Using generalized additive models and random forests to model prosodic prominence in German

Denis Arnold[1], *Petra Wagner*[2], *R. Harald Baayen*[1]

[1] Quantitativ Linguistics, University of Tübingen, Germany
[2]Faculty of Linguistics and Literature, University of Bielefeld, Germany
denis.arnold@uni-tuebingen.de, petra.wagner@uni-bielefeld.de, harald.baayen@uni-tuebingen.de

Abstract

The perception of prosodic prominence is influenced by different sources like different acoustic cues, linguistic expectations and context. We use a generalized additive model and a random forest to model the perceived prominence on a corpus of spoken German. Both models are able to explain over 80% of the variance. While the random forests give us some insights on the relative importance of the cues, the general additive model gives us insights on the interaction between different cues to prominence.
Index Terms: prosody, prominence, gam, random forests

1. Introduction

Among the plethora of factors influencing prosodic prominence, the most important ones have been identified as being duration, F0, and various measures of spectral intensity. There is quite a dispute about which acoustic measure contributes the most to the impression of prominence. [1, 2, 3] There are implementation of automatic prominence detection algorithms [4, 5] that are purely based on acoustic input, typically using a parameterization of the F0 contour. The algorithms reached between 76 and 81 % correct classification. Important linguistic predictors for prominence perception are part-of-speech and phonologic accentuation. [6] manipulated expectations of prominence using priming and showed that these expectations modulate perceived prominence of the same acoustic signal. [2, 7] predicted prominence categories for syllables purely based on linguistic information. There is substantial agreement in the literature that the impression of prominence is influenced by several sources (see for example [8]). However, an approach to integrate the influences of these various sources is missing in the literature.

2. Method

2.1. Bonner Prosodische Datenbank

The *Bonner Prosodische Datenbank* (hence forth BPD) is described in [9]. It features recordings from 3 speakers in different speaking styles and has a total of 10587 syllables. The database was annotated for prominence by three annotators. Table 1 shows the correlations between the three raters as well as the correlation with the median of the three annotations, indicating good inter-rater agreement. In addition to the prominence annotation the BPD features a broad array of linguistic annotations for each syllable as well as acoustic features like syllable duration and a parametrization of F0 [10]. The material contains different speaking styles such as questions, answers, commands and read stories. The BPD has served as a data basis for serval studies [9, 5, 7, 11].

Table 1: *Correlations between the prominence ratings of the three raters ACA0 - ACA2 as well as to the median of their ratings ACAT*

	ACAT	ACA0	ACA1	ACA2
ACA0	0.91			
ACA1	0.87	0.74		
ACA2	0.93	0.82	0.77	

2.2. Modeling

First we would like to describe the cues we use to predict prominence. In the BPD ACAT referres to the median prominence ratings, with values ranging from 0 (not prominent) to 31 (very prominent). This is the response variable in our analyses. PACAT and FACAT denote the median of the rating prior (PACAT) and following (FACAT) the current syllable. The range is the same as for ACAT. WORTART is a part-of-speech tag that can take one of eight different values. ACCE is a binary variable specifing whether a syllable can be accented. SILBE provides the identity of the syllable and NUCLEUS the identity of its nucleus. There are 848 different syllables and 31 different nuclei in the corpus. SYLDUR it the duration of the syllable in ms. We log transformed SYLDUR (SYLDURL). AMPLI, ANSTIEG, FALL and DELAY are parameters that describe the F0 contour as described in [10]. AMPLI is the amplitude of the F0 peak normalized to the speakers' maximum and minimum F0 baselines. AMPLI is speaker dependent and can take values between 0 and 100. We made use of two complementary statistical technics; generalized additive models and random forests.

2.2.1. GAMs

We used the mgcv package [12] for R [13]. For an introduction to using mgcv see [14]. A generalized additive model (GAM) uses smooth functions to model non linear functional relations between predictors and the response. Multiple predictors may be combined with the help of tensor product smooth to model wiggly regression (hyper)surfaces. We used ACCE as a factorial predictor, random slopes for SILBE and WORTART, and tensor smooth to model the interactions between SYLDURL, AMPLI and PACAT, AMPLI and FACAT, PACAT and FACAT, SYLDURL and PACAT, and SYLDURL and FACAT. Other predictors did not reach significance.

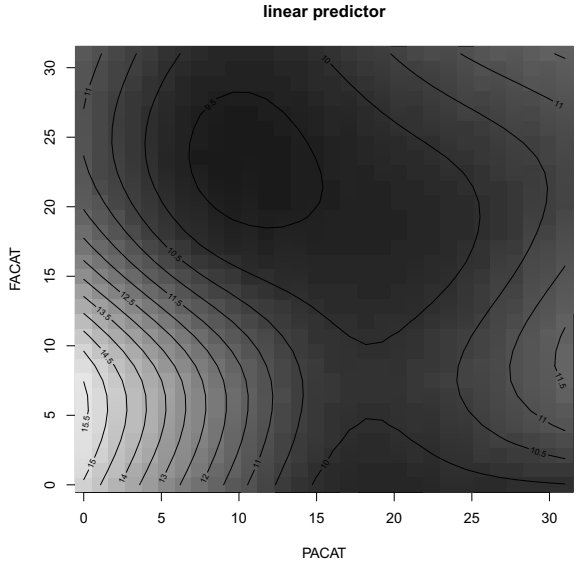

Figure 1: *Contour plot for the fitted linear predictor for prominence as a function of the preceding prominence value (PACAT) and the following prominence Value (FACAT).*

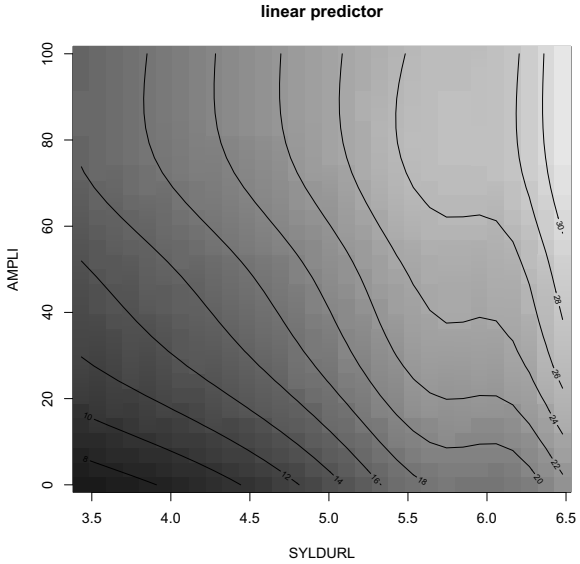

Figure 2: *Contour plot for the fitted linear predictor for prominence as a function of the speaker normalized F0 amplitude (AMPLI) and the logarithmic transformed syllable durations (SYLDURL).*

2.2.2. Random forest

We used the implementation of random forests in the party package [15, 16, 17] for R [13], which implements forests of conditional inference trees. The package provides variable importance measures. The models predicted values are obtained using a voting scheme defined over the trees. We used NUCELUS, WORTART, ACCE, AMPLI, SYLDURL, PACAT, and FACAT as predictors to prominence. Other variables were considered but turned out not to contribute. We used the default settings to train the random forests. We computed AUC-based variables importance since they seem to be more robust [18].

3. Results

3.1. GAMs

The fitted model accounted for 82 % of the variance in the data. Table 2 shows the coefficients for the parametric predictors in the GAM. Table 3 lists the numbers of degrees of freedom invested in the smooth terms of the GAM, and associated F-statistics. Figure 1 shows the fitted prominence surface as a function of FACAT and PACAT. Lighter shades of gray predict a higher prominence value while darker shades predict lower prominence values. One can see that syllables in a context of low prominent syllables are more likely to get high prominence values. Figure 2 presents the interaction between AMPLI and SYLDURL. While higher values of both predictors result in higher prominence, the effect of AMPLI is strongest for lower values of SYLDURL. The interaction of PACAT and AMPLI are presented in figure 3 and interaction between PACAT and SYLDURL in figure 4. Comparing figures 3 and 4 we find that preceding prominence modulates the contribution of AMPLI more strongly than the contribution of SYLDURL. Figure 3 shows that the effect of AMPLI vanishes for higher values of PACAT. Figure 4 suggests a u-shaped effect of PACAT that becomes stronger for longer syllables. Figures 5 and 6 present the smooth surfaces for two more interactions involving FACAT. Figure 5 indicates high prominence for high values of AMPLI and lower values of FACAT. Figure 6 indicates that the effect of FACAT is somewhat reduced for shorter syllables.

3.2. Random forest

The correlation between the predicted prominence of the random forest and the median of the prominence ratings ACAT is 0.922. The model explains 85% of the variance in the data. This correlation is higher than the inter-rater correlations and within the same range as the correlations between the individual ratings and the median rating (ACAT see table 1). Figure 7 summarizes the AUC-based variable importances [18]. The amplitude of the F0 peak receives the highest importance value. The next highest rating predictors represent linguistic properties, followed by syllable duration. The context variables PACAT and FACAT are slightly more important than the predictor specifying whether the syllable can carry an accent. Figure 8 shows the prominence ratings for the first phrase in the corpus "Lauter bitte" - "Louder, please". The first three panels show the observed ratings of the three raters ACA0 - ACA2. The next panel shows the median (ACAT). The last panel shows the prominence as predicted my the random forest. The differences between the three raters are

Table 2: *Coefficients for the linear predictors in the generalized additive model fitted to the median prominence ratings. All p-values < 0.000001.*

	Estimate	Std. Error	t value
Intercept	7.5808	0.4053	18.7
ACCE	5.4562	0.3326	16.4

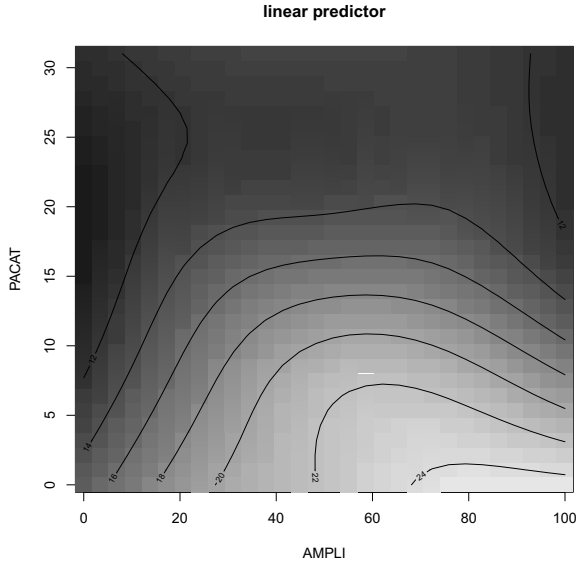

Figure 3: *Contour plot for the fitted linear predictor for prominence as a function of the preceding prominence value (PACAT) and speaker normalized F0 amplitude (AMPLI). The black lines connect identical linear predictors to syllable prominence. Dark areas indicate a lower contribution to prominence than lighter areas.*

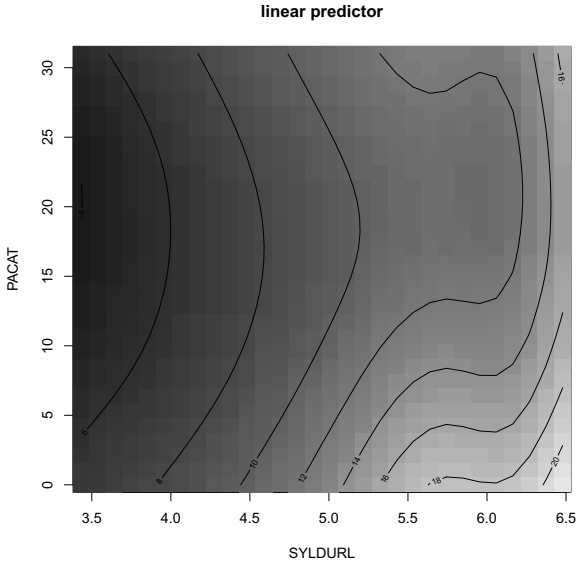

Figure 4: *Contour plot for the fitted linear predictor for prominence as a function of the preceding prominence value (PACAT) and the logarithmic transformed syllable durations (SYLDURL). The black lines connect identical linear predictors to syllable prominence. Dark areas indicate a lower contribution to prominence than lighter areas.*

greater than the difference between the ACAT ratings and the prediction of the random forrest.

4. Discussion

The random forest gives us the best fit to the data with an explained variance of 85 %. Our model accounts for 17 % more explained variance than other reported models in the literature. The AUC-based variable importance suggests that F0 is the most important cue to prominence for the corpus. This result fits well with other studies reporting F0 as the most important cue to prominence [2]. However the BPD does not provide intensity measures, and it is possible that intensity is a more powerful predictor to prominence than F0 [1]. For future work we recommend the addition of intensity measures to the BPD. Of the linguistic predictors NUCLEUS and WORTART are well supported in contrast to the variable ACCE, the indicator for wether a syllable can be accented. The GAM supplements the random forests by clarifying how the different numerical predictors interact. This is especially important for the interactions involving preceding and following context (PACAT and FACAT). The influence of the context has been observed before [19]. The present study is the first to clarify that preceding and following context codetermines the effects of F0 (AMPLI) and syllable duration (SYLDURL).

Table 3: *Coefficients for the smooth terms and tensors in the generalized additive model fitted to the median prominence ratings. All p-values < 0.000001.*

	edf	Ref.df	F
SILBE	590.675	846.00	11.276
WORTART	18.287	20.00	419.222
SYLDURL,AMPLI	6.347	6.92	7.025
AMPLI,PACAT	10.518	20.00	13.242
AMPLI,FACAT	10.250	20.00	18.245
FACAT,PACAT	8.316	20.00	56.808
SYLDURL,PACAT	6.062	16.00	85.368
SYLDURL,FACAT	11.742	13.04	9.058

5. Conclusions

The present study shows recent statistical methods offer more precise insights into the role of acoustic, linguistic and context factors on the perception of prominence. With both models we were able to explain more variance than any model in the literature. Random forests offer excellent prediction accuracy as well as a means for evaluating variable importance. GAMS are a good choice for the modeling of non linear interactions. The present study documents such nonlinear interactions — the reduction in AIC compared a standard linear model with multiplicative interactions is no less than 1917 — most of which involve the prominence of the preceding and following syllables. Further research will have to verify whether the observed interactions are robust and replicable. A further challenge for future research is to clarify why the nonlinear regression surfaces arise.

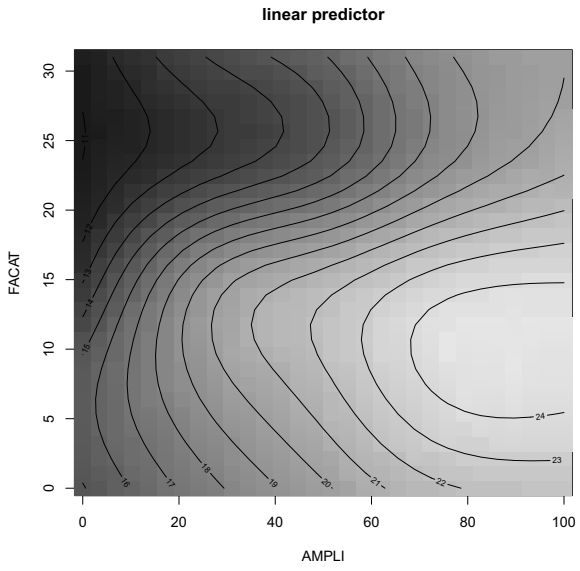

Figure 5: *Contour plot for the fitted linear predictor for prominence as a function of the following prominence value (FACAT) and speaker normalized F0 amplitude (AMPLI). The black lines connect identical linear predictors to syllable prominence. Dark areas indicate a lower contribution to prominence than lighter areas.*

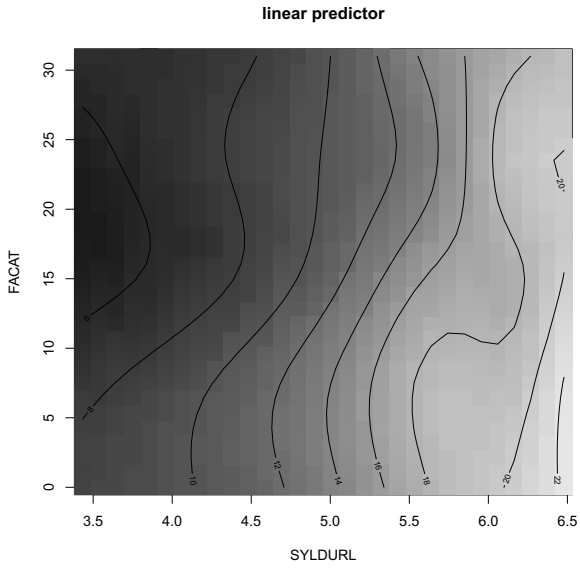

Figure 6: *Contour plot for the fitted linear predictor for prominence as a function of the following prominence value (FACAT) and the logarithmic transformed syllable durations (SYLDURL). The black lines connect identical linear predictors to syllable prominence. Dark areas indicate a lower contribution to prominence than lighter areas.*

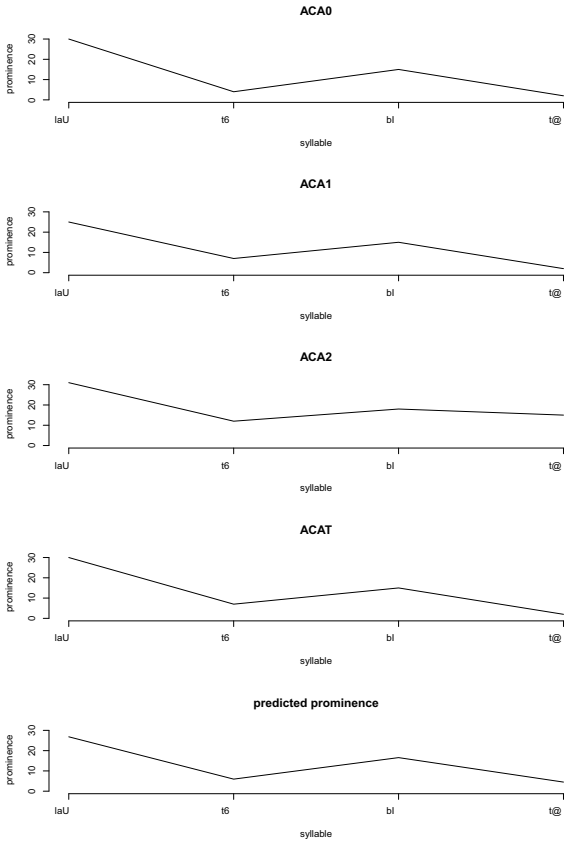

Figure 7: *Prominence ratings of the three raters, the median and the predicted prominence of the random forrest for the first phrase in the BPD. The phrase is transcribed in sampa: "Lauter bitte" - "Louder, please"*

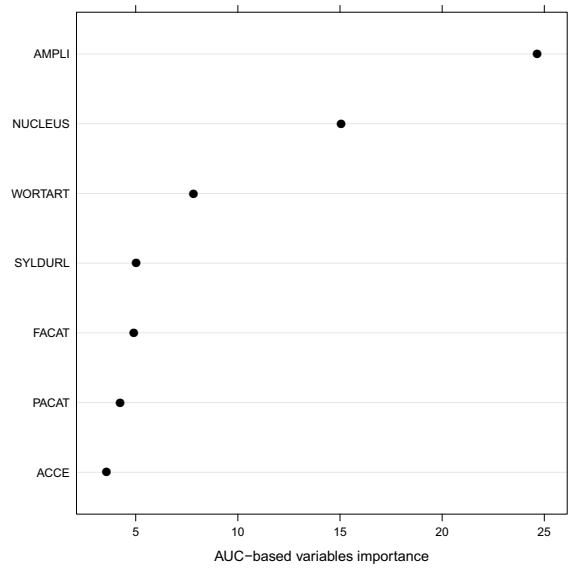

Figure 8: *AUC-based variables importance for the random forest. Higher values show more important variable importance.*

6. References

[1] Kochanski, G., Grabe, E., Coleman, J., and Rosner, B., "Loudness predicts prominence: fundamental frequency lends little". Journal of the Acoustical Society of America 118, 1038-1054, 2005.

[2] Streefkerk, B., "Prominence - Acoustic and lexical/syntactic correlates", Utrecht: LOT, 2002.

[3] Eriksson, A., Thunberg, G. und Traunmüller, H. (2001), Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing. Proceedings of Eurospeech 2001, 399-402.

[4] Wang, D. and Narayanan, S.; , "An Acoustic Measure for Word Prominence in Spontaneous Speech", IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.2, 690–701, 2007.

[5] Tamburini, F. 'Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system", Proceedings of Eurospeech 2003, 129–132, 2003.

[6] Arnold, D., Wagner, P. and Möbius, B., "The effect of priming on the correlations between prominence ratings and acoustic features", Proceedings of Speech Prosody 2010, Chicago, 2010.

[7] Wagner, P., "Wahrnehmung und Vorhersage deutscher Betonungsmuster", Universität Bonn. PhD-Thesis, 2002. Online: http://hss.ulb.uni-bonn.de/2002/0054/0054.htm, accessed on 29 Mar 2011.

[8] Watson, D. G., "The many roads to prominence: Understanding emphasis in conversation", The Psychology of Learning and Motivation, 52, 163-183, 2010.

[9] Heuft, B. "Eine prominenzbasierte Methode zur Prosodieanalyse und -synthese", Frankfurt: Peter Lang, 1996.

[10] Heuft, B., Portele, T., Höfer, F., Krämer, J., Meyer, H., Rauth, M., and Sonntag, G., "Parametric description of F0-contours in a prosodic database", Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm KTH, 378-381, 1995.

[11] Wagner, P. Breuer, S., and Stöbe, K., "Automatische Prominenzetikettierung einer Datenbank für die korpusbasierte Sprachsynthese", DAGA - Fortschritte der Akustik, Oldenburg, 2000.

[12] Wood, S.N., "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models". Journal of the Royal Statistical Society (B) 73(1):3-36, 2011.

[13] R Development Core Team. 'R: A language and environment for statistical computing.", R Foundation for Statistical Computing, Vienna, 2012.

[14] Wood, S. "Generalized additive models: an introduction", R. Chapman& Hall/CRC, 2006.

[15] Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro A., and Van Der Laan, M., "Survival Ensembles", Biostatistics, 7(3), 355–373, 2006.

[16] Strobl, C., Boulesteix, A-L., Zeileis, A., and Hothorn, T., "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution", BMC Bioinformatics, 8(25), 2007.

[17] Strobl, C,, Boulesteix, A-L., Kneib, T., Augustin, T., and Zeileis A,, "Conditional Variable Importance for Random Forests". BMC Bioinformatics, 9(307), 2008.

[18] Janitza S. and Boulesteix A-L.,"An AUC-based Permutation Variable Importance Measure for Random Forests for Unbalanced Data", Technical Report 130, Institut fuer Statistik, LMU Muenchen, 2012.

[19] Arnold, D., Möbius, B., and Wagner, P., "Comparing word and syllable prominence rated by naïve listeners", Proceedings of INTERSPEECH 2011, Florence, 2012.