Beat it! – Gesture-based Prominence Annotation as a Window to Individual Prosody Processing Strategies

Petra Wagner[1,2], *Aleksandra Ćwiek*[1], *Barbara Samlowski*[3]

[1]Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld
[2]Cluster of Excellence Cognitive Interaction Technology (CITEC)
[3] Amazon Development Center, Aachen

petra.wagner@uni-bielefeld.de, aleksandra.cwiek@uni-bielefeld.de

Abstract

In recent work [1], we have suggested a novel approach for fine-grained and fast prominence annotation by naïve listeners. Our approach relies on annotators' "drummed" replications of a perceived utterance, modulating their drumming velocity in accordance with the perceptual prominence of consecutive linguistic units (syllables, words). The drumming velocity is then used as a fine-grained operationalization of prosodic prominence. This intuitive method exploits the established link between prominence and speech-accompanying gesture [2, 3]. Due to its speed and ease, it allows for the rapid annotation of large amounts of data and yields results that are comparable to fine-grained expert annotations of prominence.

In the present study, we evaluated our method further by (1) comparing the intra-sentential prosodic variation as measured with traditional annotations and the drumming method. Our results show that "drummed" prominences capture speaking-style related variability similarly to conventional annotation methods. Additionally (2), we examined whether individual listener strategies can be identified with the help of Random Forests. This method allows for estimating the individual impact of established prominence correlates on prominence impressions. Our analyses unveil individual listener strategies for blending and integrating top-down, bottom-up and context cues into impressions of prosodic prominence.

1. Introduction

Recently, [1], we introduced a novel approach for fine-grained and fast prominence annotation by naïve listeners, relying on annotators' "drummed" replications of a perceived utterance. Instead of training annotators to make fine-grained judgements of prosodic strength in a time-consuming and cumbersome way, this approach asks listeners to "repeat" a previously heard utterance in a drumming task and to modulate their drumming velocity in accordance with the perceptual prominence of consecutive linguistic units (syllables, words). Drumming velocity (measured by the MIDI output of an electronic drum pad) is used as a fine-grained operationalization of prosodic prominence. This intuitive method exploits the established link between prominence and speech-accompanying gestures [2, 3]. The method has been shown to work with naïve annotators after a very short training phase (10 sentences) and can be used to assess impressions on the level of syllables and words. Due to its speed (close to real-time) and ease, it allows for the annotation of large amounts of data. Another interesting finding was that the prominence patterns yielded by the drumming task show a high correspondence to experts' fine-grained prominence impressions, when they are averaged across several naïve annotators. Individual naïve "drumming annotators" typically correlate only moderately among each other, especially when drumming "syllable prominences" (cf. Figure 1). This finding points to a lot of individual variation of listeners' interpretations of linguistic and acoustic cues to prosodic prominence. Obviously, expert annotations are not helpful to comprehend these individual strategies, as these are based on our existing knowledge of how prominences are signaled. The popular and fast method where rapid binary prominence annotations by naïve listeners are cumulated into a fine-grained prominence profiles [4] appears to be too coarse to provide a detailed picture of the individual listener.

In this study, we set out to examine "drummed prominence impressions" as a window to individual prominence processing strategies. To this day, a lot of research has shown a myriad of cues to be influential in prominence perception, e.g. acoustic cues such as fundamental frequency excursion and shape [5, 6], duration, intensity, [7], linguistic cues such as word order or lexical class [8, 9] and context cues such as metrical priming or the presence of a nearby pitch accent [10, 11]. It is also known that both acoustic "bottom-up" and linguistic "top-down" cues are somehow integrated when processing prominence [4, 12, 13]. However, we still know little about the presence of individual processing strategies when weighing the many prominence correlates that have been identified.

Our analysis consists of two steps: First, we evaluate our drumming method further to find out whether speaking-style related prosodic production variability is identified similarly in conventional (auditory) and drummed prominence annotations. This should shed light on the question of whether the blending of top-down and bottom-up processing works in a comparable way in both approaches. Second, we build Random Forest Models predicting individual listeners' "prominence drumming behavior". We use these to assess individual prosody perception strategies by weighing the individual importance of well-established prominence correlates (acoustic, linguistic, contextual) in the prediction models.

2. Is prosodic production variation reflected similarly in expert annotations and drumming?

If the same sentence is uttered in a prosodically different way across speakers or styles, e.g. due to a different prosodic focus or rhythmic pattern, this variation ought to be reflected

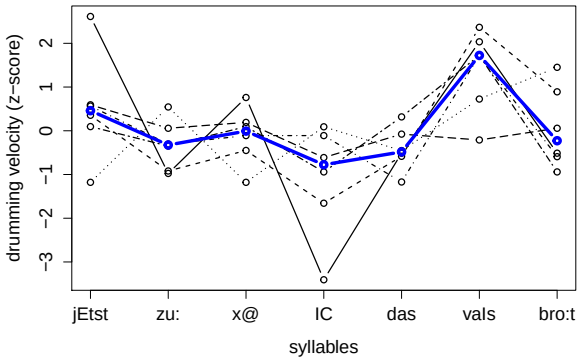

Figure 1: Syllable-based drummed prominence patterns of 6 annotators for the same sentence. The thick blue line illustrates the median drumming velocities ("average drummer")

in prominence impressions and consequently yield different prominence annotations as well. Due to the influence of top-down expectations on prominence perception (cf. section 1), such fluctuations in prosodic structure and style may be somewhat neutralized in perception: to some extent, we "perceive what we expect". It is possible that certain annotation methods cause a stronger or less strong reliance on such top-down expectations than others, e.g. as they induce listeners more to rely "on their inner voice". In order to test whether drummed or conventional prominence annotations behave differently or similarly in this respect, we compared the extent to which annotations varied across identical sentences produced by different speakers. If the balance of top-down expectation and bottom-up processing is similar in both annotation procedures, then a set of (orthographically) identical sentences perceived as prosodically rather different with a conventional annotation method, should also yield in a stronger perceptual variation in a drumming task and vice versa.

2.1. Methods

The data used in the drumming task contained a set of 20 sentences, each of which was produced by three different speakers, i.e. 20 sentence triples. Within each triple, the individual productions are likely to differ to some extent, e.g. due to different reading styles or different linguistic interpretations of the read material. These variations ought to be reflected in the prominence annotations. To test, whether the two compared annotation methods (drumming and conventional auditory prominence perception) indeed measure a very similar quality of prominence, we calculated the intra-sentential ICC-variability of "drummed" and "perceived" prominence patterns within the three productions in each triple. The"drummed prominences" are based on median velocities across 6 individual drummers ("average drummer"), the perceived prominence was based on the fine-grained (31 levels) median prominence annotations of 3 prosodic experts ("average expert listener"). As the conventional (expert) annotations were only available on the syllable level, we also used syllable level drummed annotations for the comparison. All analyses were carried out with the help of the irr-package available for the statistical software package "R" [14].

2.2. Results

When plotting the intra-sentential ICC statistics based on variability in both perception and drumming, it becomes evident that the perceived variability is indeed similar in across perceptions in both modalities, albeit not perfectly aligned (cf. Figure 2). A correlation analysis confirms this visual impression ($cc = 0.62, df = 18, p < 0.01$). The ICCs are significantly higher ($t = 4.6$, $df = 37.2$, $p < 0.0001$) for the conventional annotation method ($M = 0.78$, $SD = 0.15$) compared to the drumming method ($M = 0.55$, $SD = 0.17$).

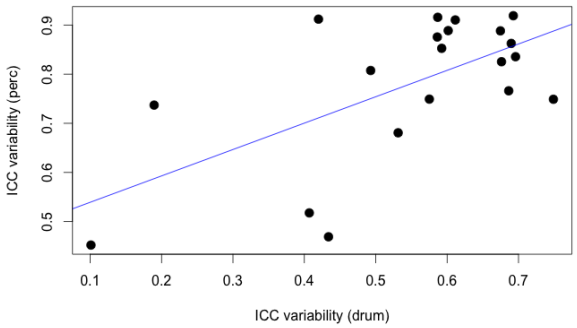

Figure 2: Relationship between conventionally perceived (y-axis) and drummed (x-axis) intra-sentential prosodic variability (ICC)

2.3. Discussion

The analyses reveal that if the same sentence uttered by different speakers receives similar prominence annotations in the drumming task, this is likely to be the case in the prominence annotation task as well and vice versa. This supports the assumption that prosodic variation on the is taken into account by both approaches similarly. However, compared to the conventional annotation, the drumming task yields overall a stronger variability across orthographically identical sentences. This may be interpreted tentatively in such a way that drumming is guided less strongly by top-down expectations, as its annotations are comparatively less uniform across linguistically identical sentences. However, a likely alternative explanation for this finding may be that the stronger overall variability in drumming is caused by the naïve annotaters who may be less stable and following more individual strategies than trained experts. At this point, further conclusions are difficult to make, especially as the analyses rely on comparatively few data points.

3. Identifying listener strategies in "prominence drumming"

In our previous study [1] we found that the drumming approach to prominence annotation reveals much individual variation. We stated above that this individual variation may provide a window to unveil individual strategies of prosodic interpretation. In this section, we therefore evaluate whether individual listener profiles can be estimated based on the drummed annotations. As these strategies may differ depending on the level of linguistic prominence annotation, we will analyze both syllable-based and word-based prominence drumming. These analyses will reveal

whether there are listeners who are guided more by bottom-up or top-down strategies than others. Also, it may reveal the influence of contextual cues to prominence perception.

3.1. Methods

60 sentences annotated by 12 annotators (6 syllable drummers, 6 word drummers) serve as material for our study. For each annotator, we trained a Random Forest regression tree to model the prominence annotations (= drumming velocities) based on a set of acoustic, linguistic and contextual prediction variables that have been shown to influence the impression of prosodic prominence (cf. Table 3.1). For both the syllable and the word drumming, an identical set of predictor variables was chosen, with the following exceptions: The syllable's stressability was not used to predict word drumming velocities, as each German word contains at least one stressable syllable, making this feature redundant. The predictor variable "Clash" refers to the accentuation status of the previous syllable in the syllable-based annotations, but to the previous word in the word-based annotations. Training and subsequent analysis was carried out with the *randomForest* R-package [14] using the standard settings and 3000 training cyles. In order to weigh the impact of the individual parameters on the drumming velocities, the importance for all predictor variables was computed as part of the training procedure. This importance measure captures the mean decrease in classification accuracy (MSE) after the predictor variable has been permuted across all trees.We use importance measure (z-score normalized) to weigh all predictor variables' influence on the dependent variable "drumming velocity".

prominence correlate	**description**	**type**
F0	a normalized value between 0 (F0-min) and 1 (F0-max))	acoustic
SyllDur	syllable duration	acoustic
POS	Lexical class	linguistic
Acce	phonological accentability status of syllable	linguistic
Clash	pitch accentedness of previous syllable/word	contextual
AccentDist	distance to previously pitch accented syllable (in syllables)	contextual

Table 1: Overview of predictor variables the RandomForests are trained on. In the word drumming task, the syllable-based measurements relate to the lexically stressed syllable.

3.2. Results for Syllable Drumming Task

When comparing the importance of the various predictor variables across all drumming annotators, it becomes evident that they relied predominantly on the F0 excursion when modulating their drumming velocity. However, the remaining prominence cues were used to different degrees. Two out of five annotators used syllable duration as the second most important cue to modulate their drumming, while for two others, POS-based information was the second-best predictor. The more "duration oriented" drummers can be regarded as being slightly more guided by bottom-up cues, the "POS-oriented" drummers as more guided by top-down cues. One of the "POS"-drummers used POS information to an equal degree as F0-based information, thus relying rather heavily on linguistic information. Contextual (Clash, AccentDist) and phonological (Acce) information was used somewhat less by most annotators. However, one drummer relied mostly on these contextual cues in combination with local F0 height and but paid little attention to duration or POS information. The overall prediction accuracy for the individual drumming behavior based on the chosen predictor variables differs vastly, and explaines practically none of the behavioral drumming variance for two anntotators (2 and 5), while explaing more than a third of the drumming variation in annotator 4. When pooling the individual annotators' impressions to an "average annotator" (cf. Section 2), the RandomForest model is able to account for 47% of the variance in the velocity data. Not surprisingly, the limited set of variables taken into account in our study is obviously not entirely sufficient to account for the collected velocity data, especially when based on a small data set of 60 sentences. An overview of the results is presented in Table 3.2

annotator	*F0*	*SyllDur*	*POS*	*Acce*	*Clash*	*Accent-Dist*	*% Variance Explained*
1	**72**	19	11	**37**	27	**29**	14
2	**84**	**46**	**50**	18	23	26	26
3	**33**	**21**	**20**	17	10	10	0
4	**83**	**47**	**55**	41	37	27	34
5	**39**	**26**	**23**	2	13	4	1
6	**68**	**34**	**67**	11	17	24	21
average annotator	**100**	**71**	**77**	34	33	39	47

Table 2: Importance (%) of the various predictor variables per annotator in the syllable drumming task. The three most important predictor variables per annotator are shown in boldface.

3.3. Results for Word Drumming Task

The results show comparatively more individual variation than the syllable drumming. Some of them rely predominantly on fundamental frequency excursion, others more on duration, linguistic cues such as lexical class or contextual cues such as the distance to the previous accented syllables. Annotators 2, 3 and 5 relied heavily on a combination of duration, lexical class and context, annotators 1, 4 and 6 on the combination likewise favored by the syllable annotators: F0 excursion, duration and lexical class. As for syllable drumming, the overall prediction accuracy for the individual drumming behavior based on the chosen predictor variables differs vastly, and explaines practically none of the behavioral drumming variance for two anntotators (4 and 6), while explaing up to two thirds of the drumming variation in the remaining annotators. When pooling the individual annotators' impressions to an "average annotator" (cf. Section 2), the RandomForest model is able to account for 67% of the variance in the velocity data, which is considerably more than what was achieved for the syllable data. Interestingly, for the average annotator, the predictor variable "clash" reaches very high importance, while it plays little role for the individual annotators. However, in this condition all predictor variables have a stable and important influence on the drumming velocity result. Not surprisingly, the limited set of variables taken into account in our study is obviously not entirely sufficient to account for the collected velocity data, especially when based on a small data set of 60 sentences only. An overview of the results is presented in Table 3.3

annotator	*F0*	*SyllDur*	*POS*	*Clash*	*Accent-Dist*	*% Variance Explained*
1	**59**	**48**	**62**	22	22	44
2	39	**61**	**60**	8	**55**	58
3	37	**49**	**64**	8	**39**	36
4	**14**	**14**	**12**	11	10	0
5	29	**41**	**46**	24	**54**	29
6	**34**	**12**	**15**	0	4	7
average annotator	**97**	77	**91**	**87**	80	76

Table 3: Importance (%) of the various predictor variables per annotator in the word drumming task. The three most important predictor variables per annotator are shown in boldface.

3.4. Discussion

In line with the previous analyses, the syllable drumming shows higher inter-annotator variation and a less clear correspondence with well-established top-down, bottom-up or contextual correlates of prosodic prominence. In both word- and syllable drumming, listeners blend both top-down and bottom-up cues when modulating drumming strength. F0 appears to be the strongest predictor of drumming velocity in the syllable drumming task across annotators, while lexical class and syllable duration appear to be more reliable cues across annotators in the word drumming task. When drumming syllables, context cues appear to be be less influential for most annotators compared to word drumming. Other than syllable drumming performance, word drumming can be explained quite well based on a very small data set and a limited set of acoustic, linguistic and contextual predictor variables. This may be partly due to the fact that out of 6 word drumming annotators, 2 (annotators 1 and 2) had prosodic training. The non-experts' performance variability is explained in a similar range as the syllable drumming performance. Interestingly, those annotators whose performance was explained least by the set of predictor variables were also the ones with the least inter-annotator agreement (annotators 3 and 5 for syllable drumming, annotator 4 for word drumming, [1]). This seems to support the fact that inter-annotator agreement can be traced to well-established cues of prosodic prominence. Interestingly, the word drumming task shows considerably more individual strategies, which may be a consequence of the stronger cognitive processing necessary for this task [1]: In word drumming, annotators deliberately choose to rely on a certain set of cues in order to fulfill the task, while the more intuitive syllable drumming appears to rely on similar cues across annotators, despite them showing more individual variability.

4. General Discussion

We found that drumming-based annotation method reflects prosodic variability present on the signal level similarly as more conventional prominence annotations. This is encouraging as it indicates a comparability of research results gathered with two rather different methods. What is still unclear is whether the drumming-based method and the fine-grained auditory prominence annotations capture the same impressions as the very popular cumulative approach to prominence annotation. We feel that the currently existing methodological pluralism is a problem and there is a need to investigate the comparability of the various annotations schemata in more detail [15]. With respect to the investigation of listener strategies, we feel that our method appears to be fruitful and could verify the importance of already well-established cues to prosodic prominence. Also, the models show that most listeners rely on a blend of top-down and bottom-up cues in their prominence interpretation. Interestingly, the word level prominence drumming revealed more individual strategies compared to the syllable-based method, perhaps pointing to a higher degree of linguistic awareness. It is difficult to say at this point which method (intuitive syllable drumming, linguistically informed word drumming) is most adequate to get to the core of prominence processing in everyday communicative interaction. For a fuller understanding of individual listening strategies, further established prominence correlates (e.g. information structure, predictability, phrasal position, F0 shape, intensity, spectral emphasis) have to be included in the models as a next step, and more data needs to be annotated.

5. Bibliography

[1] B. Samlowski and P. Wagner, "Promdrum — exploiting the prosody-gesture link for intuitive, fast and fine-grained prominence annotation," in *Proceedings of Speech Prosody 2016*, 2016, p. p5.06.

[2] P. Wagner, Z. Malisz, and S. Kopp, "Speech and gesture in interaction: an overview," *Speech Communication*, vol. 57, pp. 209–232, 2014.

[3] B. Parrell, L. Goldstein, S. Lee, and D. Byrd, "Spatiotemporal coupling between speech and manual motor actions," *Journal of Phonetics*, vol. 42, pp. 1—11, 2014.

[4] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation based factors in the perception of prosodic prominence," in *Journal of Laboratory Phonology*, vol. 1, 2010, pp. 425–452.

[5] J. Terken, "Fundamental frequency and perceived prominence," *Journal of the Acoustical Society of America*, vol. 89, no. 4, pp. 1768–1776, 1991.

[6] S. Baumann and C. T. Röhr, "The perceptual prominence of pitch accent types in german," in *Proceedings of ICPhS 2015*, Glasgow, Scotland, 2015.

[7] K. de Jong, "The supraglottal articulation of prominence in english: Linguistic stress as localized hyperarticulation," *Journal of the Acoustical Society of America*, vol. 97, pp. 491–504, 1995.

[8] M. Vainio and J. Järvikivi, "Tonal features, intensity, and word order in the perception of prominence," *Journal of Phonetics*, vol. 34, no. 3, pp. 319–342, 2006.

[9] C. Widera, T. Portele, and M. Wolters, "Prediction of word prominence," in *Proceedings of Eurospeech*, vol. 2, Rhodes, Greece, 1997, pp. 999–1002.

[10] C. Gussenhoven and A. Rietveld, "Fundamental frequency declination in dutch: testing three hypotheses," *Journal of Phonetics*, vol. 16, pp. 355–369, 1988.

[11] D. Arnold, P. Wagner, and H. Baayen, "Using generelized additive models and random forests to model prosodic prominence in german," in *Proceedings of Interspeech 2013*, 2013, pp. 272–276.

[12] A. Eriksson, G. Thunberg, and H. Traunmüller, "Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing," in *Proceedings of EUROSPEECH*, Aalborg, Denmark, 2001, pp. 399–402.

[13] P. Wagner, "Great Expectations - Introspective vs. Perceptual Prominence Ratings and their Acoustic Correlates," in *Interspeech 2005, September, 4-8, Lisbon, Portugal*, 2005, pp. 2381–2384.

[14] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org/

[15] P. Wagner, A. Origlia, C. Avesani, G. Christodoulides, F. Cutugno, M. D'Imperio, D. Escudero Mancebo, B. Gili Fivela, A. Lacheret, B. Ludusan, H. Moniz, A. Ní Chasaide, O. Niebuhr, L. Rousier-Vercruyssen, A. C. Simon, J. Simko, F. Tesser, and M. Vainio, "Different parts of the same elephant: A roadmap to disentangle and connect different perspectives on prosodic prominence," in *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, Scotland, 2015.