manual labelling, exception, non-relevancy of the acoustic parameters).

6. Conclusions

We presented a promising attempt to detect prominences with segmentation and pitch stylisation tools and a strict methodology for prominence manual annotation that was used to tune the acoustically-based prominence detection tool. A discriminant analysis showed 84.1% of human-machine agreement, very comparable to the best trained system in [8]'s review. Several other studies can now be considered as a follow up of the one just presented, such as using of Prosogram's enhanced plot to diagnose errors and get rid of them, optimizing the various parameters of our tool, testing other acoustic parameters and other corpora, categorizing the prominences, adapting the methodology and the tools to other languages.

7. Acknowledgments

This publication was supported by the FNS (subside n°100012-113726/1) and FRFC (project n° 2.4523.07).

8. References

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

[2] Morel, M., Lacheret-Dujour, A., Lyche, C. and Poiré, F. "Vous avez dit proéminences ?". JEP: 183-186, 2006.

[3] Beller, G. Hueber, T., Schwarz, D. and Rodet, X. "Speech Rates in French Expressive Speech", Speech Prosody 2006 (Dresden), 2006.

[4] t'Hart, J. Collier, R. and Cohen, A. A Perceptual Study of Intonation, an Experimental-Phonetic Approach to Speech Melody, Cambridge, University Press, 1991

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

[6] Mertens, P. "Local prominence of acoustic and psychoacoustic functions and perceived stress in French", Proc. 12th ICPhS, vol. 3: 218-221, 1991.

[7] Martin, P. "La transcription des proéminences accentuelles : mission impossible ?" Bulletin PFC 6: 81-87, 2006

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

[9] Avanzi, M., Goldman, J.-Ph., Lacheret-Dujour, A. and Simon, A-.C. "Comment outiller la transcription prosodique ? Quelques pistes d'implémentation automatique", to appear in Cahiers de l'AFLS, Workshop les français parlés des corpus, 12-13 janvier Bristol, 2007.

[10] Buhman, J et al. "Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the spoken Dutch corpus", Proceedings of LREC, s.l. 779-785, 2002.

[11] Goldman, J.-Ph. "EasyAlign: a semi-automatic phonetic alignement tool under Praat" available at http://latcui.unige.ch/phonetique

[12] Boersma, P., Weenink, D., "Praat: doing phonetics by computer (Version 4.5)", www.praat.org

[13] Young, S. et al. *The HTK book*. Cambridge University.

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