Pitch Accents as Multiparametric Configurations of Prosodic Features – Evidence from Pitch-accent Specific Micro-rhythms in German

Oliver Niebuhr
University of Southern Denmark

Abstract
Pitch accents are typically described in terms of the alignment and shape of their F0 peaks. However, some studies suggest that pitch-accent peaks also create systematic duration and intensity changes in the triplet of pre-accented, accented, and post-accented syllable. The present study examines this phenomenon in detail for three rising-falling German pitch accents: the early, medial, and late peak. A production experiment with 4 speakers finds clear acoustic evidence for these systematic duration and intensity changes. In addition, these changes also manifest themselves in a parallel dataset of syllable-synchronous finger tapping. In combination, the changes of two prominence-relevant acoustic parameters, i.e. syllable duration and intensity, and the reflection of these changes in a rhythmical finger-tapping task suggest that nuclear pitch-accent categories in German are not purely intonational phenomena but multiparametric prosodic configurations (i.e. "prosodic constructions") that include, besides their F0-peak characteristics, a pitch-accent-specific micro-rhythm in the triplet of pre-accented, accented, and post-accent syllable. The implications of this conclusion for intonational modeling are discussed.

1. Introduction

It is 30 years ago now that Kohler (1987) published his seminal paper on the categorical perception of F0-peak alignment. Kohler shifted a constant sharply rising-falling nuclear pitch-accent peak in 11 steps

Anne Mette Nyvad, Michaela Hejná, Anders Højen, Anna Bothe Jespersen & Mette Hjortshøj Sørensen (Eds.), ***A Sound Approach to Language Matters – In Honor of Ocke-Schwen Bohn*** (pp. 321-351). Dept. of English, School of Communication & Culture, Aarhus University.

across the sentence "Sie hat ja gelogen" (She's been lying, with the relevant nuclear pitch accent on [lo:] of "gelogen" [ɡ̊ɪlo:gŋ]). For each of the 11 equidistant 30-ms steps of the F0 peak-shift continuum, a stimulus was resynthesized. The resulting 11 stimuli were included in a serial discrimination test, a 2AFC AX pair discrimination test, and a 2AFC indirect identification test, in which listeners judged the stimulus sentence as either matching or not matching with a constant preceding context utterance (see also Nash & Mulac, 1980 further explanations on this test paradigm and the discussion of semantic tasks in Gussenhoven, 1999). Based on the integrated results of these experiment series, Kohler found a categorical change in perception for those stimuli whose F0 peak maximum was no longer located before but inside the accent vowel of [lo:] in "gelogen" and another, slightly weaker categorical change in perception as soon as the F0 peak was shifted with is maximum out of the accented vowel (Kohler, 1991a).

It is probably not an exaggeration to state that the finding of a categorically perceived F0-peak alignment continuum marked an important milestone for the development of phonological models of intonation and for the linguistic analysis of intonation in general. Categorical perception showed to researchers of those days that the gap between segmental elements and melodic elements (like F0 peaks) was not as big as had been commonly assumed, and that intonation would thus be open to linguistic approaches and phonetic analyses in a similar way as sound segments are. Other milestones were undoubtedly the works of Bruce (1977) and Pierrehumbert (1980), whose frameworks and conclusions also shaped Kohler's expectation about the perceptual organization of the F0 peak-shift continuum in German and, thus, about the basic principles of Kohler's Kiel Intonation Model, KIM. With reference to the perceptual tripartition of his F0 peak shift continuum and the alignment characteristics of each perceptual category relative to the boundaries of the accented vowel, Kohler (1987, 1991a) called the three identified pitch accents the early peak (i.e. the F0 maximum is before the vowel), the medial peak (i.e. the F0 maximum is inside the vowel), and the late peak (i.e. the F0 maximum is after the vowel). In terms of their communicative function, early peaks are used to mark a piece of information as being settled or unchangeable. Medial peaks signal new information and openness to discussing this new information with the interlocutor. Late peaks also signal new information, but additionally mark this new information as being in contrast to the speaker's expectation (Kohler, 2005). Depending

on the context, the early-peak meaning can also express resignation. The late-peak meaning can express either positive surprise or indignation (Niebuhr, 2007a). In major autosegmental-metrical (AM) models of German intonation, such as GToBI, the three pitch-accent categories correspond to H+L* (or H+!H*), H*, and L+*H, see Grice et al. (2005). And since the accents consist of rising-falling intonation peaks, GToBI would also add a L-% to each label in phrase-final (nuclear) position, which is the position relevant to the present paper.

The historic experimental genesis of the early, medial, and late peak as well as their acoustic definitions by Kohler are probably well known among most intonation researchers. Perhaps less well known, however, is the fact that Kohler encountered slightly different results in a later replication of his peak-shift experiment (Kohler, 1991b). For example, for the same F0 peak-shift continuum in the stimulus sentence "Sie hat ja gejodelt" (She's been yodeling, the relevant nuclear accent was on the final word and its syllable [jo:]) the perceptual change from early-peak to medial-peak perception occurred significantly earlier than in the original "gelogen"-sentence. That is, replacing the accent syllable "-lo-" [lo:] by "-jo-" [jo:] had a decisive influence on the category boundary. Kohler (1991b) explains this different outcome by the less sharp segment boundary between the accented vowel and the preceding approximant in [jo:] as compared to [lo:]. Unlike in [lo:], the continuous movement of the tongue in [jo:] does not create an inherent articulatory and spectral discontinuity landmark. On this basis, Kohler argues that the listeners were unable to detect the segment boundary accurately.

Niebuhr (2006, 2007b) countered this argument by pointing out the fact that a blurred segment boundary would have also led to a blurred, i.e. less categorical change in perception from early to medial peak. However, there is no evidence for such a weaker category boundary in Kohler's data. So, in order to find an alternative explanation for the fact that the pitch-accent boundary is closer to the accented-vowel onset in "Sie hat ja gejodelt" than in "Sie hat ja gelogen", it was necessary to look for further aspects, in which "gejodelt" differed from "gelogen". Niebuhr (2006, 2007b) focused on the intensity contour. Due to the approximant at the syllable onset of [jo:] in "gejodelt", the intensity increase into the accented vowel begins earlier and at a higher level than in the case of the alveolar lateral in [lo:] of "gelogen". Due to its higher level, the intensity increase is also shorter and culminates earlier in the accented vowel than in [lo:] of "gelogen".

On this basis, Niebuhr (2006, 2007b) hypothesized that the key factor in the alignment characteristics of early, medial, and late peaks is not the position of the F0 peak maximum relative to the spectrally defined segment boundary of the accented vowel (the vowel onset in the transition from early to medial; and the vowel offset in transition from medial to late). Rather, Niebuhr assumed that the actual key factor in the transition from early to medial and from medial to late peak perception would be the positioning of F0 and intensity movements or their maxima relative to one another.

Niebuhr gained experimental-empirical evidence for this hypothesis in a perception experiment whose methodology is largely adopted from the seminal study of Kohler (1987). Two stimulus series were created. The first series resulted from shifting a pointed rising-falling F0 peak in 11 steps from an early to a medial position into the accented vowel of "Ma-" in "Sie war mal Malerin" (She was once a painter, with the nuclear pitch accent being on [maː] of "Malerin" [maːləʁɪn]). The second series was derived from the first series such that each stimulus showed exactly the same F0 and intensity patterns as in the first series. Only the segmental string was removed and replaced by a constant schwa-like sound ('hum' in PRAAT). The stimuli of the two series were judged by different groups of German native speakers. The judgments for the first stimulus series were made on the basis of an indirect-identification test. The stimuli of the second series were presented in AXB triplets, with A and B being the first or the last stimulus of the series. Similar to the indirect-identification test in which the speech stimuli were judged on a semantic basis as either matching or not matching with a given context utterance, the frame of A and B in the AXB triplets also provided a constant context frame against which the individual 'hum' stimuli (X) could judged - on a melodic basis – as either matching or not matching (with A or B, respectively). In this sense, the listeners' tasks in the two experiments were designed to be as closely related as it was possible for a comparison of speech and non-speech stimuli.

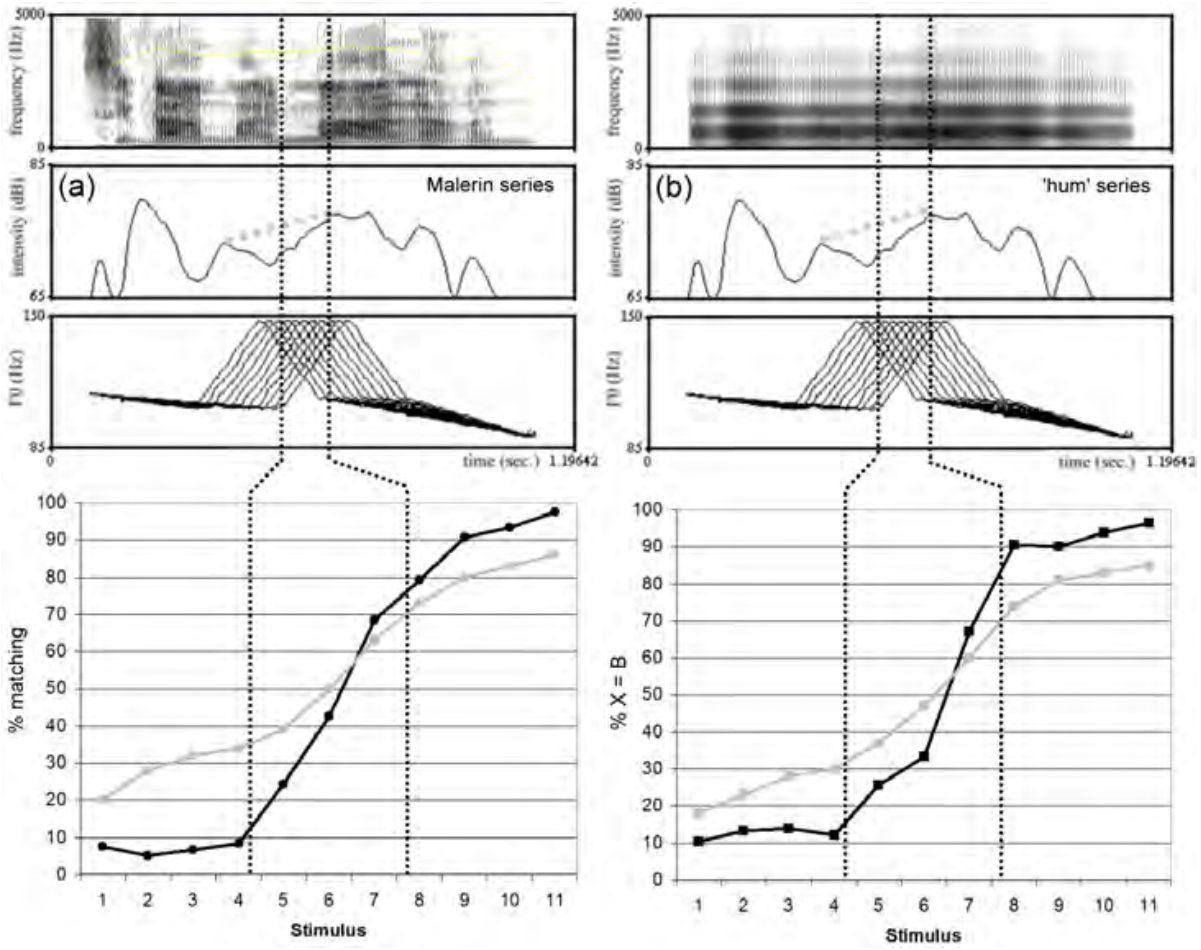

Figure 1. Top: the 11 stimuli of the 'Malerin' (a) and 'hum' (b) series. Bottom: percentages (n=140) of medial-peak intonations in terms of 'matching' (a) or (b) 'X=B' judgments; grey lines in top and bottom panels refer to the repeated experiments, in which the intensity increase into the accented vowel was more gradual.

As is shown in Figure 1, the first stimulus series yielded an abrupt change from the early-peak to the medial-peak category, just as in Kohler's original "gelogen" series. The crucial new finding is, the second series ('hum') was able to replicate this perceptual change from early to medial peak perception solely on the basis of the F0 and intensity patterns of the first series. Moreover, the change from early to medial peak intonation exactly coincided with the intensity increase from the low level of the consonant to the high level of the vowel of the accented syllable. This coincidence made Niebuhr (2007b) repeat the experiment of Niebuhr (2006) with a single modification: the steep intensity increase across the CV boundary in the stimuli was turned into a more gradual one by means of the intensity-course manipulation procedure in Adobe Audition, cf.

the gray lines in Figure 1. As a result, the less dynamic intensity increase was clearly paralleled by a less dynamic perceptual transition from early to medial peak intonation across both the original and the delexicalized 'hum' stimulus series. Niebuhr (2007b) also applied the same experimental procedure to a F0 peak-shift continuum from medial to late and gained similar results. That is, the intonation judgments for the 'hum' stimuli were statistically identical to those of the original speech stimuli, and the perceptual change from medial to late was the more gradual the more gradual the intensity decrease was after the accented-vowel offset.

However, it is not just that the intensity contour characteristics underlying a F0-peak pattern influence its pitch-accent identification. The intensity characteristics are also systematically varied by speakers in the production of pitch accents. In selecting and creating base stimuli for his perception experiments, Kohler (1991c, p. 144) already noted a "natural parallelism" of the F0 and intensity patterns in the production of pitch accents. Moreover, own informal listening made Kohler assume that these "coupled time courses [of F0 and intensity] are expected by listeners" (Kohler 1991c, p. 188), because breaking this natural parallelism (e.g., by manipulating the F0-peak alignment) seems to have, in Kohler's ears, negative consequences for the naturalness of the respective sentences and the clarity with which the pitch accents inside these sentences are perceived.

Niebuhr & Pfitzinger (2010) took up Kohler's idea and investigated the production and perception of the F0 and intensity contours of pitch accents in more detail. An acoustic analysis of read-speech material showed, not surprisingly, that the accented syllable always had higher duration and intensity levels than the two surrounding syllables. However, in addition, the read-speech material revealed pitch-accent-specific intensity levels in the triplet of pre-accented, accented, and post-accented syllable. Moreover, the variation in the intensity patterns was linked with a variation in syllable duration, a phenomenon which was already briefly pointed out by Gartenberg & Panzlaff-Reuter (1991). As is displayed in Figure 2 (upper panel), the early peak was consistently produced with high duration and intensity levels in the pre-accented syllable, whereas the duration and intensity levels in the post-accented syllable were both relatively low. The late peak had an opposite effect on the duration and intensity levels in the pre- and post-accented syllables. That is, the post-accented syllable was consistently realized with higher duration and intensity levels than the pre-accented syllable. Compared to both the early and the late peak, the medial peak was characterized by a

roughly symmetrical duration and intensity pattern across the triplet of pre-accented, accented, and post-accented syllable. While the accented syllable clearly stood out in terms of duration and intensity, the pre-accented and post-accented syllables were both produced at similarly low duration and intensity levels relative to the accented one.

Based on these production patterns, Niebuhr & Pfitzinger (2010) conducted a perception experiment with a semantic differential. They used two types of stimuli: (1) PSOLA resyntheses of original "Eine Malerin" (a painter) utterances being produced with early-, medial- and late pitch-accent peaks on the nuclear-accent syllable [ma:] of "Malerin"; and (2) PSOLA resyntheses of these original "Eine Malerin" productions but with interchanged F0 contours. The condition-(1) stimuli were only resynthesized in order to impair their sound quality in the same way as for the condition-(2) stimuli. The interchanged F0 contours in the condition-(2) stimuli had the same proportional F0-peak alignments relative to the vowel boundaries as in the condition-(1) stimuli. Naturally produced differences in F0-peak shape between the early, medial, and late peak categories were also retained.

The stimuli were presented with multiple repetitions and in differently randomised orders to native speakers of German. However, the stimuli of condition (1) were judged in a separate block after those of condition (2). The results are perfectly in accord with Kohler's (1991c) assumptions, see Figure 2, lower panel. Firstly, the stimuli with interchanged F0 contours sounded significantly more artificial than the original stimuli. Secondly, for the stimuli with interchanged F0 contours there was a separate effect of the original duration and intensity pattern. It biased judgments towards the semantic profile of that pitch-accent category with which the duration and intensity pattern was naturally produced. Thus, the findings suggest that listeners are sensitive to the duration and intensity patterns that co-occur with a certain pitch accent, and that duration and intensity make a separate contribution to identifying or conveying the communicative functions of that pitch accent. Later, follow-up experiments by Niebuhr (2011) suggest further that listeners have an internal representation of the pitch-accent-specific duration and intensity patterns shown in the upper panel of Figure 2. The "Eine Malerin" utterances were resynthesized with completely flattened F0 course and listeners were asked, in one experiment, to repeat the corresponding utterance in a more melodic fashion (and with stress on "Malerin") or, in another experiment, to draw the speech melody that they consider appropriate for the corresponding stimulus utterance on a sheet of paper. Both

experiment tasks yielded a significant correlation between the originally produced but then flattened and hence removed pitch accent in the stimulus utterance and the pitch accent that was reinserted into the repeated stimulus utterance or drawn by the participant. The only possible acoustic cues that had been able to guide these performances were the retained duration and intensity patterns in the stimuli.

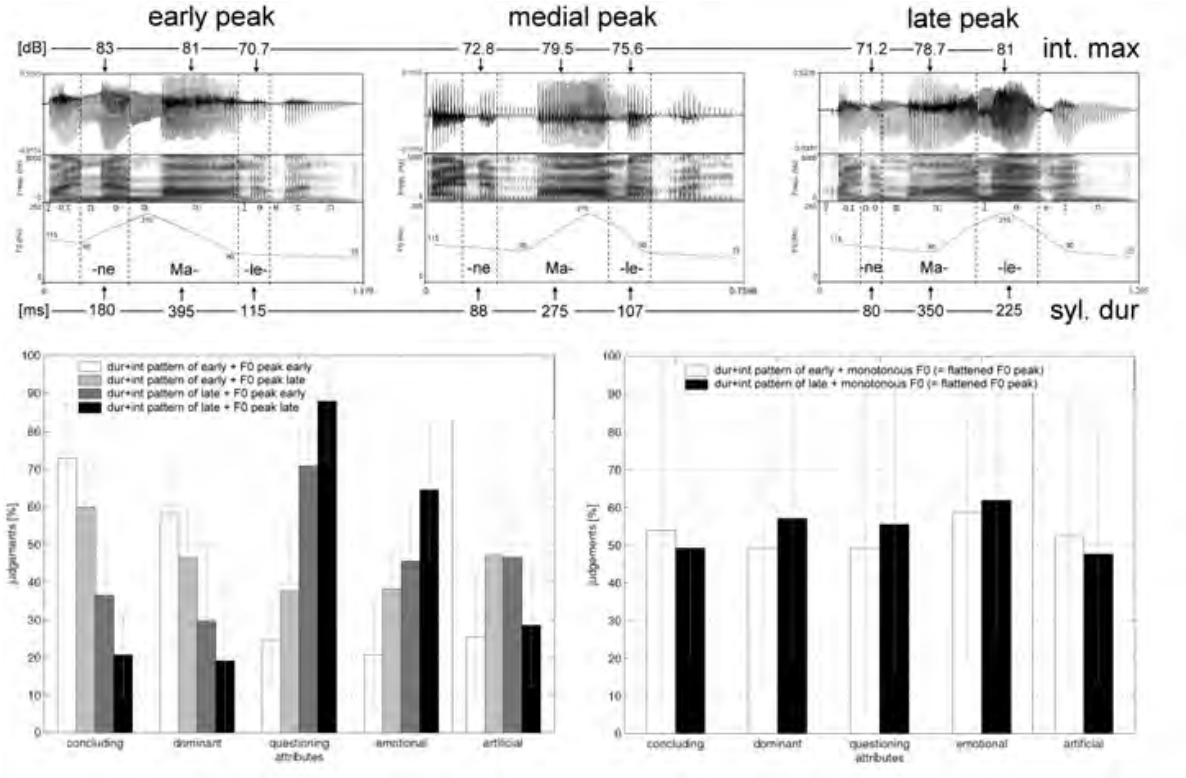

Figure 2. Top panel: Acoustic characteristics of pre-accented, accented, and post-accented syllables in "Eine Malerin" (a painter), produced with the early (left), medial (mid), and late (right) pitch accent on the accented syllable "Ma-". Bottom panel: Listener judgments on the key meaning dimensions of early and late pitch accents when being presented in original and exchanged duration and intensity contexts (left) and with flattened F0 peaks (right).

Since both duration and intensity are also important triggers of perceptual prominence in German, Niebuhr & Pfitzinger (2010) decided to refer to these characteristic duration and intensity levels that co-occur with German pitch accents and shape the triplet of pre-accented, accented, and post-accented syllable as *pitch-accent-specific micro-rhythms*. The term "rhythm" was used because Niebuhr & Pfitzinger indeed noted, on an informal auditory basis, a characteristic tri-syllabic prominence sequence for the early, medial, and late pitch accent; and a rhythm is nothing else than a sequence of different syllable-based prominences.

However, Niebuhr & Pfitzinger also noted that this rhythm is relatively subtle in perception and embedded in a larger sequence of higher-level syllable prominences, determined by the lexical-stress realizations and pitch-accent distributions in utterances. Therefore, and with respect to the restricted three-syllable time domain in which these prominence differences occurred, Niebuhr & Pfitzinger used the term "micro-rhythm".

Against the outlined body of empirical evidence on the role of duration and intensity patterns in the production and perception of German early, medial, and late pitch accents, the point of departure of the present study is as follows: Although Niebuhr & Pfitzinger coined the term pitch-accent-specific micro-rhythm for the duration and intensity patterns they found, there is no direct evidence as yet, that the characterization as "rhythm" is actually appropriate. That is, the question addressed here is whether the duration and intensity patterns across the triplet of pre-accented, accented, and post-accented syllable in fact represent a rhythm in the performance-oriented, prominence-related, and cognitive sense of the term.

In order to shed light on this issue, a formal speech-production experiment was performed. The experiment makes use of a method that has been successfully applied for "tapping into naïve listeners' intuitions about speech rhythm" (Cumming, 2010, p. 158) for more than a century: syllable-based finger tapping (cf. the historic experiments by Brücke, 1871; Meyer, 1898; and Scripture, 1902).

Rhythm is a phenomenon that, as dancing shows impressively, can connect the beats or prominence structures that listeners perceive with their physical body movements. This cross-modal mechanism is used for the purpose of the present study. Note that it is controversial in phonetic studies to what extent and how exactly a participant's individual finger tapping is time-aligned with the acoustic and perceptual boundaries of the syllables s/he perceives (see Wagner, 2008 and Cumming, 2010 for summaries). But, this potential problem is irrelevant to the present production experiment, because the present production experiment is not about *when* exactly relative to a syllable the participant's finger hits the targeted object (such as the tabletop or the push-button of a technical device). Rather, the present experiment is about *how strongly* and *for how long* the participant's finger hits the targeted object. Recent results of Samlowski & Wagner (2016) support the assumption underlying this method that there is a positive correlation between the perceived prominence of a syllable and the power and duration of the finger tapping

for that syllable (see also Parrell et al., 2014). Moreover, finger tapping (or "drumming" in the case of Samlowski & Wagner, 2016) proved to be a "a fast, intuitive and exact method" that yields fine-grained prominence patterns "for experienced and naive subjects alike" (Samlowski & Wagner, 2016, pp. 1,5). On this basis, we expect the following results to emerge for our production experiment.

- Irrespective of the pitch accent, the finger tapping for the accented syllable is always stronger and longer than for the two surrounding syllables.
- An early peak leads to an asymmetrical finger tapping pattern with an overall declining strength and duration. That is, the finger tapping is stronger and longer for the pre-accented syllable than for the post-accented syllable.
- A late peak results in an asymmetrical finger tapping pattern with an overall increasing strength and duration. That is, the finger tapping is weaker and shorter for the pre-accented syllable than for the post-accented syllable.
- A medial peak leads to a symmetrical finger tapping pattern. That is, the finger tapping is similarly weak and short for both the pre-accented and the post-accented syllable, and only strongly pronounced in terms of power and duration for the intervening accented syllable.

In addition, we analyze the finger-tapped sentences acoustically and expect that the results from Niebuhr & Pfitzinger (see Figure 2) will be replicated. This means that

- duration and intensity levels are higher for the accented syllable than for the two framing unaccented syllables.
- For the early peak, the duration and intensity levels of the pre-accented syllable are higher than those of the post-accented syllable.
- For the late peak, the duration and intensity levels of the pre-accent syllable are lower than those of the post-accented syllable.
- For the medial peak, the duration and intensity levels of the pre- and post-accented syllables are similarly low relative to those of the the accented syllable.

If the pitch-accent-specific micro-rhythm is not an integral part of the representation and production of pitch accents in German, but, for example, an epiphenomenon of another F0-related factor (perhaps of a different magnitude or onset of increased effort in the coordination of glottal and supraglottal gestures), then we still expect that the

speakers of the speech production experiment are able tap the syllables of the utterances in parallel to their production. However, under these circumstances, we would expect the tapping to be either homogeneous in the relevant triplet of pre-accented, accented, and post-accented syllable, i.e. each of the three syllables should be tapped with the same duration and intensity; or we would expect that only the macro-rhythm of the utterances would manifest itself in the finger-tapping. In this case, the accent syllable would always be tapped stronger and longer than the two surrounding syllables, while the latter two do not differ, regardless of the category of the pitch accent on the accent syllable. Only if a pitch accent is represented, planned and executed as a sequence of specifically varying syllable prominences should this be reflected in a pitch-accent-specific finger tapping.

2. Method

2.1 Participants

The study is based on realizations of target sentences by four native German speakers. The number of speakers was small but deliberately chosen with respect to validity considerations. That is, instead of recruiting a large number of naive speakers and then trying to elicit the early, medial, and late peaks on target utterances in a consistent fashion by means of specifically tailored semantic-pragmatic context precursors (Niebuhr & Michaud, 2015; Kohler, 2017), we used experienced intonation researchers as participants who were well trained in the contrastive production and perception of early, medial, and late peaks. In pilot studies, this solution proved to be better for several reasons.

For example, it turned out to be problematic for naive speakers to produce target sentences with the intended (i.e. context-matching) intonation contours, while simultaneously tapping syllable by syllable the rhythm of these target sentences. Under this condition of double cognitive workload, naive participants produced, virtually independently of context precursors, the same nuclear pitch accent, namely the medial peak, which is considered the default pitch-accent category in German. Medial peaks account for 53 % of all nuclear pitch accents the Kiel Corpus of Spontaneous Speech (Peters, 1999; Peters et al., 2005). In addition, the use of trained speakers prevented the sentences and, thus, the relevant pitch-accent patterns from being realized in an exaggerated enacted speech style, which typically occurs when naive speakers are asked to realize target sentences in expressive-emotional contexts (such

as those that would have been required for eliciting early and late peaks). The consequences of such a speech style for the external validity of the findings would have been difficult to estimate. Another reason that argued against the use of naive speakers was the internal validity of the data, more precisely the avoidance of circular reasoning. Previous studies showed that especially early and late peaks cannot be triggered and elicited with 100 % reliability by semantic-pragmatic contexts alone (Niebuhr, 2007c). However, it would have also been difficult to re-classify pitch-accent patterns with reference to acoustic or auditory criteria, because, as was pointed out by Pfitzinger & Niebuhr (2010), pitch-accent-specific micro-rhythms are in an acoustic and perceptual interplay with the alignment of F0 peaks. Thus, any post-hoc re-classification of pitch-accent patterns would have involved the risk that we either take into account this interplay and, in this way, make the actual object under investigation the criterion according to which we organize our sub-samples; or that we deliberately ignore this interplay and, thus, bias our samples and results. By using fewer speakers, but speakers who were trained in the natural production of early, medial and late peaks, these problems can be circumvented. That is, every pitch-accent pattern that these speakers produced and approved (after possible self-correction) was simply accepted as a successful rendering of the intended early, medial, or late peak.

The 4 speakers were between 25 and 41 years old. Two speakers were female and two were male. All 4 grew up in Northern Germany and were native speakers of Standard Northern German. All had normal hearing and speech skills and worked as PhD students or belonged to the scientific staff of the Dept. of Linguistics at Kiel University.

2.2 Reading Material

The reading material consisted of 20 target sentences, which were realized in isolation without any pitch-accent supporting context. All target sentences had 6-7 syllables. The syllables were embedded in a syntactic structure of personal pronoun (she/they), verb, and noun (direct object), like, for example, “Sie war mal Malerin” (She was once a painter), “Sie leben in Sambia” (They live in Zambia), or “Sie haben Sonnenbrand” (They have a sunburn). The noun elicited the relevant nuclear pitch accent on its initial CV(C) syllable. Thus, the nuclear accent was always in the second half of each target sentence and occurred 2-3 syllables before the end of the sentence. The syllable preceding the noun was always unaccented and represented a so-called “weak form”. That is, it was either a particle, a preposition, or a grammatical suffix morpheme.

All target sentences were phonetically largely voiced, especially in the triplet of pre-accented, accented and post-accented syllable.

The set of 20 target sentences was completed by 6 syntactically and phonetically similar sentences, half of which were placed as dummies before and after the 20 target sentences. The frame of three initial and final dummy sentences was to avoid the prosodic list effects that occur "almost invariably" when speakers read sequences of isolated target sentences (Ladefoged, 2003, p. 7).

Overall, the reading material comprised 26 individual sentences.

2.3 Procedure

The recording of the reading material took place for each of the 4 speakers in individual sessions. At the beginning of the session, the speaker was given the list of 26 sentences and asked to familiarize him/herself with the sentences for about 3-5 minutes.

Subsequently, the speaker was instructed to realize the sentences as separate (i.e. context-free) statements at a normal speaking rate and with a normal, clearly pronounced reading style and intonation (see Mixdorff & Pfitzinger, 2005 and Barbosa, 2015 for the prosodic characteristics of read as compared to spontaneous speech). Furthermore, the speaker received the information that there would be three rounds of recording. In the first round, each statement was to be produced with a medial peak as the nuclear pitch accent, in the second round with an early nuclear pitch accent, and in the third round with a late nuclear pitch accent. The pitch-accent elicitation order represented the subjective difficulty level with which the three accent categories can be produced (from less to more difficult). The order was constant across all 4 speakers (a complete permutation of the elicitation order would not have been possible with only 4 speakers anyway).

Then the speaker was told that, in addition to producing the sentence, s/he should in parallel tap the rhythm of each sentence with his/her index finger in syllable-by-syllable fashion. To that end, the experiment used an innovative device - a modified DJ drum pad AKAI MPD18 - that recorded the onset, offset, and power (reflected in the amplitude of the sound that it generates) of the speaker's finger tapping in parallel to his/her speech signal, in a way similar way as did the "Sentograph" device that had been developed by Manfred Clynes in 1925 (see Kopiez & Lehmann, 2013). The drum-pad button, which was to be used for finger tapping, was on the top right corner of the device, where it was most convenient to reach for the speaker's index finger. The button was also marked in red color.

The drum pad was placed on the table to the right of the speaker (all 4 speakers were right-handed). The speaker again received 3-5 minutes in order to familiarize him/herself with this simultaneous speaking-and-tapping task, using sentences of his/her choice from the list of 26. The speech tempo was initially slowed down significantly by this dual-task paradigm. However, at the end of this second familiarization phase, it had returned to the normal level of each speaker, i.e. about 4 syllables per second.

After the two familiarization phases, the actual speech recording began. The 26 sentences were presented to the speaker individually on a PC screen in font size 38 (Times New Roman, see Berger et al., 2016 for the advantages of the chosen typeface in speech-elicitation tasks). The speaker received the sentences in a constantly re-randomized order, i.e. an order that was always new in each round of recording and for each speaker. A separate re-randomization was also performed for the 6 dummy sentences. However, they remained consistently placed in triplets at the beginning and end of the list. The participant spoke into a gooseneck microphone (Sennheiser MD 421-U) that was positioned to the left of the screen. Figure 3 shows a sample photo of the recording setting.

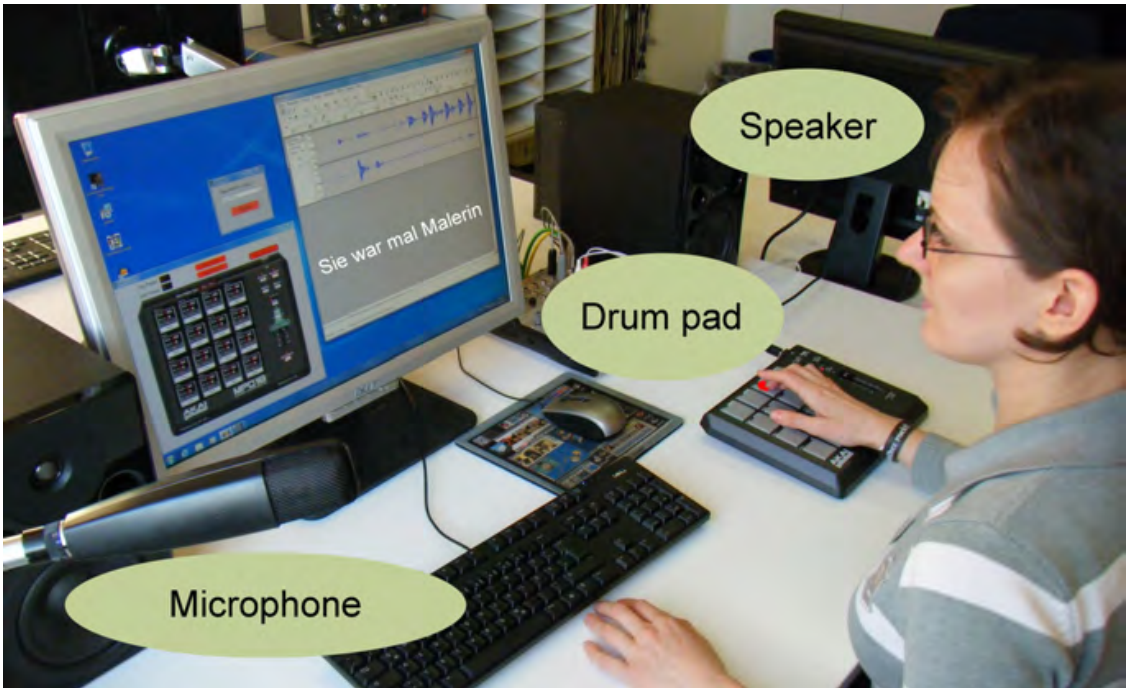

Figure 3. Edited photograph showing the recording setting with the drum pad being placed to the right of the speaker and the gooseneck microphone being located to the left of the speaker, next to the screen in the center on which the 26 target/dummy sentences were displayed individually.

The recording was carried out in a self-paced fashion. That is, the speaker pressed a button and moved on to the next sentence on the screen whenever s/he was satisfied with the production of the current sentence (especially regarding its nuclear pitch-accent pattern). On average about 15 % of the sentences (4 out of 26), were realized several times by speakers, either because the speakers corrected a slip of the tongue, or because the nuclear pitch accent was not implemented well or clearly enough in the ears of the speaker. Some sentences were also re-read because of a miscoordination between the tapping of the finger and syllables in speech production.

2.4 Data Analysis

The sound files of the recording sessions were stored as stereo signals. On the left channel was the speech signal, and on the right channel was the time-aligned finger-tapping signal. The latter signal was recorded in the form of a sinusoid (with a constant frequency). The beginning and the end of the sinusoid marked those points in time at which the speaker had touched and released the button of the drum pad; and the amplitude of the sinusoid indicated the maximum power with which the button of the drum pad was pressed down by the finger. Figures 4(a)-(b) presents two examples of recorded stereo signals. The upper example shows the tapping-and-speaking signal of "Sie mögen Blumen sehr" (They like flowers very much) produced with a nuclear late-peak accent on "Blu-" [blu:]. The lower example shows the tapping-and-speaking signal of "Sie war mal Lehrerin" (She was once a teacher) produced with a nuclear medial-peak accent on "Leh-" [le:].

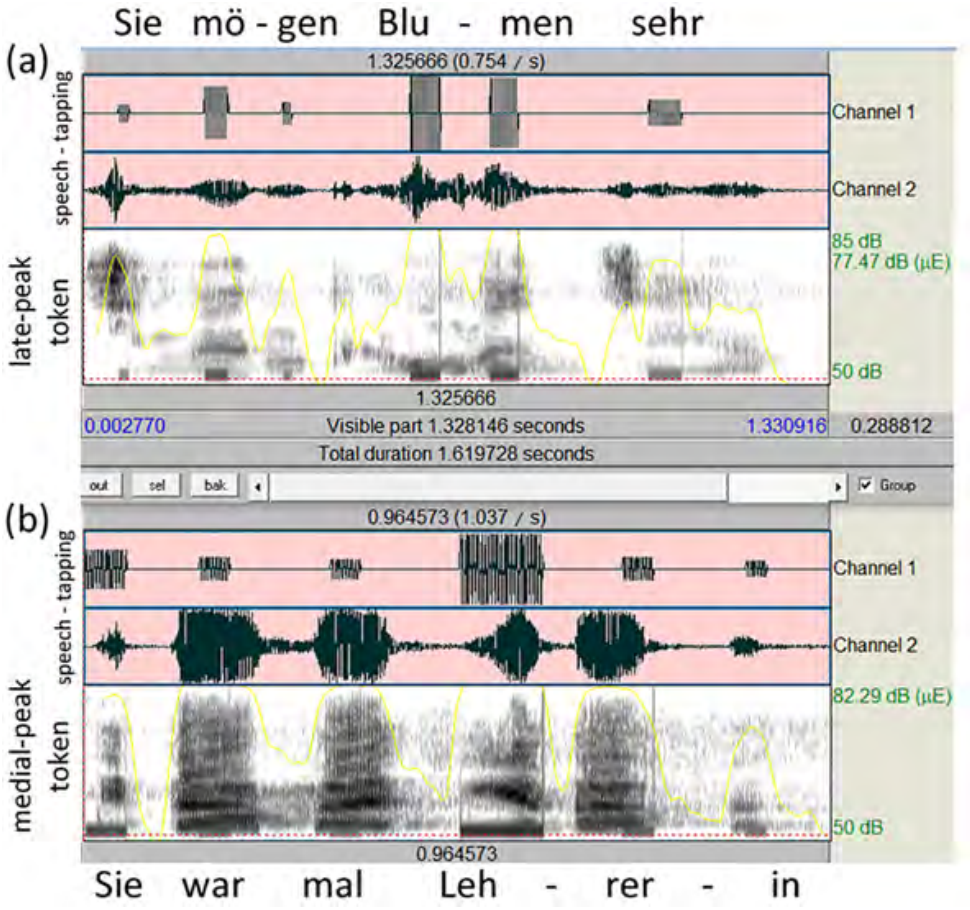

Figure 4. Recorded stereo files integrating the finger-tapping signal and the speech signal; (a) shows an example of a target sentence realized a late peak (Sie mögen Blumen sehr; they like flowers a lot), (b) shows an example of a target sentence realized a medial peak (Sie war mal Lehrerin; she was once a teacher).

The stereo signals of the 20 target sentences per participant were annotated with the Textgrid function in PRAAT (Boersma & Weenink 2018). Marked intervals were, firstly,

- the durations of the pre-accented, accented, and post-accented syllables, segmented on the basis of the acoustic speech signal (through a combined visual inspection of waveform and spectrogram representations);
- and the durations of the button presses on the drum pad, segmented on the basis of the acoustic sinusoid signal (through visual inspection of the waveform only).

The Textgrid files were used to measure (in ms) the durations of the syllables and button presses automatically by means of a PRAAT script. A similar PRAAT script was also used to determine the intensity maxima

of all annotated syllables and button presses (RMS, in dB, length of analysis window 40 ms, mean pressure subtracted). Prior to the intensity measurements, all speech files were intensity normalized (in Adobe Audition) by boosting the largest signal elongation to the maximum of the recording's dynamic range and then upscaling all other signal elongations proportionally. In this way, we removed differences due to speaker-individual loudness levels from the analysis. It was not possible to compensate, in a similar post-processing step, also for possible head or body movements of a speaker during the recording. However, given the constant contact of the speaker's arm and hand to the table and the drum pad, and because of the speaker's constant focus on the target sentences on the screen, each speaker maintained a fairly stable posture during the recording, and head movements were minimal. In relation to the mouth-to-microphone distance of about 70 cm, changes in this distance of a few centimeters represented a negligible and in any case randomly fluctuating variable.

Altogether 1,440 duration and intensity measurements were taken in the acoustic analysis, 720 for the speech data (240 per pitch accent), and 720 for the finger-tapping (i.e. drum pad) data.

3. Results

For statistical analysis of the measurements, we conducted repeated-measures MANOVAs based on the two within-subjects factors Pitch-Accent Category (3 levels: early, medial, late) and Syllable (3 levels: pre-accented, accented, post-accented). The factor Speaker was included as a covariate. One MANOVA was conducted for each type of analyzed data, i.e. the speech data and the finger-tapping data. The two dependent variables were in both MANOVAs the measured duration and intensity values. Each MANOVA was supplemented by a pair of univariate repeated-measures ANOVAs. They were based on the same two within-subject factors as the MANOVA, but looked separately at the duration and intensity parameters. Moreover, multiple post-hoc comparisons (t-test series with Bonferroni corrections of significance levels) were carried out between the levels of the two within-subject factors in each ANOVA.

Results for the speech data are depicted in Figures 5(a)-(b). The figures show along the vertical axis the mean duration and intensity values, with which the speakers have realized the triplet of pre-accented (blue), accented (red), and post-accented syllable (green) in combination with each pitch-accent category. For example, for the early-peak category in

Figure 5(a), we can see that the post-accented syllable was on average 223.4 ms long (green). The pre-accented syllable was 246.4 ms long (blue) and thus slightly longer. The accented syllable (red) was the longest of the three with an average duration of 288.4 ms. Arrows in between the vertically displayed green, blue, and red mean values indicate significant differences between mean values ($p<0.05$) , as determined in the multiple post-hoc t-test comparisons. Along the horizontal axis, it is shown how the mean values for each of the three syllables (pre-accented, accented, and post-accented syllable) changed over the pitch-accent categories of the early, medial, and late peak. For example, Figure 5(b) shows that the mean intensity of the pre-accented syllable (blue) decreases from the early peak (80.6 dB) through the medial peak (74.1 dB) to the late peak (68.8 dB). Analogous to the arrows along the vertical axis, continuous lines along the horizontal axis indicate significant differences between the PA categories ($p<0.05$). Dashed lines indicate that a difference between early and medial or medial and late peak is not significant.

The MANOVA on the speech data yielded significant main effects of Pitch-Accent Category ($F[4,630]=77.5$, $p<0.001$) and Syllable ($F[4,630]=63.3$, $p<0.001$), as well as a significant interaction of the two within-subject factors ($F[8,1262]=114.6$, $p<0.001$). According to the separate univariate ANOVAs, the two dependent variables Duration (Pitch-Accent Category: $F[2,316]=28.9.$, $p<0.001$; Syllable: $F[2, 316]=37.0$, $p<0.001$; Pitch-Accent Category * Syllable: $F[4,632]=59.4$, $p<0.001$) and Intensity (Pitch-Accent Category: $F[2,316]=19.7.$, $p<0.001$; Syllable: $F[2,316]=38.4$, $p<0.001$; Pitch-Accent Category * Syllable: $F[4,632]=45.4$, $p<0.001$) made comparably strong contributions to the MANOVA's overall main effects and their interaction. The additionally conducted multiple post-hoc comparisons yielded significant differences between all factor levels, except for the duration and intensity levels of the medial peak. For this pitch-accent category, there were no differences between the pre- and post-accented syllables on both sides of the accented one.

It can clearly be seen in the Figures 5(a)-(b) that the measured duration and intensity levels were consistently higher in the accented syllable, irrespective of pitch-accent category. More specifically, accented syllables were on average about 40-50 ms longer (284-292 ms) and had 6-8 dB higher intensity maxima (86-87 dB) than the surrounding pre- and post-accented syllables. In contrast, pre- and post-accented syllables differed strongly in their duration and intensity characteristics depending on the pitch-accent category with which they co-occurred. For early-peak

productions, pre-accented syllables were about 20 ms longer and 12 dB higher in intensity than their post-accented counterparts. An inversely asymmetrical duration and intensity pattern emerged for the late-peak productions. Here, it was the post-accented syllable that exceeded the mean duration and intensity levels of the pre-accented syllable; on average about 50 ms in duration and 8 dB in intensity.

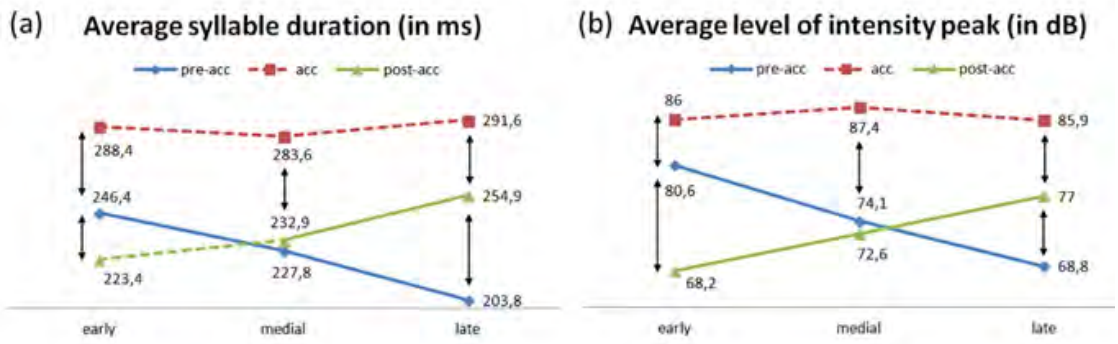

Figure 5. Average syllable durations (a) and average intensity maxima (b) across all 4 speakers, displayed separately for triplet of pre-accented, accented, and post-accented syllables produced in combination with early, medial, and late pitch accents. Continuous lines and vertical arrows show significant differences ($p<0.05$) between Pitch-Accent Category or Syllable conditions. Each data point represents 80 tokens.

The results summary of the finger-tapping data is provided in Figures 6(a)-(b). Like in Figure 5(a)-(b) above, the vertical axes in Figures 6(a)-(b) show mean value differences across the triplet of pre-accented, accented, and post-accented syllable (significant ones being marked by vertical arrows). The horizontal axes show how mean values vary across the triplet of early, medial, and late peak (with continuous lines indicating significant and dashed lines indicating not-significant differences between pitch-accent categories).

The overall results pattern closely resembles that of the speech data. The MANOVA yielded significant main effects of Pitch-Accent Category ($F[4,630]=52.9$, $p<0.001$) and Syllable ($F[4,630]=77.2$, $p<0.001$). Moreover, there was a significant interaction of Pitch-Accent Category and Syllable ($F[8,1262]=95.1$, $p<0.001$). The supplementary ANOVAs showed that the two main effects and their interaction rely to a similar degree on both dependent variables, i.e. duration (Pitch-Accent Category: $F[2,316]=33.5.$, $p<0.001$; Syllable: $F[2,316]=26.4$, $p<0.001$; Pitch-Accent Category * Syllable: $F[4,632]=82.7$, $p<0.001$) and intensity (Pitch-Accent Category: $F[2,316]=40.4$, $p<0.001$;

Syllable: $F[2,316]=38.6$, $p<0.001$; Pitch-Accent Category * Syllable: $F[4,632]=66.2$, $p<0.001$). Like for the speech data, the multiple post-hoc comparisons for the finger-tapping data yielded significant differences between all factor levels, except for the comparison of the medial peak's pre- and post-accented syllables. Their duration and intensity levels did not differ from each other.

Figure 6(a) shows that the finger-tapping durations are overall 30-60 ms shorter than the actual syllable durations (we assume that this is because our 4 speakers coordinated the entire downward and upward movement of their index finger with the speech syllables, not just the time during which the index finger pressed the button of the drum pad). Nevertheless, significant relative duration differences among the produced syllables emerged also in the finger-tapping data. That is, speakers pressed the button on the drum pad about 30 ms longer for the pre-accented or post-accented syllable, depending on whether they realized an early or a late pitch-accent peak, respectively. The longest button presses were consistently measured on the accented syllable, irrespective of pitch-accent type.

Likewise, Figure 6(b) shows that the power with which speaker pressed the button on the drum pad (i.e. the intensity of the sinusoid generated by the drum pad) was for all three pitch-accent categories significantly stronger on the accented syllable than on the two surrounding syllables. Besides this comparability of the three accents, the drum-pad button was pressed with greater power by the speakers (i.e. the drum pad generated a higher intensity level) on the pre-accented syllables in early-peak productions and on the post-accented syllables in late-peak productions.

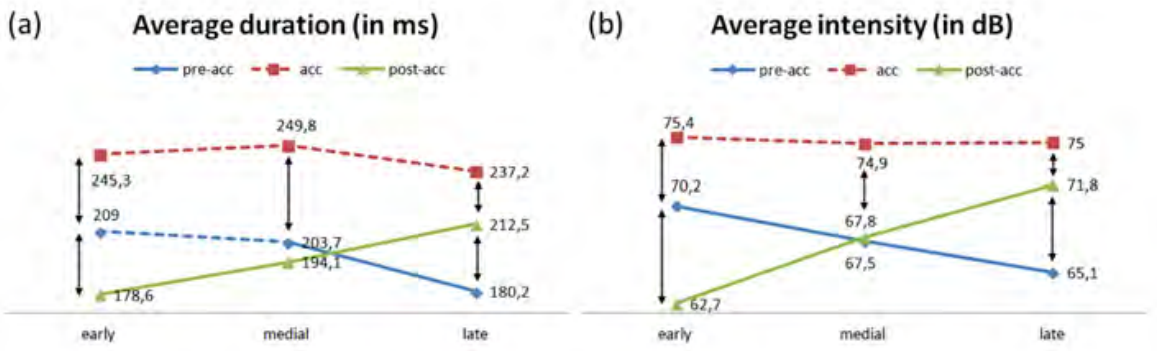

Figure 6. Average duration (a) and power (intensity) of button presses across all 4 speakers, displayed separately for triplet of pre-accented, accented, and post-accented syllables produced in combination with early, medial, and late pitch accents. Continuous lines and vertical arrows show significant differences ($p<0.05$) between Pitch-Accent Category or Syllable conditions. Each data point represents 80 tokens.

Finally, note that the effect of the covariate Speaker came out highly significant in both MANOVAs (Acoustics: $F[1,158]=120.8$, $p<0.001$; Finger tapping: $F[1,158]=98.6$, $p<0.001$). That is, there were strong speaker-specific differences in how the target syllables were tapped and acoustically realized. Many of these differences were gender-related. For example, both syllable and finger-tapping durations were longer for the female than for the male speakers. Intensity levels were on average also higher for female speakers. In contrast, power levels of button presses were higher for the male than for the female speakers. The longer duration values measured for female speakers match with the longer word and sentence durations that were found for female speakers in other studies (across languages) and that are associated with females having a slower speaking rate than males (everything else being equal), see Van Borsel & De Maesschalck (2008) as well as Weirich & Simpson (2014) for a critical discussion of gender-specific speaking rates. That female speakers produced higher intensity levels is consistent with previous studies on different languages as well, see Klatt & Klatt (1990) and Hwa Chen (2007). Also the higher finger-tapping power of male speakers replicates findings in previous studies (Aoki et al., 2005).

In addition, there seemed to be some speaker-specific trade-offs in the extent to which duration and intensity/power differences are realized between pitch accents. One female speaker seemed to focus more on duration than on intensity when creating pitch-accent-specific differences in the triplet of pre-accented, accented, and post-accented syllable, whereas one male speaker seemed to prefer intensity over duration. However, based on only 4 speakers, we refrain from making any general statements about possible trade-offs between non-F0 parameters in pitch-accent production. It is interesting to keep in mind the possibility of such trade-offs for future studies, though.

4. Discussion

Niebuhr & Pfitzinger (2010) found in an acoustic analysis of nuclear German pitch accents that the three accent categories of early, medial, and late peak (nowadays established phonological categories across models of German intonation, Grice et al., 2005; Mayer, 1995; Kohler, 1991a) involve systematic changes in the duration and intensity levels of their coinciding pre-accented, accented, and post-accented syllables. With reference to the relevance of duration and intensity for perceived syllable prominence in German (see, for example, Kohler, 2008), Niebuhr & Pfitzinger called

these effects pitch-accent-specific micro-rhythms. The attribute "micro-" takes into account the fact that the actual macro-rhythm (i.e. what is typically meant by the term speech rhythm, cf. Kohler, 2009; Cumming, 2010) is, firstly, a matter of larger prosodic domains like the intonation phrase and, secondly, a matter that relies on the relatively strong perceptual prominences of accented and/or stressed syllables, not on relatively weak perceptual prominence differences between unstressed and/or unaccented syllables.

So far, Niebuhr & Pfitzinger's idea of a pitch-accent-specific micro-rhythm was only supported by the fact that duration and intensity are prominence-related factors. There was no direct empirical evidence that the triplet of pre-accented, accented, and post-accented syllable actually forms a rhythmic pattern. Such evidence could, for example, have come from a perception experiment in which listeners judge the prominence levels of individual syllables. Previous studies showed that such judgments are possible to make for listeners with the necessary fine grading and for sequences of consecutive syllables (Jensen & Tøndering, 2005; Arnold et al., 2011). Nevertheless, the present study took an alternative approach, which was assumed to be still easier to implement and still more direct in reflecting speech rhythm: syllable-synchronous finger tapping. While speaking, participants pressed a button in a drum pad, once for each syllable they produced. These motor reflexes of speech rhythm were then analyzed in terms of the duration and power of the individual button presses (on the relevant syllable triplet) and additionally set in relation to the acoustic duration and intensity values of the coinciding syllables.

The acoustic analysis of 240 target sentences (80 tokens per pitch-accent category) replicated the findings of Niebuhr & Pfitzinger (2010) and, thus, was in accord with the hypotheses that were put forward on this basis in the present study.

- Duration and intensity levels are higher for the accented syllable than for the two framing unaccented syllables.
- For the early peak, the duration and intensity levels of the pre-accented syllable are higher than those of the post-accented syllable.
- For the late peak, the duration and intensity levels of the pre-accent syllable are lower than those of the post-accented syllable.
- For the medial peak, the duration and intensity levels of the pre- and post-accented syllables are equally low relative to those of the accented syllable.

Furthermore, and crucial for the present study, the pitch-accent-specific micro-rhythms derived from the acoustic prominence factors clearly also manifested themselves in the speaker's finger-tapping behavior. Thus, the corresponding hypotheses are supported.

- Irrespective of the pitch accent, the finger tapping for the accented syllable is always stronger and longer than for the two surrounding syllables.
- An early peak leads to an asymmetrical finger-tapping pattern with an overall declining strength and duration. That is, the finger tapping is stronger and longer for the pre-accented syllable than for the post-accented syllable.
- A late peak results in an asymmetrical finger-tapping pattern with an overall increasing strength and duration. That is, the finger tapping is weaker and shorter for the pre-accented syllable than for the post-accented syllable.
- A medial peak leads to a symmetrical finger tapping pattern. That is, the finger tapping is similarly weak and short for both the pre-accented and the post-accented syllable, and only strongly pronounced in terms of power and duration for the intervening accented syllable.

Expressed in prominence patterns, the early peak seems to be characterized by a slight increase in prominence towards the accented syllable, followed by a strong drop in prominence after the accented syllable. In contrast, the late peak is associated with a strong increase in prominence towards the accented syllable and only a slight prominence decrease after the accented syllable. In other words, for early peaks, two approximately equally strong prominent syllables follow a weakly prominent syllable, and for late peaks, a weakly prominent syllable is followed to two approximately equally prominent syllables. The medial peak is characterized by a strong prominence contrast between the pre- and post-accented syllables and the accented syllable in the center of the triplet that clearly stands out against its two neighbors.

Initial experimental data from Niebuhr & Pfitzinger (2010) and Niebuhr (2011) suggest that these pitch-accent-specific micro-rhythms are perceptually relevant. This applies both to the identification of the pitch accents and to the perception of their communicative meanings. This perceptual relevance is not sufficiently represented in current intonation models and phonologies as they are all focused on F0 alone.

Note, however, that there is an interesting parallel between the pitch-accent-specific micro-rhythms determined here and the representations of the early, middle, and late peaks in the major autosegmental-metrical (AM) model of German intonation, GToBI (Grice et al., 2005). The early peak is conceptualized in GToBI as H+L* (or H+!H*), the medial peak as H*, and the late peak as L*+H. That is, the position of the leading or trailing tone relative to the starred tone is the same as the position of the more prominent pre- or post-accented syllable relative to the accented syllble in the pitch-accent-specific micro-rhythms. H* does not have a training or leading tone in GToBI and neither did we find a prominent pre- or post-accented syllable for this pitch accent category. However, in GToBI (as in the original AM framework of Pierrehumbert, 1980), the leading and trailing tones are not separately associated with (pre- or post-accented) syllables, and they also need not coincide with particular syllables or syllable boundaries. Thus, in order to explain and represent pitch-accent-specific micro-rhythms by means of trailing or leading tones in the AM framework, auxiliary phonological concepts such as the secondary-association concept would be required (Prieto et al., 2005); and even on this basis the complex interaction of F0, duration, and intensity in the signaling of pitch accents can probably not be adequately and fully covered. For example, the F0 peaks themselves can show also considerable variation in peak shape and alignment (Niebuhr, 2007a,c), and trailing or leading tones cannot represent both F0 shape characteristics and pitch-accent-specific rhythm characteristics at the same time. In addition, there are the indications in the present data for speaker-specific trade-offs in the extent to which duration and intensity/power differences are realized between pitch accents. Except for the fact that tonal targets like leading and trailing tones are only two-dimensional descriptor units, which are unable to represent continuous prosodic variation beyond the F0 alignment and scaling dimensions (syllable association is a third but categorical or binary variable), modeling duration and intensity interactions by means of F0-related units seems in general to be at best a preliminary solution; provided that these interactions (trade-offs) are supported and further substantiated by follow-up studies with a larger speaker sample.

Overall, empirical evidence suggests that F0 on the one hand and syllable duration and intensity (i.e. patterns of prominence or rhythm) on the other are connected but conceptually independent signaling systems of pitch accents. In combination, these signaling systems form what

Ward & Gallardo (2017) call a "prosodic construction", i.e. a coherent multiparametric configuration of prosodic features (see the corresponding special session at the International Congress of Phonetic Sciences, ICPhS, Melbourne, 2019: http://www.cs.utep.edu/nigel/pconstructions/icphs-configs.html). The system of syllable duration and intensity does not seem to be an epiphenomenon of a F0 system controlled by tonal targets and their primary or secondary association.

An alternative framework may be more suitable for explaining and modeling the present findings: the perception-based Contrast Theory of Niebuhr (2007b, see also Niebuhr, 2013). The Contrast Theory is based on similar ideas and concepts as the Tonal Center of Gravity (TCoG) theory of Barnes et al. (2012). It too showcases the complex interplay of seemingly disparate aspects of the acoustic signal in the domain of perception, but its focus is more strongly on perceived prominence. The Contrast Theory's basic assumption is that all different realization strategies that are observed for pitch-accent categories at the level of acoustics boil down to making some sections of F0 peaks or movements stand out more prominently in perception than others. For differentiating between early and medial peaks, for example, the final low F0 section and the middle high F0 section of the rising-falling F0 peaks must achieve maximum prominence respectively. The typical alignment differences between early and medial peaks (see Figures 1-2), according to the Contrast Theory, are so widespread across speakers and languages because they represent the simplest way to achieve this prominence goal, namely by moving the corresponding section of the F0 peak into an area in which its prominence is inherently enhanced by a high acoustic energy level: the accented vowel.

In the Contrast Theory, the duration and intensity differences between the pre- and post-accented syllables would only be an additional strategy to make especially the early and late peak categories phonetically and phonologically more dissimilar. Unlike the medial peak, both the early and the late peak are prosodically constructed around a low-pitched prominence. Therefore, both pitch accents additionally have a secondary high-pitched prominence. While in the early peak pattern this secondary high-pitched prominence precedes the major low-pitched prominence, it follows the major low-pitched prominence in the late peak pattern. The similarity of this concept to the GToBI representations H+L* and L*+H is obvious, but the essential difference between the GToBI representation and the conceptualization of the pitch accents in the Contrast Theory

is that the latter theory views pitch accents as multiparametric prosodic configurations ("constructions") that are inseparably constituted of a pitch Gestalt and a prominence Gestalt (Niebuhr, 2007c, 2013). In addition, the Contrast Theory sees the phonologically distinctive elements of all three pitch accents not in the pitch Gestalt but in the prominence Gestalt.

The Contrast Theory also explains why, in speech production, early, medial, and late peaks show specific F0-peak shape and range properties that are also relevant in pitch-accent perception. For example, characteristic of the early peak is a shallower rise towards the F0 peak maximum (Niebuhr, 2007a). In the case of the late peak, it is an expanded F0 peak range that is characteristic of this category (Niebuhr & Ambrazaitis, 2006; Niebuhr, 2007c). Both strategies are also suitable to further enhance the secondary high-pitched prominence before or after the major low-pitched prominence on the accent syllable. For the medial peak, it is characteristic and perceptually advantageous when both the rising and the falling F0 slope are steep. This can be understood as the avoidance of any secondary high-pitched prominences on the surrounding syllables.

The Contrast Theory, however, is not yet a fully developed intonation model. Nevertheless, it shows, in a similar way as the TCoG theory of Barnes et al. (2012), a possible way of reducing and understanding the complex acoustic signaling of pitch accents in terms of a manageable number of perceptual variables, also with respect to a trade-off between acoustic parameters, for which some indications were found in the present study. Additional trade-offs of a different kind are included in the Gestalt-based Functional Contour superposition model (SFC) of Bailly & Holm (2005) and its further developed variant, the Variational Prosody Model (VPM) of Gerazov et al. (2018). These AI-driven models take into account the possibility that each prosodic configuration at each point in time reflects not a single communicative function, but a number of simultaneously coded functions. The SFC and VPM models use a separate set of (hyper)parameters that represent how prominently each communicative function is present in the speech coded at each point in time, i.e. how strongly the corresponding signal features are pronounced by the speaker. In defining these signal features and the meaningful Gestalt-like signal configurations that they form, the SFC and VPM models go beyond the Contrast Theory and the TCoG theory in that they include also visual features, i.e. a speaker's mimic, head, and body movements, in the overall signal configuration (which is therefore not a

mere prosodic configuration but a multi-modal signal configuration). The rich and sophisticated models like SFC and VPM already are, the less insightful they are when it comes to understanding and explaining the actual links between speech production and speech perception, i.e. why certain prosodic and visual signals are used and how the listener's ear determines their configurational combination and inter-individual as well as cross-linguistic interactions. To that end, computer-based models like SFC and VPM will have to be combined with psycho-phonetic concepts as they are represented in the Contrast Theory and the TCoG theory.

While this is still a future task, the success of SFC and VPM in modeling empirical data beyond the auditory modality – and beyond individual syllables and even rhythmic feet – further stresses the fact that pitch accents are no simple F0 patterns, and certainly no individual local target points. The present study was only a small contribution to emphasizing the actual nature of pitch accents as coherent configurations of multiple prosodic features. Many more studies have to follow, especially those with a comparative cross-linguistic perspective, as the early, medial, and late peaks are melodic elements that also occur in many other languages (but often with different communicative functions). This includes internationally used and taught languages such as English (Kleber, 2006) and Scandinavian languages such as Swedish (Ambrazaitis et al., 2012) and Icelandic (Dehé, 2010).

In addition, follow-up studies should investigate, in a cross-linguistic perspective, to what extent the micro-rhythms outlined here are actually really more "micro" than "macro" in terms of perceptual prominences, given that the pitch-accent-specific duration and intensity differences measured between pre- and post-accented syllables are not much smaller as those measured between each of these syllables and the accented one. As was mentioned above, Jensen & Tøndering (2005) and Arnold et al. (2011) have tested and shown that listeners can apply 31-point scales to represent in a sensible way perceived prominence differences between syllables in stimulus sentences. Such scales seem sensitive enough to quantify (i) how big or small the prominence gap is between the accented syllable and its pre- and post-accented syllables (especially pre-accented syllables in early-peak and post-accented syllables in late-peak contexts), (ii) how much the pre- and post-accented syllables vary in their perceived prominence levels depending on the pitch-accent category, and (iii) how much the prominence levels of pre- and post-accented syllables increase for specific-pitch accents relative

to other adjacent unaccented syllables. These experiments are currently ongoing and will be followed by speech-production experiments with a larger speaker sample before we turn to the questions of prosodic modeling outlined above.

5. Acknowledgments

I would like to thank Allard Jongman and the other reviewers of my paper for their constructive and insightful comments. I am also greatly indebted to Lene Boysen for her excellent work and help in collecting and analyzing the data of the present study (as part of her BA thesis at Kiel University, 2015). Furthermore, my thanks are due to Nigel Ward, Gérard Bailly, Yi Xu, Benno Peters, and Ernst Dombrowski as well as to all participants of the Symposium on Speech and Language on the occasion of Ocke-Schwen Bohn's 65th birthday at Aarhus University (May 2018) for the many inspiring discussions with me on the nature and investigation of multiparametric prosodic configurations. Last but not least, I am indebted to the editors of the Festschrift for the honor and the opportunity to make a small contribution to their collection of papers and for their motivating and patient correspondence with me.

References

Ambrazaitis, G., J. Frid, & G. Bruce. (2012). Revisiting Southern and Central Swedish intonation from a comparative and functional perspective. In: O. Niebuhr (ed.), *Understanding Prosody – The role of context, function, and communication* (pp. 138-158). Berlin/New York: de Gruyter.

Aoki, T., S. Furuya, & H. Kinoshita. (2005). Finger-Tapping Ability in Male and Female Pianists and Nonmusician Controls. *Motor Control 9*, 23-39.

Arnold, D., P. Wagner, & B. Möbius. (2011). Evaluating different rating scales for obtaining judgments of syllable prominence from naive listeners. *Proc. 17th International Congress of Phonetic Sciences, Hong Kong, China,* 252-255.

Bailly, G. & B. Holm. (2005). SFC: a trainable prosodic model. *Speech Communication 46*, 348-364.

Barbosa, P. A. (2015). Temporal parameters discriminate better between read from narrated speech in Brazilian Portuguese. *Proc. 18th International Congress of Phonetic Sciences, Glasgow, UK*, 1053-1057.

Barnes, J., A. Brugos, S. Shattuck-Hufnagel, & N. Veilleux. (2012). On the nature of perceptual differences between accentual peaks and plateaux. In O. Niebuhr

(Ed.), *Understandig prosody – The role of context, function and communication* (pp. 93-118). Berlin/New York: de Gruyter.

Berger, S., C. Marquard, & O. Niebuhr. (2016). INSPECTing read speech : How different typefaces affect speech prosody. *Proc. 8th International Conference of Speech Prosody, Boston, USA,* 513-517

Boersma, P. & D. Weenink. (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.39, retrieved 3 April 2018 from *http://www.praat.org/*

Bruce, G. (1977). *Swedish word accents in sentence perspective*. Lund: Gleerup.

Brücke, E. (1871). *Physiologische Grundlagen der neuhochdeutschen Verskunst*. Vienna: Gerold.

Cumming, R. E. (2010). *Speech rhythm: the language-specific integration of pitch and duration*. PhD thesis, Downing College, UK. https://doi.org/10.17863/CAM.16499.

Dehé, N. (2010). The nature and use of Icelandic prenuclear and nuclear pitch accents: Evidence from F0 alignment and syllable/segment duration. *Nordic Journal of Linguistics 33*, 31-65.

Gartenberg, R. & C. Panzlaff-Reuter. (1991). Production and perception of f0 peak patterns in German. *AIPUK 25*, 29-115.

Gerazov, B., G. Bailly, & Y. Xu. (2018). A Weighted Superposition of Functional Contours model for modelling contextual prominence of elementary prosodic contours. *Proc. 19th International Interspeech Conference, Hyderabad, India*, 1-5.

Grice, M., S. Baumann & R. Benzmüller. (2005). German Intonation in Autosegmental-Metrical Phonology. In: Jun, Sun-Ah (ed.) *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Gussenhoven, C. (1999). Discreteness and Gradience in Intonational Contrasts. *Language and Speech 42*, 283-305.

Hwa Chen, S. (2007). Sex Differences in Frequency and Intensity in Reading and Voice Range Profiles for Taiwanese Adult Speakers. *Folia Phoniatrica et Logopaedica 59*, 1-9.

Jensen, C. & J. Tøndering. (2005). Choosing a Scale for Measuring Perceived Prominence. *Proc. 5th International Interspeech Conference, Lisbon, Portugal*, 2385-2388.

Kleber, F. (2006). Form and function of falling pitch contours in English. *Proc. 3rd International Conference of Speech Prosody, Dresden, Germany,* 61-64.

Kohler, K. J. (1991a). A model of German intonation. *AIPUK 25*, 295-360.

Kohler, K. J. (1991b). Terminal intonation patterns in single-accent utterances in German: phonetics, phonology and semantics. *AIPUK 25*, 117-185.

Kohler, K. J. (1991c). The interaction of fundamental frequency and intensity in the perception of intonation. *Proc. 12th International Congress of Phonetic Sciences, Aix-en-Provence, France,* 186-189.

Kohler, K. J. (2005). Timing and functions of pitch contours. *Phonetica 62*, 88-105.

Kohler, K. J. (2008). The perception of prominence patterns. *Phonetica 65*, 257-269.

Kohler, K. J. (2009). Rhythm in Speech and Language – A New Research Paradigm. *Phonetica, 66*, 29-45.

Kohler, K. J. (2017). *Communicative Functions and Linguistic Forms in Speech Interaction* (Cambridge Studies in Linguistics 156). Cambridge: Cambridge University Press.

Klatt, D. H. & L. C. Klatt. (1990). Analysis, synthesis and perception of voice quality variations among male and female talkers. *Journal of the Acoustical Society of America 87*, 820-856.

Kopiez, R. & A. C. Lehmann (2013). Der Sentograph und seine Anwendung in der musikalischen Ausdrucksforschung – Erkenntnisse aus einer Einzelfallstudie. In: V. Busch, K. Schlemmer, C. Wöllner (eds), *Wahrnehmung – Erkenntnis – Vermittlung. Musikwissenschaftliche Brückenschläge. Festschrift für Wolfgang Auhagen zum sechzigsten Geburtstag* (pp. 121-130). Hildesheim: Olms.

Ladefoged, P. (2003). *Phonetic Data Analysis. An Introduction to Fieldwork and Instrumental Techniques*. Oxford: Blackwell.

Mayer, J. (1995). *Transcription of German Intonation: The Stuttgart System*. Technischer Bericht, Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung.

Meyer, E. (1898). Die neueren Sprachen (Vol. 6).

Mixdorff, H. & H. R. Pfitzinger. (2005). Analysing fundamental frequency contours and local speech rate in map task dialogs. *Speech Communication 46*, 310-325.

Nash, R. & A. Mulac. (1980). The intonation of verifiability. In L. R. Waugh & C. H. van Schooneveld (eds), *The melody of language: Intonation and prosody* (pp. 219-241). Baltimore: University Park Press.

Niebuhr, O. (2006). The role of the accented-vowel onset in the perception of German early and medial peaks. *Proc. 3rd International Conference Speech Prosody, Dresden, Germany,* 109-112.

Niebuhr, O. & G. Ambrazaitis. (2006). Alignment of medial and late peaks in German spontaneous speech. *Proc. 3rd International Conference Speech Prosody, Dresden, Germany,* 161-164.

Niebuhr, O. (2007a). The signalling of German rising-falling intonation categories – The interplay of synchronization, shape, and height. *Phonetica, 64*, 174-193.

Niebuhr, O. (2007b). Categorical perception in intonation: a matter of signal dynamics? *Proc. 7th Internaitonal Interspeech Conference, Antwerp, Belgium,* 109-112.

Niebuhr, O. (2013). The acoustic complexity of intonation. In E-L. Asu, & P. Lippus (eds), *Nordic Prosody XI* (pp. 25-38). Frankfurt: Peter Lang.

Niebuhr, O. & A. Michaud. (2015). Speech data acquisition: the underestimated challenge. *KALIPHO (Kieler Arbeiten in Linguistic und Phonetik) 3*, 1-42.

Parrell, B., L. Goldstein, S. Lee, & D. Byrd. (2014). Spatiotemporal coupling between speech and manual motor actions. *Journal of Phonetics 42*, 1-11.

Peters, B. (1999). Prototypische Intonationsmuster in deutscher Lese- und Spontansprache. *AIPUK 34*, 1-173.

Peters, B., K. J. Kohler, & T. Wesener. (2005). Melodische Satzakzentmuster in prosodischen Phrasen deutscher Spontansprache. *AIPUK 35a*, 7-54.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. PhD Diss, MIT, USA.

Prieto, P., M. D'Imperio, & B. Gili-Fivela. (2005). Pitch accent alignment in Romance: primary and secondary associations with metrical structure. *Language and Speech 48*, 359-396.

Samlowski, B. & P. Wagner. (2015). Promdrum – exploiting the prosody-gesture link for intuitive, fast and finegrained prominence annotation. *Proc. 8th International Conference of Speech Prosody*, Boston, USA, 1-5.

Scripture, E. W. (1902). *The Elements of Experimental Phonetics*. New York: Charles Scribner's Sons.

Wagner, P. (2008). *The rhythm of language and speech: Constraining factors, models, metrics and applications*. Habil. thesis (Habilitationsschrift), University of Bonn, Germany.

Van Borsel, J. & D. De Maesschalck. (2008). Speech rate in males, females, and male-to-female transsexuals. *Clinical Linguistics & Phonetics 22*, 679-685.

Ward, N. G. & P. Gallardo. (2017). Non-Native Differences in Prosodic-Construction Use. *Dialogue and Discourse 8*, 1-30.

Weirich, M. & A. P. Simpson. (2014). Differences in acoustic vowel space and the perception of speech tempo. *Journal of Phonetics 43*, 1-10.