Text-based and Signal-based Prediction of Break Indices and Pause Durations

Hartmut R. Pfitzinger & *Uwe D. Reichel*

Institute of Phonetics and Speech Communication
University of Munich, Germany
{hpt;reichelu}@phonetik.uni-muenchen.de

Abstract

The relation between symbolic and signal features of prosodic boundaries is experimentally studied using prediction methods. Text-based break index prediction turns out to be fairly good, but signal-based prediction and pause duration prediction perform worse. A possible reason is that random signal feature variations, as usually produced by humans, are hard to predict.

1. Introduction

Speakers divide their utterances into prosodic phrases separated by weak or strong prosodic boundaries. These boundaries can be perceptually classified according to the ToBI break indices, or they can be described by measurements of prosodic features of the speech signal, i.e. the pause duration and the local contours of F0, speech rate, amplitude, and voice quality in the vicinity of the prosodic boundary. The goal of the present study is to investigate the relationship between symbolic and signal representations of prosodic boundaries by means of three prediction methods (see Fig. 1).

Prosodic phrase break prediction can be decomposed into two tasks: break localization and prediction of its phonetic realization. For the latter this study concentrates on pause duration.

Rule-based or statistical approaches tackle both tasks by exploiting part of speech (POS), syntactical, and rhythmical information. For rule-based break localization Liberman & Church [9] use the fact, that certain POS classes occur preferably phrase initially (*chinks*) while others do not (*chunks*). Gee & Grosjean [7] and Bachenko & Fitzpatrick [2] impose syntactically motivated performance structures on an utterance. The distance of two adjacent words in these structures given in the level of their common node determines the break strength and thus the pause duration between these words. The rule-based Keller-Zellner algorithm [19] also incorporates rhythmical constraints.

In some statistical approaches phrase breaks are predicted from a given part of speech sequence by a modified Markov Tagger. For example, Black & Taylor's tagger [3] is based on conditional emission probabilities for POS sequences co-occurring with the break type given at the corresponding state and conditional transition probabilities of this break type for a given break type history. Others (e.g. [1]) use classificators as CART [4] in order to predict phrase breaks and pause lengths from a set of linguistic features as part of speech and word and syllable distances. CARTs are well suited for this task due to their ability to cope with categorical and continuous types of dependent and independent variables.

In 2002, Mixdorff [11] investigated, among other prosodic properties, pause durations in the IMS Radio News Corpus [15]. Mean values and standard deviations of pauses at sentence boundaries were 716 ms and 336 ms, respectively, and of pauses within sentences were 327 ms and 132 ms, respectively.

2. Method

We investigated the predictability of break indices and pause durations by text-based and by signal-based features (see Fig. 1). In order to account for human duration perception, pause durations were represented on a logarithmic scale. For break index prediction we used C4.5 decision trees [14], for pause durations CARTs (classification and regression trees [4]) provided by R. Additionally, we analysed the correlation between prosodic measures and pause duration.

2.1. Data

The present study is based on the IMS Radio News Corpus [15] which consists of German news texts read by professional speakers. Originally, the data is automatically segmented into phonemes according to the German SAM-PA inventory followed by some manual refinements. Prosody of the data was manually labelled following the GToBI conventions [10].

Prior to analysis of the data a substantial correction was necessary. We omitted news repetitions and manually corrected the phone boundaries in order to achieve reliable pause duration and PLSR measures (perceptual local speech rate [12, 13]). Particularly, we replaced canonical transcriptions by their actual phonetic realizations, inserted glottal stops, which have unfortunately been absent, and replaced /Vowel/-/R/-pairs by their corresponding /Vowel/-/6/-diphthongs, especially /@/-/R/ by /6/.

Finally, our data comprises 28 news articles read by one male speaker. It consists of 16285 phones, 2384 word tokens, and 970 word types, which is approximately half of the original data. For text-based prediction all final words of each signal were excluded, because subsequent pause duration naturally could not be measured. 80% of this data were used for training and the remainder as test set. The data for n-gram and collocation modeling comes from diverse written news corpora and consists of 327821 word tokens and 42741 types.

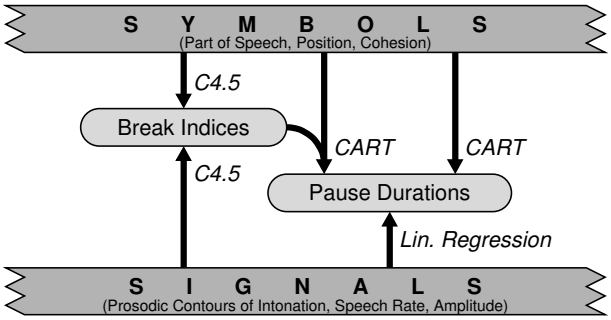

Figure 1: *Text-based (top-down) and signal-based (bottom-up) predictions of break indices and pause durations in this study. Categorical data are predicted by C4.5 decision trees, and continuous data by CARTs or by linear regression.*

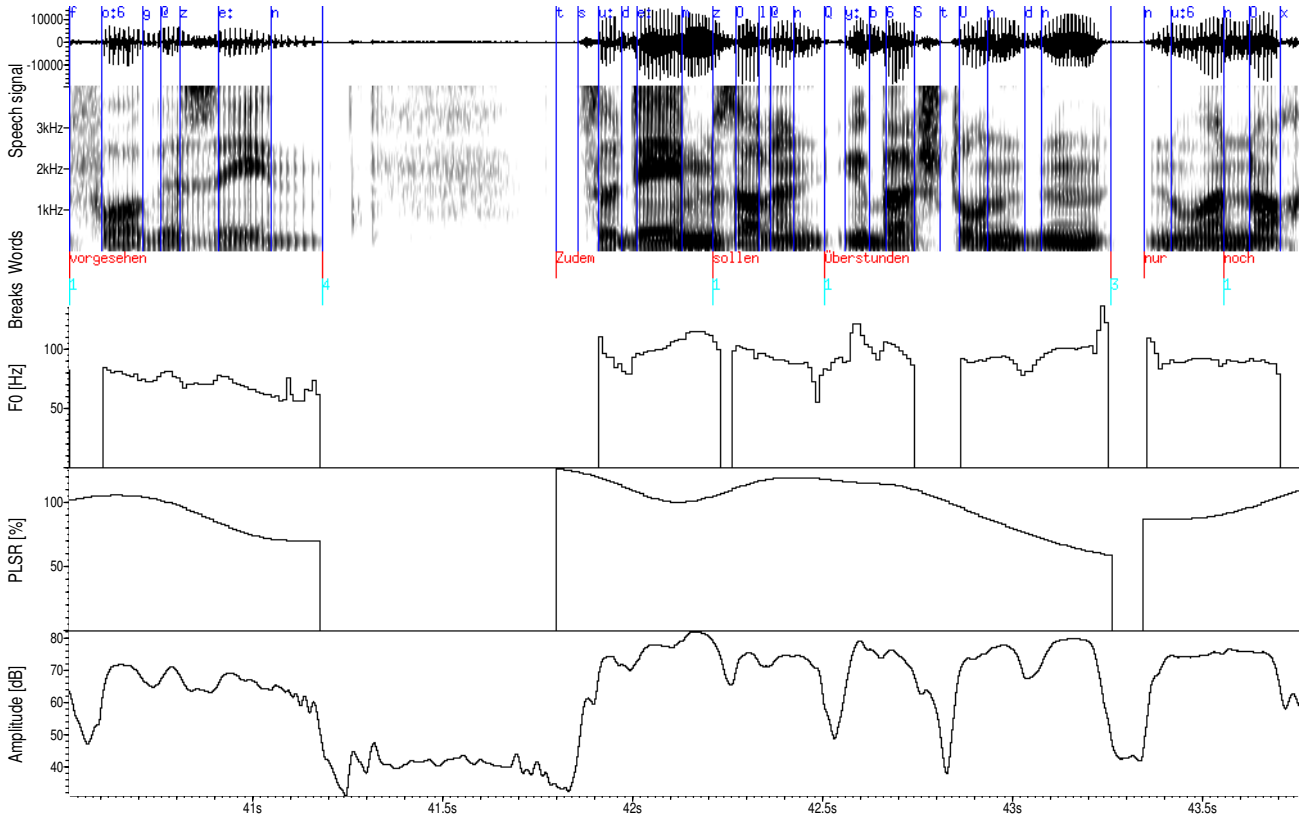

Figure 2: *Speech signal with phone segmentation, word segmentation, break indices, and prosodic contours of F0, PLSR, and amplitude of the utterance* `dlf951121.1200.n7.wav`*: "[...] vorgesehen. Zudem sollen Überstunden nur noch [...]" — English translation: "[...] planned. Furthermore, overtime shall only [...]". Notice the different prosodic behaviour in the vicinity of the two speech pauses.*

2.2. Text-based Features

2.2.1. Part of speech

Due to a restricted number of factor levels in R's 'tree'-package POS labels are clustered heuristically with respect to their syntactic behaviour into 32 classes. Analogous to Liberman & Church's *chinks* and *chunks* [9] we additionally classified POS tags with respect to their likelihood to occur in phrase final position into *kinches* and *kunches*.

2.2.2. Position

For each word we determined the syllable and word distances to the next preceding and following punctuation marks as well as the relative position between these markers and within the sentence. We also noted the kind of punctuation following the word in question (including *no punctuation*).

2.2.3. Cohesion

Opposed to the approaches of e.g. Gee & Grosjean [7] and due to the ease of automatic treatment, our notion of cohesion is statistical and not syntactical. We use linearly interpolated n-gram probabilities as well as likelihood ratios for stems of adjacent words (see [16] for the morphologic method). Likelihood ratios compare the goodnesses of fit of two models to the occurrence probabilities observed in the training data. The numerator-model assumes dependence, the denominator model independence of the words in question. Thus, we hypothesize that the higher the n-gram probability and the likelihood ratio for adjacent words are, the stronger these words are connected, and the less probable is a prosodic break separating these words.

2.3. Signal-based Features

F0 and pitch reset are important features for prosodic boundaries [6]. But speech rate [8] seems to be comparably important, particularly when revealing pre-boundary lengthening [18].

The IMS Radio News Corpus provides raw fundamental frequency contours which have been extracted with the "Entropics ESPS/waves get_f0" [17]. We smoothed this data by means of Median and Hanning filtering to reduce detection errors and micro-prosodic effects (Fig. 2 shows the raw F0 data).

Next, we estimated perceptual local speech rate (PLSR) contours from manual phone and syllable segmentations using a prediction model based on perception tests and linear regression [12, 13]. Generally, it is able to predict perception results of local speech rate with a mean deviation of 10%, where 100% represent a typical average speech rate.

Finally, we extracted amplitude contours using a moving 30-ms-Kaiser window with approx. 50 dB sidelobe suppression. Because the underlying speech corpus is recorded after radio broadcasting [15, p. 102], the speech signal is likely to be dynamically processed which considerably reduces prosodic amplitude variations and might deteriorate prediction performance.

Even though prosodic voice quality features are supposed to play a significant role in prosodic boundary detection, we excluded them from the present study because of the still inconsistent and unreliable automatic extraction methods.

Fig. 2 displays all extracted prosodic contours of a short stretch of speech containing a strong and a weak prosodic boundary. Obviously, the strong boundary is accompanied by a pitch reset while the weak boundary is not. However, both boundaries show a PLSR reset. The amplitude is nonuniform.

3. Results

3.1. Text- and Signal-based Prediction of Break Indices

The signal-based approach has to discriminate four break levels: 1 for default word boundary, 2 for irregular boundaries, 3 for weak prosodic boundaries and 4 for strong prosodic boundaries. However, in the text-based prediction task the number of categories was reduced to three, because irregular boundaries are generally not motivated by the contents of the underlying text. Therefore, in the text-based task irregular boundaries were set to break level 1.

Evaluation results in Table 1 show that text-based prediction of break indices using C4.5 decision trees achieved an accuracy of 87.72%. Both precision (the number of correct identifications of a class divided by the total number of predictions of this class) and recall (the number of found class instances divided by the number of all instances of this class) turned out to be clearly lowest for the break level 3 (the weak boundary or intermediate level).

Signal-based prediction yields a lower accuracy than text-based prediction (compare Table 2 with 1). Its prediction bias towards the 'no prosodic boundary' category is reflected in the low recall values for any other break level than 1. Considering break level 2, the weak performance can obviously be attributed to its rare occurrence in training (29) and test data (13). Besides, again prediction results for break level 3 were poorest.

The confusion matrix in Table 4 shows that almost half of the strong prosodic boundaries are miss-classified as default word boundaries. Finally, it is remarkable that both confusion matrices in Tables 3 and 4 reveal a tendency to treat break level 3 rather as no prosodic boundary than a strong one.

break index	accuracy [%]	recall [%]	precision [%]
all	87.72		
1		94.03	92.98
3		47.06	57.14
4		84.44	76.00

Table 1: *Evaluation results of text-based break index prediction.*

break index	accuracy [%]	recall [%]	precision [%]
all	83.22		
1		96.86	85.88
2		0.0	0.0
3		43.10	54.35
4		46.58	85.00

Table 2: *Evaluation results of signal-based break index prediction.*

	classified as		
break index	1	3	4
1	331	16	5
3	20	24	7
4	5	2	38

Table 3: *Confusion matrix of text-based break index prediction.*

	classified as			
break index	1	2	3	4
1	432	1	9	4
2	11	0	2	0
3	31	0	25	2
4	29	0	10	34

Table 4: *Confusion matrix of signal-based break index prediction.*

3.2. Prediction of Pause Durations using CARTs

Correlations for both text- and signal-based predictions of pause durations are rather low as can be seen in Table 5, which might be explained by the fact that CARTs need a much bigger corpus than chosen for the present study to perform well.

Considering correlation (due to the lack of normal distributions Spearman's correlation coefficient was used) and mean deviation factor, text-based prediction led to better results than signal-based prediction. The performance was further improved by adding the break index values, which were predicted by C4.5 decision trees in the previous section, to the feature pool:

	Spearman	mean deviation factor
text-based, without BI	.591	1.72
text-based, with BI	.668	1.63
signal-based	.510	2.36

Table 5: *Results of the text- and signal-based predictions of pause durations. Conditions: 'with BI': predicted break indices incorporated into the feature pool, 'without BI': not added.*

3.3. Prediction of Pause Durations using Linear Regression

Prosodic features are measured in the centers of preceeding and succeeding syllables of a pause. As shown in Fig. 2 prosodic contours in the vicinity of weak versus strong boundaries differs considerably. Thus, a single linear combination of prosodic features cannot be expected to sufficiently predict pause durations of both weak and strong boundaries. Instead, we use two different linear models for break level 3 and 4, respectively. When excluding prosodic boundaries with no speech pause, the remaining pause durations are normally distributed on a logarithmic scale, and the Pearson correlation coefficient is appropriate.

Table 6 shows that while break level 3 prediction is best when including all features, F0-features dominate break level 4 prediction. Fig. 3 shows prediction details and reveals outliers.

	break level 3		break level 4	
c_1F0 + $c_2$log. F0 reset	.207	1.71	.505	1.40
c_1PLSR + c_2PLSR reset	.260	1.70	.132	1.51
c_1Ampl. + c_2Ampl. reset	.280	1.69	.440	1.40
all signal features	.481	1.57	.535	1.38

Table 6: *Pearson r and mean deviation factor between measured and predicted pause durations using linear regression.*

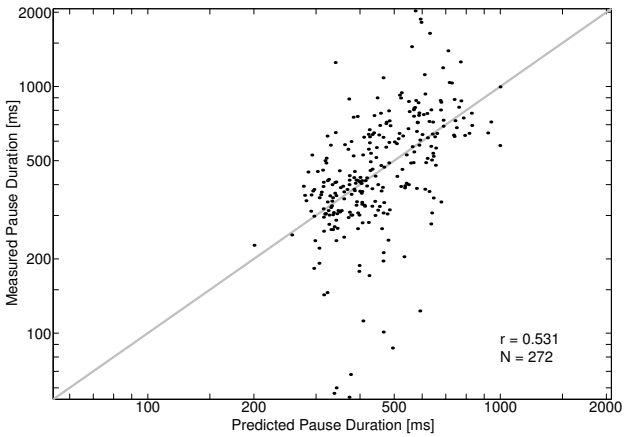

Figure 3: *Scatter plot of predicted and measured pause durations of break index 4. Predictions are performed by linear regression based on Amplitude, F0, and logarithmic F0 reset.*

4. Discussion

In this paper we examined the relationship between the symbolic layer and the signal layer of speech with regard to prosodic boundaries. The question where the prosodic boundaries are, is answered with break indices which represent the perceptual equivalent. On the symbolic layer prosodic boundaries are reflected by punctuation and the syntactic as well as the semantic structure. On the signal layer, the known acoustic manifestations of prosodic boundaries are pause durations as well as contours of F0, speech rate, amplitude, and voice quality.

The results showed poor performances of predicting intermediate break levels. This suggest that either the size of training data or the features used in our study are not sufficient to achieve the classification quality of human transcribers. But generally, the prediction accuracy of perceptual break indices purely by means of text-based features is fairly good (87.72%).

On the contrary, any signal-based prediction or prediction of actual pause durations performs worse. The reason is twofold: On the one hand outlier durations (e.g. caused by additional mute phases triggered by the speaker to hide clearing his throat) as well as random pause duration variations, which are typically produced by humans, are not predictable, on the other hand signal properties of prosodic boundaries are to a significant extent speaker-specific.

The weak correlation between signal-based prediction and measures of pause duration clearly indicates that pause duration is not just a redundant parameter reflected by the other boundary signals. Instead, to some extent it carries non-redundant information, and it is controlled independently by the speaker.

The IMS Radio News Corpus provides some word-for-word repetitions of news which enable the quantitative analysis of random pause duration variation. Unfortunately, only 40 pauses are repeated two to five times in identical sentence context. This allows for only informal conclusions:

Short pauses with about 200 ms duration show standard deviations between 40 and 180 ms, long pauses with durations of approx. 800 ms have standard deviations between 300 and 550 ms. This indicates a tendency but no clear evidence of a linear relation between mean pause durations and standard deviations. Obviously, further investigations regarding sources and limits of speech pause duration variation are necessary.

5. Acknowledgements

We are very grateful to Bernd Möbius from Stuttgart University for making available to us the IMS Radio News Corpus, and to BMW Group Research and Technology Pty Ltd, Munich for partly supporting this work.

6. References

[1] Apel, J.; Neubarth, F.; Pirker, H.; Trost, H. (2004). Have a break! Modelling pauses in German speech. In Buchberger, E., ed., *KONVENS*, pp. 5–12. OEGAI, Vienna; Austria.

[2] Bachenko, J.; Fitzpatrick, E. (1990). A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16(3): 155–170.

[3] Black, A. W.; Taylor, P. (1997). Assigning phrase breaks from part-of-speech sequences. In *Proc. of EUROSPEECH '97*, vol. 2, pp. 995–998, Rhodes; Greece.

[4] Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1984). *Classification and regression trees*. Wadsworth and Brooks/Cole, Monterey, CA.

[5] Butcher, A. (1981). Aspects of the speech pause: Phonetic correlates and communicative functions. Arbeitsberichte (AIPUK) 15, Inst. für Phonetik und digitale Sprachverarbeitung der Univ. Kiel.

[6] de Pijper, J. R.; Sanderman, A. A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *J. of the Acoustical Society of America*, 96(4): 2037–2047.

[7] Gee, J. P.; Grosjean, F. (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15: 411–458.

[8] Hansson, P. (2002). Prosodic phrasing and articulation rate variation. In *Proc. FONETIK 2002*, vol. 44 of *TMH-QPSR*, pp. 173–176, Stockholm, Sweden.

[9] Liberman, M. Y.; Church, K. W. (1992). Text analysis and word pronunciation in text-to-speech synthesis. In Furui, S.; Sondhi, M. M., eds., *Advances in speech signal processing*, pp. 791–831. Marcel Dekker, New York, Basel, Hong Kong.

[10] Mayer, J. (1997). Intonation und Bedeutung. Arbeitspapiere (phonetikAIMS) 3(4), pp. 1–210, Inst. für Maschinelle Sprachverarbeitung, Lehrstuhl für experimentelle Phonetik der Univ. Stuttgart.

[11] Mixdorff, H. (2002). Syntax and prosodic phrasing in news readings. In Hoffmann, R., ed., *Elektronische Sprachsignalverarbeitung (ESSV), 13. Konferenz*, pp. 282–288, Dresden. w.e.b. Universitätsverlag.

[12] Pfitzinger, H. R. (1999). Local speech rate perception in German speech. In *Proc. of the XIVth Int. Congress of Phonetic Sciences*, vol. 2, pp. 893–896, San Francisco.

[13] Pfitzinger, H. R. (2001). Phonetische Analyse der Sprechgeschwindigkeit. Forschungsberichte (FIPKM) 38, pp. 117–264, Inst. für Phonetik und Sprachliche Kommunikation der Univ. München.

[14] Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann, San Mateo, CA.

[15] Rapp, S. (1998). Automatisierte Erstellung von Korpora für die Prosodieforschung. Arbeitspapiere (phonetikAIMS) 4(1), pp. 1–167, Inst. für Maschinelle Sprachverarbeitung, Lehrstuhl für experimentelle Phonetik der Univ. Stuttgart.

[16] Reichel, U. D.; Weilhammer, K. (2004). Automated morphological segmentation and evaluation. In *Proc. of the fourth Int. Conf. on Language Recources and Evaluation (LREC '04)*, vol. 1, pp. 503–506, Lisbon; Portugal.

[17] Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In Kleijn, W. B.; Paliwal, K. K., eds., *Speech coding and synthesis*. Elsevier, New York.

[18] Wightman, C. W.; Shattuck-Hufnagel, S.; Ostendorf, M.; Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *J. of the Acoustical Society of America*, 91(3): 1707–1717.

[19] Zellner, B. (1994). Pauses and the temporal structure of speech. In Keller, E., ed., *Fundamentals of speech synthesis and speech recognition*, pp. 41–62. John Wiley, Chichester.