How Much Prosody is Needed? -
Perception of Prosodic Grouping in Gated Ambiguous Name Sequences

Marie Hansen*, Valentina Aivazian*, Clara Huttenlauch, Carola de Beer, Sandra Hanne & Isabell Wartenburger (* equal contribution)
SFB 1287, Department of Linguistics, University of Potsdam
marie.hansen@uni-potsdam.de

Background: Prosodic cues successively help to disambiguate incoming information in spoken language perception [1, 3]. In structurally ambiguous coordinate structures, such as name sequences, intonation-phrase boundaries (IPB, see example (I)) can indicate the intended grouping of constituents [e.g., 2, 4]. Studies on coordinate structures often focus on the realization of the prosodic cues directly at the IPB, as the strongest prosodic cues are located there (Fig. 1). For perception, however, the global prosodic contour matters [e.g., 2, 3] and speakers also modulate prosodic cues located earlier, that is before the IPB in the utterance [2, 4]. So far, it is less clear to what extent perception of constituents is influenced by prosodic cues that are located before the IPB.

Aim: The current study makes use of a gating paradigm to test if it is possible to detect an intended grouping in a coordinate structure by exploiting prosodic information already before the IPB. On this account, human listeners are compared to machine learning (ML) models (see p. 3) to elucidate whether prosodic cues before the IBP are in principle informative and to describe differences and similarities of humans' decisions and ML models' predictions.

Method: Stimuli consisted of six different name sequences with three disyllabic, trochaic German names, respectively, that were coordinated by und ("and"). Stimuli appeared in two conditions, one with an intended internal grouping of the first two names (I) and one without grouping (II):

(I) (Name1 and Name2)$_{IPB}$ (and Name3) (see Fig. 1)
(II) (Name1 and Name2 and Name3) (see Fig. 2)

Recordings of four different speakers were used with 24 recordings per condition, yielding a total number of 192 stimuli, for which the identifiability to the respective condition had been verified in a prior study. Human listeners: Each of the stimuli was split into seven parts (gates) and presented to naïve participants (n= 43) successively with increasing length of utterance and amount of prosodic information (Gate 1: "Le" | Gate 2: "Leni" | Gate 3: "Leni und"… etc., see Tab. 1 and Fig. 1 and 2). Participants performed a forced-choice decision task and assigned stimuli to conditions (with (I) or without (II) grouping) via button press. Response accuracy and reaction times were measured and analyzed.
ML models: Acoustic measures of prosodic cues regarding duration and f0 each on Name1 and Name2 (f0-rise, final lengthening, and pause) were introduced to the linear models incrementally (Tab. 1). Hence, the incremental steps of information given to humans and ML models were mostly similar, but the order of adding them to the models was slightly different.

Results and Discussion: Preliminary results of the human listeners show that accuracy was above chance already at gate 3 on average. However, only 28 of 43 participants contributed to this result. The number of participants scoring above chance rises only at gate 5, after the IPB. This was mirrored by the incremental ML models: the model that incorporated all cues on Name1 as well as on Name2 (ML model 6) could best discriminate between conditions (Tab. 1). Subgroups among human listeners might have used different strategies that lead to differences in accuracy scores in earlier gates. We will compare results from the human population to incremental ML models and further investigate performance by means of a subgroup-analysis, an analysis of reaction times, and by modeling the decision process with Drift Diffusion Models [5].

References

[1] Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience*, 30, 1–31.

[2] Kentner, G., & Féry, C. (2013). A new approach to prosodic grouping. *The Linguistic Review*, 30 (2), 277–311.

[3] Frazier, L., Carlson, K., & Clifton, C. Jr. (2006). Prosodic phrasing is central to language comprehension. *Trends in Cognitive Science*, 10, 244–249.

[4] Huttenlauch, C., de Beer, C., Hanne, S., & Wartenburger, I. (submitted). Production of prosodic cues in coordinate name sequences addressing varying interlocutors.

[5] Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20(4), 873–922.

[6] Pedregosa, F., Varoquaux, G., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Stimulus	Name1				and	Name2				and	Name3	
	Le	ni			und	Mim	mi			und	Ma	nu
Prosodic cues		f0	leng	pau		leng	f0	leng	pau			
Gates presented to human listeners	1	2			3	4	5			6	7	
ML models	1	2		3		4	5		6			

Table 1 Gated stimulus presentation and corresponding incremental ML models. Gates were similar for participants and incremental models. leng= duration of the second syllable; f0= f0-rise; pau= pause; IPB containing the strongest prosodic cues in the condition with grouping (II); Gate at which accuracy was above chance level in most (n= 42 of 43) human listeners; ML model that best discriminated between conditions

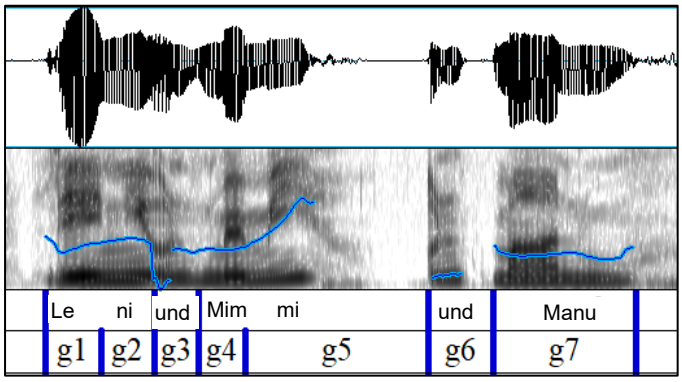

Figure 1 Spectrogram and gates (g1 – g7) of an example stimulus with grouping (I)

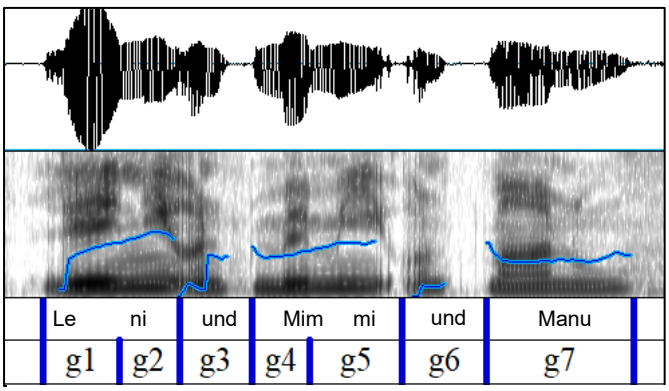

Figure 2 Spectrogram and gates (g1 – g7) of an example stimulus without grouping (II)

ML models characteristics: Linear classification models were used to classify the stimuli into name sequences with or without a grouping. Logistic regression was chosen as the optimality criterion. In particular the logistic loss was minimized.

$$Risk = \sum_i log\left(1 + exp(-y_i w^T x_i)\right) + w^T w$$

Here, y_i denotes the outcome of the *i*-th trial, $\boldsymbol{x_i}$ the features of the *i*-th trial and $\boldsymbol{w}$ the selected vector of weights. The prediction of the model is 1 (with grouping) if $w^T x_i > 0$ and -1 (without grouping) otherwise.
This was implemented using the class `LogisticRegressionCV` of the Python library Sklearn [6]. The parameter λ is determined by this class using cross validation. To verify the quality of this classifier, repeated k-fold cross validation was used with *k* = 10-fold splits and 2 repetitions. This type of cross validation splits the sample set into 10 subsets (folds) and each of the subsets is used once for testing while the remaining 9 subsets are used for training the model. This process is repeated 2 times and the average is calculated.
Variables (measured prosodic cues) were added to the models incrementally: with each new model more and more variables were introduced to imitate how they were introduced to the human participants with gates, resulting in the creation of 6 models with increasing accuracy reaching 98% in the last model which uses all the available cues (Fig. 3).

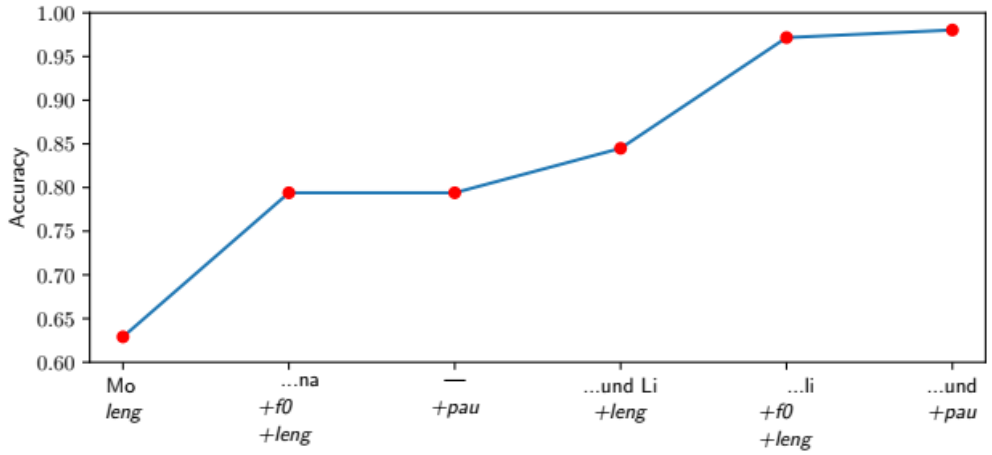

Figure 3 Incremental models' accuracy increases with more cues, reaching 98% in the last model. An example sentence is used to illustrate the position of added variables in the sentence. These models correspond to the ML models row in Tab. 1.