PromDrum — Exploiting the prosody-gesture link for intuitive, fast and fine-grained prominence annotation

Barbara Samlowski[1], *Petra Wagner*[1,2]

[1]Bielefeld University, Faculty of Linguistics and Literary Studies, Germany
[2]Center of Excellence for Cognitive Interaction Technology (CITEC), Germany
petra.wagner@uni-bielefeld.de, barbara.samlowski@uni-bielefeld.de

Abstract

Most prominence annotation methods have certain drawbacks. Simple binary scales may be too coarse to capture fine-grained prominence differences, and multi-level annotation schemes have been shown to be time-consuming and difficult to use for non-expert annotators. This study proposes a novel method for fine-grained and fast prominence annotation by exploiting the prosody-gesture link. On a sentence-by-sentence basis, native German participants were instructed to listen to audio recordings and reiterate them by beating on an electronic drum pad either once per syllable (experiment 1) or once per word (experiment 2), modulating the strength of each beat according to how strongly the syllable or word stood out in the sentence. The velocity profiles of MIDI outputs were then interpreted as correlates of perceived prominence and compared with fine-grained prominence ratings by three expert annotators. While word-level drumming showed high correlations to conventional ratings for some of the subjects, inexperienced participants often had considerable difficulty performing the task. Syllable-level drumming, on the other hand, proved to be a time-efficient and intuitive method for experienced and naive subjects alike. Especially by pooling velocity results from several participants to create mean values, it was possible to maintain high levels of correlation with expert prominence ratings.

Index Terms: prominence, gesture-prosody link, drumming, annotation

1. Introduction

Although research on prosodic prominence has recently regained attention, there exists no standard or consensus approach for its annotation. Instead, a variety of annotation schemata have been proposed or used in the past, differing in (i) the level of annotation or prominence domain, (ii) the scale used for prominence annotation and (iii) the way of how prominence judgments are averaged and normalized across several listeners[1]. [2] suggest a multilevel scale of 31 levels of syllabic prominence. In a related approach, [3] introduce a continuous scale for prominence ratings, using GUI-based sliders to assess the prominence impressions for individual syllables. Other researchers have used fewer levels of prominence annotation, e.g. 11 [4], 4 [5], or 3 [6]. [7, 8] operationalize continuous prominence annotations as binary impressions of word prominence cumulated across several listeners. For an illustration of the most popular approaches to prominence analysis, cf. Figure 1. To this day, the majority of prominence studies rely on binary impressions of prominence (e.g. [9]). These simplistic approaches constrain any investigations of more fine-grained aspects of prominence, e.g. differences of word vs. sentence level stress, lexical class specific prominence, word-internal prominence relations or fine-grained aspects of prominence related to pragmatic functions.

Despite this striking heterogeneity within the field, comparatively few studies have been specifically dedicated on the evaluation of these competing approaches. [10] report on a high agreement between expert annotators trained on an annotation approach using 31 levels of prominence (Spearman-ρ between 0.7 and 0.8). [6] find a good agreement for expert annotators using 3 levels of prominence, while [11] argue that for word-based prominence distinctions, cumulating binary prominence impressions across several naive listeners will reach similar results as more fine-grained expert annonations. [12] systematically compared the efficiency of various multi-level prominence annoatation schemata (4, 7, 11 or 31 levels, continuous scale), gathering the prominence impressions using a slider-based GUI-approach. They concluded that multi-level schemata reflect richer impressionistic details (e.g. caused by contextual priming) and are not considerably more time-consuming than approaches with slightly fewer annotation levels. [13] showed that word-level prominence judgements are more reliable and in higher agreement with acoustic prominence correlates syllable-based ones, indicating that word level prominence judgements may be comparatively easier to annotate. However, for certain research prominence-related questions, e.g. sublexical prominence relations, word-level prominence annotations are inadequate. In their crowdsourcing study on Polish syllable prominence, [4] found a very high variation of prominence judgements across participants for a multi-level scale (11 levels). Their naïve participants also reported the task to be difficult and cumbersome, which is in line with the conclusions by [11] in favor of a simpler approach for naïve listeners. These investigations can be summarized as such:

1. Multi-level annoations may reveal more relevant phonetic detail than those using fewer levels.
2. Naïve annotators may have difficulties using these fine-grained scales.
3. Expert annotators reach good inter-annotator agreement.
4. Word and syllable prominence may be in need of different annotation procedures.

In this paper, we suggest a novel approach to prominence annotation which is (i) able to reflect the rich phonetic detail of a multi-level prominence impression, (ii) can be used by naïve listeners as well as expert annotators, (iii) allows for a quick and intuitive way of assessing prominence impressions, thus enabling an annotation of large amounts of recordings.

In the following, we are testing a method that aims to satisfy these quality criteria by exploiting the link between prosodic

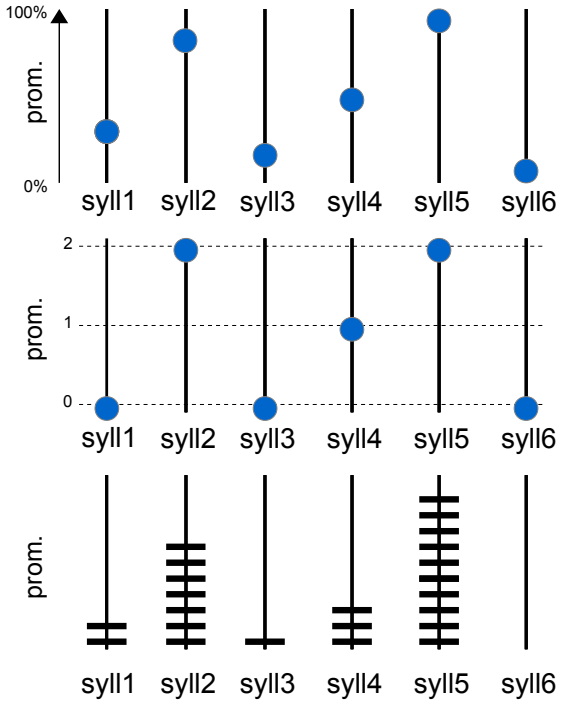

Figure 1: *Schematic overview of three popular prominence annotation methods (from top to bottom): (1) Fine-grained continuous prominence annotation. (2) Less fine-grained prominence annotation using a discrete (here: three) amount of distinct levels. (3) Prominence annotation based on cumulative binary impressions across several annotators.*

prominence and speech-accompanying gestures [14]. We know about a strong parallelism in prominence production and simultaneous manual beat gestures [15, 16] which develops early in language-acquisition [17]. [18] show that this speech-motor coupling is not constrained to temporal alignment or movement duration, but that verbal emphasis also influences the magnitude of manual movement in a co-speech tapping task. We therefore hypothesize that in a task where participants are asked to "repeat" a previously heard utterance by drumming, the resulting drumming intensity may provide a good indicator of their fine-grained prominence perceptions. We furthermore hypothesize that this task will be easy to carry out for naïve annotators without time-consuming prosodic training. Lastly, we hypothesize that due to the strong temporal speech-motor coupling, syllable prominence drumming is easier to perform than word prominence drumming, as at least in a language allowing for polysyllabic words, words provide considerably more temporal variation than syllables.

2. Methods

2.1. Material

The material for the drumming experiments was taken from the Bonn Prosodic Database [10], which contains annotated audio recordings of sentences read by three different speakers. The annotations for each recording include syllable prominence ratings on a 31-point scale performed by three expert annotators. Twenty sentences read by each of the three speakers were extracted from the database for the main experiment, with ten additional sentences serving as training material for the participants. In order to compare the prominence annotations with the experiment results on word level as well as syllable level, the maximum syllable prominence estimate of each word was interpreted as the word-level prominence rating.

When choosing which sentences to use for the experiments, care was taken to ensure that on the one hand that there was no strong disagreement across annotators concerning how syllable prominence should be rated, and that on the other hand the three realizations of each sentence differed somewhat from one another in terms of prominence patterns. This way, it was possible to use the sentences to examine influences of different sentence realizations as well as of the top-down expectations participants had concerning how they thought the sentence in question should be produced.

2.2. Participants

Ten native German speakers took part in this study (3 men, 7 women, ages ranging between 20 and 58). Of these, five were presented with the syllable drumming task while the other five were instructed to drum once per word. Although nearly all participants came from a linguistic background, only two of them had some training in prosody annotation.

2.3. Procedure

The experiments were performed with an electronic drum pad (Alesis SamplePad) in a sound-treated studio at Bielefeld University. Participants were presented first with the ten training sentences and afterwards with the sixty sentence recordings of the main experiment. The order of the training and test sentences was randomized for each participant, taking care that repetitions of the same sentence by different speakers were maximally far apart from each other. The participants were instructed to listen to each sentence over headphones and then beat on the electronic drum pad once per perceived syllable (experiment 1) or once per perceived word (experiment 2), using a standard rock music drum stick (Maple, 5B). They were allowed to listen to sentence recordings again and/or repeat their drumming if they were unhappy with their performance. While drumming, participants listened to their drumming performance via headphones. The two prosodically trained participants were both assigned to the word annotation task, as a result of the random distribution.

Audio and MIDI output of the drum pad were recorded as well as the sentence stimuli which were played to the participants. The drummed sentences were semi-automatically annotated using the audio analysis and segmentation program Praat [19]. By extracting the information encoded in the MIDI output[20] and comparing the MIDI time stamps with the relevant drum sounds in the audio file, it was possible to determine for each of the drum beats the velocity information stored in the MIDI file, i.e. the speed with which the drum was hit and on which the intensity of the output sound was based.

3. Results

3.1. Coverage and Time Consumption

As drum beats had to be attributed to the individual syllables / words, it was only possible to interpret participants' responses

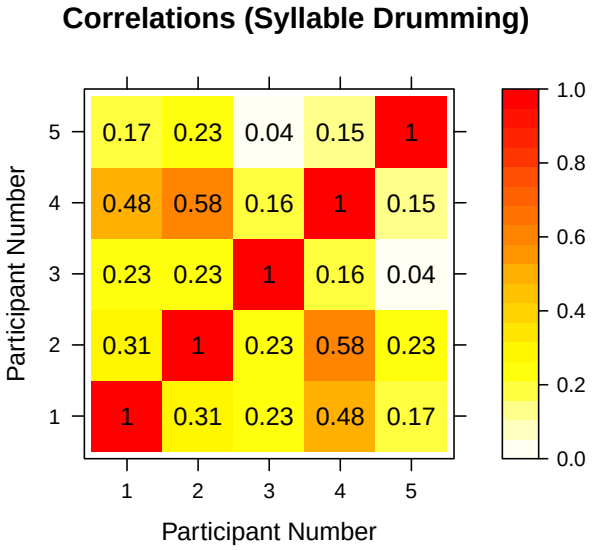

Figure 2: Median Pearsson correlation between velocity results from different participants (experiment 1)

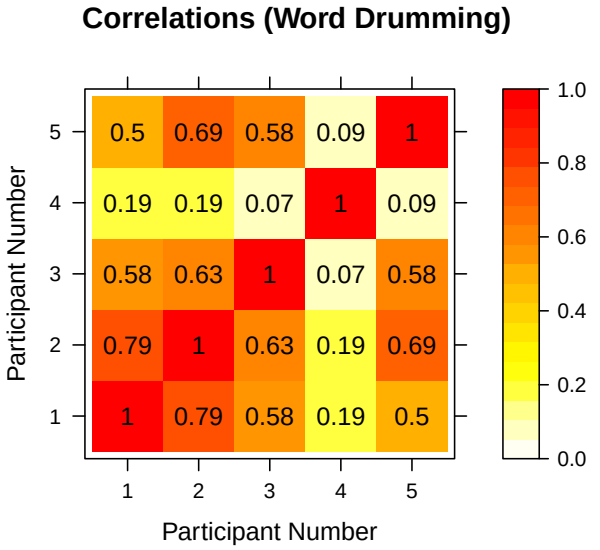

Figure 3: *Median Pearsson correlation between velocity results from different participants (experiment 2)*

if the number of drum beats matched the number of words or syllables in the sentence. For the syllable drumming task, 16 of the 300 items (5%) had to be left out of the analysis for this reason. The word drumming task resulted in a smaller output. Here, 38 of the 300 items (13%) could not be interpreted. Word level drumming appeared to be considerably more difficult to perform using this task.

For both conditions, participants were able to go through the drumming task very quickly and only occasionally demanded to hear or drum a particular sentence again. Consequently, the average time consumption was considerably shorter than for fine-grained prominence annotation by conventional annotation methods. Including the time used for the training sentences as well as repetitions or sentences which had to be discarded from analysis, the average time consumption per analyzed word varied between 1.8 and 3.6 seconds for the syllable drumming task, and between 2.5 and 4.6 seconds for the word drumming task, indicating a higher cognitive load for word drumming.

3.2. Inter-Participant Correlation

The average strength of the drum beats as well as their variability can differ across individual sentences and participants. For this reason it is useful to normalize prominence ratings before comparing them. Following the suggestions made by [21], we z-score normalized the drumming data by subtracting the mean velocity for each item from the average values and dividing the result by the mean absolute deviation across all items produced by the same participants. The similarity between normalized velocity results from different participants was calculated by determining Pearsson correlations for each sentence separately and then computing their median value. Similarities between normalized velocity results from different participants were calculated by determining Pearsson correlations for each sentence separately and then computing median values for each comparison.

The analysis shows comparatively low correlation values between participants who were presented with the syllable drumming task (see Figure 2). Median values varied between 0.04 and 0.58. This suggests that individual participants payed attention to different cues when interpreting how strongly syllables stood out in the sentence. Although the word drumming experiment appeared to be more difficult and error-prone than the syllable drumming experiment, the correlations between participants who did manage to perform this task were higher than the results for syllable drumming, reaching values of up to 0.79 in a comparison between the two expert annotators (see Figure 3).

3.3. Correlations to prominence annotations

Apart from the issue of how consistent the drumming results were across participants, it is important to examine how well they correspond to conventional fine-grained syllable prominence ratings. This question was investigated by comparing the normalized velocity results with the prominence ratings presented in the database. In preparation of the analysis, these ratings were normalized in a similar manner as the velocity results were, by subtracting the mean rating for each sentence from the absolute values and then dividing by the absolute mean deviation of all investigated items which were annotated by the person in question. Since annotators may not always focus on the same prominence cues, mean estimates from multiple participants may in fact show a more representative picture than ratings by individual participants. For this reason, our investigation included comparisons with mean estimates calculated from the normalized prominence ratings and velocity values. Correlations were computed for each sentence recording separately, and the median of each comparison group was reported.

In the syllable drumming task, median correlations between participants and annotators varied from 0.24 to 0.59 (see Table 1). As the sentences chosen for the experiment by design showed a high inter-annotator agreement in terms of prominence ratings, correlations with experiment participants did not

	Participants (syllable drumming)					
	1	2	3	4	5	Mean
Annotator 1	0.50	0.56	0.29	0.68	0.25	**0.73**
Annotator 2	0.55	0.56	0.31	0.68	0.24	**0.72**
Annotator 3	0.48	0.59	0.29	0.68	0.32	**0.71**
Mean	0.54	0.60	0.27	0.73	0.36	**0.74**

Table 1: *Pearsson correlations between prominence annotations and velocity results (experiment 1)*

	Participants (word drumming)					
	1	2	3	4	5	Mean
Annotator 1	0.79	0.75	0.62	0.27	0.53	**0.78**
Annotator 2	0.74	0.75	0.66	0.26	0.59	**0.81**
Annotator 3	0.83	0.84	0.64	0.31	0.58	**0.86**
Mean	0.83	0.82	0.64	0.23	0.60	**0.83**

Table 2: *Pearsson correlations between prominence annotations and velocity results (experiment 2)*

differ much across the three annotators. Comparisons of mean estimates based on multiple annotators or participants tended to lead to higher correlations than comparisons between individual annotators or participants. The highest median correlation (0.74) was found for a comparison between mean prominence rating on the one hand and mean drumming velocity on the other.

Just as correlations between results from individual participants were higher for the word drumming task than for the syllable drumming task, the word drumming task also tended to result in higher correlations to conventional prominence estimates (see Table 2). Again, comparisons with mean prominence or velocity values often resulted in higher correlation values than comparisons between individual participants and annotators, which varied between 0.27 and 0.79. In sentence-by-sentence comparisons, half of the items had a correlation between mean velocity and mean prominence which was higher than 0.83.

4. Discussion

As hypothesized, we found that the drumming task allows for a very fast and intuitive way to gather listeners' impressions of previously heard utterances. In fact, the procedure allows for an annotation speed sufficiently close to real time and with a training phase of a few minutes only. Due to this speed, the proposed method seems even suitable for the fine-grained manual prominence annotation of large corpora.

As shown previously for conventional prominence annotations by laypersons, the word-level prominence annotations reach higher inter-annotator agreement, even with a drumming task [13]. This contradicts our hypothesis of the syllable drumming task to be not only easier to do but also being more consistent. Still, the word drumming was considerably more error prone and time consuming than the syllable drumming, leading to believe that the coordination of hand movements and words was more difficult and needed more cognitive resources. The reason for this may be that speakers may be used to a prosody-gesture alignment within highly prominent words, but within less prominent words (cf. Introduction and references therein). In the syllable task, the temporal alignment may be easier as the perceptual center appears to be a good candidate for prosody-gesture alignment below the word [22], thus providing a suitable temporal scaffold for speech-motor coordination.

Another possibly reason for the word task being more consistent across participants would be that it is subject to more top-down processing than syllable drumming. However, this claim would need further empirical substantiation. A possible criticism against the method would be the comparatively low correlations across individual participants, especially in the syllable drumming task. However, since the method does not try to define a standard for prosodic expert annotation with a set of well-defined annotation criteria, we do not consider this a major disadvantage. Rather, our approach welcomes the inter-individual variety, as different listeners may pay attention to various aspects of phonetic detail, all of which may contribute to the overall impression. As the averaged impressions thus gathered showed high correlations with expert annotations and similar correlations as can be reached between expert annotators likewise we still consider our approach as a very promising one. Certainly, a corpus annotated this way may not rely on one or two annotators only, as it needs to take into account several listening strategies. Possibly, a longer training of experts would ultimately yield higher inter-rater correlations, but possibly, this type of a "trained ear" is not at all advantageous as it may highlight certain phonetic features known for prominence-lending effects, while ignoring those of which we are not yet aware. Thus, having naïve annotators may actually be a way to circumvent the problem of circular reasoning in the investigation of prosodic function-signal relationships. Further research should investigate the listening strategies across individual participants by measuring their relationship to signal- or expectancy-based correlates of prosodic prominence.

Another issue that needs further investigation lies in the definition of how many annotators are minimally needed in order to gain a suitably reliable annotation using this approach. Possibly, fewer than five annotators may even be sufficient for gaining impressions comparable with results relying on more time consuming methods. Besides, it needs to be explored how our approach relates to cumulative binary impressions of multiple naïve annotators [7].

We believe our method to be interesting not only for laboratory investigations: As the MIDI-baed velocity output practically models the drumming intensity, it should be possibly to transfer the method to an analogue instrument, e.g. a wooden drum stick and a table serving as a "drum pad". This would enable fine-grained prominence investigations in field work situations where access to electricity, complex laboratory settings and expert annotators is limited.

5. Conclusions

By exploiting the prosody-gesture link, drumming provides a fast, intuitive and exact method for fine-grained prominence annotations which are difficult and time-consuming to gather using conventional annotation settings. The annotations can be easily performed by naïve annotators, with word prominence annotations being more error prone than annotations of syllable prominence. However, word prominence annotations reach higher inter-annotator correlations, which corroborates previous findings. Due to its simplicity and speed, the method is useful in larger annotation tasks and may be of use in prosodic field work.

6. Acknowledgements

The authors would like to thank the participants of this study.

7. References

[1] 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.

[2] G. Fant and A. Kruckenberg, "Preliminaries to the study of swedish prose reading and reading style," *STL-QPSR*, vol. 30, no. 2, pp. 1–80, 1989.

[3] 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.

[4] Z. Malisz, A. Cwiek, and P. Wagner, "The perception of prominence by polish native speakers: a crowdsourcing study," in *Poznan Linguistics Meeting 2015*, 2015.

[5] F. Kügler, B. Smolibocki, D. Arnold, S. Baumann, B. Braun, M. Grice, S. Jannedy, J. Michalsky, O. Niebuhr, J. Peters, S. Ritter, C. T. Röhr, A. Schweitzer, K. Schweitzer, and P. Wagner, "Dima - annotation guidelines for german intonation," in *Proceedings of the 18th International Congress of Phonetic Sciences*, 2015, p. 317.

[6] A. Lacheret, A. C. Simon, J. Goldman, and M. Avanzi, "Prominence perception and accent detection in french: from phonetic processing to grammatical analysis," *Language Sciences*, vol. 39, pp. 95–106, 2013.

[7] 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.

[8] C. Whightman, "Perception of multiple levels of prominence in spontaneous speech," in *ASA 126th Meeting*, Denver, 1993.

[9] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 690–701, 2007.

[10] T. Portele, B. Heuft, C. Widera, P. Wagner, and M. Wolters, "Perceptual prominence," in *Speech and Signals*, W. Sendlmeier, Ed. Hektor, Frankfurt a. M., 2000, pp. 97–115, festschrift for Wolfgang Hess on the occasion of his 60th birthday.

[11] C. Jensen and J. Tøndering, "Choosing a scale for measuring perceived prominence," in *Proceedings of Interspeech*, Lisbon, Portugal, 2005, pp. 2385–2388.

[12] D. Arnold, P. Wagner, and B. Möbius, "Evaluating different rating scales for obtaining judgments of syllable prominence from naïve listeners," in *International Congress of the Phonetic Sciences*, 2011, pp. 252–255.

[13] D. Arnold, B. Möbius, and P. Wagner, "Comparing word and syllable prominence rated by naive listeners," in *Proceedings of Interspeech 2011*, 2011, pp. 1877–1880.

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

[15] D. Loehr, "Temporal, structural, and pragmatic synchrony between intonation and gesture," *Laboratory Phonology*, vol. 3, no. 1, 2012.

[16] N. Mendoza-Denton and S. Jannedy, "Semiotic layering through gesture and intonation: A case study of complementary and supplementary multimodality in political speech," *Journal of English Linguistics*, vol. 39, no. 3, pp. 265–299, 2011.

[17] N. Esteve-Gibert and P. Prieto, "Infants temporally coordinate gesture-speech combinations before they produce their first words," *Speech Communication*, vol. 57, pp. 301–316, 2014.

[18] 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.

[19] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[20] J. Walker, "Midicsv: Convert midi files to and from csv," 2008. [Online]. Available: http://www.fourmilab.ch/webtools/midicsv/

[21] C. Sappok and D. Arnold, "More on the normalization of syllable prominence ratings," in *Proceedings of Interspeech 2012*, Portland, OR, 2012, pp. 2418–242.

[22] T. Leonard and F. Cummins, "The temporal relation between beat gestures and speech," *Language and Cognitive Processes*, vol. 26, p. 1295Á1309, 2010.