ProsoDyn: a graphical representation of macroprosody for phonostylistic ambiance change detection

Jean-Philippe Goldman

Department of Linguistics, University of Geneva, Switzerland

Jean-Philippe.Goldman@unige.ch

Abstract

Following the ProsoReport, we present a new prosodic tool in Praat allowing a dynamic analysis of prosodic dimensions such as temporal, intonational and accentual parameters, to graphically represent the macro-prosodic variations in *Praat*. It helps the user to visually detect "ambiance changes" that are supposed to be significant. Two different graphical modes can be activated at the same time: parameters averaged relative to an external macro-segmentation (e.g. utterance, speaker's turns, etc.) and relative to a sliding window for a representation of local speech rate and register. Interactive buttons allow the adjustment of scales and help the user to explore data a macroprosodic level.

Index Terms: prosody analysis tool, macroprosody, ambiance change, Praat

1. Introduction

According to [1], discourse is an object that is built and develops in time. In addition, it may be considered as a subjective experience since speaker and hearer continuously synchronize to one another [2][3]. The discourse atmosphere or ambiance might be captured thanks to certain prosodic aspects. By talking about several *prosodies*, [4] tackles the idea of independent prosodic dimensions that interact with each other, as well as with the verbal chain, and with the context of situation, the interaction between those dimensions creating the discourse itself defined as complex and multimodal subjective experience .

In a sense, the *prosodies* can be associated with the notion of *phonostyle* thinking of all the prosodic components of speech that are continuously managed and modified in order to convey cues to discourse interpretation [5]. Those prosodic features can be represented by a great number of explicit prosodic acoustic parameters in domains such as

- Intonation domain: height, span and variation of pitch at various scopes, or any representation of the distribution of pitch values throughout the speech recordings, allowing the representation, for example, of changes in register.
- Temporal domain: duration of speech units such as phones, syllables, intonation and breath groups, pauses, among others, but also, one can address rhythmic questions about tempo and synchronicity.
- Intensity and voice quality domains, often neglected but very communicative dimensions.
- Phonological domain with a special focus on accent position, surface form, and alignment

A *phonostylistic picture* of a sample of speech can be captured by a collection of the many explicit parameters listed above and used in [6] and [7]. This helps to compare several *phonogenres* (through their rendering as a collection of individual *idiostyles*) as in [8] and [9].

But this global approach clearly ignores dynamic variations across a speech event, which can be progressive or, more likely, sudden, and affect one or more prosodic dimensions. As the utterance circumstances may vary during the oral production, the *phonostyle* may be affected by dynamic variations throughout the temporal development of discourse. These variations may be due to, or contextualize:

- contextual changes: a modification in the emotional state of the speaker (e.g. acceleration of tempo in a sport commentary due to a goal), new speaker, etc...
- discourse change: quotation (direct vs. reported speech press review accompanied with a register modification), emphasis, speaker turn, topic change.

Moreover, these contrasts may span over various linguistic units, as short as an emphasized word and more generally changes in prosodic ambiance can be perceptually captured at a sentence / utterance level. Global changes can of course affect much larger scopes (like the paratone, the spoken equivalent of paragraphs). Finally, beyond the differences between two consecutive ambiances, the transition from one ambiance to another could be studied by itself in terms of convergence or divergence (creating delay) between prosodic parameters [10]. This delay can be compared to transition delays between articulatory features of segmental level.

Observing the contrasts within sequences [5][11] that are created by dynamic prosodic variations, can lead to discovering convergence/divergence effects between prosodic dimensions or between prosodic and linguistic ones. These contrasts can initiate some interpretative effect in discourses such as iconicity [12][13], polyphony [14][15] and enaction of communicational roles and places [16].

The goal of this communication is to propose a graphical tool called *ProsoDyn* that represents the mentioned temporal variations of macro-prosody at various scopes, and helps the users to detect (or ideally detects automatically) the ambiance changes that would be significant at a discourse level. The objective is to provide an intermediate solution between a prosodic picture made of dozens of explicit parameters (*ProsoReport*) and usual detailed studies looking at prosodic parameters as syllabic of accentual group level with topics such as initial phrase accent or penultimate lengthening. One should mention a similar attempt in which clustering algorithms automatically detect variations in register and tempo [17].

2. Method

The methodology described here enumerates the successive tools that have been developed to achieve this approach. After segmentation at the syllabic level, some intermediate tools do pitch stylization and prominence detection and lead to a global phonostylistic picture. All these cascading preliminary steps gather enough information about the prosody of the speech recording. Some of them may be bypassed but robustness of the whole process would be altered.

2.1. Syllable segmentation

The segmentation (or alignment) is performed semi-automatically with *Easyalign* [18]. This tool is distributed as a plugin of *Praat* [19]. The result is a multi-level (or tiers) annotation providing sound-to-text alignment at phone, syllable and word level. The segmentation is done in several automatic steps with some minor manual verification and adjustments to achieve better quality: 1) a macro-segmentation is done from the orthographic transcription; 2) a grapheme-to-phoneme conversion tool produces a phonetic transcription; a manual listen-and-correct step is required to adjust the phonetic transcription to the utterances; 3) a HMM-based speech recognizer is used in forced-alignment mode to provide tiers at phoneme and word levels, and finally 4) a syllable tier is derived from the phone level.
This tool is currently available for French and Spanish, whereas other languages such as English, Italian and Portuguese are under development.

2.2. Robust two-pass pitch detection

Even if *Praat* provides numerous pitch detection algorithms, with numerous adjustable parameters, the pitch detection is always questionable, especially in noisy or multi-speaker recordings. Some other detectors claim to be more robust [20]. We give credit to the simple but robust method described in [21]. The possible errors in pitch tracking are largely avoided by estimating the pitch floor and pitch ceiling in a first pitch estimation on the basis of the distribution of pitch targets. Then these parameters are used in a second pitch detection that yields less octave errors.

2.3. Pitch Stylization

The *ProsoGram* [22]**Error! Reference source not found.** is a Praat script that provides a tonal stylization of vocalic nuclei. By using intensity and voicing parameters within each syllable, the script first determines the intense and salient part of the vocalic nucleus. Then it stylizes the F0 curve of each nucleus based on a perceptual approach and provides a simplified tonal representation of intonation corresponding to the perceived height and possibly slope of each syllable.

2.4. Detection of prominent syllables

Eventually, *Prosoprom* [23][24] detects among the syllables, the prominence on the basis of syllabic F0 height and movement as well as syllabic duration. For both F0 height and syllable duration, the parameter is relativized to the adjacent syllables (more precisely to the previous two and the following one). If one or more parameter(s) exceed(s) a predetermined threshold, the syllable is considered as prominent.

3. Prosodyn

On the basis of this robust prosodic information for each syllable, the tool *ProsoDyn* proposes a dynamic graphical representation of temporal evolution of the main prosodic parameters:

- Speech rate in syllable /second including the pauses
- F0 mean in semitones relatively to 1 Hz
- F0 range in semitones relatively to 1 Hz
- Prominence density (proportion of prominent syllables within the speech sequence by the user)

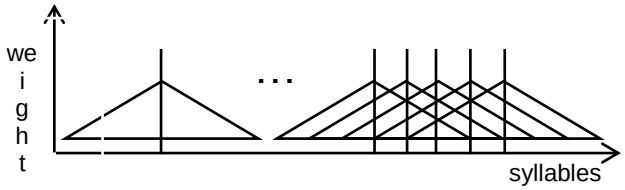

Figure 1 Syllabic synchronized sliding analysis window

To capture the macroprosodic variations of the speech data, two ways of dynamically averaging these parameters are offered. The first one uses an external macrosegmentation and the second one a sliding window.

3.1. External macrosegmentation windowing

An external *ad hoc* macro segmentation can be used to put in evidence contrasts between speech units. The segmentation into macro-units (MU) can be of any kind:

- speaker turns
- dependency unit or clause
- intonational groups, rhythmic groups
- discourse unit: direct vs. indirect discourse, speech acts…

The F0 mean and range, as well as the rate, and the prominence density are computed for each MU, stored in a table and represented graphically. From this table, any statistical hypothesis can be verified. The visual part within *ProsoDyn* represents the MU on a time axis according to their duration together with their label and the value of the prosodic parameters.

3.2. Sliding window

The sliding window (SW) technique smoothes the parameter curves by doing a weighted mean of F0, rate and prominence density. The aim is to somehow represent the evolution of local rate and register. Some settings are adjustable on the fly:

- The width of the averaging window
- The window analysis step (default: 1 syllable)

Figure 1 shows the principle of the weighted sliding window. Some attempts on various speaking styles led to a default averaging window of 15 syllables.

3.3. ProsoDyn functionalities

As can be seen on Figure 2, *ProsoDyn* has numerous buttons to adjust the time axis, the visibility of windowing techniques (MU and/or SW) and the parameters:

- the 4 buttons at top-left corner modify the main panel as they toggle the appearance of (from top to bottom):
  - the representation from the sliding window (SW), with internal buttons (+/-) to adjust the span of the SW
  - the global mean and standard deviation for the visible prosodic parameters in horizontal lines
  - the representation from the segmentation in MU
  - the segmentation in MU, like Praat TextGrids
- the 4 buttons on the left toggle the prosodic parameters: rate, F0, F0 range, prominence density
- the 4 buttons at the bottom add even more functionalities:
  - modify the scale on vertical axis with 4 modes: (from form , +/- 4sd, +/- 2 sd, min-to-max)
  - Play the full recording, the selected MU or the SW
  - Paste the current graph to Picture window
  - Exit ProsoDyn

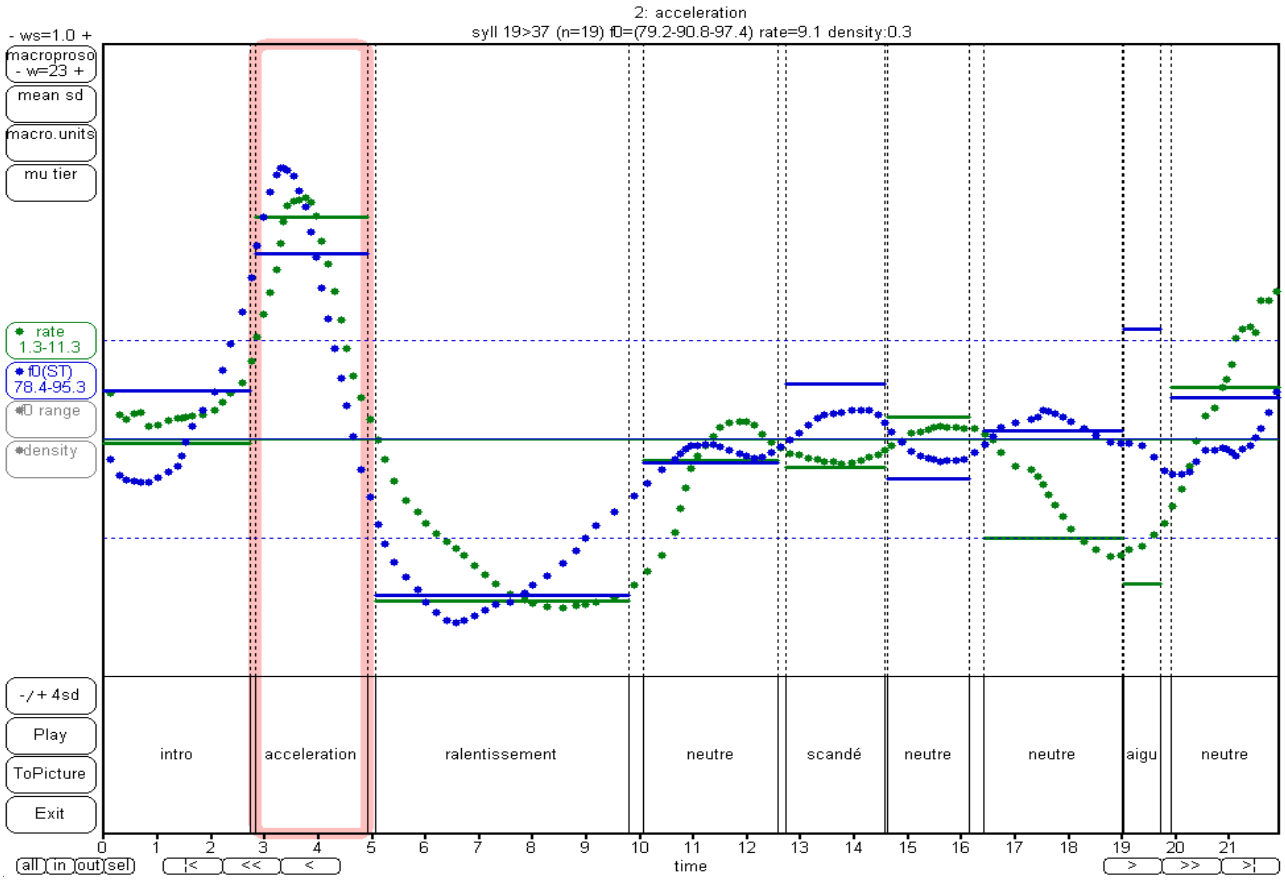

Figure 2 ProsoDyn pane with both MU (vertical lines) and SW (smoothed curves) windowing and showing F0 (blue) and rate (green). The successive discourse macro-units are manually labeled as *introduction, acceleration, slow, neutral, pounded, neutral, high-pitched*, as they correspond to perceived changes in prosody

- the lower buttons adjust the time axis with 6 "arrow" buttons and 4 Praat like buttons : all (view all), in (zoom in), out (zoom out), sel (narrow to selection)

Finally, if a MU is selected, the label and prosodic parameters of this MU are mentioned above the figure, and it can be played individually.

4. Examples

The example in Figure 2 is a 22 second recording of a radio announcement for a bank advertising a very flexible loan taken from [25]. The speaker utters his text very iconically since each sentence creates a maximal contrast with the previous one. The second MU (highlighted) has a high speech rate, whereas the third one is very slow. The fifth one is uttered with a scansion and finally the penultimate is very high-pitched. This deliberately exaggerated and expressive example was chosen to illustrate the *ProsoDyn* abilities. In the figure, both MU parameters (blue and green horizontal segments) and SW curves are shown. Several remarks can be made on the basis on this figure:

1. the second sequence was perceived as accelerated but it becomes clear that this perception comes from the combination of speech rate and pitch height.
2. in the 2nd and 3rd sequences, the rate and pitch range seem to be correlated as they evolve in the same way.
3. nevertheless one can notice that the ambiance change at the third sequence occurs with a delay for speech rate, whereas the pitch drop happens earlier, even if the main feeling about this sequence was *slow*.
4. in the following sequences, these two parameters show a kind of opposite phase, which could be explored in detail.

In sum, one could see here various ways of combining modification in speech rate and F0 registers that produce specific prosodic ambiances. For sake of clarity, the prominence density is not shown here, but a clear contrast appears between the *slow* sequence (in which melodic variation contribute to perceived prominences) and the *accentuated* one (where the syllables have the same height and duration and produce scansion).

The next example (Figure 3) is a 44-second excerpt of radio news taken from [26]. The MU segmentation is at sentence level. Two things are striking: 1) the declination phenomenon clearly appears with a melodic resetting at the MU beginning. 2) the added bold line represents a topic change within the discourse of the speaker. One can notice that the melodic resetting is more pronounced at that transition.

In the last example (Figure 4), a 330-second interview of a French actor [3] shows a climax at 140-160 seconds. A closer look also reveals some evidence of reported speech.

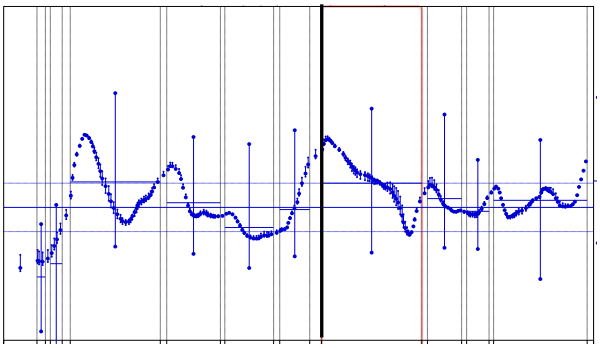

Figure 3: 44-second radio news with sentential MU segmentation. The melodic reseting is noticeable at most sentence beginnings (dotted lines) and topic change (bold line)

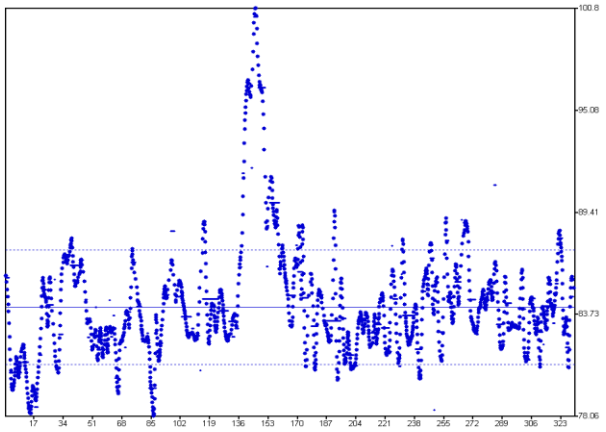

Figure 4: 330-second interview of a French actor

5. Conclusion

ProsoDyn is a promising graphical tool that allows a visual inspection of macroprosodic variations at any scope. The two windowing modes (sliding window and external macrosegmentation) help the user to have a closer look at expected or unexpected modification in macro-prosody. Thus, many hypotheses on the prosody-discourse interface can be verified or rejected. This tool is freely available as a Praat plugin at this address: http://latlntic.unige.ch/phonetic

6. Acknowledgements

This research is partly funded by The Swiss National Science Foundation - FNS Grant nr 100012_134818.

7. References

[1] Auer, P. & A. di Luzio (eds), "The contextualization of Language", Amsterdam, Philadelphia, John Benjamin, 1992

[2] Auchlin 1999. "Les dimensions de l'analyse pragmatique du discours dans une approche expérientielle et systémique de la compétence discursive", in Verschueren J. (ed.), Pragmatics in 1998: Selected papers from the 6th International Pragmatics Conference, vol. 2, Anvers, IPrA, 1-21.

[3] Simon, A.C. 2004. La structuration prosodique du discours en français. Bern : Peter Lang.

[4] Odgen, R. (2001, novembre 16). « We speak prosodies and we listen to them », J R Firth. Conference pronounced at the Grammar in Interaction Conference in Sweden, Uppsala.

[5] Gumperz J. "Contextualization and understanding". In A. Duranti and C. Goodwin (eds.), Rethinking Context: Language as an Interactive Phenomenon". Cambridge: Cambridge University Press, 229-252, 1992

[6] Goldman, J.-Ph., A. Auchlin, A. C. Simon and M. Avanzi, "Phonostylographe: un outil de description prosodique. Comparaison du style radiophonique et lu", Nouveaux cahiers de linguistique française 28, 219-237, 2007

[7] Goldman, J.-Ph., Auchlin, A. Avanzi, M. & Simon, A.C. "ProsoReport: an automatic tool for prosodic description. Application to a radio style", in Barbosa, P. A., Madureira, S., and Reis, C. (éds), Proceedings of the Speech Prosody 2008 Conference. Campinas, Brazil: Editora RG/CNPq, 701-704.

[8] Goldman J.-P., A. Auchlin, and Simon A.C., "Discrimination de styles de parole par analyse prosodique semi-automatique", in Yoo, H-Y and Delais-Roussarie, E. [eds.], Actes d'IDP, Paris, ISSN 2114-7612, 207-221, 2009

[9] N. Obin, A. Lacheret-Dujour, C. Veaux, X. Rodet, A. Simon "A Method for Automatic and Dynamic Estimation of Discourse Genre Typology with Prosodic Features", Interspeech, 2008

[10] Kern, F., "Speaking dramatically. The prosody of life radio commentary of football matches", in Barth-Weingarten, D., Reber E., and M. Selting [eds.] Prosody in interaction, John Benjamins, 217-237, 2010.

[11] Simon, A.C.; Auchlin, A. (2001). Multi-modal, multi-focal : les 'hors-phase' de la prosodie. In : Cavé, Guaïtella & Santi (eds), Oralité et gestualité. Interaction et comportements multimodaux dans la communication, Paris : L'Harmattan, 629-633.

[12] Perlman, M., "Talking fast: the use of speech rate as iconic gesture", in: Fey Parrill, Vera Tobin, and Mark Turner (2010) (eds), Meaning, Form, and Body, CSLI Publications, Stanford.

[13] Auchlin A., à paraître. « Prosodic Iconicity and Experiential Blending », in Hancil S. (ed.) Actes du colloque international Prosodie et Iconicité, Rouen, avril 2009; Serie Iconicity in Language and Literature, Amsterdam : John Benjamins

[14] Bertrand, R. "De l'Hétérogénéite de la Parole. Analyse énonciative de phénomènes prosodiques et kinésiques dans l'interaction interindividuelle". Thèse de doctorat de Sciences du Langage : Université Aix-Marseille I, 419 p.

[15] Pršir, Tea & Simon, Anne Catherine, (sous presse), Iconic interpretation of rhythm in speech, for series Iconicity in Language and Literature, Amsterdam : John Benjamins,2011.

[16] Auchlin A. et al 2004. "(En)action, expérienciation du discours et prosodie", Cahiers de linguistique francaise 26, 217-249.

[17] De Looze, C. & Rauzy, S. (2009). "Automatic Detection and Prediction of Topic Changes Through Automatic Detection of Register variations and Pause Duration", Proceedings of InterSpeech'09, Brighton, England

[18] Goldman, J.-P., "EasyAlign: a friendly automatic phonetic alignment tool under Praat", Proceedings of Interspeech Conference 2011, Florence, Italy, 2011.

[19] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer" (version 5.3), 2011, http://www.praat.org.

[20] De Cheveigné A. and Kawahara H, "YIN, a fundamental frequency estimator for speech and music J. Acoust. Soc. Am. Volume 111, Issue 4, pp. 1917-1930 (2002)

[21] De Looze, C. & Hirst, DJ. (2008). "Detecting Key and Range for the Automatic Modelling and Coding of Intonation", Speech Prosody 2008 Conference, Campinas, Brazil, 135-138

[22] Mertens, P. 2004, "Un outil pour la transcription de la prosodie dans les corpus oraux", Traitement Automatique des langues 45 (2) : 109-130, 2004.

[23] Goldman, J.-P., Avanzi, M., Lacheret-Dujour, A. Simon, A. C. & Auchlin, A. "A Methodology for the Automatic Detection of Perceived Prominent Syllables in Spoken French". Proceedings of Interspeech'2007 Antwerp, Belgium, pp. 91-120

[24] Simon A.-C. et al. "La détection des proéminences syllabiques. Un aller-retour entre l'annotation manuelle et le traitement automatique" CMLF 2008 (2008) 151

[25] Simon, A.C.; Bachy, S. 2009. Parole & Langue. Parcours et exercices de linguistique (DVD-ROM). Presses universitaires de Louvain : Louvain-la-Neuve

[26] Avanzi, M.; Simon, A.C.; Goldman, J.-Ph.; Auchlin, A. 2010. *An annotated corpus for French prominence studies*. Proceedings of Speech Prosody 2010