# Perceptually Motivated Parameters for Automatic Prosodic Annotation

*Antje Schweitzer*

Institute for Natural Language Processing, Stuttgart University, Germany

antje.schweitzer@ims.uni-stuttgart.de

This contribution presents an approach to automatic prosodic annotation which emphasizes the linguistic motivation and perceptual relevance of the features used for classifying the prosodic categories. The analyses and experiments presented here were conducted on a 2.5 hours German news-like corpus which had been manually annotated using GToBI(S) (Mayer, 1995). GToBI(S) is an adaptation of American English ToBI (Silverman *et al.*, 1992; Beckman and Ayers, 1994) to German.

As for the parameters used for classifying the GToBI(S) events, I suggest a measure of local speech rate, viz. duration z-scores, as temporal features, and tonal parameters describing the shape of F0 contours related to prosodic categories, the so-called PaIntE parameters, as tonal features. The duration z-scores are obtained by standardizing phone durations using phoneme-specific means and standard deviations. The tonal PaIntE parameters are derived by approximating the F0 contour in a three-syllable window around the syllable of interest using the PaIntE model (Möhler and Conkie, 1998). A schematic of the PaIntE approximation function is depicted in fig. 1. The accented syllable is indicated by the asterisk ($\sigma$*). The exact shape of the function is determined by six parameters which are linguistically motivated: peak height is determined by parameter *d*, amplitudes of range and fall correspond to parameters *c1* and *c2*, respectively, and peak alignment depends on the *b* parameter. Parameters *a1* and *a2*, which are not indicated in fig. 1, represent the (amplitude-normalized) steepness of the rising and falling movements.

To motivate the perceptual relevance of duration z-scores and PaIntE parameters, realizations of prosodic categories in our data are investigated by examining their distributions for each parameter. It is shown that the parameters capture well-known aspects of the realization of prosodic events, such as phrase-final lengthening related to prosodic phrases, the differences in the alignment of peaks and those between rise and fall amplitudes for the different categories as predicted by GToBI(S), the optimal alignment of peaks with syllable structure (House, 1996), but also more recent findings such as the influence of vowel height on the alignment of peaks in German H*L accents (Jilka and Möbius, 2007).

These findings will be discussed at the workshop (see also Schweitzer, 2011). In this abstract, I will only discuss House's (1996) tonal alignment and the GToBI(S) predictions. Fig. 2 shows density plots of the *b* parameter for the most frequent accents and for unaccented syllables. The syllable for which the approximation was carried out, i.e. the syllable which is possibly accented, is highlighted in gray. In the density plot, peaks appear at values that are more likely to occur for the underlying sample, whereas valleys appear at values that are less likely to occur. Fig. 2 shows that H*L accents (blue, long-dashed line) have their peak earlier in the accented syllable than H* accents (orange, dashed line) and L*HL accents (red, dot-dashed line). L*H accents (green, solid line) have their peak either on the accented or on the post-accented syllable. These are plausible findings given the GToBI(S) categories, since H* and H*L accents involve a high target on the accented syllable, i.e. the peak must be on the accented syllable. Also, in an H*L accent we expect a fall which begins already on the accented syllable, i.e. it is not surprising that peaks in H*L accents occur slightly earlier to make room for the fall. Similarly, L*H accents involve a low target on the accented syllable followed by a rise, which means that the peak can only occur toward the end of the accented syllable or on the following syllable. As stated above, the data also corroborate House's (1996) observation that peaks are optimally perceived if they occur after syllable onset, as in our data peaks in general are unlikely at syllable onsets: all distributions exhibit valleys after syllable boundaries. Confidence tests confirm that for the GToBI(S) categories, the parameter distributions differ significantly. This, and the fact that well-known aspects of prosodic realization are evident in the distributions of the parameters, supports the perceptual relevance of the parameters.

Besides their perceptual relevance, I have shown that the parameters, together with some higher-linguistic features, are useful in automatic prosodic annotation (Schweitzer, 2011): Results of 86% accuracy for pitch accents (i.e., accented vs. unaccented), and 93% for boundaries (i.e., boundary vs. no boundary) compared well to results reported in other recent studies (Rosenberg, 2009; Sridhar *et al.*, 2008; Hasegawa-Johnson *et al.*, 2005), particularly to results on German (Zeißler *et al.*, 2006; Braunschweiler, 2006). Also, the classifiers generalized well to similar data of a female speaker in that they performed equally well as classifiers trained directly on the female data. The classifiers have been integrated into a prototype of a tool for automatic prosodic labeling. Some examples of automatic prosodic annotations produced by this tool will be given at the workshop to illustrate the parameters' usefulness in automatic prosodic labeling.

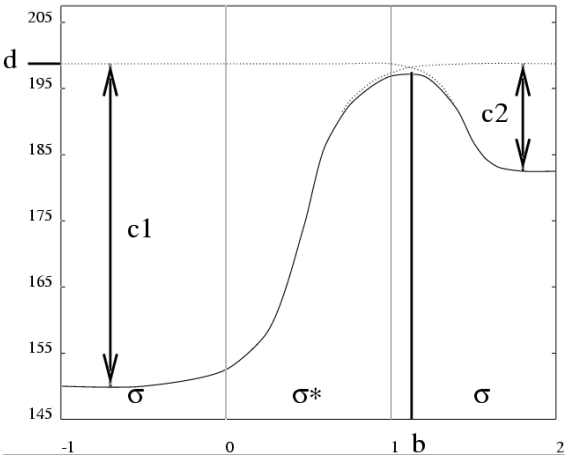


**Figure 1:** Schematic of the PaIntE approximation function reproduced from (Möhler and Conkie, 1998)

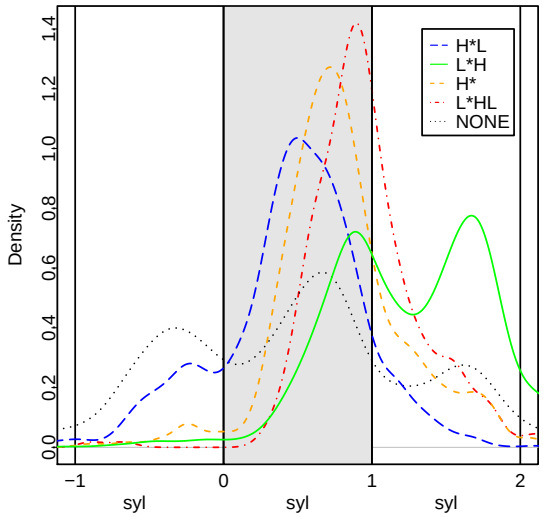


**Figure 2:** Density plots of the *b* parameter for the most frequent accents and for unaccented syllables.

# References


Beckman, M. E. and Ayers, G. M. (1994). Guidelines for tobi labelling, version 2.0.

Braunschweiler, N. (2006). The prosodizer — automatic prosodic annotations of speech synthesis databases. In *Proceedings of Speech Prosody 2006 (Dresden)*.

Hasegawa-Johnson, M., Chen, K., Cole, J., Borys, S., Kim, S.-S., Cohen, A., Zhang, T., Choi, J.-Y., Kim, H., and Yoon, T. (2005). Simultaneous recognition of words and prosody in the boston university radio speech corpus. *Speech Communication*, **46**(3-4), 418–439.

House, D. (1996). Differential perception of tonal contours through the syllable. In *Proceedings of the International Conference on Spoken Language Processing (Philadelphia, PA)*, volume 1, pages 2048–2051.

Jilka, M. and Möbius, B. (2007). The influence of vowel quality features on peak alignment. In *Proceedings of Interspeech 2007 (Antwerpen)*, pages 2621–2624.

Mayer, J. (1995). Transcription of German intonation—the Stuttgart system. Technical report, Institute of Natural Language Processing, University of Stuttgart.

Möhler, G. and Conkie, A. (1998). Parametric modeling of intonation using vector quantization. In *Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia)*, pages 311–316.

Rosenberg, A. (2009). *Automatic Detection and Classification of Prosodic Events*. Ph.D. thesis, Columbia University.

Schweitzer, A. (2011). *Production and Perception of Prosodic Events—Evidence from Corpus-based Experiments*. Doctoral dissertation, Universität Stuttgart.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). Tobi - a standard for labeling english prosody. In *Proceedings of the International Conference on Spoken Language processing (ICSLP, Banff)*, pages 867–870.

Sridhar, V. K. R., Bangalore, S., and Narayanan, S. (2008). Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. *IEEE Transactions on Audio, Speech, and Language Processing*, **16**(4).

Zeißler, V., Adelhardt, J., Batliner, A., Frank, C., Nöth, E., Shi, R. P., and Niemann, H. (2006). *The Prosody Module*, pages 139–152. Springer, Berlin.