ProsoReport: an automatic tool for prosodic description. Application to a radio style[1]

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

[1]Department of Linguistic, Geneva University; [2]Institut de Philologie Romane et de Linguistique Française, Neuchâtel University; [3]Département d'Études romanes, UCLouvain

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

Abstract

The paper has two goals: to present a toolbox for prosodic and phonostylistic description, and to use it for studying a specific radio style. This tool is quasi-automatic and modular. It consists of a set of Praat-based scripts like phonetic segmentation, melodic stylisation and prominence detection. It produces a phonostylistic report – called *ProsoReport* – on the basis of an audio file and optionally an orthographic transcript. The tool is used here to identify phonostylistic properties of French public radio France Info features (hence FIF; *chroniques radiophoniques*): three two-minute-long recordings are compared with a plain neutral reading of the same texts. Results confirm our initial hypotheses about FIF phonostylistic distinctive characteristics – leaving questions open to further study.

1. Introduction

Prosodic analysis deals with several dimensions (intonation, accents, rhythm, vocal quality). Each dimension needs a specific representation or transcription (intonation curve, accent labelling, rhythm pattern). On the one hand, a representation close to speech substance allows us to measure several parameters (like f0, duration, intensity…) and to correlate them with external variables. On the other hand, a symbolic transcription is restricted to functional prosodic variation by using a limited set of symbols (tones, contours, accents). Finally, a transcription can be done manually (by an expert) or (semi)automatically. We aim at the latter approach, by first getting broad measures from the substance and, from there, automating more and more the transcription task.

We present here a set of tools for prosodic analysis as simple and as robust as possible. These tools were developed within the Praat software [2] and allow to:

- segment the speech recording into phonetic segments, syllables, and words [5];
- stylise f0 curve, and provide a simplified representation corresponding to perception in syllable nuclei [13] ;
- detect automatically prominent syllables [1] ;
- add morpho-syntactic information to each syllable [8];
- compute parameters like speech duration, articulation duration, speech rate, syllable mean duration, mean and range of pitch register, proportion of prominent syllables, and present all of them in a table called *ProsoReport*.

These tools can combine together to produce a fold-out *ProsoReport*. The user decides which annotations he adds, depending on which information he provides with the recording, and which results he is looking for: mean tonal register, span, can be calculated without text to sound alignment, whereas speech rate or prominence detection presuppose syllabic segmentation and alignment.

More specifically, results obtained by a non-aligned corpus automatic analysis cannot account for such phonostylistic markers as French 'accent initial' distribution, neither for any linguistically anchored prosodic manifestation. Segmental or (at least) syllabic alignment, as well as grammatical annotation (like *functional/lexical words* distinction) is required for such *finer grained* phonostylistic description.

2. *ProsoReport*'s tools

We present here the four tools embedded in *ProsoReport*.

2.1. Syllabic segmentation and alignment

A phonetic, syllabic and lexical word alignment of speech signal can be obtained with *EasyAlign* [5] on the basis of an orthographic or a phonetic transcription, with minimal hand correction. This quasi-automatic tool is available for a growing number of languages.

2.2. F0 stylization

F0 stylization is a procedure that *simplifies* f0 contour: "by eliminating all details of the pitch contour that plays no communicative role, those perceptual properties of the pitch contour become apparent that are essential constituents patterns of the intonation patterns of the utterance." [9:29].

Prosogram [13] delivers a stylized representation of f0 variation calibrated by *perception thresholds*; it is a readable, objective, quantified, semi-automatic, perceptually motivated, theory- and language-independent prosodic transcription. It is grounded on an existing model of tonal perception, applied to vowel nuclei. It extracts stable and intense periodic parts of the signal, where f0 is generally best detected. Infra-liminary variations appear as flat lines, glissandos as one or several tilted segments (see Fig. 1).

Besides eliminating non communicative pitch micro-variations (due to co-articulation, e.g.), stylization prevents pitch detection errors, especially at voicing and devoicing.

Prosogram operates in two steps. First, it segments the f0 curve into nuclei. This can be done on a purely acoustic basis, from harmonicity peak extraction (no phonetic segmentation is required), or on the basis of a phonetic labelling, in which

[1] We thank the Swiss National Fund for Research (subside n° 100012-113726/1) and the Belgian National Fund for Research (FRFC n° 2.4523.07).

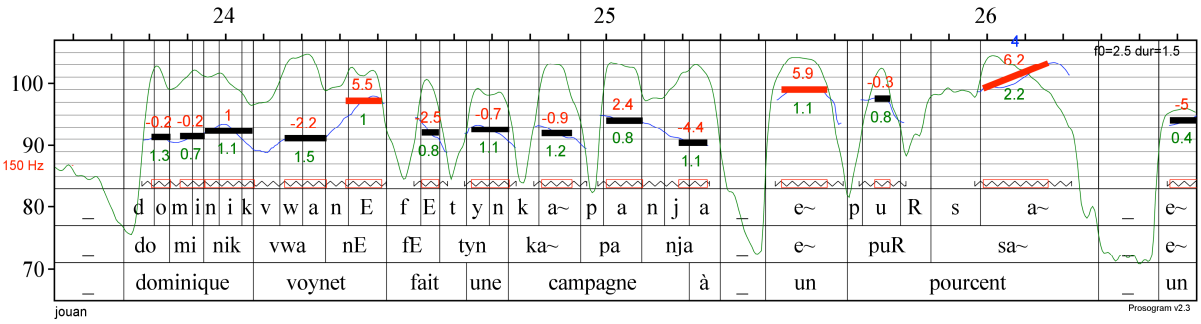

Figure 1*: Enriched Prosogram. Each syllale's nuclei is represented in bold dash with its prosodic parameters (from bottom to top: relative duration, height in ST and movement in ST). The red (gray) ones are considered as prominent tones/syllables*

nuclei are strictly constrained to vowels. We use a slightly modified version of *Prosogram* [1], which allows nuclei to be spread over the voiced part of the syllable, thus including parts of semi-vowels as well as sonorant consonants that are above intensity thresholds. Then, whatever the segmentation method, *Prosogram* transforms each nucleus pitch curve into a stylized tone.

2.3. Prominence automatic detection and stylization

Prominences play a fundamental role in accentuation and rhythm; prominence detection is therefore a requisite of many prosodic studies. Prominence may be defined as "a quantitative parameter of a syllable […] that describes markedness relative to surrounding syllables" (Portele & Heuft 1997, 63 quoted from [9]).

ProsoProm [1; 6] is a script that relies on *Prosogram*'s pitch stylization and on *EasyAlign*'s syllabic alignment for the detection of prominent syllables, based on syllable pitch and duration relatively to surrounding syllables, and on internal f0 movements.

Several information is added to the graphical output of *Prosogram*. On the stylized f0 curve, segments detected as prominent are shown in red (grey) lines; acoustic parameters are displayed for each stylized nucleus (from bottom to top: relative duration, relative height, and, for dynamic tones, intra-nucleus movement in ST).

In Figure 1, the last syllable of proper name "Voynet", as well as numeral "un", are detected as prominent due to their relative height; last syllable of "pourcent" is prominent due to its internal movement.

2.4. Morphosyntactic annotation tool (*functional vs lexical*)

A prominent syllable is potentially an accented syllable. In French, some grammatical information is needed in order to identify the accent type (primary/secondary stress [10]; final/initial stress [12], etc.), which modifies prominence values and functions.

Distinguishing lexical words from functional ones (hence clitics or not) is not easy, neither is distinguishing single words from compound-words. Our study is restricted to a gross annotation, discriminating clitics from non-clitics, following [14], [15]; and, from there, to labelling initial and final syllables of lexical words only. Part of the annotation was done manually, while fully automatic grammatical labelling is under development.

3. The *ProsoReport*

The *ProsoReport* provides a detailed prosodic report, i.e. an exhaustive set of prosodic and phonostylistic measurements for a given recording. The granularity of description depends on the user's choice to activate specific tools (see § 2). We make here a qualitative description of the *ProsoReport*. The first part depicts a basic prosodic report, presenting vital stats. Then, we show that *ProsoReport* can make filtered statistics, constrained by optional information like prominence of syllables or morpho-syntactic information.

3.1. Basic *ProsoReport*

On the basis of *EasyAlign's* and *Prosogram's* results, *ProsoReport* presents statistics on the following parameters:

Parameters		Sequences of syllables	
duration	**global**	speech duration	s. (with pauses)
		articulation duration	s. (without pauses)
		articulation ratio	% (articul./speech)
		speech rate	syll./s. (for speech)
		articulation rate	syll./s. (for articul.)
	local	syll. dur. (mean, std dev)	s.
		nuclei dur. (mean, std dev)	
		nuclei over articulation	%
f 0	**global**	mean and std dev	ST (rel. to 1 Hz)
		quantiles, interquantile range	ST
	tones	static tones ratio	%
		dynamic tones ratio	
		falling tones ratio	
		complex tones ratio	
	tone movements	mvt (dyn tones only)	ST/syll, ST/second
		mvt (all tones)	
		inter-nuclei mvt	
		melodic agitation (see text)	
intensity		mean intensity within nuclei	dB
		mean intensity outside nuclei	
		nuclei/non-nuc. intensity ratio	

3.2. Constrained *ProsoReport*

The prominence detection tool and the semi-automatic morpho-syntactic annotation tool allow getting a *ProsoReport* of a selected subset of syllables according to a specific criterion. For example, duration, melodic and intensity measures of prominent syllables can be compared to non-prominent ones. Or, word-initial syllables can be compared to word-final syllables. The two additional tiers, namely *prominence* and *initial/final syllable*, are used as a syllable filter. More precisely, if the prominence detection tool is used, *ProsoReport* gives the number and the ratio of prominent vs. non-prominent syllables, and a description of all the prosodic parameters presented as above, comparing both types of syllables.

With prominence detection tool activated:		
prominence	**prominent** syll ratio	% (of syllables)
	non-prominent syll ratio	
…and duration, f0 and intensity of **prominent** vs. **non-prom**. syllables		

In a similar way, the morpho-syntactic annotation distinguishes the lexical from the functional words, and, within the lexical words, the initial (**i**) and the final (**f**) syllables:

With morpho-syntactic annotation tool activated:		
syllable position	lexical-word **initial** syllable ratio	% (of syllables)
	lexical-word **final** syllable ratio	
…and duration, f0 and intensity of **initial** vs. **final** syllables		

Finally, these two criteria can be combined to select syllables according to their linguistic localisation and their prominence feature:

With prom. detection and morpho-syntactic annotation activated:		
syllable position and prominence	lexical-word **initial prominent** syllable ratio	% (of syllables)
	lexical-word **final prominent** syllable ratio	
…and duration, f0 and intensity of **prominent/non-prominent initial/final** syllables		

This makes available the number of prominent syllables per second (that we could associate with pace or tempo), as well as the proportion of prominent syllables at initial positions of lexical words (linked to insistent or initial accent in French). More generally, the combination of acoustic and linguistic annotation allows us to consider the development of an automatic segmentation in major prosodic units thanks to final prominent syllables.

Two limitations should be mentioned: 1. the syllable has been chosen as the main prosodic unit, thus no polysyllabic phenomena like contours can be characterized or measured yet (for example, a call contour or a focus accent over a whole word). Similarly, temporary changes of global prosodic parameters can not be detected (like low parentheticals). 2. A risk of methodology circularity has to be considered: syllables are considered as prominent if their relative height and/or duration are above defined thresholds, thus they are inevitably higher or longer than surrounding syllables. But the other non-correlated parameters, as their number or their position, are still valid for statistics.

4. Comparing radio and read speech

ProsoReport is used here to compare the phonostyle of France Info features (FIF) with read-aloud neutral speech. Some hypotheses were made in [3], and partly confirmed in [7], about features specific to FIF, especially: over-articulation; over-segmentation; melodic hyperactivity and over-exploitation of dynamics; large quantity of optional initial accents.

The radio corpus is made of three short features (a male and a female from *France Info* and a female from *France Inter*, respectively R-g, R-a, R-j), amounting to a total duration of 6'38''. The exact transcriptions of these three features were read by a female speaker. The three readings, named L-g, L-a, L-j, (L stands for Lecture) have a total duration of 7'03''. This represents 1901 syllables in the radio corpus and 1941 in the read one, despite the equality of the texts. This is explained by various phonological choices made by the speakers.

We present here some extracts of the *ProsoReport*, showing significant differences, whether they were part or not of our hypotheses, as well as non verified hypotheses.

Based on the six articulation ratios computed by *ProsoReport* as in last line of Table 1 (i.e. proportion of articulation time over speech time, the latter including pauses), a significant difference can be shown with a higher **articulation ratio** for FIF style. Deeper observations show no significant difference of the speech rate or of the articulation rate (i.e. the number of syllables per second over the total speech time, and over the articulation time respectively) but more silent pauses for the reader.

	L-j	L-a	L-g	L	R	R-j	R-a	R-g
speech time(s)	161	139	122	**141**	**128**	150	123	111
articul time(s)	129	111	100	**113**	**112**	133	111	94
speech rate	4.8	4.6	4.3	**4.57**	**4.93**	5	5.2	4.6
articul. rate	6	5.8	5.3	**5.7**	**5.63**	5.7	5.8	5.4
Artic. ratio(%)	79.2	79.8	81.8	**80.2**	**87.5**	88.8	89.8	84

Table 1: *Means of both reading (L) and radio(R) are in middle columns in bold; rates are in syll/sec.*

The mean f0 is difficult to compare as our corpus is composed of speakers of both genders. Nevertheless, a test of equality of **variance of f0** distributions by corpus over all the syllables shows a very significantly higher variance for FIF style (7.21 ST for radio vs. 5.40 ST for reading. F-test $p<0.001$). Similarly, Table 2 shows the f0 ranges for each speaker. The absolute measure (from the minimum to the maximum) seems to be unreliable as figures vary greatly even within the reader's extracts. But the more robust 5%-95% interquantile range, clearing away outliers, shows also a higher dynamics for FIF style.

range	L-j	L-a	L-g	L	R	R-j	R-a	R-g
max-min	15.8	10.9	23	**16.6**	**18.9**	23	15.4	18.2
95%-5%	8.6	8.4	8.7	**8.6**	**12.9**	13.8	12.2	12.7

Table 2: *Absolute range and smart f0 range (in ST)*

Comparing nuclei f0 dynamics also show a significant difference between radio and read speech. The former style has more dynamic tones, especially falling ones, whereas the proportion of rising tones are similar for both conditions.

	L-j	L-a	L-g	L	R	R-j	R-a	R-g
Static	89.2	85.2	85.7	**86.70**	**82.87**	82.8	79.8	86
rising	3.5	7.6	7.2	**6.10**	**6.73**	9.4	6.7	4.1
falling	7.3	7.1	7	**7.13**	**10.37**	7.8	13.5	9.8

Table 3: *Static, rising and falling tones proportion (in %)*

The so-called mean melodic movement corresponds to the melodic path covered during one second of articulation (in ST/second). This *cumulated melodic path*, due to both melodic movement within nuclei (intra-mvt) and melodic gap between nuclei (inter-mvt), is greater for FIF than read speech. The initial hypothesis of melodic agitation is associated with this greater score.

Movement	L-j	L-a	L-g	L	R	R-j	R-a	R-g
Dyn. tones	18.2	20	20	**19.4**	**25.7**	26.3	26.1	24.6
All tones	2	3	2.9	**2.6**	**4.4**	4.5	5.3	3.4
Inter-mvt	13.3	14.5	14.9	**14.2**	**17.2**	18.4	15	18.3
Agitation	15.3	17.5	17.8	**16.9**	**21.6**	22.9	20.3	21.7

Table 4: *Movement in ST of dynamic tones only, all tones, between tones and overall agitation (i.e. intra- and inter-mvts)*

The Proportion of prominent syllables greatly varies between the two conditions, in favour of FIF. Moreover, the number of prominent syllables at initial position of lexical words (i) is greater for radio, whereas, on the contrary, the final syllables (f) are equally prominent. The proportion of prominent syllables might seem overestimated. Actually, this depends on threshold settings during prominence detection. Nevertheless, comparisons between corpora are still valid.

	L-j	L-a	L-g	L	R	R-j	R-a	R-g
Prom	29.5	33.2	34.5	**32.4**	**37**	35.6	35.9	39.6
Prom/i	19.7	18.1	26.3	**21.4**	**31.4**	30.1	30.3	33.9
Prom/f	58.6	64.5	58.8	**60.6**	**59.6**	59.3	59.4	60.1

Table 5: *Proportion of prominent syllables, prominent syllables at initial and final position of lexical words (in %)*

These measurements are in the line of our hypotheses, and show that it is possible to quantitatively define differences between the two phonostyles of our corpus.

Some other measures invalidate our predictions. For instance, mean syllable durations have no difference. One reason lies in large intraspeaker variations. But the mean nucleus *duration* seems more robust as it is less responsive to intraspeaker variation, and a significant mean difference exists between FIF and read style. Radio nuclei being shorter, consonants predominate **in time** for this speaking style as shown in Table 6.

	L-j	L-a	L-g	L	Rl	R-j	R-a	R-g
syll	166	172	190	**176**	**178**	176	173	184
nuc	78	80	80	**79**	**72**	75	74	68
ratio	47	46.5	42.1	45.1	40.7	42.6	42.8	37

Table 6: *Mean syllable and nucleus duration (in ms) and mean nucleus/syllable ratio (in %)*

Comparing nuclei/non-nuclei *intensity* ratio for both speaking conditions is a less reliable measure: it is sensitive to variations in recording conditions. *Less* intensity for non-nuclei (i.e. consonantic part relative to nuclei) was found in FIF. In combination with above observations, we can conclude that FIF non-nuclei are less intense but longer. Duration clearly prevails over intensity in explaining the consonant-energy intuition.

This analysis gave some credit to our intuition on prosodic phonostylistic differenciation (here a specific public radio channel's features compared to "standard" reading). The *ProsoReport* helped in validating three hypotheses stipulating that radio style has: 1. a greater contrast for f0 variations and for a so-called "covered cumulated melodic path", confirming a greater "melodic agitation", 2. a bigger proportion of initial optional accents and 3. a greater overarticulation, partially verified by more prominent initial syllables and by longer consonants (non-nuclei) but invalidated by a lower intensity relatively to nuclei.

5. Conclusions

ProsoReport is a tool designed to track general prosodic properties; as for now, it cannot detect occasional occurrences of incident properties (singularities) - which can mark a specific phonostyle as Ch. Bally suggested (quoted in [11], [3]). It presents a global picture using many parameters which show main differences between corpora. This first attempt to build a user-friendly voice report tool is satisfactory. This intra- vs. interspeaker comparison tends to validate our measures. And some extensions are already under development, such as looking for some temporal evolution of the prosodic parameters through the recording, or, following [4], investigating in more depth spectral parameters.

6. References

[1] Avanzi, M.; Goldman, J.-P.; Lacheret-Dujour, A.; Simon, A.-C.; Auchlin, A., 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] Boersma, P.; Weenink, D., 2007. *Praat: doing phonetics by computer* (version 4.6.36), http://www.praat.org.

[3] Burger, M.; Auchlin, A., 2007. Quand le parler radio dérange : remarques sur le phono-style de France Info. In: *Le Francais parlé des médias. Actes du colloque de Stockholm 8-12 juin 2005*, Broth, M.; Forsgren, M.; Norén, C.; Sullet-Nylander, F., (eds), Stockholm: Acta Universitatis Stockholmiensis, 97-111.

[4] D'Alessandro, Ch., 2006. Voice Source Parameters and Prosodic Analysis. In *Methods in Empirical Prosody Research*, Sudhoff, S. et al. (eds), Berlin-New York: Walter de Gruyter, 63-87.

[5] Goldman, J.-P., (2007). *EasyAligner: a semi-automatic phonetic alignment tool under Praat*. Available at http://laltcui.unige.ch/phonetique/easyalign .

[6] 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.

[7] Goldman, J.-P.; Auchlin, A., 2006. Quelques observations intuitives et mesurées sur le phonostyle de France Info. Colloque international Phonologie du Français Contemporain 2006, *Approches phonologiques et prosodiques de la variation sociolinguistique: le cas du français*. Louvain-la-Neuve, 6-8 juillet 2006.

[8] Goldman, J.-P.; Auchlin, A.; Avanzi, M.; Simon, A-.C., 2007. Phonostylographe: un outil de description prosodique. Comparaison du style radiophonique et lu. *Cahiers de linguistique française* 28, 219-237.

[9] Hermes, D.J., 2006. Stylization of Pitch Contours. In *Methods in Empirical Prosody Research*, Sudhoff, S. et al. (eds). Berlin-New York: Walter de Gruyter, 29-61.

[10] Lacheret-Dujour, A.; Beaugendre, F., 1999. *La prosodie du français*. Paris: Éditions du CNRS.

[11] Léon, P., 1993. *Précis de phonostylistique. Parole et expressivité*. Paris : Nathan.

[12] Mertens, P., 1987. *L'intonation du français. De la description linguistique à la reconnaissance automatique*. Thèse de doctorat, Université de Leuven.

[13] Mertens, P., 2004. Le prosogramme: une transcription semi-automatique de la prosodie. *Cahiers de l'Institut de Linguistique de Louvain* 30/1-3, 7-25.

[14] Mertens, P., 2004. Quelques allers-retours entre la prosodie et son traitement automatique. *Le français moderne* 72 (1), 39-57.

[15] Mertens, P., 2006. A Predictive Approach to the Analysis of Intonation in Discourse in French. In *Prosody and Syntax*, Kawaguchi, Y. ; Fonagy, I. ; Moriguchi, T., (eds). Amsterdam: John Benjamins, 64-101.