FEATURE PATTERNS OF SENTENCE ACCENT IN GERMAN INTERROGATIVE SENTENCES

Fred Englert
Johann Wolfgang Goethe-Universität, Frankfurt am Main, Germany

ABSTRACT

For German, the sentence accent allegedly is an important feature of the prosodic system. It is generally described as an accent that renders a syllable to be the most prominent of a sentence. Both location and configuration of the sentence accent are supposed to contribute to the transmission of sentence structure and meaning. Thus, the term 'sentence accent' is often connected with linguistic concepts of discourse structure, particularly the so-called focus/background differentiation. Despite the assumed importance of the sentence accent, empirical evidence about its phonetic features is rarely found.

Feature patterns of sentence accent in German interrogative sentences read-aloud will be presented in this paper. Candidates for typical feature patterns were found by means of a perception experiment. Phonetic features were derived from the domains of fundamental frequency, segmental duration and signal energy.

1. INTRODUCTION

As stated by Thorsen [13], the phenomenon of sentence accent – also known as sentence stress, primary accent, primary stress, nucleus, nuclear accent, focal accent, tonic and Satzakzent – is regarded as an obligatory feature in the prosodic systems of different languages. As the many names for this phenomenon suggest, its function, perception and phonetic manifestation is a subject of scholarly discussion. Broad agreement may be found to the statement that the position of the sentence accent can affect the meaning of a sentence.

In linguistic concepts about the signalling of new/given- or focus/background-information the sentence accent is often considered as a marker of focus domains [2][6][14]. There is widespread agreement about the fact that the perceptual prominence attached to syllables is an important feature for the perception of sentence accents, and that this prominence can be evoked by melodic and rhythmic patterns as well as by spectral variations. On the other hand listeners' expectations of focus location, influenced by the predominant syntactic structure of a language, may bias the perception of a most prominent syllable or word and hence of the sentence accent. As reported by Brown et al. [3] about perception experiments addressed to the questions whether listeners "would agree about where the tonic(s) was placed in a given chunk of speech" and "what cues they were using to identify the tonic", the biasing effect of such a default location often made even experienced experts choose the last lexical word of an utterance instead of an acoustically more prominent one. Overviews given in [12] and [8] suggest that the concept of a default position for the perception of sentence accent may be valid as well for German.

The aim of this study was to investigate if typical patterns of phonetic features can be found at and near the most prominent syllables in interrogative sentences. Features were derived from the domains of fundamental frequency (F_0), segmental duration and signal energy. Spectral features which may reflect changes in the glottal excitation were omitted from the investigation. The results of a perception experiment were used in order to find utterances which can be considered as perceptually equivalent with respect to the location of the most prominent syllable. Additional motivation for this study has emerged from the field of speech technology, particularly from questions concerning the modeling of prosodic features for emphasized words in Text-to-Speech or Concept-to-Speech-systems.

In the next section, the material used in this study will be described. Section 3 reports the implementation and the results of a perception experiment in order to find sentences which are perceptually equivalent for a group of listeners with respect to the location of the sentence accent. Section 4 describes the feature extraction. Feature Patterns found in the vicinity of sentence accents are presented in Section 5. Then Section 6 concludes with a discussion of significant results and an outlook on possible applications.

2. MATERIAL

The material under investigation is a part of '*The Kiel Corpus of Read Speech Vol. I*' [9], which consists of 200 sentences in the domain of train travel inquiries. Each of these sentences was read aloud by 5 speakers (2 females: R, U and 3 males: D, H, K). The orthographic forms of the sentences presented to the speakers are based on transcriptions of spontaneous men-machine dialogues. Therefore, the syntactic structures found in these sentences may be regarded as typical for spontaneous speech. Further resemblance to spontaneous speech originates from the fact that the speakers often have preferred a rather high speaking rate. Consequently, reduction processes , for example [a:m] instead of [a:bn̩t] for the word "*Abend*", occur frequently in the material. These 1000 digitized utterances (referred to as C_{1000}) - together with the corresponding time-aligned phonetic transcriptions - were accessible from a CD-ROM in a 16 bit/16 kHz format.

77 interrogative sentences of this corpus, read by all speakers, were selected for the investigation. The sentence-length extends from two words, e. g. "*Ohne umsteigen?*", up to 21 words, like "*Können Sie mir sagen, wann ich spätestens in München losfahren muß, wenn ich noch vor zehn Uhr in Augsburg sein will?*". From this selection a total of 385 sentences results for the perception experiment.

3. PERCEPTION EXPERIMENT

In this experiment seven judges had to choose the most prominent syllable in each of the 385 sentences. All of them were stu-

dents of phonetics. Nevertheless, the listeners were asked to make their decisions intuitively. The instruction for an intuitive judgment was supported by the experimental setup. Each sentence was played only once through loudspeakers and all of the sentences were presented in one session. The interval between two sentences was 2 seconds. The sentences were grouped according to speakers. A short pause was made after each group of 77 sentences. The listeners were asked to underline the most prominent syllable of each sentence in an accompanying orthographic text.

In contrast to the experiments in Brown et al. [3] - where analytic listening was demanded – our judges reported that they found the task not a very difficult one. Unanimous decisions for the most prominent syllable occurred for 80 sentences. In all cases positions of lexical accents were marked. The decisions seem not to be heavily speaker-dependent, since 17 perceptually equivalent sentences were found for speaker D, 11 for H, 12 for K and 20 for each of the female speakers R and U. The 80 sentences (C_{80}) found in this way were selected for further investigations of phonetic features.

4. FEATURE EXTRACTION

4.1 Duration

In order to obtain useful values for the variation of segmental durations it seems advantageous to use a measure that is based on Z-scores. A measure of this kind was introduced by Wightman et al. [15]; the *normalized segmental duration*

$$\tilde{d}(i) = \left[d(i) - \alpha\mu_p \right] / \left[\alpha\sigma_p \right]$$

weights the difference between the mean duration μ_p of a phone p contained in segment i and its duration $d(i)$ by the standard deviation σ_p from μ_p. The factor α compensates for the variations in speaking rate and is estimated sentence-wise by

$$\alpha = \frac{1}{N} \sum_{i=1}^{N} \frac{d_i}{\mu_{p_i}},$$

with d_i for the duration of segment i and μ_{p_i} for the mean duration of the corresponding phone p. Standard deviations σ_p and means μ_p were estimated using all sentences from C_{1000}. Phones p and their corresponding segmental durations were adopted - with some minor changes - from the time-aligned transcription supplied together with the speech database.

In order to compensate for some of the variance caused by reduced articulation of function words and effects of preboundary lengthening, phone classes with the features "*function-word*" and "*boundary*" were added. Markings for function words were obtained directly from the transcription while the boundary-feature was attached to phones belonging to the rhyme [15] of a syllable preceding a punctuation mark. Standard deviations and means for the durations of the resulting phone classes were computed for each speaker individually.

Syllable durations were computed for all sentences contained in C_{80}. The normalized duration of a syllable S_j containing N segments was computed by

$$\tilde{D}_{S_j} = \frac{1}{N} \sum_{i=1}^{N} \tilde{d}(i).$$

Using the normalized duration leads to a representation that indicates the lengthening and shortening of syllables respectively.

4.2 Signal Energy

The energy of a segment j in a speech signal s consisting of N samples was computed by a measure of the average magnitude

$$E_j = \log \frac{1}{N} \sum_{i=1}^{N} \left| s_j(i) \right|.$$

The variation of energy for individual phones and syllables was computed using Z-scores as described in the previous section.

4.3 Fundamental Frequency – F_0

Fundamental frequency was computed in non-overlapping time-frames with a length of 10 ms. The absolute frequency values were then converted to speaker-dependent relative representations. The conversion of a frequency value f to a semitone value ST was accomplished by

$$ST = \frac{12}{\ln 2} \ln\left(\frac{f}{f_{ref}} \right)$$

where f_{ref} was a speakers mean fundamental frequency.

5. RESULTS

In order to detect typical contours for the selected features all sentences belonging to C_{80} were centered, or "synchronized", with respect to the syllable carrying the sentence accent, denoted as nucleus in the following. Subsequently, the average of feature values – calculated separately for the respective syllable positions - was used to reveal predominant feature patterns which might occur in the vicinity of the sentence accent.

5.1 Duration

As expected, the contour of average normalized syllable durations exhibits a distinct lengthening at the position of the nucleus. In Figure 1 this contour is plotted for the nucleus and its five left and right neighbours, respectively. In addition to that lengthening there seems to be a tendency for the shortening of syllables that precede the nucleus. This compression may be caused by an effort in order to enhance the following lengthening and/or to maintain an isochronous rhythm pattern.

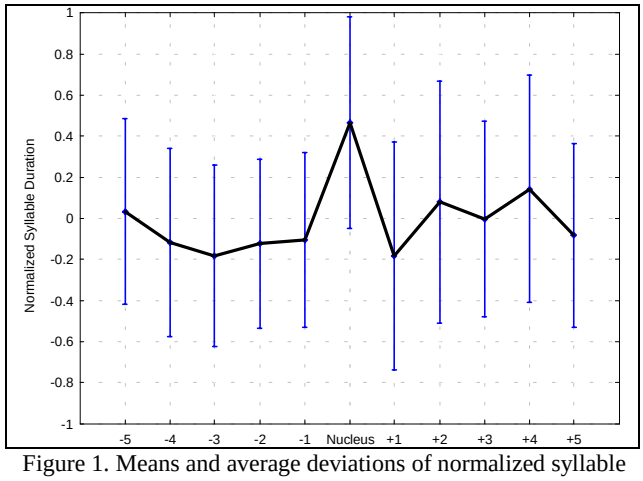

Figure 1. Means and average deviations of normalized syllable durations in vicinity of the sentence accent.

The mean values of normalized durations shown in table 1 indicate differences in the contribution of phone classes to the lengthening observed at the nucleus. The durations of fricatives were hardly changed, whereas vowels, sonorants and stop consonants show a stronger contribution to the lengthening of the nucleus.

	Sentence accent		**Other positions**	
	Mean	Ave.-Dev.	Mean	Ave.-Dev.
Vowels	0.53	0.82	-0.08	0.73
Sonorants	0.38	0.76	-0.01	0.79
Fricatives	0.09	0.82	-0.14	0.81
Stops	0.25	0.95	-0.06	0.79

Table 1. Means and average deviations of normalized durations for classes of phones.

5.2 Signal Energy

As stated by other scholars (see [11] for example), the contribution of signal energy to the perception of accent seems to be less important than that of syllable duration, at least if measured in terms of the normalized energy as defined above. The contour plotted in Fig. 2 exhibits a rather weak local maximum for the syllable following the nucleus, whereas the distribution of normalized energy values is almost perfectly centered around zero for the nucleus itself.

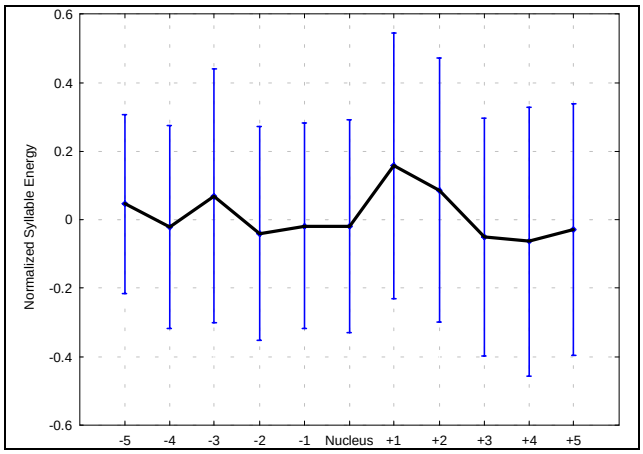

Figure 2. Means and average deviations of normalized syllable energy in vicinity of the sentence accent.

5.3 Fundamental Frequency – F_0

As for duration and energy the syllables of all sentences in C_{80} were synchronized with respect to the nucleus. In order to represent F0-trajectories by piecewise linear segments [1], mean values for the start- and endpoints of F_0-movement were calculated for each syllable position. The resulting contour reveals one dominating pattern which appears at the nucleus and its neighbour to the left. In Fig. 3 this dominating fall-rise-pattern is clearly visible. The F_0-pattern in Fig. 3a results from 25 utterances with a F_0-contour ending below the respective speakers mean F_0. The vast majority of these sentences were „w-questions“ starting with „*wie*“, „*wann*“ or „*welche*“. Other question types, mainly “verb-first-questions”, were predominant in the second group of 55 utterances, ending with an F_0-contour above the speakers mean F_0. The contour resulting for these utterances is plotted in Fig. 3b. In both of these contours the nucleus is preceded by a falling melody followed by a sharp rise at the nucleus itself. The contours plotted in Fig. 3 suggest a rather strong interaction of the melodic course of an utterance and the configuration of F_0-movement at the sentence accent.

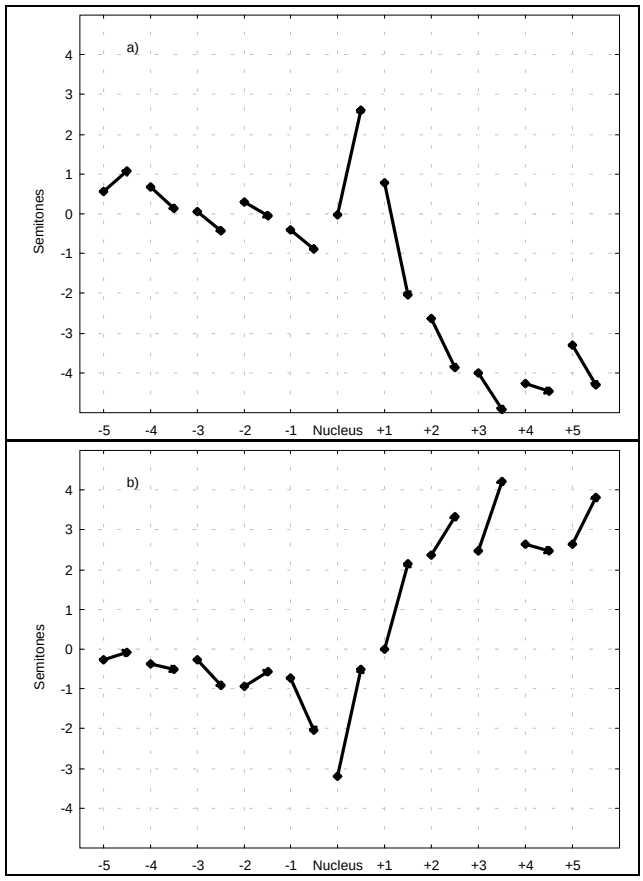

Figure 3. Averaged F_0-patterns centered at the nucleus for a) low-ending and b) high-ending F_0-contours.

Although the fall-rise-pattern was found to be predominant, it did not appear in all of the utterances. According to the F_0-values found at the respective start- and endpoints of the nucleus and its direct left neighbour, three main classes of F_0-patterns can be defined. In 55 out of 80 cases (69 %) a fall-rise-pattern (F-R) occurred at this position. Fall-fall- (F-F) and rise-rise-patterns (R-R) were found in 15 (19 %) and 10 cases (12 %) respectively. Taking into account the direction of F_0 on the syllable following the nucleus leads to the six pattern classes plotted in Fig. 4. Not included in Fig. 4 are F_0-values from four utterances with F-R patterns where the sentence accent was located at the last syllable of the respective utterances.

It should be mentioned that sentences appear not to be bound to only one pattern class. For example, two of the sentences in C_{80} were spoken by all speakers and in each group of five utterances three different patterns (e.g. F-F-R, F-R-F and F-R-R) were found.

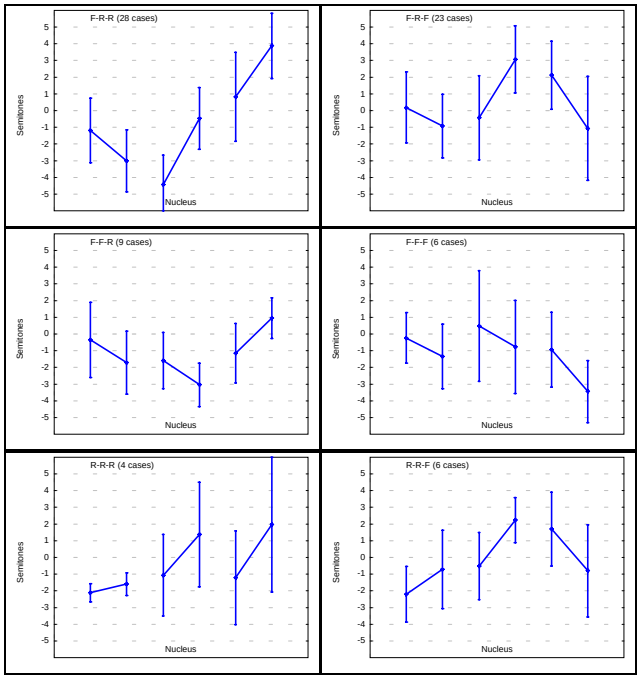

Figure 4. Means and average deviations of F_0-patterns at the position of sentence accent.

CONCLUSION

In summary, interrogative sentences which are perceptually equivalent with respect to the location of the sentence accent were found by the way of a perception experiment. Measures based on Z-scores were used to examine whether duration and signal energy of syllables contribute to the perception of the sentence accent. No such evidence could be found for the signal energy, whereas for syllable duration the global maximum (Z=0.46) of the average duration contour indicated a lengthening of the nucleus.

Using a classification of syllables according to the direction of their respective F0-movements, reveals the predominance of a fall-rise pattern in the scope of the nucleus and its left neighbour. Two less frequent patterns were “fall-fall” and “rise-rise”, whereas no rise-fall pattern was found in this position. In most cases the accent-lending F_0-movement seems to act like a switch, directing the speech melody from mid-low position to either “high” or “low”. Observing similar phenomena for Swedish, Gårding [7] denoted the location of a focal accent as a pivot of a prosodic phrase or utterance (cf. [4]).

Knowledge about the rhythmic and melodic structures used to signal the sentence accent can be used for the prediction of prosodic parameters in Text-to-Speech or Concept-to-Speech systems [10]. Experiments with neural networks for the prediction of prosodic parameters [5] revealed a tendency of the networks to produce a rather flat speech melody, whereas predicted segmental durations came closer to the ones in natural speech. To be fair, it should be mentioned that there was no information about parts-of-speech or focussed words in the input of the networks.

If the location of a sentence accent in an interrogative sentence is known or if it can be predicted, the feature patterns described above may be applied in order to signal focussed words. Examples for the “implantation” of feature patterns can be found in the accompanying sound files [SOUND 18FFR.WAV] and [SOUND 18FRF.WAV]. The speech signals therein were synthesized using the MBROLA speech synthesizer [16]. Each sound file contains two signals, the first one originated from the synthesis according to the original output of a prosody network while the second one was synthesized from a modified version of the networks output. The modifications were achieved by implanting the mean normalized duration contour (Figure 1) as well as the respective means of F_0-patterns F-F-R and F-R-F. In order to keep the pitch-level reached by the respective pattern, trailing F_0-values were multiplied by a constant.

ACKNOWLEDGMENTS

The author would like to thank the students of the experimental-phonetics-class of ‘98 for their participation in the perception experiment and their assistance with its evaluation.

REFERENCES

[1] Adriaens, L. M. H. 1991. *Ein Modell deutscher Intonation*. Diss. Technische Universität Eindhoven.

[2] Bannert, R. 1985. Fokus, Kontrast und Phrasenintonation im Deutschen. *Zeitschrift für Dialektologie und Linguistik*, 52, 289-305.

[3] Brown, G., Currie, K. L. and Kenworthy J. 1980. *Questions of Intonation*. London: Croom Helm.

[4] Bruce, G., Touati, P. 1990. On the Analysis of Prosody in Spontaneous Dialogue. *Working Papers, Dept. of Linguistics, Lund University*, 35, 37-55.

[5] Englert, F. 1999. We’ve got rhythm (but no melody) – An experiment with basic input parameters for prosody networks. In: Wodarz, H.-W. (ed.), *Papers in Phonetics and Linguistics.* Phonetica Francofortensia 7. Frankfurt am Main: Hector.

[6] Féry, C. 1992. *Focus, Topic and Intonation in German.* Arbeitspapiere des Sonderforschungsbereichs 340 “Sprachtheoretische Grundlagen für die Computerlinguistik”.

[7] Gårding, E. 1981. Contrastive prosody: a model and its application. *Studia Linguistica*, 35, 146-166.

[8] Gibbon, D. 1998. Intonation in German. In Hirst, D., Di Cristo, A. (eds.), *Intonation Systems*. Cambridge: University Press.

[9] Informationsheft zu den PHONDAT-Signalkorpora auf CD_ROM sowie zu Segment- und Labeldaten. 1993. Institut für Phonetik und Sprachliche Kommunikation, München.

[10] Prevost, S., Steedman, M. 1994. Specifying Intonation from Context for Speech Synthesis. *Speech Communication*, 15, 139-153.

[11] Sluijter, A. 1995. *Phonetic Correlates of Stress and Accent*. Dordrecht: ICG.

[12] Stock, E. 1980. *Untersuchungen zu Form, Bedeutung und Funktion der Intonation im Deutschen*. Berlin: Akademie-Verlag.

[13] Thorsen, N. G. 1987. Suprasegmental Transcription. In Almeida, A., Braun, A. (eds.) *Probleme der phonetischen Transkription.* Zeitschrift für Dialektologie und Linguistik: Beihefte, 54, Stuttgart: Steiner.

[14] Uhmann, S. 1991. *Fokusphonologie*. Tübingen: Niemeyer.

[15] Wightman, C. W. et al. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, 91 (3), 1707-1717.

[16] http://tcts.fpms.ac.be/synthesis