The Influence of F0 Contour Continuity on Prominence Perception

Hansjörg Mixdorff [1], *Oliver Niebuhr*[2]

[1] Department of Computer Science and Media, Beuth University Berlin, Germany
[2] Department of General Linguistics, ISFAS, Christian-Albrecht-University of Kiel, Germany
mixdorff@beuth-hochschule.de, niebuhr@isfas.uni-kiel.de

Abstract

The presented study concerns the influence of the syllabic structure on perceived prominence. We examined how gaps in the *F0* contour due to unvoiced consonants affect prominence perception, given that such gaps can either be filled or blinded out by listeners. For this purpose we created a stimulus set of real disyllabic words which differed in the quantity of the vowel of the accented syllable nucleus and the types of subsequent intervocalic consonant(s). Results include, inter alia, that stimuli with unvoiced gaps in the *F0* contour are indeed perceived as less prominent. The prominence reduction is smaller for monotonous stimuli than for stimuli with *F0* excursions across the accented syllable. Moreover, in combination with *F0* excursions, it also mattered whether *F0* had to be interpolated or extrapolated, and whether or not the gap included a fricative sound. The results support both the filling-in and blinding-out of F0 gaps, which fits in well with earlier experiments on the production and perception of pitch.

Index Terms: *F0*, perception, prominence, segmental intonation.

1. Introduction

It is a well-known fact that the information structure of an utterance is coded in the relative saliency of its lexical constituents. At the acoustic level we observe that accented syllables serve as anchoring points of this structure. They are emphasized or toned down by phonetic means. The perceptual correlate of this process is the so-called prominence, cf. [21]. Various segmental and suprasegmental factors have been shown to affect prominence, cf. [1,2,3].

In an earlier study [4], the first author and his co-worker investigated the relationship between perceived syllable prominence and the *F0* contour in terms of the parameters of the Fujisaki model [5]. The model was used to parameterize a subcorpus of the Bonn Prosodic Database [6], which included normalized log syllable durations. Analysis showed that prominences labeled on a scale from 0-31 strongly correlated with the excursion of *F0* movements, as represented by the amplitude *Aa* of accent commands, however, only in combination with the high log syllable durations of accented syllables. So, similarly extensive *F0* movements that spanned unaccented syllables with small log syllable durations had only little effect on prominence. The fact that the prominence-lending *F0* movement does not necessarily take place inside the accented syllable indicates that the prominence judgment is partly guided by linguistic considerations. Evidence in support of this assumption has been presented for many languages, including German [7,8,9], which is the language of the present study.

Since the Fujisaki model fits natural *F0* contours continuoously with a defined value for each speech frame, it smoothly interpolates or extrapolates *F0* gaps owing to unvoiced sounds. However, from a communicative point of view, the implicit claim of using the same underlying prosodic gesture for voiced and unvoiced sound sections is that listeners are also able to interpolate or extrapolate *F0* gaps. Recent evidence from a tonal scaling study [10] is inconsistent with this implicit claim. Subjects were presented with short resynthesized utterances and asked to rate the tonal height of accent-related *F0* rises. The rises led to a peak that was either present or absent due to an unvoiced stop consonant. Tonal height ratings were made and analyzed relative to reference utterances in which the *F0* rise was replaced by a flat F0 stretch, yielding a constant tonal height. The findings of [10] suggested that the subjective continuity of pitch contours in speech is due to the fact that the auditory system simply ignores rather than fills *F0* gaps.

It was primarily this conclusion of [10] that motivated the present study. First, we think that the task used by [10] forced listeners into an analytic, purely psychoacoustic listening mode. Comparing the scaling of local (and temporally fairly remote) pitch events within a complex utterance tune is (i) a hard task – 33% of the listeners had to be excluded from the analysis – and (ii) differs considerably from meaning-oriented speech perception. In addition, the conclusion of [10] ignores that silent gaps in the *F0* contour may differ from gaps that (partly) consist of frication. Fricative sounds are able to induce aperiodic pitch impressions, and they are actually varied in accord with this ability in whispered speech and when they interfere with nuclear-accent contours in German. The latter finding has been termed "segmental intonation", cf. [11]. Moreover, it is well known from studies with sinusoid stimuli that listeners do interpolate and perceptually fill gaps in the frequency contour, but only if the gap offers a reasonable explanation for the contour discontinuity. Such an explanation includes, among others, that the interruptor/masker of the frequency contour is frication and not silence, cf. [12,13].

Against this background, we took up the experiment of [10] with a modified methodology. This includes that we used simpler and shorter stimuli that were compared and judged in terms of prominence. Unlike remote tonal-height differences between local pitch events, prominence levels are – due to their basic role in information structure – more directly linked with communicative meaning and hence easier to handle by listeners, particularly when the prominent syllables are also temporally adjacent. Our prominence measure assumes a positive correlation between the amount and variation of pitch (associated with the accented syllable) in a word and its prominence level. This entails that perceptually filled *F0* gaps will be reflected in higher prominence ratings. We included stimuli in which the gaps do or do not contain a fricative sound. Finally, we also included two different types of intonation categories, the medial and the late peak [14], so that it was either mainly the slope of the *F0* peak (interpolation condition) or the area around the peak maximum (extrapolation condition) that coincided with the unvoiced stimulus section.

As we will show in the following, the results of our study suggest taking syllable and segmental properties into account when modelling prominence and intonation, for example, in terms of a syllable-specific weighting of the accent command amplitude *Aa* in the Fujisaki model.

2. Method

We constructed stimuli composed of real disyllabic German words. They are shown in Table 1 with their critical segments set in bold in the SAMPA transcription. All disyllables occur similarly frequent in German (to avoid intrinsic prominence biases) and are realized with lexical stress and accent on the initial syllable. The mean energy in the critical segment given in dB decreases from Rahmen to Ratten.

Table 1. *The five target words and their critical segments.*

Word	SAMPA	English	Critical Segment	energy
Rahmen	[R**a:m**@n]	frame	long vowel (LV), voiced (vcd) nasal	74.23 dB
Rasen	[R**a:z**@n]	lawn	LV, vcd fricative	72.10 dB
Raten	[R**a:t**@n]	guess	LV, voiceless (vcl) plosive	68.10 dB
Rasten	[R**ast**@n]	rest	short vowel (SV), vcl fricative+plosive	66.19 dB
Ratten	[R**at:**@n]	rats	SV, long vcl plosive	50.68 dB

Since we required the stimuli to be also phonetically maximally uniform, we decided to create them using the *MBROLA* concatenative speech synthesizer driving the German male voice *de8* [11]. As a first step we created monotonous stimuli at *F0*=100Hz. The long vowel [a:] was adjusted to a duration of 244ms and the central consonant portion to 126ms. Strictly speaking, mean durations for [z] in natural speech are typically larger than for [t] and [m], but we had to compromise in order not to incur prominence-revelant duration biases. This is especially true for [Rat@n] where we used a relatively long silent pause, in order to maintain the same distance between the onsets of [a] and [@] segment as in the other stimuli. However, informal listening showed that none of the stimuli sounded exaggerated, disfluent/emphatic or unnatural.

Using the *FujiParaEditor* and Praat PSOLA resynthesis [16,17] we created further stimuli by adding *F0* peak contours to the monotonous stimuli. The contour basis was laid by a phrase component, constant for all stimuli. One accent component with a duration of 200ms was superimposed on the base contour. As we intended to examine the effect of *F0* gaps on different portions of the accent peak, the accent component was timed such that it created medial-peak and late-peak tokens, i.e. an established phonological intonation contrast in German [14]. In the long-vowel target words, the *F0* maxima of medial peaks were aligned close to the accented-vowel offset, in line with previous findings [18] and observations in citations forms. Late-peak maxima occurred towards the end of the subsequent consonant. So, filling-in the *F0* gap in the long-vowel target word Raten with a medial peak required *F0* interpolation, whereas the late peaks also required *F0* extrapolation. The opposite was true for the short-vowel target words, since we used the same accent command timing for these words as well. Thus, for Rasten and Ratten, medial peaks represented the extrapolation and late peaks the interpolation condition. Figure 1 displays the stimuli Rahmen, Rasen and Rasten with medial and late peaks at *Aa*=0.6.

The range of the *F0* peaks was varied in the form of three different accent command amplitudes (*Aa*): 0.4 (which resulted in *F0* excursions similar to the natural recordings), 0.6 (about 3 semitones higher) and 0.8 (about 6 semitones higher).

The resynthesis yielded a total of seven tokens for each target word. Based on these tokens, we created stimulus pairs in both orders, AB and BA, and with a silent pause of 1 second between the two stimuli of a pair. From all possible pairings, only the following two types were selected for the perception test: (1) pairs with the same words and peak positions (both either medial or late) but with different peak heights (*Aa*=0.4, 0.6, 0.8); (2) pairs of different words, but with the same peak positions and heights. The latter included *Aa*=0.6, 0.8, and stimulus pairs with monotonous *F0*.

Group (1) consisted of 60 stimulus pairs and was created to ensure that participants reliably interpreted larger *F0* excursions in terms of higher prominence levels. Group (2) contained those 100 stimulus pairs that concerned the main focus of our study, namely potential prominence differences due to interruptions in the *F0* contour.

The stimulus pairs were judged in a 2AFC design. The perceptual test proceeded as follows: First, the subjects were instructed that the experiment would be about synthetic speech and the ability to convey different meanings by assigning different accent levels to a word. They would hear pairs of German words. Their task would be to listen carefully to each pair and to decide afterwards whether word A or word B had been more strongly accented. They would be allowed to replay every stimulus pair once, in order to make a final decision. The actual experiment was preceded by a training session (tutorial), which was based on 10 stimulus pairs with a maximum difference in peak height (*Aa*=0.4 and 0.8). The subjects were walked through the tutorial and received feedback whether or not their decisions had been correct (due to the clear peak height differences, all training pairs had an "expected/correct" outcome). Then, after a short break, which included the possibility to ask questions, the 160 stimulus pairs of the actual experiment were presented separately to each subject via headphones in individually randomized orders.

The experiment was programmed as a server application, accessed through a web browser, and the results were logged to an SQL database. Participants were 22 students of Media Informatics at Beuth University, of these 17 male and 5 female German native speakers between 20 and 31 years of age. The experiment took between 11 and 31 minutes. Participation was rewarded by course credits. Based on our experimental design, we put forward and tested the following five hypotheses:

(1) Monotonous stimuli will display the direct influence of the sonority of the critical segments on prominence perception, i.e. Rahmen should have the highest and Ratten the lowest prominence. Peak stimuli mirror this sonority influence.

(2) In pairs of words with different peak heights, a higher *Aa* will win out over a lower *Aa*.

(3) In pairs with medial-peak alignment, Rasten and Ratten will yield considerably lower prominence levels than Rahmen, Rasen and Raten, since the *F0* contour up to the peak maximum was intact for the latter three words, whereas the top portion was missing in the former two words.

(4) In pairs with late-peak alignment, the prominence level of Raten will fall off against those of Rahmen and Rasen, since the peak of Raten is masked by a silent pause. But Rasten and Ratten will still differ from the other words, as in (3).

(5) Listeners will perceptually fill the *F0* gap during the fricative in Rasten. However, parallel to findings with sinusoid stimuli [12], this filling-in will only occur for interpolations of *F0* slopes, i.e. peak maximum areas will not be extrapolated. Therefore, Rasten will win out over Ratten, but only for late peaks, and more clearly so in the stimulus pairs with *Aa*=0.8.

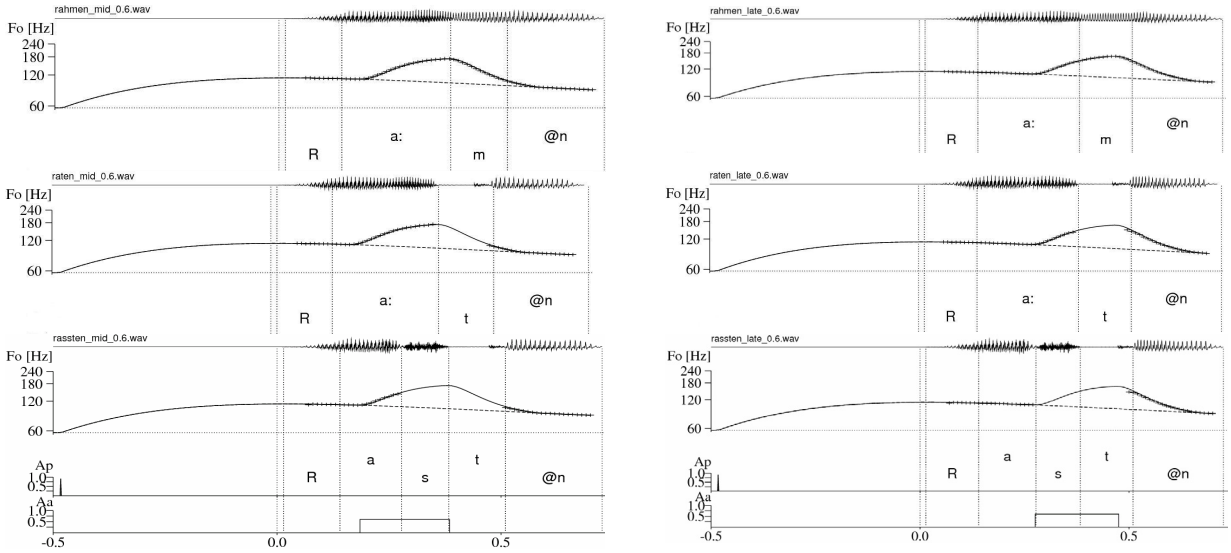

Figure 1. *Examples of stimulus words Rahmen, Raten and Rasten with medial (left) and late (right) peaks at Aa=0.6. Panels display waveform (top), F0 contour (+++extracted, —modeled, middle), and underlying phrase/accent commands (bottom).*

3. Results

The prominence judgments of our 22 German subjects were analyzed from the perspective of the second word in a stimulus pair. Analyses were performed separately for each hypothesis. The inferential statistics were based on Cochran's Q tests (the equivalent of a one-way repeated-measures ANOVA for binominal data) and included multiple comparisons between the levels of the independent variable after Sheskin [19]. The binary dependent variable was in all tests the relative prominence level of the second word (0= lower than first word; 1= higher than first word).

As regards hypothesis (1), Figure 2(a) displays that the percentages with which the second word in a pair was judged to be more prominent decrease successively from Rahmen to Ratten, in parallel to the sonority levels of the critical segments (cf. Tab.1). The corresponding Q test with Target Word as independent variable (5 levels, n=22 for each level) yielded a significant main effect (Q=17.13; *df*=4; p=0.002) and showed additionally that this main effect relied on differences between all target words except for Rasten vs Ratten.

The Q tests for hypothesis (2) were performed with Accent as independent variable. It combined peak alignment (medial, late) with *Aa* (0.4 vs 0.6 and 0.4 vs 0.8) and *Aa* order (first or second peak *Aa*= 0.4) and hence included 8 levels (n=22 for each level). The analysis included five Q tests, one for each target word. All five Q tests resulted in significant main effects of Accent: Rahmen (Q=70.62; *df*=7; p<0.001), Rasen (Q=71.99; *df*=7; p<0.001), Raten (Q=56.03; *df*=7; p<0.001), Rasten (Q= 39.55; *df*=7; p<0.001), and Ratten (Q=33.62; *df*=7; p<0.001). All main effects reflect that the word with the higher peak was perceived by our listeners to be more prominent. This difference, which is clearly represented in Figures 2(b)-(c), was significantly stronger for stimulus pairs with the extreme height contrasts (0.4 vs 0.8 or vice versa) than for pairs with the moderate height contrast (0.4 vs 0.6. or vice versa). Peak alignment (not shown in Fig. 2b-c) was irrelevant in the Q tests on Rahmen and Rasen, but it made a separate significant contribution in combination with the other three words. For Raten, the prominence differences between the height contrasts came out more clearly for medial than for late peaks. The opposite was true for Rasten and Ratten. That is, here it were the late peaks that brought out of the height-related prominence differences more clearly.

The results concerning hypotheses (3) and (4) are illustrated in Figures 2(d)-(e). The Q tests for hypotheses (3) and (4) were both based on the independent variable AccWord which was a combination of *Aa* (both peaks in a stimulus pair either 0.6 or 0.8) and target word (n= 22 for each of the resulting 10 levels). For the Q test of hypothesis (3), which concerned the stimulus pairs with medial peaks, there was a significant main effect of AccWord (Q=74.74; *df*=9; p<0.001). Taking the results of the multiple comparisons into account, this main effect was based on separate influences of *Aa* and target word. Prominence differences between the words were perceived more clearly for *Aa*= 0.8 than for *Aa*= 0.6 (Fig.2d). For example, Rahmen won out significantly more often over all other words in the 0.8 than in the 0.6 word pairs. In contrast, the prominence drop from Rasen to Raten was significantly larger in the 0.8 word pairs, which is reflected in the fact that Raten won significantly less prominence comparisons in the 0.8 than in the 0.6 condition. As can further be seen in Figure 2(d), the prominence levels decrease successively across the target words, the only non-significant difference being that between Rasten and Ratten. Here lies the only major difference between the Q tests for the medial and the late peak stimuli. The Q test for the late peak stimuli also yielded a significant main effect of AccWord (Q= 81.24; *df*=9; p<0.001). But, unlike in the Q test for the medial peak stimuli, this main effect includes a significant decrease in prominence level from Rasten to Ratten under both *Aa* conditions, however, more clearly so in the 0.8 condition than in the 0.6 condition. That is, Rasten won more and Ratten less prominence comparisons for *Aa*= 0.8 than for *Aa*= 0.6, cf. Figure 2(e). The differential effects of peak height (*Aa*= 0.6 vs 0.8) and peak alignment (medial vs late) on the perceived prominence levels of Rasten and Ratten, which were found in the Q tests on hypotheses (3)-(4), already provide an answer to hypothesis (5) so that the latter required no additional Q test.

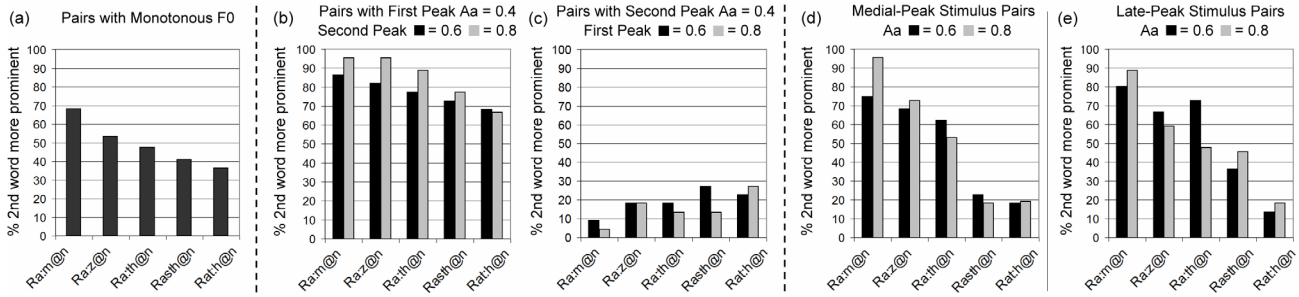

Figure 2. *Percentages of 'second word more prominent' judgments, displayed separately for each target words as against all other target words of the same stimulus condition. The data in (b) and (c) were pooled across medial and late peaks, n=22.*

4. Discussion and Conclusions

Most basically, our perception experiment provided further supporting evidence for the well-known fact that *F0* is a major cue to the perceived prominence of a syllable or word in German, cf. [3,4,7,8]. However, our findings also suggest refining the current F0-prominence link. It seems that not only *F0* movements and their excursion sizes are positively correlated with prominence. The mere presence of *F0*, including monotonous *F0*, as a supplier of acoustic energy and sonority, also affects prominence: The more *F0* is present in a linguistic unit (like a stressed syllable or word) the higher is its perceived prominence.

With regard to our hypotheses, this means that hypothesis (2) is unconditionally confirmed by our data. Those words whose *F0* movements were larger due to a larger *Aa* command were unambiguously and significantly perceived to be more prominent. Moreover, the prominence differences between the words in a pair increased when the *Aa* difference increased from 0.2 (0.4 vs 0.6) to 0.4 (0.4 vs 0.8). This was also true for word pairs based on Rasten and Ratten. The few contradicting prominence judgments (about 10% for most stimulus pairs) may be ascribed to a baseline error rate that is typical of perception experiments and, for example, results from biases of the presentation order or pressing the wrong buttons, particularly when each stimulus pair is only judged once by each subject. Hypothesis (1) can also be accepted, with one restriction though. It was found that a lower sonority level of the critical segment reduced the prominence level of that word relative to all other words with the same monotonous *F0*. However, this did not apply to the largest sonority difference between Rasten and Ratten (cf. Tab.1), which stems from a high-energy voiceless fricative in Rasten vs a long silence in Ratten. This suggests that segmental sonority differences are only relevant for prominence perception when they involve *F0*. In other words, it may be necessary to add to hypothesis (1) that *sonority differences between voiceless segments do not count for prominence*, at least not in combination with a monotonous *F0*, cf. hypothesis (5) below. This suggestion must be scrutinized in follow-up studies.

As regards the effects of *F0* contour continuity on prominence perception and the related issue of the perceptual filling-in of *F0* gaps, our findings show the following. First, when the prominence levels of two different words had to be directly compared in a stimulus pair, the winning word was significantly more often that in which more of the *F0* peak contour was physically present (Fig.2d-e). Likewise, when pairs of the same word but with different peak excursions had to be compared, prominence differences were significantly clearer for those words in which more of the *F0* peak contour was physically present (Fig.2c-d). Both effects were stronger for a larger peak excursion (i.e. *Aa*= 0.8) and when only a part of the *F0* slope rather than the top portion of the peak and hence the exact *F0* excursion size was missing in the signal (cf. long-vowel vs short-vowel target words). These findings agree with hypotheses (3)-(4). This also means that, even though we used simpler stimuli and a more meaning-oriented task, our findings support the claim of [10]: Listeners do not simply fill in *F0* gaps during speech perception.

However, two important differences between Rasten and Ratten argue against the general validity of this claim. Shifting the accent contours in Rasten and Ratten from a medial to a late position widely restored the top portion of the peak. In this interpolation condition, in which only the rising *F0* slope was absent from the signal, Rasten gained a significantly higher prominence level than Ratten for both excursion sizes, *Aa*= 0.6 and 0.8 (Fig.2e). In contrast, in the extrapolation condition of the medial-peak (Fig.2d), Rasten and Ratten yielded equal prominence levels. Additionally, when the same word was presented with different peak excursions (Fig. 2b-c), a larger excursion difference (of late peaks) also resulted in a clearer prominence difference for Rasten but not for Ratten. These two differences between Rasten and Ratten suggest – in accord with the idea of "segmental intonation" [11] – that voiceless fricatives like [s] differ from voiceless plosives like [t:] in that the former can in fact trigger a perceptual filling-in of *F0* gaps. Like for sinusoid stimuli [12], such a filling-in seems to be restricted to interpolation conditions (e.g., a missing rise), i.e. it does not occur if the missing top portion of the peak must be extrapolated. So, hypothesis (5) is confirmed by our findings on German.

In summary, concerning the perception of interrupted *F0* contours in speech, our findings call for a differentiated statement: *Listeners do not fill in all F0 gaps, but they seem to fill in some (non-silent) F0 gaps*. Thus the fact that utterance tunes are *"certainly subjectively continuous"* [20:275] seems to be due to two perceptual processes, ignoring gaps [10] and filling gaps [11]. Subsequent studies must provide further evidence for this assumption and also address the question if filling-in of gaps involves merely a fuzzy pitch impression or an actual continuation of the pitch curve. Initial evidence of [11] is in favour of the latter, which stresses that treating unvoiced fricatives and plosives differently is not just important for modelling prominence but also for modelling intonation.

5. Acknowledgements

We thank students at Beuth University, Berlin, Germany, for their participation, and especially Martin Stenzel for providing the experiment software. Thanks also go to Angelika Hönemann for supervising the tests.

6. References

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

[2] Gay, T., “Physiological and acoustic correlates of perceived stress. Language and Speech 21, 347-353, 1978.

[3] Koreman, J., Van Dommelen, W., Sikveland, R., Andreeva, B., Barry, W.J., ”Cross-language differences in the production of phrasal prominence in Norwegian and German”, in M. Vainio, R. Aulanko, O. Aaltonen (eds.): Nordic Prosody X. Frankfurt: Peter Lang, 139-150, 2009.

[4] Mixdorff, H., Widera, C., “Perceived Prominence in Terms of a Linguistically Motivated Quantitative Intonation Model”, Proc. Eurospeech 2001, Aalborg, Denmark, 403-406, 2001.

[5] Fujisaki, H., Hirose, K., “Analysis of voice fundamental frequency contours for declarative sentences of Japanese”, Journal of the Acoustical Society of Japan 5, 233-241, 1984.

[6] Heuft, B., “Eine prominenzbasierte Methode zur Prosodieanalyse und –synthese”, in W. Hess and W. Lenders (eds.): Computer Studies in Language and Speech, Vol. 2, Peter Lang, Frankfurt am Main, 1999.

[7] Niebuhr, O., “Interpretation of pitch patterns and its effects on accentual prominence in German”, Proc. Tone and Intonation in Europe 3, Lisbon, Portugal, 2008.

[8] Niebuhr, O., “F0-based rhythm effects on the perception of local syllable prominence”, Phonetica 66, 95-112, 2009.

[9] Kleber, F., Niebuhr, O., “Semantic-context effects on lexical stress and syllable prominence”, Proc. 5th Speech Prosody, Chi-cago, USA, 1-4, 2010.

[10] Barnes, J., Brugos, A., Veilleux, N., Shattuck-Hufnagel, S., “Voiceless Intervals and Perceptual Completion in F0 contours: Evidence from scaling perception in American English”, Proc. 16th ICPhS, Hong Kong, China, 108-111, 2011.

[11] Niebuhr, O., “At the edge of intonation – The interplay of utterance-final F0 movements and voiceless fricative sounds”, Phonetica 69, 7-27, 2012.

[12] Dannenbring, G.L., “Perceived auditory continuity with alternately rising and falling frequency transitions”, Canadian Journal of Psychology 30, 99-114, 1976.

[13] Bregman, A.S., “Auditory Scene Analysis: The perceptual orga-nization of sound”, Cambridge: MIT Press, 1990.

[14] Kohler, K.J., “Timing and communicative functions of pitch contours”, Phonetica 62, 88-105, 2005.

[15] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., van der Vreken, O., “The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes”, Proc. ICSLP, Philadelphia, USA, 1393-1396, 1996.

[16] Mixdorff, H., “FujiParaEditor”, http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html, 2009.

[17] Boersma, P., “Praat, a system for doing phonetics by computer”, Glot International 5, 341-345, 2001.

[18] Niebuhr, O., Ambrazaitis, G.I., “Alignment of medial and late peaks in German spontaneous speech”, Proc. 3rd Speech Prosody, Dresden, Germany, 161-164, 2006.

[19] Sheskin D.J., “Handbook of parametric and nonparametric statistical procedures”, Boca Raton: Chapman & Hall, 2004.

[20] Jones, D., “Intonation curves”, Teubner, Leipzig, 1909.

[21] Terken, J., ”Fundamental frequency and perceived prominence”, Journal of the Acoustical Society of America 89, 1768-1776, 1991.