Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Using a replication task to study prosodic highlighting

Rémi Godement-Berline

Univ. Paris Diderot, Sorbonne Paris Cité
LLF (UMR 7110), 75013, Paris, France
remi.godement@linguist.univ-paris-diderot.fr

Abstract

It is hypothesized that prosodic highlighting (the accenting of a constituent for reasons other than rhythmic, i. e. focus marking or expressive emphasis) can be best studied by asking subjects to act out the script of a conversation, either reading it aloud or performing it from memory. Four subjects "replicated" a previously recorded and transcribed spontaneous conversation in French, thus allowing for a three-way comparison between spontaneous, read and interpreted speech using the exact same text. A group of prosody experts annotated the occurrences of prosodic highlighting in each recording. The results confirm the hypothesis on one count but not on the others. The frequency of occurrence of prosodic highlighting is, as expected, highest in interpretation, followed by reading. However, mean F0 and mean syllable duration of annotated words do not follow the same gradation. On the phonological side, there are no differences in the distribution of prosodic contours present on annotated words, as well as a few other features.

Index Terms: elicitation task, prosodic highlighting, read speech versus spontaneous speech, memory

1. Introduction

The "replication" task (RepTask), devised by Laurens, Marandin, Patin and Yoo 2011 [1], consists in asking subjects to "reenact a conversation that has been recorded beforehand and turned into a script."[1] This allows for a comparison between laboratory speech and spontaneous speech on the basis of the exact same text. Laurens et al. concluded from their study that, from the point of view of semantically or pragmatically motivated prosodic phenomena, "subjects' choices in the lab converge with speakers' choices in everyday interactions." They based their conclusion on two case studies: the rising "list" contour (cf [3]) and one special kind of topic shift. This paper addresses the following question: are there differences between laboratory speech and spontaneous speech regarding the phenomenon of "prosodic highlighting", or the accenting of a constituent for reasons other than rhythmic, i. e. the various subfunctions of focus marking and of expressive emphasis? In French, it has already been shown by [4] that the frequency of word-initial accents (which in this language are non-lexical and either semantic-pragmatic or rhythmic) increases with the degree of "preparation" of a given speaking style (from improvised, to rehearsed but not read aloud, to rehearsed and read aloud). My previous study [5] showed that the speech from two theatrical performances displayed a relatively high frequency of prosodic highlighting as compared to other styles. The present study aims to further explore the influence of the choice of elicitation task on the frequency, as well as the phonetic and phonological forms, of prosodic highlighting. Unlike the original replication experiment by Laurens and al., it also asks whether the additional requirement that the subjects memorize the text and then perform it by heart produces different results than simply asking them to read it. After they read aloud the text that was transcribed from a previously recorded spontaneous conversation in French, the subjects of this experiment (four actors, amateur and professional) were asked to come back a few days later after having memorized the text, and give an "interpretation" of it. The recordings were then given to ten prosody experts who annotated the occurrences of prosodic highlighting (henceforth PH). It is put forward here that PH is a crucial differentiating feature between the three types of speech (spontaneous, read, and interpreted). Based on [4] [5] and some intuitions, three hypotheses are tested:

1) the frequency of occurrence of PH increases from spontaneous speech, to reading, to interpretation;
2) mean F0 and mean syllable duration of highlighted words increase from spontaneous speech, to reading, to interpretation;
3) the diversity of phonological accent categories of occurrences of PH increases from spontaneous speech, to reading, to interpretation (the features that are considered are the type of prosodic contour, the contour's syllabic span and the presence of a word-initial secondary accent).

2. Methods

2.1. Elicitation of data

The original spontaneous data are taken from two recordings, one from the CID corpus [6] and one realized for this experiment. The former is a conversation between two men from the Provence-Côte d'Azur region, and the latter a conversation between two women from Paris. The speakers are between 25 and 45 years old and are all researchers or PhD students in a linguistics department. Both recordings took place in a quiet room with minimum background noise. A head-mounted microphone was used for the first recording and a Zoom H2 portable microphone (16 bit/44.1 kHz WAV format) for the second one. In each case, the type of speech activity is

[1] A similar protocol is described in Mixdorff and Pfitzinger 2005 [2], although it seems that it should produce less "natural" data since the original recording is obtained through a map task (and not a spontaneous conversation) and the subjects don't see each other, both in the map task and its reenactment. On the other hand, the protocol has the advantage of using the same speakers in both tasks, which removes any individual differences.

 doi: 10.21437/SpeechProsody.2016-108

small talk. The sequences that were selected for the experiment consist in one speaker telling an unusual and funny anecdote, the other speaker being relatively silent. This allows to record subjects of the subsequent replication task individually and not in pairs, and thus to have more control over the recording settings; it also helps them to perform (and memorize) the text as it has an interesting content.

The read and interpreted versions were elicited by asking subjects to reenact the spontaneous sequences. The sequences were orthographically transcribed. Punctuation was added to improve readability (it was attempted to match the original prosody as well as possible). Disfluencies were preserved as much as possible, but not so much that they prevented a good restitution of the text. The subjects are two women and two men from Paris between 20 and 40 years old. Each of them recorded the text that was produced by a speaker of the same sex in the spontaneous versions. All subjects have experience in acting, to varying degrees: one is an amateur actor, one is an acting student, one is a semi-professional actor and one is a professional actress. The reason why layperson speakers were not chosen is that they might not have been able to memorize the text well enough for the purposes of the experience. Each speaker was paid 75 euros for their participation. The recordings took place in a soundproof room in Paris Diderot University, using a Rode NT1-A studio microphone, a Roland Quad-Capture audio interface and the Audacity software (16 bit/44.1 kHz WAV format). The subjects read aloud, and a few days later performed from memory, the lines from the main speaker in the sequence. In order to prevent any prosodic influence on their production, they were not "fed" the lines from the other speaker in the sequence, but instead mentally "imagined" them. They were asked to play the text as if they were the character and really participated to the conversation, and to respect as much as possible the exact words of the text. For the read version, they discovered the text in the recording room but were allowed to read it to themselves before the recording started. For the subsequent interpreted version, they were given a few days to memorize the text as closely as possible, and were not allowed to keep the text with them during the recording.

The corpus is available along with text-to-speech alignment files at "http://www.llf.cnrs.fr/reptask".

2.2. Annotation of prosodic highlighting

The recordings (two spontaneous, four read and four interpreted) were given to a group of ten prosody experts for them to annotate the occurrences of prosodic highlighting. The experts are researchers, PhD students or graduate students in phonetics, phonology or pragmatics who all have experience in the study of prosody. Each expert annotated four recordings, so that each recording was annotated by four experts. The fact that each recording was not annotated by the same experts constitutes a bias to the experiment. The experts knew what type of speech the recordings that they annotated belonged to, which constitutes a second bias. The experts listened to the recordings without the assistance of a speech analysis software and annotated all occurrences of PH, referring themselves to the following acoustic features (alone or in combination), taken from the literature on PH in French, e. g. [7] [8] [9] [10]: i) increase in pitch, duration or intensity on a part or the whole of the highlighted constituent; ii) initial secondary accent at the beginning of the highlighted constituent; iii) terminal prosodic contour at the end of the highlighted constituent; iv) pitch register compression on the constituents before and/or after the highlighted constituent. The experts were asked not to annotate purely rhythmic accents (marking the right or left boundary of a prosodic phrase). Only the cases where at least three out of four experts recognized the presence of PH were considered to be occurrences of PH.

2.3. Prosodic analysis

The recordings were segmented into words, syllables and phones with the help of Praat plugin EasyAlign [11]. They were then analyzed with Praat plugin Prosogram [12], which gives a series of prosodic measures for each syllable, using the previous segmentation. Based on a psycho-acoustic tonal perception model, Prosogram measures F0 on relevant voiced portions only, in this case determined by the program within the syllable rhyme. F0 was converted in semitones (relative to 1 Hertz) in order to enable cross-speaker comparison. Duration was normalized with respect to syllable structure by dividing syllable duration by the number of phones in the syllable, and with respect to speakers' speech rate by converting the previous values into z-scores for each speaker.

Each occurrence of PH was analyzed auditorily and visually (by myself) on Praat in order to determine its prosodic contour, using the ToBI transcription system for French (cf [13]). The syllabic span of the contour was also determined (i. e., what syllable(s) of the highlighted word the contour spreads over). Finally, the presence of a secondary accent on the initial boundary of the highlighted word or phrase was determined using the automatic prominence detection function of the Analor software [14] (the detection is based on prosodic parameters calculated for French).

3. Results

The corpus contains a total of 5644 words and 7800 syllables. There is 54% of lexical words. There is 56.6% of CV type syllables, 14.8% of V type, 13.6% of CVC type, 9.8% of CCV type and 5.2% of other types.

3.1. Agreement rate

The inter-rater reliability scores for each type of speech and for the entire corpus were obtained by computing Fleiss' Kappa for each recording and by calculating the mean for each group (Table 1). The agreement rate across types is relatively fair. Noticeable differences can be observed between types: the agreement rate for spontaneous speech is higher than that for interpretation, itself higher than that for reading.

Table 1. *Inter-rater reliability (Fleiss' Kappa).*

	Spon-taneous	Read	Inter-preted	All
Kappa (mean)	0.372 z=25.30 p=0	0.217 z=15.01 p=0	0.280 z=19.31 p=0	0.273 z=18.79 p=0

3.2. Frequency of occurrence

The frequency of occurrence of PH is low across types (Table 2). It is higher for reading than for spontaneous speech, and higher still for interpretation; the differences are significant using 95% confidence intervals.

Table 2. *Percentage of highlighted syllables.*

	Spon-taneous	Read	Inter-preted	All
Highlighted syllables (%)	6.59	10.59	14.11	11.22

3.3. Mean F0 and mean syllable duration

A clear difference can be observed in the entire corpus between mean F0 and mean syllable duration of occurrences of PH and that of other words (Fig. 1). The data was analyzed using a linear mixed effects model, with presence of PH as fixed effect and speaker, group of prosody experts and type of speech as random effects. P-values were obtained by likelihood ratio tests of the full model against the model without the fixed effect. The differences are significant for both F0 ($\chi^2(1) = 532.8$, $p < 0.01$) and duration ($\chi^2(1) = 71.8$, $p < 0.01$).

Between types of speech (Fig. 2), mean F0 of PH is higher in spontaneous speech than in interpretation and reading, and mean syllable duration of PH is higher in reading than in spontaneous speech and especially interpretation. A linear mixed effects model was run with style as fixed effect and speaker and experts group as random effects, using a model reduction and likelihood ratio tests to obtain P-values. None of the differences proved to be significant, for both F0 ($\chi^2(2) = 0.137$, $p > 0.05$) and duration ($\chi^2(2) = 3.617$, $p > 0.05$).

The overall differences in mean F0 and mean syllable duration between types of speech were also calculated, in order to determine their possible influence on previous results. Anovas were realized and revealed a significant difference for mean F0, which is higher in spontaneous speech than in the two other styles ($F(2) = 7.147$, $P < 0.001$); however there is no significant difference for mean syllable duration ($F(2) = 0$, $P = 1$).

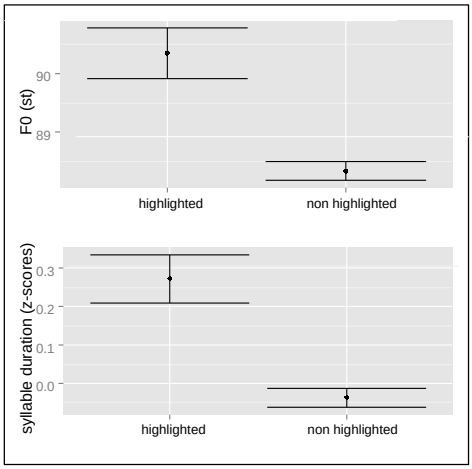

Figure 1: *F0 and syllable duration of highlighted versus non highlighted syllables in the entire corpus.*

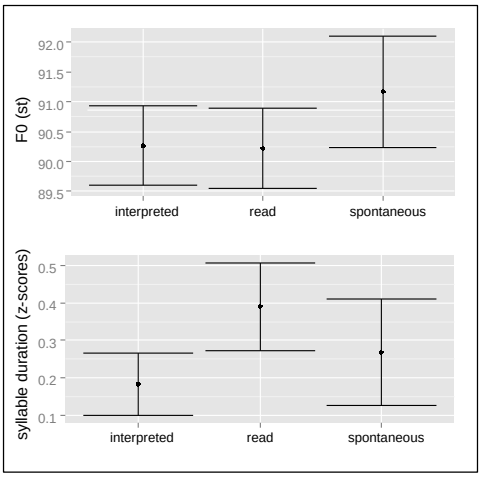

Figure 2: *F0 and syllable duration of highlighted syllables in each type of speech.*

3.4. Accent categories

There are no significant differences between the three types of speech regarding the diversity of phonological accent categories of occurrences of PH (Fig. 3, 4 and 5). This was confirmed by Pearson's chi-squared tests for the prosodic contour type ($\chi^2(12) = 13.20$, $p > 0.05$) and the contour's syllabic span ($\chi^2(6) = 7.14$, $p > 0.05$). For the presence of an initial secondary accent, a linear mixed effects model was run with style as fixed effect and speaker and experts group as random effects, using a model reduction and likelihood ratio tests to obtain P-values, and again showed no difference ($\chi^2(2) = 4.07$, $p > 0.05$).

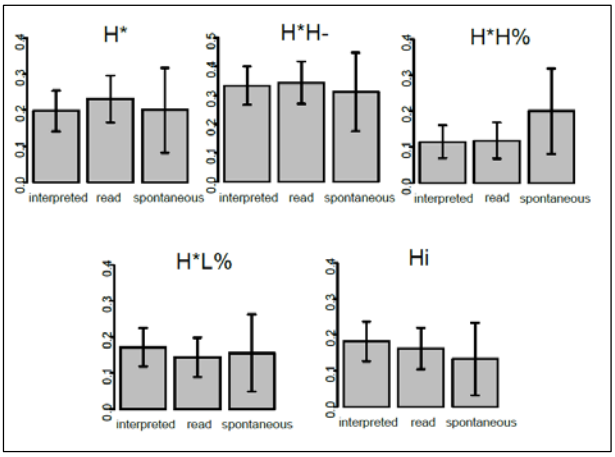

Figure 3: *Percentages of occurrence of the main types of prosodic contour on highlighted words in each type of speech (cf files "contour.wav" and "contour.Textgrid").*

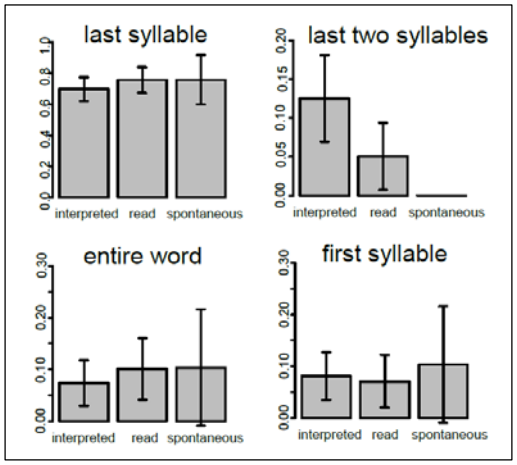

Figure 4: *Percentages of occurrence of the main types of contour syllabic span on polysyllabic highlighted words in each type of speech (cf files "span.wav" and "span.Textgrid").*

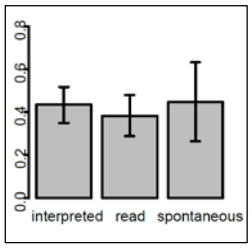

Figure 5: *Percentages of occurrence of initial secondary accent on polysyllabic highlighted words in each type of speech (cf file "initial.wav").*

4. Discussion

The results draw a mixed picture of the role of prosodic highlighting in differentiating spontaneous speech, read speech and interpreted speech.

The first hypothesis is confirmed. Frequency of occurrence of PH is higher in read than in spontaneous speech and is higher still in interpretation, which shows that some element in this condition (memorizing and/or the act of performing) does trigger, or favor, the occurrence of PH. This result cannot be attributed to an overall speech rate difference between types of speech (considering that it would be easier for the experts to detect accents in slower speech) since no such difference was observed (in the future, we will also determine the possible influence of articulation rate). However, the result may also be due to one bias mentioned in §2.2: since the experts knew what type of speech they were annotating, their perception may have been influenced by their expectations about each type, i. e. they may have annotated more occurrences of PH in reading and interpretation because they expected these types to contain more occurrences[2].

The second and third hypotheses are not confirmed. Mean F0 of occurrences of PH does not increase from spontaneous speech, to reading, to interpretation. On the contrary, it is slightly higher in spontaneous speech, although this is just a tendency and is likely due to the independent fact that mean F0 was higher in this type of speech. The only observable tendency for mean syllable duration is that it is higher in reading than in interpretation, a fact that cannot be attributed to a speech rate difference between types of speech. Finally, PH has the same distribution of accent categories in each type of speech.

It is interesting to note that agreement rate between the experts is highest for spontaneous speech and lowest for read speech, a result that goes against the plausible assumption that, due to its normativity, reading aloud would be the type of speech in which it would be easiest to annotate accents.

Some confounding factors may have played a role in the results. The fact that the speakers of read and interpreted speech all have some experience in acting (cf §2.1) may have introduced an independent stylistic element to these types of speech. Also, the fact that the transcription of the original versions was slightly simplified for the replicated versions (cf §2.1; however, the texts were not syntactically altered) may have created some differences, for instance by producing longer phrases in read and interpreted speech.

5. Conclusion

Prosodic highlighting was investigated using two different elicitation tasks: spontaneous speech production, and replication of the same speech (by reading aloud or by performing from memory). Prosodic highlighting was found to be more frequent, but not more marked, in speech elicited through the replication task than in spontaneous speech. Future research will focus on the prosodic marking of the various semantic-pragmatic and expressive functions of prosodic highlighting, and its potential correlation with the choice of elicitation task.

6. Acknowledgements

This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR-10-LABX-0083). Many thanks to Philippe Martin, Jean-Marie Marandin, Fabián Santiago Vargas, Barbara Hemforth, Georges Boulakia, Hiyon Yoo and Elisabeth Delais-Roussarie for their help at various stages, and to all the participants.

[2] The activation state of discourse referents may also play a role here. The spontaneous speech data may contain a lot of given information because it is taken from long conversations where a lot of ideas are exchanged over time, and given information is typically not accented. By contrast, the speakers in read and interpreted speech do not have access to the information background and therefore they may consider a lot of information to be new, and produce more accents. On the other hand, since the type of speech activity in the spontaneous speech data is small talk, it should be expected (cf [1]) that discourse topics change rapidly and that therefore there is not a lot of given information at any point in the conversation (the spontaneous speech samples were selected for just that reason, so that speakers in read and interpreted speech would be able to reenact them without previous knowledge of the information background). In addition, one may expect speakers in read and interpreted speech to have a tendency to consider more information to be given than spontaneous speakers, since they are either reading a text or have memorized it and therefore know in advance (to different extents) what is going to be said.

7. References

[1] F. Laurens, J.-M. Marandin, C. Patin, and H. Yoo, "The Used and the Possible. The Use of Elicited Conversations in the study of Prosody," in *IDP 2009 (Prosody–Discourse Interface), September 9–11, Paris, France, Proceedings*, 2011, pp. 239–257.

[2] H. Mixdorff and H. R. Pfitzinger, "Analysing Fundamental Frequency Contours and Speech Rate in Map Task Dialogs," *Speech Communication*, vol. 46, pp. 310–325, 2005.

[3] C. Portes, R. Bertrand, and R. Espesser, "Contribution to a grammar of intonation in French. Form and function of three rising patterns in French," *Nouveaux cahiers de linguistique française*, vol. 28, pp. 155–162, 2007.

[4] J.-P. Goldman, A. Auchlin, and A.-C. Simon, "Discrimination de Styles de Parole par Analyse Prosodique Semi-Automatique," in *IDP 2009 (Prosody–Discourse Interface), September 9–11, Paris, France, Proceedings*, 2011, pp. 207–21.

[5] R. Godement-Berline, "L'emploi de la focalisation prosodique dans le jeu d'acteur," *Nouveaux Cahiers de Linguistique Française*, vol. 31, pp. 129–139, 2014.

[6] R. Bertrand, Ph. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, and S. Rauzy, "Le CID – Corpus of Interactional Data – Annotation et Exploitation Multimodale de Parole Conversationnelle," *Traitement Automatique des Langues*, vol. 49, no. 3, pp. 1–30, 2008.

[7] M. Rossi, *L'intonation: Le système du français*. Paris: Ophrys, 1999.

[8] A. Di Cristo, "Le cadre accentuel du français contemporain : essai de modélisation," *Langues*, vol. 2, no. 4, pp. 258–267, 1999.

[9] S.-A. Jun and C. Fougeron, "A Phonological model of French intonation," *Intonation: Analysis, Modeling and Technology*. Dordrecht: Kluwer Academic Publishers, pp.209-242, 2000.

[10] Ph. Martin, *Intonation du français*. Paris: Armand Colin, 2009.

[11] J.-Ph. Goldman, "EasyAlign: an automatic phonetic alignment tool under Praat," in *INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association, August 28–31, Firenze, Italy, Proceedings*, 2011, pp. 3233–3236.

[12] P. Mertens, "The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model," in *Speech Prosody 2004, March 23–26, Nara, Japan, Proceedings*, 2004.

[13] E. Delais-Roussarie, B. Post, M. Avanzi, C. Buthke, A. Di Cristo, I. Feldhausen, S.-A. Jun, Ph. Martin, T. Meisenburg, A. Rialland, R. Sichel-Bazin, and H. Yoo, "Intonational Phonology of French: Developing a ToBI System for French," *Intonation in Romance*. Oxford: Oxford University Press, pp. 63–100, 2015.

[14] M. Avanzi, A. Lacheret-Dujour, and B. Victorri, "ANALOR. A tool for semi-automatic annotation of french prosodic structure," in *Speech Prosody 2008, May 6–9, Campinas, Brazil, Proceedings*, 2008, pp. 119–122.