Rhythmic Alternations in German Read Speech

Petra S. Wagner

Institut für Kommunikationsforschung und Phonetik (IKP) , Universität Bonn, Germany

wagner@ikp.uni–bonn.de

Abstract

In this study, rhythmic alternations in German read speech have been examined on prosodic phrase level. Based on statistical examinations and phonological theory, several hypotheses concerning the status of rhythm rules and word–class related prominence have been formulated, formalized and implemented. The predictions made by these rules were compared with speaker's judgements on perceptual prominence patterns in a prosodic database of read speech. The basic outcomes of this study are the following: If the predictions are based upon different levels of prominence depending on lexical class, the correlation between predicted and perceived prominences lies within the range of different speakers' correlations. Stress clashes hardly ever occur on the basis of this (successful) method. Speakers of German tend to fill rhythmical gaps, especially in polysyllabic words. An FST modelling the filling of rhythmical gaps improves the predictions the algorithm.

1. Motivation

Graphical models used in Metrical Phonology are supposed to reflect the perceptual prominence of utterances. The more beats (illustrated by columns of x's) a syllable carries, the higher is its perceptual prominence. In this study, the beats will be represented by numbers. For German, no empirical basis existed on which the insights of Metrical Phonology could be tested. An approach based upon intuition concerning wrong or right stress patterns may lead to significant problems, because speakers differ in their intuitions. These differences may be due to regional variation, different semantic/pragmatic interpretations of the situation/text, or be simply individual variants. Due to these variations and a lack of empirical data on prominence patterns, the predictive power of phonological rules is still unclear. Especially the status of so–called euphonic or eurhythmic rules is an issue of debate. Previous analyses on sentence level lead us to believe that the semantic load of a word is of higher significance for a correct prediction of prominence than the syntactic phrasing of the utterance[1]. The knowledge about the exact correlations between word and syllable prominence and the linguistic structure of an utterance is not only important in order to falsify phonological hypotheses. It may also have a practical application by making synthetic speech less monotonous and rhythmically more interesting. In order to achieve this goal, the relationship between accessible linguistic structure of a text and the perceptual prominence of its parts needs to be examined [9].

2. Previous Studies

2.1. Parts–of–Speech, syntactic structure and semantic load

The fact that the part–of–speech (POS) type is related to word prominence is a well–known issue in speech synthesis. Metrical phonologists suggested a simple differentiation into function vs. content words in order to capture the different prominence relationships between words on utterance level [10]. Sometimes, the differentiation is also made on grounds of the semantic or syntactic focal structure of an utterance. Such a focus–based procedure would not capture any further differences between specific parts–of–speech, but mark those in focus as particularly prominent. However, until today, a reliable determination of focal structure is very difficult, unless one has access to sentences designed for this purpose. Even though a simple differentiation into content vs. function words is not sufficient for a correct prediction of prominence, it is clearly superior to a purely syntactic approach [12], as long as the

[1] Probably, the knowledge about syntactic phrasing is more important for a correct prediction of prosodic phrase boundaries than of syllable and word prominence.

well–known Nuclear Stress Rule (NSR) [2] is obeyed. Furthermore, it could be found that for a correct prediction of prominence on utterance level (and for speech synthesis), not only the knowledge about the part–of–speech is important but also the subtle interaction between contiguous words [13]. In this work, typcial levels of word prominences were attached to specific word classes. For a successful prediction, position of the word in the utterance and the POS of contiguous words played a significant role as well.

2. 2. Rhythmic Alternations of Syllables

According to the work by Selkirk [11], an ideal metrical grid should obey the "principle of rhythmic alternation". This principle expresses a preference for binary stress patterns in English. The rules creating such patterns (euphonic or eurhythmic rules) result in two phonological processes: Beat Addition (in order to prevent rhythmical gaps or longer sequences of unstressed syllables) or Beat Deletion (in order to prevent 'stress clashes'). The so–called 'Beat Movement' can be regarded as a combination of Deletion and Addition. Féry [6] claims that in German, binary stress is as frequent as ternary stress, Wiese [14] believes binary stress patterns to be fundamental. All these claims are lacking empirical evidence which is independent of native speaker intuition. An extensive examination of our prosodic database [12] lead us to believe that even if only a very simple algorithm for prominence prediction is employed, stress shifts hardly ever provide the reason for wrong predictions. At least in the observed data, stress shift does not seem to play a major role in German. A substantial number of cases could be detected, where predicted sequences of unstressed syllables created a problem. Therefore, finding a strategy for the filling of rhythmical gaps when predicting prominence patterns appeared to be worth for further investigation.

3. Method

Perceptual prominence judgements taken from a prosodic database of German [7] were related to prominence predictions based upon phonological hypotheses. The database was labelled by three subjects for perceived prominence according to the method introduced by Fant and Kruckenberg [4]. According to this procedure, perceived prominence is annotated on a free grid, and grid is later transformed into a scale between 0 and 31. The database contains 227 sentences and three stories and was read by three speakers (one professional speaker). The database contains 10661 syllables in total. High correlations between listeners were found [8]. However, to even out listener–specific effects in prominence ratings, subsequent analyses were based on the median prominence ratings of all three listeners.

Since the realisation of prominence patterns differs depending on the speaker and his or her semantic interpretation of the utterance, dialect etc. [1,3,5], a baseline for a correct prediction was based on the average correlation of speakers in a database [12]. The inter–speaker correlation ranged between =0.72 and =0.82. The average interspeaker–correlation and resulting baseline was found to be high (=0.78, $p<0.001$).

4. Part–of–Speech Driven Prominence Prediction

In a first step towards prominence prediction (Rule #1) , schwa–syllables and syllabic consonants were assigned the prominence value of "0", all other syllables the value of "1". Next, the different parts–of–speech were assigned different inherent prominence levels primarily based on Widera et al.'s results [13] (Rule #2). They were slightly simplified and led to the lexical prominence hierarchy illustrated in Table 1. Based on this hierarchy, prominence values were assigned to the syllable carrying lexical stress in each word. According to this procedure, lexical prominence is part of the phonological form of a lexical entry. Thus, each noun would be inherently more prominent than a lexical entry of an adjective and so forth. Not surprisingly, it becomes apparent, that words which presumably are more informative in an utterance – namely content words – are also inherently more prominent.

Another interesting fact reported by [13] is, that the higher the inherent prominence value of a word, the less affected this prominence value is by contiguous words. Thus, nouns show a relative stability in their prominence whereas determiners appear to vary a lot depending on both position in the utterance and contiguity. The tendency of function words to be comparatively more prominent is most clearly visible

phrase initially. This factor is taken into account by the next step in prominence assignment, where each function word receives an additional beat in (prosodic) phrase initial position (Rule #3).

By assigning different prominence values to different parts–of–speech, stress clashes are prevented in most cases. However, if two words of the identical prominence class are uttered consecutively, this may create a typical stress clash situation and may be avoided by the speaker. Indeed, evidence for such effects are reported [13]. Thus, a simple rule was implemented to prevent such cases: Function words receive an additional beat when followed by a function word of the same prominence class, whereas content words receive an additional beat when preceded by a content word of the same prominence class (Rule #4).

In a last step, the NSR is modelled (Rule #5). The last content word (except verbs) in each prosodic phrase receives one more beat than the most prominent word in the remaining phrase. Verbs receive this beat unless they are preceded by a noun.

Part–of–Speech		***Prominence Value***
Content Word	*Function Word*	
Nouns, Numerals, Proper Names		5
Adverbs, Adjectives		4
Verbs	Demonstrative Pronouns, WH–Pronouns	3
	Modal/Auxiliary Verbs, Affirmative/Negation Particles	2
	Determiners, Conjunctions, Subjunctions, Prepositions	1

Table 1: Prominence values specific for different lexical classes

4.1. Results

The results are very encouraging. A prediction based solely on very simple phonological classifications (accentuability) and taking into account POS–specific prominence levels, results in a correlation between predicted and perceived prominence that already comes rather close to the defined baseline. Previous analyses [12] that took into account syntactic phrasing, NSR and a differentiation into content vs. function words, were slightly worse. The phrase initial prominence strengthening of function words improved the results even further, so did the NSR. However, no visible effect of the simple rule for a prevention of stress clash on a lexical level was detectable. It rather damaged the predictive power of the algorithm, probably by irritating the subtle POS–specific prominence relationships within the utterances. The highest correlations between predicted and perceived prominences were achieved for the data of the professional speaker ($\rho=0.805$, $p<0.01$). Table 2 shows an overview of the different rules and the resulting correlation between predicted and perceived prominence values.

Applied Rules for Prediction	*Correlation between perceived and predicted prominence (baseline: ρ=0.78)*
1 (Accentuability of syllable)	ρ=0.64 p<0.01
1 + 2 (POS−specific prominence assignment)	ρ=0.77, p<0.01
1 + 2 + 3 (additional beat to initial function words)	ρ=0.78, p<0.01
1 + 2 + 3 + 5 (NSR)	ρ=0.784, p<0.01
1 + 2 + 3 + 5 + 4 (prevention of lexical 'stress clash')	ρ= 0.77, p<0.01

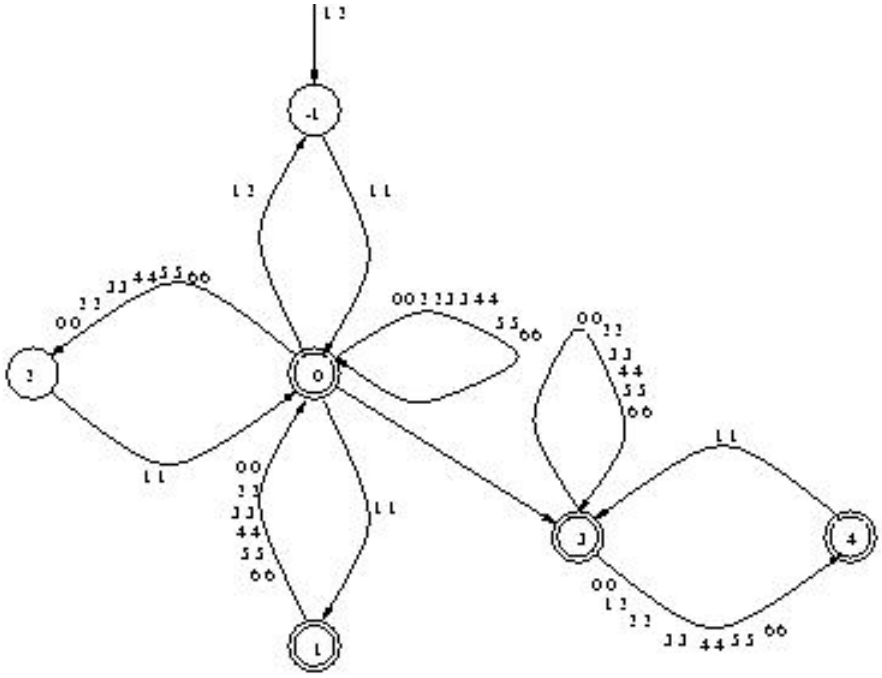
Table 2: Correlations between predicted and perceived prominence for different rule combinations

5. Rhythmical Gaps

The algorithm presented in the previous section achieves results which lie well within the defined baseline. Previous quantitative analyses [12] indicated that even though stress clashes are not obviously avoided by human speakers, rhythmical gaps are. When looking at a small part of our database (300 phrases) which contains a lot of polysyllabic words, the truth of this becomes evident. Here, the correlation between predicted and perceived prominence lies clearly lower ($\rho = 0.76$, $p<0.01$).

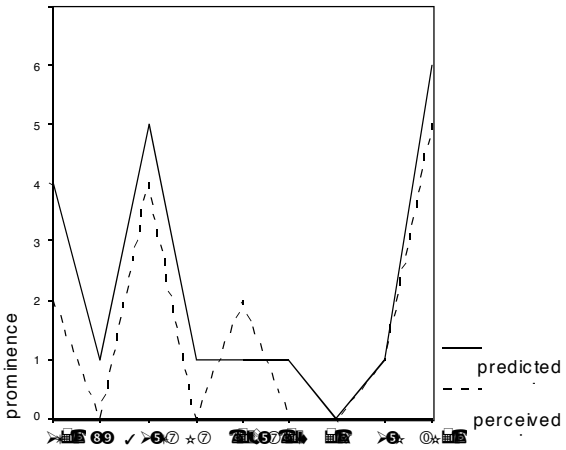
In order to model the filling of rhythmical gaps, a simple Finite−State−Transducer (FST) was implemented which prevents longer sequences (more than two) of syllables which are low in prominence (prominence level 1). Input to the FST were the stress patterns using the most successful algorithm described in section 4. In most cases, the transducer creates binary stress patterns, ternary ones in some cases. The automaton is illustrated in Figure 1.

Figure SEQ Figure * ARABISCH 1: The FST preventing rhythmical gaps.



5.1 Results

After application of the FST to the specified part of the database, correlations between predicted and perceived prominence have reached the defined baseline again (ρ=0.78, $p<0.01$). This effect is clearly visible for all speakers. Again, the best results were achieved for the professional speaker (ρ=0.79, $p<0.01$). Figure 2 shows a comparison of perceived and predicted prominence before and after the application of the FST for the prosodic phrase “Ist der einundzwanzigste August”. For purposes of illustration, the scale of perceived prominence (0–31) was mapped onto the 7–level prominence scale (0–6) resulting from the prominence prediction algorithm.



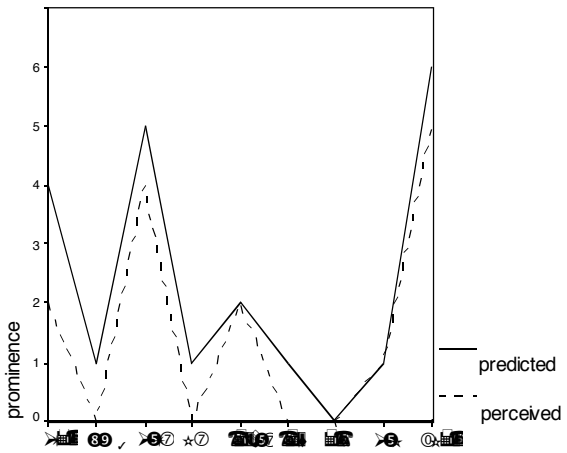


Figure 2: Predicted and perceived prominence before (left) and after (right) filling of rhythmical gaps.

6. Conclusions

In the data on hand, no evidence for a prevention of stress clashes by Beat Addition or Beat Deletion on the level of perceptual prominence could be detected. However, for a further confirmation of this insight, more material needs to be examined, because the prediction algorithm inherently prevents most stress clashes by assigning different prominence levels to different lexical classes. According to our implementation, prominence is a feature of the phonological form of a lexical entry and not a purely syntactic or sentence semantic phenomenon.

Concering the question of whether and how rhythmical gaps are prevented in German, the outcome of the study is clear: It is a strategy employed by all speakers in our database.

The predictions of the presented algorithm lie within the upper range of correlations between different speakers. Some speaker–specific outcomes are apparent. The professional speaker is more in accordance with our prediction algorithm than the others. Since the algorithm works best for the professional speaker in the database ($\rho=0.81$, $p<0.01$ after FST application) , it can be concluded that the predictions are relatively close to a kind of "standard" stress pattern.

Of course, the use of read monological speech in this study paints a somewhat simplified picture. It does not pay any attention to the influence of focal structure, givenness or other phenomena affecting the prosodic pattern of an utterance. However, we believe that a good prediction of a "default prosodic pattern" provides a valuable basis for further research of these much more difficult issues.

Since the interface between perceptual prominence and acoustic features such as pitch accents, duration and intensity have been extensively studied for German [8], the presented research may prove directly useful for both rule and corpus–based speech synthesis applications.

7. Literature

[1] Wilbur Benware. Accent variation in German nominal compounds of the type (A(BC)). *Linguistische Berichte* 108, 1987:102–127.

[2] Noam Chomsky and Morris Halle. *The Sound Pattern of English*. New York: 1968.

[3] Ursula Doleschal. Zum deutschen Kompositionsakzent. Tema con variationi. *Wiener Linguistische Gazette* 40/41, 1988:3–28.

[4] Gunnar Fant and Anita Kruckenberg. Preliminaries to the study of Swedish Prose Reading and Reading style. *STL–QPSR* 2/1989, 1989:1–68.

[5] Caroline Féry. Metrische Phonologie und Wortakzent im Deutschen. *Studium Linguistik* 20, 1986:16–43.

[6] Caroline Féry. Rhythmische und tonale Struktur der Intonationsphrase. In: H.Altmann (ed.). *Intonationsforschung*. Tübingen: Niemeyer, 1988:41–64.

[7] Barbara Heuft, Thomas Portele, Florian Höfer, Jürgen Krämer, Horst Meyer, Monika Rauth and Gerit Sonntag. Parametric description of F0–contours in a prosodic database. *Proceedings of ICSLP ´95*, Stockholm, Sweden, vol. 2, 1995:378–381.

[8] Barbara Heuft. *Eine prominenzbasierte Methode zur Prosodieanalyse und –synthese*. Computer Studies in Language and Speech 2. Peter Lang, Frankfurt: 1999.

[9] Thomas Portele. JUstCONcatenation. A corpus–based approach and its limits. Selected *Papers of the 3rd Speech Synthesis Workshop at Jenolan Caves (Working Title)*. To appear.

[10] Susanne Uhmann. *Fokusphonologie. Eine Analyse deutscher Intonationskonturen im Rahmen der nicht–linearen Phonologie*. Linguistische Arbeiten 252. Niemeyer, Tübingen:1991.

[11] Selkirk, E.O. Phonology and Syntax. *The relation between sound and structure*. Cambridge University Press, 1984.

[12] Petra S. Wagner. Evaluating Metrical Phonology. A Computational–Empirical Approach. *Proceedings of KONVENS 2000*, Ilmenau, Germany.

[13] Christina Widera, Thomas Portele and Maria Wolters. Prediction of Word Prominence. *Proceedings of EUROSPEECH ´97*. Rhodes, Greece, 1997:999–1003.

[14] Richard Wiese. *The Phonology of German*. Clarendon Press, Oxford:1996.