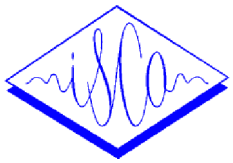

Speech Prosody 2008
Campinas, Brazil
May 6-9, 2008

# The effect of global rhythms on local accent perceptions in German

*Oliver Niebuhr*

Laboratoire Parole & Langage, Université de Provence, Aix-en-Provence, France
Oliver.Niebuhr AT lpl-aix.fr

## Abstract

Accent perception is only partly due to local phonetic cues. Moreover, global phonetic patterns as well as signal-external top-down knowledge play an important role. The presented experiment continues this line of research and provides evidence for German that the perceived local accent position is affected by a global factor which is central for speech, but which has not been addressed so far: the rhythmic pattern.

## 1. Introduction

In perception research prominence refers to a concrete property of perceptual entities. It describes how conspicuous a certain entity appears in relation to the surrounding ones. In speech, the syllable is an established prominence unit. Accented syllables or accents, respectively, are generally characterized by a high prominence. From a simplified functional point of view, accents are related in different ways to lexical units. For instance, on the lexical level the presence and/or the place of an accent distinguish words in many languages, cf. [1]. On the utterance level, the perceptual highlighting of words by the accentuation of (specific) syllables within the words marks the corresponding pieces of information as new and/or crucial for the hearer or contrasts them with preceding information, cf. [2,3]. While accent is thus a structurally discrete (although not necessarily binary) feature of the speech code (cf. [4,5]), the prominence associated with the accent can vary gradually (which may then be relevant for other prominence-based functions like, e.g., emphasis, cf. [6]).

In languages like German or English pitch or pitch changes, respectively, are primarily responsible for the higher prominences of accented syllables, in addition to other phonetic factors like duration, energy, and spectral energy distribution, cf. [1,5]. On the other hand, as already implied by the language-specific effectives of phonetic factors, perceived prominence is a *cognitive construction* that is only in parts determined directly by local cues like pitch (changes). Additionally, global phonetic patterns as well as signal-external expectations and (language-specific) top-down knowledge contribute substantially to the perceived degree of prominence.

As regards relations between pitch and accentual prominence, studies generally come up with a positive correlation between F0 peak height and the degree of prominence. That is, the higher the peak, the more prominent the accented syllable is perceived, cf. [8,9,10,11,12]. Comparable conclusions can be drawn from production experiments like in [13]. Furthermore, it was found that shifting the peak contour or maximum across the utterance is paralleled by abrupt changes in the perceived accent position. However, in order to cause such a perceptual change, the F0 peak (maximum) need not be located on the corresponding accented syllable (cf. [5,7]). Such dissociations are accounted for in the AM phonology ([4]) by regarding the phonetic events more or less as 'guides' and not as immediate cues to prominence, which is the view of the Kiel Intonation Model ([5]), taken as basis for the present study. However, irrespective of this conceptual difference, accentual prominence perception is even more complex.

For Dutch it was shown by [14] that the perceived prominence of accented syllables compensates for declination, i.e. the successive lowering and narrowing of F0 in the course of an utterance. So, for creating a constant prominence impression, the rising-falling F0 peaks of accented syllables may show smaller F0 movements and relatively lower peak heights the later the corresponding syllable occurs in the utterance. This general effect has been observed across different experimental settings and for listeners with different language backgrounds (e.g., [10,12]). As for the latter, the declination effect on prominence was also replicated for Japanese by [15] who investigated lexical accent perception. Moreover, superimposed on this effect, evidence for an 'accent boost subtraction' was found. That is, the increased peak height marking syllables of words with lexical accents in Japanese is not reflected in the prominence impressions of the Japanese listeners. Since this subtractive effect requires signal-external lexical knowledge, it demonstrates the relevance of top-down processes in accentual prominence perception.

Further evidence for top-down influences is provided by [16]. In this study, native and non-native listeners diverge in their prominence judgements for the syllables of different German sentences. While the prominence judgements of the non-native group could well be correlated with phonetic cues (like F0 and duration), the judgements of the native group did not show a comparable predictability. Instead, their perceptual impressions were more strongly influenced by the linguistic structures and semantics of the sentences and the resulting *expected* prominence patterns. The findings of [17,18,19] point to similar contributions of linguistically-based expectations to perceived prominence patterns based on comparisons between native and non-native listeners and/or between auditory and visual stimuli. Finally, experiments of [20] suggest that, e.g., verbs evoke less prominence than adjectives in similar contexts. Regarding syntactic orders, it is argued that such word-class effects are partly responsible for the cross-linguistically observed V-like prominence patterns of utterances.

In view of the sketched background, the present paper aims at widening the scope of the constructive nature of accentual prominence. It investigates for German whether the perceived positions of local accented syllables are influenced by a global variable which is central in speech, but which has not been taken into account so far: the rhythmic pattern. Since the study is based on accents, it focuses on F0 peaks.

In the field of music psychology, it is a well-investigated phenomenon that the position and the degree of prominence of local melodic accents are influenced by the global rhythmic pattern. This effect shows up most clearly in so-called subjective rhythms (cf. [21]). Similarly, in the prosodic labelling of German spontaneous speech corpora it was frequently observed by the author (on) that prosodic phrases are perceived with different rhythmic patterns and hence with different ac-

10.21437/SpeechProsody.2008-74

cent positions, when the complete phrase or just a section is played. Moreover, if in such a section the F0 pattern of a single accented syllable is manipulated (e.g., shifted or extended in range), the rhythmic pattern and hence the position of the other local accented syllables can change. With regard to these informal observations and to the commonalities in the cognitive processes of speech and music concerning the creation of prominence and the rhythmic organization (cf. [21]), it is assumed that also for speech stimuli the perceived position of a local accented syllable is influenced by the preceding rhythmic pattern, created by the preceding accented syllables.

## 2. Method

The latter assumption is tested on the basis of a short utterance consisting of monosyllabic words. In this, the accentual structure was manipulated in two ways. First, the perceived position of the last accented syllable is varied by shifting the corresponding rising-falling F0 peak in equal-sized steps from the centre of a syllable (A) into the centre of the following syllable (B). Second, the F0 peak shift was integrated into two preceding accent-based rhythms: one dactyl without upbeat, i.e. condition D, and another one with upbeat, i.e. condition Dup, each of them comprised two accent groups or accent-related feet, respectively. So, an accent on syllable (A) fits into rhythmic condition D, whereas an accent on syllable (B) continues the dactylic rhythm of Dup. Therefore, provided that the rhythmic conditions influence the perceived position of the last accent, condition Dup should support the perception of an accent on syllable (B). Dactyls were selected since sequences of one accented and two unaccented syllables are frequent in German, more frequent than, e.g., accents separated by a single unaccented syllable, cf. [11].

The stimulus generation was based on the utterance "mit mir geht's nur heut' um punkt neun Uhr" ('for me, it is only possible today at 9 o'clock sharp'). It was produced by the author (on) with the two accent-based rhythms D and Dup. Since the two utterances D and Dup consisted of 9 monosyllabic words, the rhythmic pattern in each of them is based on three accents and hence on three feet. They were perceptually equidistant. However, compared with the utterance D, the accents and their corresponding feet in the utterance Dup are all shifted to the following word. On the other hand, in both utterances all accents were realized with comparable prominences and the same intonation unit, a medial peak or H*, respectively (cf. [4,5]). So, the overall F0 courses of D and Dup comprised three rising-falling F0 patterns, superimposed on a slight declination slope ending in a terminal fall. The F0 peaks reached their maxima around the end of the vowel of the accented syllable; and each peak maximum was slightly lower than the preceding one. Both is characteristic for chaining medial peaks or H* in German (cf. [5,22]).

The overall F0 courses of the utterances D and Dup were stylized in 'praat' so that each rising-falling F0 peak was represented by three contour points: rise onset, maximum, and fall offset. The F0 between all contour points was interpolated on a linear Hz scale. The positioning of the points was guided by the naturally produced declination slopes and peak shapes. Finally, the utterances D and Dup were resynthesized with the modified F0 courses, using the 'PSOLA' provided by 'praat'.

Additionally, the final utterance section comprising the four monosyllabic words "um punkt neun Uhr" ('at 9 o'clock sharp') was produced in isolation by the speaker (on) with comparable overall speech rate, voice quality, and energy level. In this production, the two words "punkt" and "neun" were comparably 'stressed'. That is, while the words showed durations and intensity peaks similar to the ones of the accented words in the utterances D and Dup, F0 was held approximately constant on a middle level. The same level F0 also marked the adjacent words "um" and "Uhr". In a following step, 'praat' was used to delete the small F0 variations between the initial and final value and to raise/lower the latter two values so that they corresponded as close as possible to the initial and final F0 values measured in the "punkt neun Uhr" sections of utterances D and Dup. Secondly, three further F0 contour points were integrated into the resulting declination slope and a symmetrical F0 peak was constructed with rising and falling movements of 150ms. The frequency values of the three contour points as well as their temporal distances were again orientated towards the natural peak productions on "punkt" and "neun" in the utterances D or Dup, respectively. On this basis, a peak shift continuum was created. Starting from a position at the vowel onset of "punkt", the complete rising-falling F0 peak was shifted in 11 equal-sized steps of 50ms to the centre of the preceding diphthong of "neun", and for each step a stimulus was resynthesized by means of PSOLA in 'praat'. With regard to previous studies the created range of F0 peak positions should be sufficient for a shift of the accent position from "punkt" to "neun", which will be paralleled by a clear change in the perceptual prominence relation between the two words, cf. [5,7,10].

In the final step of the stimulus generation, the last four words "um punkt neun Uhr" of the utterances D and Dup were cut off after the /t/ aspiration of "heut'" and replaced by each of the 11 resynthesized utterance sections, hence yielding 11 stimuli based on utterance D and another 11 stimuli based on the utterance Dup. In the two resulting stimulus series D and Dup the sections "um punkt neun Uhr" are physically *identical* for the same stimulus numbers, while the preceding sections *constantly differ* at each stimulus number with the regard to the accent positions and the resulting rhythmic pattern. Figure 1 illustrates the output of the stimulus generation by means of the Dup series.

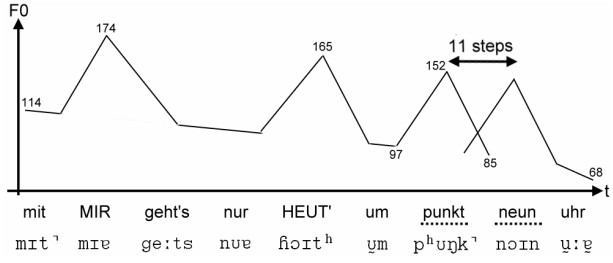


Figure 1: *F0 courses in the 11 stimuli of the Dup series. Orthographic and phonetic transcriptions are synchronized with F0. F0 values refer to utterance onset and offset as well as to peak maxima. For the third, shifted F0 peak additional values specify rise onset and fall offset. Dotted lines indicate the assumed accent shift from "punkt" and "neun", all preceding constant accented syllables are in capital letters.*

Regarding Figure 1, the reasons for selecting the stimulus utterance become obvious. First, the exclusive use of monosyllabic words made it possible to create the two rhythmic patterns as well as the intended change in the final accent position within a single utterance and hence on a constant segmental string. Since the syllable that can receive an accent is fixed for each word in German, both would not be possible

for utterances consisting of polysyllabic words (cf. [5]). Secondly, from a functional point of view, the semantic content of the utterance is composed so that the accents in both rhythmic conditions, D and Dup, are possible and equally compatible with accents on "punkt" and "neun". So, the perceptual judgements of the subjects concerning the final accent position as well as possible effects of the two experimental conditions D and Dup on this position should not be biased by functional interpretations and can therefore be interpreted with regard to local phonetic cues and global rhythmic patterns.

For each of the two stimulus series, a separate perception test was created. In this, the stimuli were repeated five times in a randomized order. Furthermore, each of the 55 stimuli of the two tests was introduced by a bleep and followed by a pause of four seconds. Blocks of 11 stimuli were separated by an additional bleep.

The perception tests were presented in a silent room via loudspeaker to 30 native speakers of German: 11 male and 19 female, all undergraduate students from the University of Kiel. At the beginning of the experimental session, the subjects were presented examples of different utterances. The terms 'perceptually standing out' and 'stressed' were used to describe accented syllables and their prominence impressions to the subjects. After this metalinguistic and phenomenological introduction, the subjects were instructed to listen to the German utterance "mit mir geht's nur heut' um punkt neun Uhr" and to judge, whether they perceive either the word "punkt" or "neun" as standing out. The term 'standing out' represents a very transparent metaphor and was therefore regarded to be intuitively accessible and more suitable than potentially ambiguous and/or metalinguistic terms like 'emphasized' or '(more) prominent'. Judgements should be given on prepared answer sheets during the four seconds pause after each stimulus. Prior to each of the two experimental perception tests, the subjects received a shorter variant as pilot test, which should be performed in the same way. After the whole experimental session, personal data were collected from the subjects, e.g., concerning their age and language background. Moreover, it was asked whether they would judge their selves as musical or not. Altogether, the perception experiment took about 40 minutes.

## 3. Results

For each of the 11 stimuli of the two experimental conditions D and Dup, it was counted in how many of the possible cases the subjects judged "punkt" as standing out compared with the adjacent "neun". Thus, for each stimulus the "punkt" judgements were summed up across all 30 subjects. Since each stimulus was judged five times by the 30 subjects, the values of these sums could vary between 0 and 150. On this basis, percentages were calculated for each stimulus of the two conditions D and Dup.

The percentages concerning the total sample are given at the top of Table 1. They show a change in the judgement behaviour across the peak shift continuum for both experimental conditions. While for stimuli 1-4 the word "punkt" was judged as standing out in more than 80% of the cases, the opposite holds for the last four stimuli 8-11. In between, there is a short transition phase concerning stimuli 5-7. In these stimuli, the F0 peak is shifted with its maximum out of the nasal in the syllable coda of "punkt" and into the syllable-initial nasal of "neun". So, in those cases in which the F0 peak was located with its maximum in the voiced portions of "punkt" or "neun", the corresponding word was judged as standing out in (almost) all presentations. However, apart from a clear perceptual effect of peak position or peak maximum location, respectively, there was no influence of the two rhythmic patterns D and Dup on the subject's decision which of the two words ("punkt" or "neun") is perceived as standing out. This is also confirmed by inferential statistics based on the summed "punkt" judgements of the individual subjects (i.e. n=30). A t test of paired samples was used to compare these sums between the two conditions D vs. Dup. It yielded no significant difference. However, a significant trend was found (mean sum D=25.4; mean sum Dup=26.7; t=-1.998; df=29; p=0.055).

With regard to this trend, the total sample was split up into the 17 musical and the 13 non-musical subjects (according to self-assessment). On this basis, a clear effect of the experimental conditions D and Dup was found for the 17 musical subjects (12 female, 5 male). The corresponding paired percentages are presented at the bottom of Table 1. As can be seen, the perceptual bipartition of the stimulus series found for the total sample holds in a comparable way for the sub-sample of musical subjects as well. In addition to this positional effect Table 1 reveals a small but consistent influence of the preceding rhythmic patterns, in the way that the dactyl rhythm with up-beat (Dup) supported and accelerated perceptions in which "neun" represents the word that stands out in the F0 peak shift continuum from left to right. The divergence of the judgements between the experimental conditions is strongest for stimulus 5 in the perceptual transition phase. Thus, for an accent shift defined by a change in the majority of judgements, the Dup stimuli already cause the shift from "punkt" to "neun" between stimuli 4 and 5, whereas for the D stimuli the boundary is located between stimuli 5-6. That is, the F0 peak has to be positioned at least 50ms more to the right.

A t test for paired samples corroborated the observed differences. The individual sums of "punkt" judgements differ very significantly between the two rhythmic patterns D and Dup (mean sum D=23.4; mean sum Dup=26.1; t=-2.985; df=16; p=0.008**). On the other hand, an analogues inferential statistics based on the individual sums of "punkt" judgements of the remaining 13 non-musical subjects yielded no significant effect of the independent variable D vs. Dup (mean sum D=28; mean sum D=27.6; t=0.452; df=12; p=0.659).

Table 1: *Percentages across the subjects judging "punkt" as standing out (as against "neun") for the 11 stimuli within the two rhythmic conditions D and Dup. The pairs of values at the top refer to the total sample (n=30), the ones at the bottom refer to the sub-sample of musical subjects (n=17).*

| **Stim** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **D** | 99 | 97 | 93 | 85 | 61 | 47 | 19 | 13 | 8 | 8 | 5 |
| **Dup** | 99 | 98 | 87 | 80 | 56 | 40 | 23 | 9 | 8 | 4 | 3 |
| **D** | 99 | 98 | 91 | 85 | 58 | 37 | 15 | 13 | 11 | 11 | 7 |
| **Dup** | 100 | 98 | 87 | 79 | 42 | 34 | 13 | 1 | 4 | 5 | 0 |

## 4. Discussion

Shifting the complete rising-falling F0 peak from the onset of the vowel in "punkt" to the centre of the following diphthong in "neun" changed the judgements for the word that is standing out clearly and rapidly from "punkt" to "neun". Based on the assumption that the subjects indirectly judged the perceptual prominence relations between "punkt" and "neun", the results may be interpreted as a change in the accent position from "punkt" to "neun" across the peak shift continuum. Fur-

thermore, since this accent shift was solely caused by F0, the results are generally in line with [4,5,7]. Finally, it goes well with [4] that the accent position was closely connected with the peak maximum location of H* in either "punkt" or "neun".

In addition to the accent shift itself, the present experiment revealed differences between the judgements for two series D and Dup. In the case of Dup, the accent was significantly more frequently perceived on "neun" than on "punkt", in particular for stimulus 5 in which the F0 peak (maximum) was located at the offset of the nasal of "punkt". These peaks turned out to be a less clear indicator for the accent position. Since from a functional point of view accents on "punkt" and "neun" were equally possible in the selected utterance and also equally compatible with the two preceding rhythmic (accent) patterns (cf. Fig.1), the differences in the perceived accent positions may be interpreted as being due to global influences of accent-based rhythms. As such, the findings match with the initial assumption. That is, the two dactyl rhythms created by the preceding patterns of accented and unaccented words support the perception of an accent on the word that continues the global rhythm. Thus, the perception of local accents is not only determined by local syllable-related, but also by global phrase-related phonetic factors. As stated before, it is a cognitive construction based on bottom-up as well as on top-down information. Therefore, it is impossible to reliably soak up local prominences – and hence also accent positions and levels – solely from phonetic cues like F0 peak height or range, i.e. by neglecting the cognitive component. Correspondingly, [21] states: *"There is no automatic correspondence between the perceived stress and any acoustic measure. This is true for speech as it is for music […] finally, stress and accent are in the 'head'"*.

However, the rhythmic effect primarily showed up for the sub-sample of the 17 musical subjects. This raises the questions, if the difference between the two sub-samples of subjects reflects that musical ones perceive speech, in particular prosody, different from non-musical ones, or if it is just an experimental artefact. Both may be true. On the one hand, there are quite a few studies demonstrating that musical subjects outperform non-musical ones in their ability to perceive and to interpret prosodic patterns of speech, e.g., [23]. On the other hand, it is likely that this is due to the conscious, analytic way of listening required in experimental tasks. So, e.g., prominence patterns and the created rhythms should be more transparent for musical subjects and hence they can include a wider rhythmic perspective in their judgements, while non-musical subjects focus on the relevant words (which is even encouraged in the used 2AFC task). However, this need not mean that non-musical subjects are worse in producing and perceiving speech rhythm in natural *meaning-directed* communication.

On this basis, a follow-up experiment is performed which tests the rhythmic effect on local accent perceptions of subjects via changes in the meaning of the stimuli. Preliminary results suggest that (a) non-musical subjects in fact do not differ from musical ones under these circumstances and that (b) both groups show much more pronounced rhythmic effects than the ones revealed above. After this, the next step will be to investigate whether the less regularly timed rhythms of real speech utterances have comparable effects on the perceived local accent position both qualitatively and quantitatively.

## 5. References


[1] Fry, D. B. 1958. Experiments in the perception of stress. *Language and Speech* 1, 126-152.

[2] Bolinger, D. 1972. Accent is predictable (if you are a mind-reader). *Language* 48, 633-644.

[3] Baumann, S. 2006. *The intonation of givenness. Evidence from German*. Tübingen:Niemeyer.

[4] Ladd, D.R. 1996. *Intonational Phonology*. Cambridge: CUP.

[5] Kohler, K.J. 1991. Prosody in speech synthesis: the interplay between basic research and TTS application. *Journal of Phonetics* 19, 121-138.

[6] Kohler, K.J. and O. Niebuhr. 2007. The phonetics of emphasis. *Proc. 16th ICPhS, Saarbrücken*, 2145-2148.

[7] Shattuck-Hufnagel, S., L. Dilley, N. Veilleux, A. Brugos, R. Speer. 2004. F0 peaks and valleys aligned with non-prominent syllables can influence perceived prominence in adjacent syllables. *Proc. 2nd conference of speech prosody, Nara*, 89-92.

[8] Pierrehumbert, J. 1979. The perception of fundamental frequency declination. *JASA* 66, 363-369.

[9] Rietveld, A.C.M. and C. Gussenhoven. 1985. On the relation between pitch excursion size and prominence. *Journal of Phonetics* 13, 299-308.

[10] Terken, J. 1991. Fundamental frequency and perceived prominence of accented syllables. *JASA* 89, 1768-1776.

[11] Niebuhr, O. 2005. Sequenzen phonologischer Intonationsgipfel. Theoretische Möglichkeiten und empirische Realisierungen. *AIPUK* 35a, 55-95.

[12] Gussenhoven, C, B.H. Repp, A. Rietveld, H.H. Rump, and J. Terken. The perceptual prominence of fundamental frequency peaks. *JASA* 102, 3009-3022.

[13] Liberman, M. and J. Pierrehumbert. 1984. Intonational invariance under changes in pitch range and length. In M. Aronoff and R. Oehrle, *Language Sound and Structure*. Cambridge: MIT Press, 157-233.

[14] Gussenhoven, C. and A.C.M. Rietveld. 1988. Fundamental frequency declination in Dutch: testing three hypotheses. *Journal of Phonetics* 16, 355-369.

[15] Shinya, T. 2006. Lexical accent status affects perceived prominence of intonational peaks in Japanese. *Proc. 3rd conference of speech prosody, Dresden*, 89-92

[16] Wagner, P. 2005. Great expectations – introspective vs. perceptual prominence ratings and their acoustic correlates. *Proc. Interspeech 2005, Lisbon*, 2381-2384.

[17] Peperkamp, S., E. Dupoux, N. Sebastián-Gallés. 1999. Perception of stress by French, Spanish, and bilingual subjects. *Proc. 7th Eurospeech, Budapest*, 2683-2686.

[18] Fant, G. and A. Krukenberg. 1989. Preliminaries to the study of Swedish prose reading and reading style. *STL-QPRS 2, Stockholm*, 1-83.

[19] Eriksson, A., E. Grabe and H. Traunmüller. 2002. Perception of syllable prominence by listeners with and without competence in the tested language. *Proc. 1st conference of speech prosody, Aix-en-Prov.*, 275-278.

[20] Jensen, Ch. 2006. Are verbs less prominent? *Lund University, Centre for Language & Literature, Working Papers 52*, 73-76.

[21] Handel, S. 1986. *Listening – An introduction to the perception of auditory events*. Cambridge: MIT Press.

[22] Niebuhr, O. and G.I. Ambrazaitis. Alignment of medial and late peaks in German spontaneous speech. , *Proc. 3rd conference of speech prosody, Dresden*, 161-164.

[23] Thompson, W.F., E.G. Schellenberg, and G. Husain. 2004. Decoding speech prosody: do music lessons help? *Emotion* 4, 46-64.