9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Eliciting extra prominence in read-speech tasks: The effects of different text-highlighting methods on acoustic cues to perceived prominence

Stephanie Berger[1], *Oliver Niebuhr*[2], *Kerstin Fischer*[2]

[1]Dept. of General Linguistics, ISFAS, Kiel University, Germany
[2]Innovation Research Cluster Alsion, University of Southern Denmark, Sonderborg, Denmark

mail.stephanie.berger@gmail.com, olni@sdu.dk, kerstin@sdu.dk

Abstract

The research initiative **I**nnovating **Sp**eech **E**li**C**itation **T**echniques (INSPECT) aims to describe and quantify how recording methods, situations and materials influence speech produc-tion in lab-speech experiments. On this basis, INSPECT aims to develop methods that reliably stimulate specific patterns and styles of speech, like expressive or conversational speech or different types emphatic accents. The present study investigates if and how different text highlighting methods (yellow background, bold, capital letter, italics, and underlining) make speakers reinforce the level of perceived prominence of pitch-accented German target words. Analyzed prominence parameters were F0 level, F0 range, normalized intensity level, and word duration. Results show that text highlighting in fact caused prominence parameters to increase. Based on the prominence sensitivity of the affected parameters and the magnitude of their increase, the tested highlighting strategies form the following order of (descending) effectiveness: (i) italics, (ii) yellow, (iii) bold, (iv) capital letters, and (v) underlining.

Index Terms: Text highlighting, prominence, lab speech.

1. Introduction

1.1. INSPECT

For many years, phonetic research has relied on laboratory speech to discover (morpho)phonological patterns and describe their phonetic detail. While the so-called “lab speech” is often the issue of controversy and discussion, it offers the opportunity to study specific acoustic and prosodic phenomena in a controlled manner. Xu [1] argues that “the quality of lab speech is a design issue rather than a matter of fundamental limitation” (p. 329). In a lab experiment, many factors connected to the speakers, the environment, and the recording task have to be considered in order to design a conclusive and successful experiment. Aspects like interpersonal relationships between speakers, skills like musical training, the recording space itself as well as the visible technical equipment and its specifications can influence experimental results. However, probably the most important decision of the researcher is the choice of the speech-elicitation task. There are many tasks to choose from. Speakers can be asked to produce isolated words, sentences, monologues and dialogues (scripted or unscripted) – each serving their own purpose and differing in the degree of spontaneity and formality; see [2] for a detailed account of the pros and cons of different tasks and features related to phonetic recordings and measurements in the laboratory.

For instance, read monologues can end up being prosodically more formal and further away from spontaneous speech than read dialogues are from spontaneous dialogues, cf. [2,3]. INSPECT works to develop, compare, and standardize methodological procedures in experimental acoustic-phonetic studies that use lab speech. The aim is to develop methods that reliably stimulate desired phonetic patterns and/or support the elicitation of specific speaking styles without having to give explicit instructions to speakers. Studies connected to IN-SPECT have so far dealt with factors like time of day, presence and properties of the dialogue partner, characteristics of the elicitation task, musical training, and the use of virtual-reality speaking environments. Recently, Berger et al. [4] also studied how prosodic characteristics of specific speaking styles (expressive speech, fluent speech, speech with hesitations, etc.) can be elicited and controlled in the laboratory solely by using different typefaces for the displayed texts.

Continuing this line of research on the outer appearance of alphabetic characters, the present study deals with the influence of text highlighting on the reinforcement of acoustic cues to the prosodic prominence of German target words in a read-monologue task. Can text highlighting be used to make the corresponding words more prominent relative to a reference set of non-highlighted words? And are the many different ways of highlighting text equally effective in making speakers produce extra prominence? To the best of our knowledge, this is the first time this question is addressed in phonetic research. However, there are a few previous studies that have addressed similar questions on the visual effects or interpretations of different text-highlighting methods. We briefly summarize these studies below and derive our hypotheses from their findings.

1.2. Previous research

Previous studies have tested and compared the visual effects or interpretations of text highlighting within research areas like text analysis [5,6] or the design of study material [7,8]. Probably the largest number of studies is concerned with effects of text highlighting on the readability and recollection of text content. For example, dyslexic children have a hard time seeing the difference between text in italics and regular text, whereas italics is one of the most preferred and effective highlighting methods for children without dyslexia [9]. Studies by [10-13] show that highlighted text content is better recollected by readers that non-highlighted text.

With regard to visual saliency of highlighted text, the area which is most relevant to the present study, [7] found that words in bold capital letters – a combination of two highlighting methods used in our present study – are interpreted by viewers/readers as being “spoken loudly or emphasized and words printed in italics as having a softer emphasis” (p. 28). Yet, [7] also points out the effectiveness of italic (and bold) print for stressing important words in a text, especially in contrast to underlining. Strobelt et al. [5] created a visual hierarchy of highlighting

10.21437/SpeechProsody.2018-15

methods based on “the strength of their pop-out effect” (p. 491). They investigated nine different highlighting techniques, four of which were also used in our study: yellow background, bold print, underlining, and italics. Of these four techniques, [5] ranked yellow background highest (in the top quarter of the ranking) followed by bold print. Underlining ranked third, while italics occupied the lowest rank. The empirical study of [6] came up with a similar ranking. The only two relevant differences to that of [5] were that italics ranked in the middle of the hierarchy and capital letters (not tested in [5]) ranked highest.

From the point of view of a reinforcement effect on acoustic-prosodic prominence cues, we would not expect capital letters to rank highest, though. Despite their undoubtedly strong visual pop-out effect, "reading of upper-case words [...] occurs in a character-by-character order, thereby reducing the speed of reading" [8:13]. It is reasonable to assume that this reduced readability of capitalized words interferes with eliciting greater perceived prominence of words in a reading task. The specific hypotheses tested in our study are derived from the combined results on visual saliency by [5,6,7,8].

1.3. Hypotheses

Four hypotheses are tested in the present acoustic-prosodic experiment:

1. The visual pop-out effect of highlighted as compared to regularly printed text translates into a reinforcement of acoustic-prosodic cues to perceived prominence. In other words, text highlighting is a suitable means to increase the prominence of target words in a reading task.
2. The general effect of (1) is stronger for target words highlighted with more visibly salient methods (yellow, bold, capital letters) than for target words highlighted with more subtle methods (italics, underlined).
3. As the visual pop-out effect is independent of speaker gender and word class, we also expect its effect on prosodic prominence to be independent of these factors.
4. The ranking of text-highlighting effectiveness put forward in [5] is reflected in prosodic-prominence reinforcement (except for capital letters, which are arbitrarily put in the middle of the hierarchy for the reason explained at the end of 1.2). That is, the expected ranking is: yellow > bold > capital letters > underlined > italics.

2. Methods

2.1. Speakers

Thirty speakers were recorded for this study. The speakers were between 16 and 48 years old and recruited such that speaker-gender was balanced (i.e., 15 males, 15 females). All speakers resided in Northern Germany. The speakers did not receive any compensation for their participation.

2.2. Reading material

The basis for the current study were two text sections from the book “Mindset – The new psychology of success” by Carol Dweck [14] in its German translation [15]. The sections used were the preface as well as the initial section of the first chapter. The two sections were typed up using Times New Roman in 14 pt with 1.5 line spacing and margins of 2.5 cm (top and left), 2 cm (bottom), and 5 cm (right), following the speech-elicitation materials used in the preceding typeface study of Berger et al. [4]. Both text sections were together about 1,000 words long and extended over 1.5 pages.

Twenty-two target words as well as 14 non-highlighted reference words located in similar prosodic environments as the target words were chosen across the two text sections; "prosodically similar" means here that both target and reference words consisted of 2-3 syllables and occurred at nuclear pitch-accent positions in the text. Moreover, none of the lexically stressed and pitch-accented (i.e. prominent) syllables coincided with a prosodic-phrase boundary. The target words belonged to different word classes (noun, verb, pronoun). Their frequencies were also proportionally mirrored in the set of reference words. Distributing the 22 highlighted words over about 1,000 words resulted in a target-word density that was sufficiently low for speakers to return to their normal prosodic settings before arriving at the next highlighted target word. The reference words were taken from other parts of the text so they could be analyzed for all speakers to make sure that all speakers were comparable to each other.

Five versions of the two text sections were created, each with a different highlighting method, see Figure 1. That is, in one text version, all target words were highlighted in bold face, in another text version the same target words were all highlighted in italics etc. Each text version represented one experimental condition. The 30 speakers were split up into five groups of six speakers (three males and females), and each group read the text version of a different experimental condition. This way, we kept the reading task short enough to avoid artifacts of fatigue and at the same time precluded speakers from discovering that our actual aim was to compare different highlighting methods.

Before proceeding with the acoustic analysis, we carefully checked on an auditory basis that all target words were indeed realized with a nuclear pitch accent and did not contain or occur next to any disfluency phenomena. After excluding some items for this and other reasons, a total of 395 reference words and 1,057 target words (with repetitions resulting from re-reads of misread sentences) were submitted to analysis.

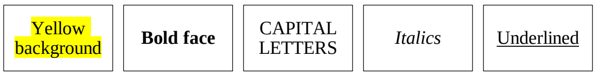

Figure 1: *The five compared highlighting methods.*

2.3. Analyses

The recordings were annotated in Praat [16] on four tiers. For this study, only the interval tier of the target/reference words was analyzed prosodically using the ProsodyPro script [17]. The following ProsodyPro measurements were selected and further processed for this study: pitch level (recalculated as semitones [st] in relation to the speaker's average F0); pitch range (in st); intensity level (normalized against the average dB level of the entire text), and duration (in seconds). These parameters were chosen because they all correlate strongly with perceived prominence in German and other Western Germanic languages [18,19,20]. The correlations are positive, i.e. higher parameter levels are associated with higher levels of perceived prominence. F0 makes the biggest contribution to cueing perceived prominence, followed by duration, which, in turn, is superior to intensity in terms of cueing power [18-19].

Note that we took our measurements for the entire target words, although we expected that text-highlighting effects would primarily manifest themselves on the words' stressed and pitch-accented syllables. However, firstly, it was reasonable to assume that effects concerning the stressed and pitch-accented syllables would still be sufficiently reflected at the word level; and, secondly, we considered our study to be methodologically more sound if the domain of highlighting is identical to the domain of prosodic analysis (in this way we also include potential effects that concern the word as a whole).

The statistical analysis of the acoustic measurements included two tests. First, we tested in a repeated-measures MANOVA based on the two within-subject factors Text Condition (target vs reference) and Word Class (N vs V vs PN) whether text highlighting methods generally increased prosodic prominence as compared to no text highlighting, and whether there are any interactions between prominence effects and word class. Second, we tested with a further MANOVA based on the between-subject factors Highlighting Type (5 levels), and Word Class (N vs V vs PN) how effective the individual highlighting methods are with respect to which prosodic prominence parameters they increase and how much and whether these effects additionally differ as a function of target-word type. Within-factor comparisons were Sidak-correct-ed for multiple testing. Note that we did not include Speaker Gender as a factor in the MANOVA as the most gender-dependent pitch and intensity measurements were normalized. However, in order to anticipate questions on that matter, a re-run MANOVA showed that Speaker Gender had no separate main effect on the prosodic prominence measurements and did also not interact with any of the other factors.

2.4. Recording procedure

The speech recordings were made as individual sessions in silent rooms. Adopting the successfully tested speech-recording innovation from [4] (mimicking modern digital everyday conversation behavior), we used smart phones that the speakers held to their ear while performing the speech task and that were set to a sampling rate of 44.1 kHz and a 16-bit quantization. Speakers stated after the reading task that they were at ease with speaking into a cell phone.

The speakers were sitting on a comfortable chair during the recordings, with the printed text sections in front of them. They held the phone in a constant, comfortable position while reading the printed text sections out loud. At the beginning of the recording session, they were given time to familiarize themselves with the text. The experiment started with written instructions the speakers had to read, whilst having the possibility to ask questions to the experimenter. The speakers were instructed to read as spontaneously and vividly as possible, and to re-read a sentence when they stumbled on its words.

The recording procedure took around 30 to 45 minutes per speaker, including answering a metadata questionnaire, familiarization with the texts, and an explicit speaker de-briefing.

3. Results

3.1. Effects of text highlighting in general

The repeated-measures MANOVA yielded strong significant main effects of Text Condition. That is, independently of which text-highlighting method was used, the highlighted target words were realized across speakers with more strongly pronounced prosodic prominence cues than the non-highlight-ed reference words. This difference involved all prominence-relevant parameters, see Figure 2. Compared to the reference words, target words were characterized by a higher pitch level (F[1,1421]=51.9, p<0.001), a larger pitch range (F[1,1421]= 26.5, p<0.001), a higher intensity level (F[1,1421]=19.0, p<0.001), and a longer duration (F[1,1421]=27.4, p<0.001).

The main effects of Word Class were not significant, neither were the interactions between Text Condition and Word Class. Thus, the way speakers implemented the text highlighting with respect to prosodic prominence cues was independent of the highlighted word or its word class.

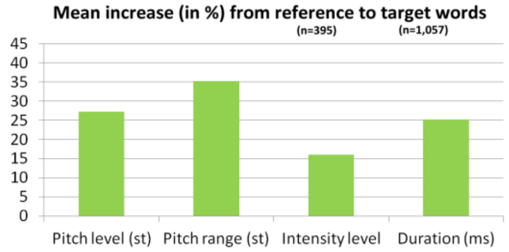

Figure 2: *Within-subjects comparison of effects (%) of Text Condition on acoustic-prosodic cues to perceived prominence.*

3.2. Differences between highlighting methods

The second MANOVA revealed clear differences between the five compared text-highlighting methods. The corresponding significant main effects concerned each prosodic-prominence parameter (pitch level: F[4,977]=3.1, p=0.013; pitch range: F[4,977]=3.4, p=0.009; intensity: F[4,977]=99.4, p<0.001; duration: F[4,977]=4.5, p=0.001) and showed no interactions with Word Class. Thus, the different prosodic-prominence patterns can be described as the sole consequence of reading the target words in combination with a different highlighting. Figure 3 provides a summary of how the five compared text-highlighting methods performed in terms of their reinforcement of acoustic-prosodic cues to perceived prominence.

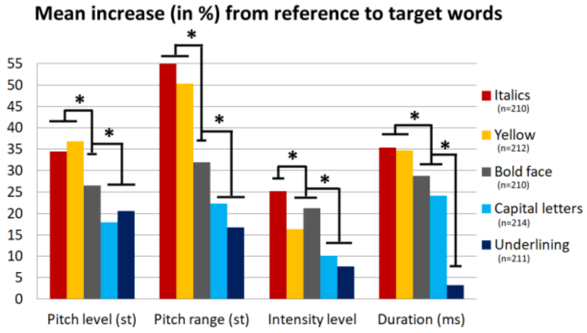

Figure 3: *Text-Condition effects (%) on prominence cues, broken down by the individual text-highlighting methods that represent between-subjects comparisons (asterisks = p<0.05).*

The overall smallest increases in prominence parameters relative to the non-highlighted reference words were found for the underlined target words. Italics, on the other hand triggered the strongest effects on target-word prominence. When highlighted in italics, speakers significantly increased (according to multiple post-hoc t-test comparisons, see asterisks in Fig.3) a target word's intensity level on average by about 9 dB (p<0.01), its average pitch level and range by about 2 or 3 st each (p<0.05, p<0.01), and its average duration by about 75 ms (p<0.01). For yellow highlighting, the increases in prosodic prominence cues (relative to non-highlighted reference words) were equally strong as for italics, except for the intensity level, which was the only dimension on which italics significantly outperformed yellow highlighting and on which yellow highlighting performed statistically on a par with the smaller effect of bold face.

Bold face yielded an intermediate performance in terms of its increase of prominence parameters relative to non-highlighted reference words. Bold face had significantly smaller effects on pitch level and range than italics and yellow highlighting, but these effects were still stronger than those of capitalized letters and underlining (p<0.05; p<0.01).

Using capital letters for text highlighting turned out to be almost as ineffective as underlining text. Capital letters only outperformed underlined text in that they triggered a significantly higher duration of target items (p<0.05). This increase in duration was as strong as that of bold face. Underlining takes the last place in all statistical comparisons and is, moreover, the only highlighting condition for which one prominence parameter (i.e. duration) showed no significant increase relative to the reference condition of no highlighting.

4. Conclusions

Two hypotheses were supported by our results. Compared to non-highlighted text, all of the text highlighting methods used in this study elicited stronger acoustic-prosodic cues to perceived prominence in a reading task (Hypothesis 1). Furthermore, the effects of highlighting on prominence were independent of Speaker Gender and Word Class (Hypothesis 3). Both male and female speakers behaved the same in their reactions to the highlighting methods, and these reactions were moreover not affected by the individual word and its grammatical category. Note that the latter also means that the words' syntactic position, which is correlated with grammatical category in German, was not important for the occurrence and strength of the highlighting effect either. So, it seems that text highlighting methods can be used without much caution and contextual control in phonetic lab-speech reading tasks.

Hypotheses 2 and 4 – a greater influence of highlighting methods with a greater visual pop-out effect, as well as the expected ranking (yellow > bold > capital letters > underlined > italics) – were partly supported with one exception: Italics ranked highest on all acoustic parameters in this study, suggesting that italic print is most suitable for eliciting extra prominence on target words. Why did italics have such a strong effect? We think that this could be due to the typeface we used: Times New Roman. The salient serifs of Times New Roman change considerably when set in italics, and this has probably amplified the prominence effect of the italic print.

The rest of the ranking of Hypothesis 4 came out just as expected from previous findings on the visual saliency of the different text highlighting methods [5-8]. This suggests that the phonetic pop-out effect is at least to some degree mediated by the visual one. Yellow background turned out to be almost as effective as italics for eliciting more prominent pitch accents or words based on increases in pitch level, pitch range, and duration. The effect the yellow background on intensity was smaller than for italics, yet intensity is the parameter that has the lowest influence on perceived prominence according to [18] and [19], which is why we rank the yellow background right after italics but clearly before bold face, see Figure 4.

Bold face was almost as effective as italics and yellow highlighting for increasing a target word's duration, but performed less well in increasing the target word's F0 parameters, which are the most powerful acoustic cues to perceived prominence [18-20]. Therefore, bold face is put behind the yellow background in our ranking.

Capital letters only had the effect of slightly increasing the duration of a target word, which is perhaps not even directly related to the speaker's intention to add extra prominence to the word, but to the generally reduced readability of upper-case text [8]. In all other parameters, capitalization proved to be just as ineffective for eliciting increased prosodic prominence as underlining. Yet, given that duration is the second most powerful cue to prominence in German [20], we placed upper-case highlighting before underlining in our ranking.

A reason for the very low influence of underlined text on prominence could be that the underlining is very hard to localize in a text and to associate with a particular word. Thus, returning to the link between the visual pop-out effect of a highlighting method and its effect on making speakers add extra prominence to target words, visual saliency seems to be not the only relevant factor. The degree to which a highlighting method can visually be clearly associated with a particular word, and the degree to which a highlighting method interacts with the visual appearance of the word's letters could also be important (e.g., underlining and bold face both pop out visually but do not directly interfere with the words visual appearance as much as the yellow background and italics in Times New Roman).

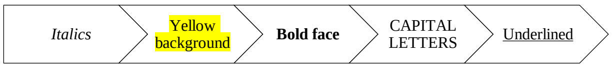

Figure 4: *The hierarchy of the highlighting methods from the most effective method italics (left) to the least effective method underlining (right).*

In summary, our results show that different highlighting methods can be used in phonetic lab-speech experiments in order to prompt a speaker to read with extra prominence without giving him/her the explicit instruction to do so. Presenting words in front of a yellow background seems to be very effective for that purpose, as well as setting words in italics – at least when the typeface used is Times New Roman. In addition, our study has opened up further research perspectives on the relation between text highlighting methods and prosodic cues to perceived prominence: (i) Are italics and the yellow background effective enough to change the regular pitch accent on a target word into an emphatic accent in the definition of [21]? (ii) Can we control which prominence parameters are affected through the choice of the text-highlighting method? (iii) Is there an effect of highlighting beyond the stressed and pitch-accented syllable, i.e. at the level of the entire target word? (iv) Is highlighting also effective at the level of entire prosodic phrases?

All questions are worth pursuing. Initial observations and pilot data suggest that eliciting emphatic pitch accents through text highlighting is indeed possible, and that highlighting effects are restricted to stressed/accented syllables and generally less effective at the phrase than at the word level. Future studies will put these indications on solid empirical grounds.

5. Acknowledgements

We would like to thank Carina Marquard, Nele Kiupel, Jana Voße, and Gunnar Gerstenkorn for their help with this project.

6. References

[1] Y. Xu, In defense of lab speech. In *Journal of Phonetics*, vol. 38, pp. 329–336, 2010.

[2] O. Niebuhr, and A. Michaud, Speech data acquisition – The underestimated challenge. In *Kieler Arbeiten zur Linguistik und Phonetik (KALIPHO)*, vol. 3, pp. 1–42, 2015.

[3] H. Mixdorff, and H.R. Pfitzinger, Analysing fundamental frequency contours and local speech rate in map task dialogs. In *Speech Communication*, vol. 46, pp. 310–325, 2005.

[4] S. Berger, C. Marquard, and O. Niebuhr. INSPECTing read speech – How different typefaces affect speech prosody. In *Proceedings of Speech Prosody 8*, pp. 513–517, 2016.

[5] H. Strobelt, D. Oelke, B.C. Kwon, T. Schreck, and H. Pfister, Guidelines for effective usage of text highlighting techniques. In *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 489–498, 2016.

[6] D. Simard, Differential effects of textual enhancement formats on intake. In *System*, vol. 37, pp. 124–135, 2009.

[7] C. Acrey, C. Johnstone, and C. Milligan, Using universal design to unlock the potential for academic achievement of at-risk learners. In *TEACHING Exceptional Children*, vol. 38, no. 2, pp. 22–31, 2005.

[8] A. Degani, *On the typography of flight-deck documentation*. Moffett Field, California: Ames Research Center, 1992.

[9] R. Ismail, A. Jaafar, Important features in text presentation for children with dyslexia. In *Journal of Theoretical and Applied Information Technology 63*, pp. 694-700, 2015.

[10] J. White, *An input enhancement study with ESL children: effects on the acquisition of possessive determiners.* Doctoral Dissertation, McGill University, Montréal, Québec, Canada, 1996.

[11] D.J. Shook, FL/L2 reading, grammatical information, and the input to intake phenomenon. In *Applied Language Learning 5*, pp. 57–93, 1994.

[12] R. Jourdenais, M. Ota, S. Stauffer, B. Boyson, C. Doughty, Does textual enhancement promote noticing? A think-aloud protocol analysis. In *Attention and Awareness in Second Language Learning 9*, pp. 183–216, 1995.

[13] J. Ling, P. van Schaik, The influence of font type and line length on visual search and information retrieval in web pages. In *International Journal of Human Computer Studies 64*, pp. 395–404, 2006.

[14] C. Dweck, *Mindset – The new psychology of success*. London: Random House, 2006.

[15] C. Dweck, and J. Neubauer, *Selbstbild – Wie unser Denken Erfolge oder Niederlagen bewirkt.* München/Berlin: Piper Verlag GmbH, 2009.

[16] P. Boersma, and D. Weenink, *Praat: doing phonetics by computer, version 6.0.17*, 2016. (Retrieved 22 March 2016 from http://www.praat.org/)

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

[18] K.J. Kohler, Rhythm in speech and language. In *Phonetica, vol. 66*, pp. 29–45, 2009.

[19] D.B. Fry, Experiments in the perception of stress. In *Language and speech, vol. 1, no. 2*, pp. 126–152, 1958.

[20] O. Niebuhr, J. Winkler, The Relative Cueing Power of F0 and Duration in German Prominence Perception. In *Proc. 18th International Interspeech Conference, Stockholm, Sweden*, pp. 611-615, 2017.

[21] S. Baumann, O. Niebuhr, B. Schroeter, Acoustic Cues to Perceived Prominence Levels – Evidence from German Spontaneous Speech In *8th International Conference of Speech Prosody, Boston, USA*, pp. 711-715, 2016.