10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

ProsoBox, a Praat Plugin for Analysing Prosody

Jean-Philippe Goldman[1], *Anne Catherine Simon*[2]

[1]Department of Linguistics, University of Geneva, Switzerland
[2]Institute for Language and Communication, UCLouvain, Belgium
jean-philippe.goldman@unige.ch, anne-catherine.simon@uclouvain.be

Abstract

This contribution presents a Praat plugin for automatic prosodic analysis called ProsoBox. It consists of a collection of tools ran successively: stylization of f0, automatic segmentation of the sound into speech segments, automatic detection of prominent syllables, prosodic report. It produces various output formats: modification of the TextGrid with additional tiers; tables; dynamic or static graphical visualizations. ProsoBox is compared to other automatic tools for prosodic analysis to make clear the specificity of each tool.

Index Terms: automatic prosodic analysis, prominence detection, prosodic features extraction, Praat, Prosogram

1 Introduction

Alluding to Praat's well-known 'doing phonetics by computer' motto [1], we present a toolbox for prosodic analysis. ProsoBox consists of a sequence of Praat scripts that allows for automatic prosodic analysis.

The scripts in ProsoBox allow the user to:

- Segment a sound file (or a collection of sound files) into inter-pausal units (or any kind of speech segments);
- Detect prominent syllables, either with a threshold strategy or with a cumulative strategy;
- Get a series of temporal and fundamental frequency measures at the level of syllables, inter-pausal units or entire sound file;
- Get raw results formats (in tables) usable for further statistical analysis;
- Get graphical representations of prosodic data, either dynamic or static.

For the user's convenience of use, each step of analysis applies to all the files in a directory, which makes it possible to apply manual correction and evaluation after each step.

The plugin function in Praat does not require the user to manipulate scripts directly. After the ProsoBox plugin has been installed, the scripts are called directly from the Praat menu.

The algorithms gathered together into the ProsoBox plugin have been developed incrementally since 2008. The reason for publishing this tool-oriented paper is twofold. First, ProsoBox is now openly distributed on GitLab [2]. Second, a comprehensive and detailed manual has been added to the plugin (using the Praat Manual format), making it easier to use. ProsoBox can now be disseminated among the speech prosody community and be easily used by non-specialists of prosodic studies and, indeed, further developed.

ProsoBox scripts have been created to address our own research needs. First, we developed scripts aiming at automatic prominent syllables detection [1],[4], as prominences are highly relevant for studying prominent (stressed) syllables and intonation. Later on, we developed scripts for extracting prosodic features in large corpora to investigate speaking styles [5],[6]. We ensured that ProsoBox analyses would allow for modeling between-speaker and within-speaker variation, by distinguishing between global prosodic measures (i.e stable for the whole sound file) and dynamic prosodic measures (i.e changing throughout the sound file). Finally, we added scripts for creating graphical representations of prosodic analyses, either in a static way or in an interactive way (i.e. redoable without going back to the initial form of the script) [7].

The need for automatic tools in the speech prosody community has been constantly growing in the last decades. Each new script or tool first meets specific expectations of the researcher or team. Then, it sometimes gets to reach a larger audience if the tool is well documented, is ergonomic enough and has a more generalized usage. We look forward to being part of a large community collaborating and exchanging tools. None of the existing tool can meet all the expectations. In our view, tools must complement each other and provide new instruments for excellent research.

The structure of this contribution is as follows: we present the technical requirements for using ProsoBox, such as file formats and pre-processing (section 2). In the main part of the paper, we introduce the five major steps of analysis (section 3). We then explain the various output formats of analyses (section 4) and conclude by comparing ProsoBox with other tools of automatic prosodic analysis (section 5).

2 Preprocessing of files

The two main prerequisites to use ProsoBox are sound segmentation and fundamental frequency detection and stylization.

2.1 Sound segmentation and alignment

The data to be analyzed lies in a directory of the user's working directory. All sound files (i.e. recordings) are in the same directory along with a TextGrid file having exactly the same name as the corresponding sound file. The TextGrid must contain a few annotation tiers, including a segmentation into syllables. This is for two reasons. First, analyzing temporal variables of speech is based on the duration of speech units (e.g. phones, syllables). Second, relying on vowel nuclei reduces the risks of erroneous estimation of fundamental frequency (see [8] and section 2.2).

Aligned segmentation requires a tier with orthographic transcription (automatic or manual). Phonetization and

10.21437/SpeechProsody.2020-206

segmentation is then easily obtained using programs like EasyAlign [9], SPASS [10] or Train&Align [11].

ProsoBox scripts are ran in Praat. Annotation files are TextGrids that minimally display the following tiers:

- *phones* (or any other tier name): phonetic transcription using SAMPA phonetic symbols;
- *syll* (or any other name): syllabic segmentation and transcription using SAMPA phonetic symbols;
- *words* (or any other tier name): word or multi-word segmentation in orthographic spelling.

At the three annotation levels of *phones*, *syll* and *words*, silent pauses are annotated in separate intervals with a specific symbol (e.g. _).

2.2 Stylization of f0 using Prosogram

The extraction of an accurate f0 track frequently proves to be a complex task. In addition to the f0 detection in Praat [1], ProsoBox relies on f0 detection and stylization by Prosogram [6], a tool for the analysis and transcription of pitch variations in speech. Prosogram stylization simulates the auditory perception of pitch by the listener and restricts f0 extraction to stable, nucleic part of the syllables.

More specifically, a version of the Prosogram script (version 2.4f) has been integrated as the very first step of ProsoBox analyses and adapted to our needs. The script

- extracts f0 curve (Praat pitch file);
- stylizes every nucleus of syllable with level or dynamic tones according to perceptive threshold [12].

3 Steps of analysis and application to illustrative corpora

Each step of analysis is illustrated by data from the C-PROM corpus developed by [13]. C-PROM is an annotated corpus for French prominence studies, including different regional varieties of French (Belgian, Swiss and metropolitan French) and 7 speaking styles (from oral reading to spontaneous conversations) for a total duration of 70 minutes. The manual correction of transcription and segmentation as well as the annotation for prominence and disfluent events were conducted by two phoneticians in parallel.

The following sections depict the five successive steps of analysis within ProsoBox. Each step is performed with a script of the same name.

3.1 *MakeSSTier*: segmentation in speech segments (SS)

Segmenting a sound file into smaller speech segments may be valuable for a number of reasons. First, for prosodic analysis, it makes it possible to identify units (linguistic, interactional, informational, etc.) for which one would like to get specific prosodic measures, for comparison purposes. For example, one could compare the prosodic features of turn-initial sentences with turn-final sentences. Secondly, taking repeated prosodic measures across a sound file (for each speech segment) helps determining whether variables like speech rate of f0 register are constant or not. This is also feasible using a fixed-size sliding window of analysis (see section 3.4).

The speech segment tier may be done either manually or automatically. With the manual method, the user inserts boundaries in an interval tier according to his/her own rationale and may annotate each interval; or they can use any existing tier (e.g. a tier with orthographic transcription of utterances). With the automatic method, the *Make Speech Segment Tier* script retrieves silent pauses durations (from the syllable tier) and segments a TextGrid (or a collection of TextGrids) into intervals that correspond to inter-pausal units (or interpause chunks, see [14], [15]). The pause duration threshold is by default 0.250 sec. and is adjustable by the user.

3.2 *ProsoProm:* automatic detection of prominent syllables

The *ProsoProm* script automatically detects prominent syllables based on acoustic cues (duration and f0). Prominent syllables can be interpreted as accents, stresses or boundaries, according to the prosodic phonology model one adopts. Once prominence detection is completed, the script:

- duplicates the syllabic tier into a new tier called *promauto*. Each syllabic segment is evaluated as prominent (marked with P) or not prominent (nothing);
- produces a table of syllables (filename*_syllsheet.txt*) with all the prosodic information;
- optionally adds three tiers with acoustic values of relative duration, relative height, pitch movement; the tiers are called 'relative parameters' (see Figure 1);
- creates a new point tier called *midvowel*, pointing to the middle of each vowel;
- returns a short report on detection results (absolute and relative frequency of prominent syllables) and allows for comparing the resulting automatic prominence detection with any other available annotation.

The algorithm behind prominent syllables detection takes acoustic features of the syllable into account, namely the relative length, relative pitch and internal (intra-syllable) pitch movements. Acoustic features of each syllable are compared to surrounding syllables. By default, the scope of comparison is 2 syllables before and 1 syllable after the target syllable. Based on previous work on expert-annotated corpora [2], a syllable is detected as prominent when it is longer or higher than the surrounding syllables, or when it has a dynamic pitch movement. Duration and f0 thresholds are set by default and can be adapted by the user.

The detection procedure is rule-based, explicit. It is highly parametric and can be adjusted with input forms: the scope of acoustic analysis, the threshold values for detection, the kind of strategy (binary detection or gradual detection), etc.

3.3 *MakeSSTable:* speech segment-based extraction of prosodic features

The *MakeSSTable* script creates a new table (.csv format) with the following acoustic measures for each speech segment, among others: label of the speech segment (e.g. orthographic transcription); duration of the speech segment; number of articulated syllables, of (internal) silent pauses; duration of the pause before and after the SS; duration of articulation time and pause time; articulation time to pause time ratio; articulation rate, speech rate; high, low and mean pitch; start and end pitch; pitch range; total melodic path; number and proportion of prominent syllables; proportion of level, rising, falling tone syllables; etc.

The SSTable with acoustic measures by speech segment is useful for analyzing within-speaker prosodic variation and can

easily be uploaded into a statistical software for statistical analysis and graphs.

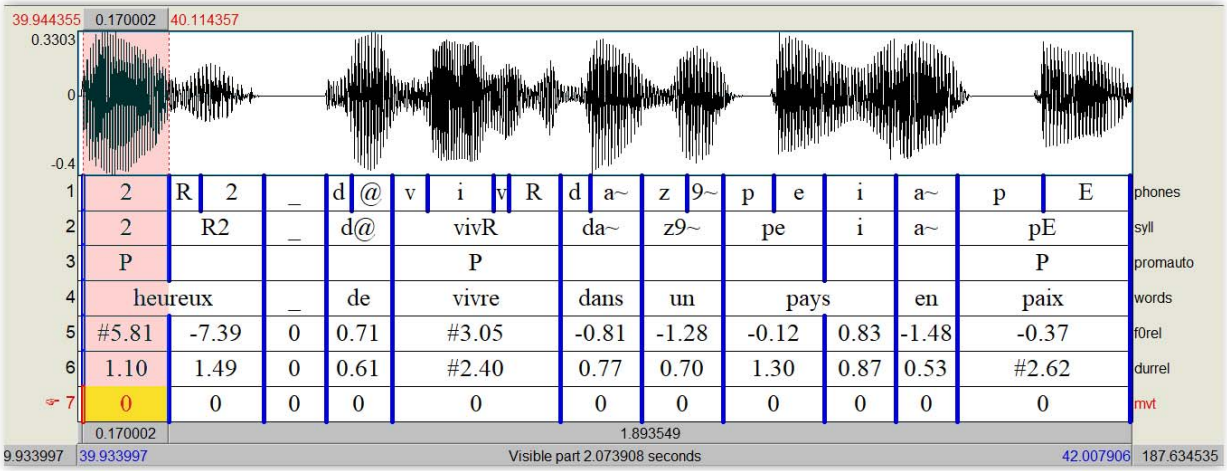


Figure 1. *TextGrid with promauto tier (3) and relative prosodic parameter tiers (5, 6, 7)*

3.4 *ProsoDyn:* dynamic visualization tool for prosodic variation

The *ProsoDyn* script produces a graphical representation of the evolution of prosodic parameters across a sound file. This tool is dynamic (time-based) and interactive (adjustable online).

ProsoDyn plots prosodic parameters taken from the Speech Segment Table or runs online prosodic analysis with the help of a fixed-size sliding window (e.g. 15-syllable width with a one-syllable step). The following prosodic parameters are displayed:

- in green (dotted line): articulation rate (shown in Figure 2);
- in blue (curve): mean f0 (in semitones);
- in blue (boxplot): f0 range (in semitones);
- in purple: density (ratio of prominent syllables).

The panel can play the corresponding sound file, entirely, by speech segments or by fixed-size fragments. It also offers to export a Praat Picture.

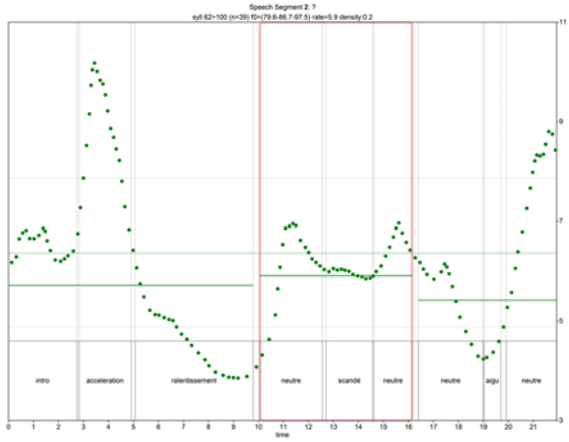


Figure 2: *Example of smoothed speech rate in ProsoDyn*

ProsoDyn helps the user detect (or, ideally, automatically detects) the prosodic changes that would be significant at a discourse level (similar to contextualization cues in Gumperz's interactional approach [16]). The objective is to provide an intermediate solution between a prosodic picture made of dozens of explicit parameters and the usual detailed studies looking at prosodic parameters at a syllabic of accentual group level, with observed cues such as initial phrase accent or final lengthening.

In the example above (see Figure 2), depicting the speech rate of a 20-second advertisement aired on radio, one can see some rising and reduction of the speech rate.

3.5 *ProsoReport:* extraction of prosodic features

The *ProsoReport* script produces a complete description of prosodic parameters in a sound file or a collection of sound files. The report is calculated from the table of syllables of each sound file. The output format is a table (.csv file) with one column per sound file and about sixty lines of prosodic measures. In case several sound files are taken together, ProsoReport adds two columns with mean and standard deviation values. (See Table 1).

This is a tool suited to describe the phonostylistic use of prosody, in the sense of a bundle of prosodic features typical of a style (formal or informal), a professional voice (journalism, teaching) or a regional dialect.

ProsoReport documents temporal variables (proportion of articulated time, articulation rate, syllable mean duration, etc.), melodic features (f0 min and max, f0 dynamic, etc.) and accentuation (proportion of prominent syllables, etc.).

Optionally, the ProsoReport can be extended with additional measures. The script duplicates measures for subsets of syllables (e.g. word-initial syllables).

4 Output formats

Overall, the outputs of ProsoBox can take various forms:

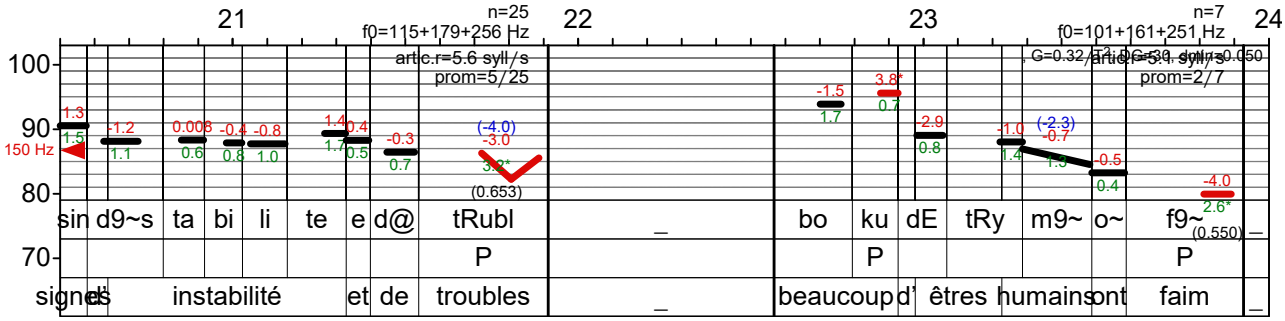
Tables: the three tools *ProsoProm*, *MakeSSTable* or *ProsoReport* produce tables with various prosodic measurements for each syllable, respective speech segment and full recording. These tables are meant to be easily imported into statistical softwares in order to produce graphics and statistical analysis.

TextGrids: the original TextGrid files are enhanced with additional tiers such as the *promauto* tier (for automatic prominence), relative parameter tiers (relative parameters used for prominence detection), or the *speech segment* tier.

Table 1. *ProsoReport for 6 sound files (3 narratives and 3 political speeches) with mean values and standard deviation.*

	nar-be	nar-ch	nar-fr	nar_sd	nar_mean	pol_mean	pol_sd	pol-be	pol-ch	pol-fr
Number of articulated syll	949	949	776	100	891	725	296	420	1011	744
Number of pauses	72	61	71	6	68	93	5	89	98	91
Recording time (s)	206.5	217.5	197.8	9.8	207.3	211.4	21.6	187.5	229.6	217.1
Articulation dur (s)	170.8	184.9	151.9	16.5	169.2	145.7	41.4	106.3	188.9	141.9
Pause duration (s)	35.6	31.9	43.4	5.86	37	64.7	21	78.8	40.5	74.9
Pause ratio (%)	17.3	14.7	22.2	3.8	18.1	31.6	12.7	42.6	17.7	34.6
Speech rate (syll/s)	4.6	4.4	4.0	0.3	4.3	3.4	1.1	2.3	4.4	3.4
Articulation rate (syll/s)	5.6	5.1	5.1	0.3	5.3	4.8	0.8	3.9	5.4	5.2
Speech segment (n)	73	62	72	6.1	69	93.7	4.7	90	99	92
Speech segment duration (s)	2.3	3	2.1	0.5	2.5	1.5	0.4	1.2	1.9	1.5
Speech segment length (syll)	13	15.3	10.8	2.3	13	7.7	2.8	4.7	10.2	8.1
Syllable duration mean (s)	0.18	0.195	0.196	0.009	0.19	0.21	0.037	0.253	0.187	0.191
Syllable duration dev (s)	0.115	0.122	0.099	0.012	0.112	0.1	0.033	0.138	0.075	0.088
F0 mean	92.7	91.8	94.6	1.4	93	86.2	2.2	84.8	88.7	85
F0 range (1>99.ST)	15.5	9.9	9	3.5	11.5	14.8	0.5	15.3	14.8	14.4
F0 narrow range (5>95.ST)	9.7	6.3	6.2	2	7.4	10.5	1.2	11.9	9.8	9.8
Static (%)	87.6	88.5	90.6	1.5	88.9	80.7	15.4	63.1	91.1	88
Rising (%)	5.1	4.7	4.9	0.2	4.9	6.7	4	10	2.3	7.8
Falling (%)	7.4	6.7	4.5	1.5	6.2	12.6	12.5	26.9	6.6	4.2
promauto=0 %	72.9	80.3	78.9	3.9	77.4	71.7	8.7	62.9	80.2	71.9
dur (ms)	0.15	0.164	0.175	0.013	0.163	0.175	0.021	0.2	0.167	0.16
f0_mean (ST)	91.9	91.5	94.4	1.6	92.6	85.4	2.6	83.7	88.4	84.2
promauto=1 %	23.6	17.1	14.7	4.6	18.5	26.4	8.2	34.5	18.2	26.5
dur (ms)	0.278	0.346	0.305	0.034	0.31	0.297	0.047	0.35	0.265	0.274
f0_mean (ST)	95.2	93.3	95.6	1.2	94.7	88	1.9	86.8	90.1	87

Figure 3: *Enriched Prosogram with prosodic parameters for each syllable and speech segment. Prominences are in red.*



Graphics: ProsoBox is able to produce rich graphics depicting the stylized intonation curve, including the values of some prosodic parameters at both syllabic and speech segment levels, for qualitative interpretation (see Figure 3).

5 Comparison with existing tools

There is a tendency for automatic tools to claim they are as 'neutral' as possible, as far as prosodic theory, prosodic phonology or prosodic models are concerned. On the contrary, we think that every tool rests upon theoretical claims, to some extent. What are the prosodic premises underlying ProsoBox plugin?

First, the syllable is considered a primitive unit for prosodic analysis, not only regarding temporal variables (such as speech rate), but also regarding rhythmic and accentuation variables (relying on prominent syllables distribution and features), and even intonation (apprehended as a sequence of tonal targets combining into contours, rather than as a continuous curve). Our syllable-based approach is in line with the Prosogram plugin [8] and distinguishes itself from tools like Momel-INTSINT, where melodic curves are represented as continuous and smooth [17]. Second, we avoid restricting the analysis of prosody to specific phonological domains or primitives like stressed syllables or boundary tones. In our view, the phonetic suprasegmental features play an important part in the identification of phonostyles [19], the analysis of discourse functions [20] or the iconic interpretation of prosody [21].

ProsoBox shares some features with the ProZed plugin, a speech prosody analysis-by-synthesis tool for linguists developed by Hirst [22], although ProZed deals with the symbolic representation of rhythm, tones and intonation events in more depth. A similar attempt to represent dynamic variation in prosody is De Looze and Rauzy's approach, in which clustering algorithms automatically detect variations in register and tempo [20]. Finally, one should mention ProsodyPro [23], which uses f0 trimming and time-normalization algorithms.

6 Conclusion

We presented here a versatile plugin for Praat in order to apply various prosodic tools oriented toward prominence detection and speaking style comparison on large corpora, with quantitative tables for further statistical description and with graphic outputs for visualization. Disseminating this tool within the community and encouraging its use will foster feedback from users as well as further developments.

7 Acknowledgements

The authors would like to thank theirs friends and colleagues Mathieu Avanzi and Antoine Auchlin. They were part of the SAGA team when developing, testing and developing again the scripts in ProsoBox.

8 References

[1] P. Boersma & D. Weenink, "Praat: doing phonetics by computer" [Computer program], Version 6.1.08, retrieved 5 December 2019 from http://www.praat.org/, 2019.

[2] J.P. Goldman, ProsoBox plugin for analyzing prosody, https://gitlab.com/praatplugins [Computer program], retrieved 19 December 2019.

[3] J.-P. Goldman, A. Auchlin, S. Roekhaut, A.C. Simon, and M. Avanzi, "Prominence perception and accent detection in French. A corpus-based account", *Proceedings of the Speech Prosody 2010 Conference*, Chicago, Illinois, 100575, 2010.

[4] J.-Ph. Goldman, M. Avanzi, A. Auchlin, and A.C. Simon, "A Continuous Prominence Score Based on Acoustic Features", *Proceedings Interspeech*, Portland, Oregon, 2012.

[5] J.-P. Goldman, A. Auchlin, M. Avanzi, and A.C. Simon, A.C., "ProsoReport: An automatic tool for prosodic description. Application to a radio style", *Proceedings of the Speech Prosody 2008 Conference*, Campinas, Brazil, 2008, pp. 701-704.

[6] T. Prsir, J.-P. Goldman, and A. Auchlin, "Prosodic features of situational variation across nine speaking styles in French", *Journal of Speech Sciences*, 4(1), 2014, pp. 41-60.

[7] J.P. Goldman, Jean-Philippe. (2012). "ProsoDyn: A graphical representation of macroprosody for phonostylistic ambiance change detection", *Proceedings of the 6th International Conference on Speech Prosody 2012*, Shanghai, China, 2012, pp. 75-78.

[8] P. Mertens, "The prosogram: Semi-automatic transcription of prosody based on a tonal perception model", *Proceedings of Speech Prosody 2004*, Nara, Japan, 2004, pp. 23–26.

[9] J.P. Goldman, "EasyAlign: An automatic phonetic alignment tool under Praat", *Proceedings Interspeech 2011*, Florence, Italy, 2011, pp. 3233-3236.

[10] Bigi, B., "SPPAS: Segmentation, phonétisation, alignement, syllabation", in L. Besacier, H. Blanchon, and G. Sérasset (ed.), *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012*, volume 5 : Software Demonstrations, Grenoble, France, June 4-8, 2012, pp. 9–10.

[11] S. Brognaux, S. Roekhaut, T. Drugman and R. Beaufort, "Train&Align: A new online tool for automatic phonetic alignment". *2012 IEEE Spoken Language Technology Workshop* (SLT), 2012.

[12] C. d'Alessandro, and P. Mertens, P. "Automatic pitch contour stylization using a model of tonal perception", *Computer Speech & Language*, 9(3), 1995, pp. 257–288.

[13] M. Avanzi, A.C. Simon, J.-P. Goldman, and A. Auchlin, "An annotated corpus for French prominence studies", *Proceedings of Prosodic Prominence: Perceptual and Automatic Identification*, Chicago, Illinois, 2012, 102005.

[14] H. Quené, "Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo", *The Journal of the Acoustical Society of America*, 123(2), 2008, pp. 1104-1113.

[15] R. Bertrand, and R. Espesser, "Co-narration in French conversation storytelling: A quantitative insight", *Journal of Pragmatics*, 111, 2017, pp. 33-53.

[16] J. Gumperz, "Contextualization and Understanding", in C. Goodwin, and A. Duranti (ed.), *Rethinking Context : Language as an Interactive Phenomenon,* Cambridge: Cambridge University Press, 1992, pp. 229-252.

[17] D. Hirst, A Di Cristo & R. Espesser. "Levels of representation and levels of analysis for the description of intonation systems", in M. Horne (ed.), *Prosody: Theory and experiment. Studies presented to Gösta Bruce*, Dordrecht: Kluwer Academic Publisher, 2000, pp. 51-87.

[18] D. Hirst, "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation", *Proceeding ICPhS XVI*, 2017, ID 1443.

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

[20] C. De Looze, and S. Rauzy, "Automatic Detection and Prediction of Topic Changes Through Automatic Detection of Register variations and Pause Duration", *Proceedings of InterSpeech 2009*, Brighton, England, 2009.

[21] A. Auchlin, "Prosodic iconicity and experiential blending", in S. Hancil and D. Hirst (ed.), *Prosody and Iconicity,* Amsterdam / Philadelphia: John Benjamins, 2013, pp. 1-32.

[22] D. Hirst, "ProZed: A Speech Prosody Editor for Linguists, Using Analysis-by-Synthesis", in K. Hirose and J. Tao (ed.), *Speech Prosody in Speech Synthesis : Modeling and generation of prosody for high quality and flexible speech synthesis*, 2015, pp. 3-17.

[23] Y. Xu, "ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis". *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, Aix-en-Provence, France. 7-10.