The acoustic complexity of intonation

Niebuhr, Oliver

Published in:
Nordic Prosody XI

Publication date:
2013

Document version
Peer reviewed version

Citation for published version (APA):
The acoustic complexity of intonation

Oliver Niebuhr

Abstract
The acoustic differences that underlie the production and perception of the elements of intonation in German are much more complex than predicted by the simple synchronisation concept of the Kiel Intonation Model, which relies on the timing of local peak maxima or valley minima relative to the boundaries of the accented vowel. While this relative timing is undoubtedly important, signalling intonation elements is also based on contour shape. Additionally, it goes beyond F0 and involves syllable duration and intensity, as well as variation in the quality of speech sounds. This quality variation was called “segmental intonation”, as it is suitable to support the perception of intonational forms (i.e. pitch contours) and their functions. The implications of the acoustic complexity of intonation for phonological modelling are outlined.

1. Introduction

It is about 25 years ago that Klaus Kohler set up a phonological framework for Standard German intonation, the Kiel Intonation Model, KIM (cf. Kohler, 2006 for a summary). The key postulations of the KIM are:
- The phonology of German intonation is embodied in F0 features.
- Holistic contours rather than separate tonal events are the basic elements.
- The holistic contours are attached to accented syllables.
- There are two contour classes, rising-falling peaks and falling-rising valleys.
  - Within each contour class, phonological contrasts result from the synchronisation of F0-peak maxima or F0-valley minima relative to the boundaries of the vowel of the accented syllable.
  - Every peak or valley category can occur on every accented syllable in an utterance. The final accent – the intonational nucleus – has no special status with respect to the number of paradigmatic contrasts.

The KIM distinguishes five intonational categories: early, medial, and late peaks on the one hand, and early and late valleys on the other. The peak categories are synchronised with the accented vowel such that their F0 maxima occur before the vowel onset (early), within and typically towards the end of the vowel (medial), or after the vowel offset (late). Valley minima are synchronised either before (early) or after the accented-vowel onset (late).

Since the inception of the KIM, its approach to intonation is characterised by combining production and perception experiments, using intonational function as a constant point of reference. This means for production studies that recording situations and/or the semantic-pragmatic contexts of monologues or dia-
logues are designed such that they elicit the desired intonation category on a target word without any further metalinguistic instruction (e.g. Niebuhr, 2012). Perception studies in connection with the KIM are typically based on semantic differentials or indirect-identification tests. The semantic differential paradigm is an effective way to screen a stimulus continuum for the number and types of intonation categories it contains. Each intonation category creates a unique imprint on the semantic scales (e.g. Dombrowski, 2003; Ambrazaitis, 2005). Indirect-identification tests ask listeners to judge whether or not target stimuli match with a preceding or following context stimulus (e.g. Kohler, 1991). Wording and intonation of the context stimulus are selected to only match with the target stimuli, if these convey a specific communicative function, represented by a specific intonation category. Indirect-identification tasks are particularly suitable to scrutinise the phonetic encoding of an intonational contrast. In a more communication-oriented variant of the indirect-identification task, listeners are asked to reply to the target stimuli. Conclusions about the intonation categories in the target stimuli can then be drawn from the phonetic or semantic content of the answers or from the time interval that listeners need to reply.

Using the experimental methods outlined above, studies in the past decade have identified a number of shortcomings in the original phonological framework of the KIM and thus suggested considerable refinements. A first type of refinement concerns the phonological structure. The KIM turned out to be underspecified. For instance, a fourth peak category, the late-medial peak, was introduced. Its peak-maximum alignment is similar to that of the late peak, but the rise is less steep and sets in earlier relative to the accented-vowel onset (Kohler, 2005, 2006; Niebuhr & Ambrazaitis, 2006). Furthermore, Ambrazaitis & Niebuhr (2008) corroborated the assumption of Kohler (1991) that the hat-pattern vs. dip-pattern contrast in sequences of two accents has a separate function and is not a matter of contextually determined phonetic variation (cf. also Peters et al., 2005). This fact gave rise to the idea of a third contour class: the concatenation contour. Niebuhr & Zellers (2012) presented evidence in favour of this idea and conclude additionally that there are two functionally different types of late peaks: one with an immediate steep fall and one with a delayed/shallower fall or no fall at all (as at the onset of a hat pattern). As regards valley contours, Dombrowski & Niebuhr (2005, 2010) showed that the difference between convex and concave rises is phonologically contrastive for early valleys and presumably also for late valleys, although not with the same communicative function.

In parallel to adding further intonation categories to the KIM, the second type of refinement challenged and revised the key postulations on which the KIM categories rely. In particular, the production and perception of intonation categories was found to be much more complex than predicted by the KIM,
whose simple synchronisation concept solely draws on the relative timing of a single F0 turning point. Focussing on the intensively studied early, medial, and late peaks, the aim of the present paper is to give the reader an impression of the actual acoustic complexity of intonation categories in Standard German and to outline on this basis, how this complexity may be taken into account in an advanced version of the KIM. Audio examples for sections 2–4 are available at http://www.linguistik.uni-kiel.de/downloads/Pres_Niebuhr_NordPros2012.ppt.

2. Shape of F0 contours

The pivotal role that the synchronisation concept plays for the KIM is not least reflected in the category labels ‘early’, ‘medial’, and ‘late’. As long as two peak contours have the same synchronisation, i.e. as long as their F0 maxima are located at equal distances from the accented-vowel onset, they are considered to be instances of the same intonation category. The shape of the peak contours, for example in terms of the durations of their rising and falling slopes, is supposed to be phonologically irrelevant. Starting from perception experiments, Niebuhr (2003, 2007a, 2007b) presented evidence that undermined this synchronisation concept of the KIM. He created four stimulus continua. In each of them, a rising-falling F0 peak contour was shifted in the same equal-sized steps across the accented vowel of “Ma-“ in the utterance “Sie war mal Malerin” (She was a painter). The resulting stimulus continua were identical, except for the shape of the shifted peak. In two continua the shifted peaks had a symmetrical, blunt or pointed shape. The other two continua contained clearly asymmetrical peak shapes built from the fast or slow rising and falling F0 slopes of the blunt and pointed peaks. The stimuli of all four continua were presented in randomised orders and with multiple repetitions to native speakers of German, who judged them in indirect-identification and AX discrimination tasks.

As for the early vs. medial contrast, the results showed in accordance with Kohler (1991) that a pointed F0 peak with fast rising and falling slopes (f/f) must be shifted with its maximum into the vowel in order to clearly trigger medial-peak identification. The same applied to the asymmetrical F0 peak with the slowly rising, fast falling shape, s/f (Figure 1a–b). However, in order to be functionally and perceptually equivalent medial-peak indicators, it was sufficient for the blunt (s/s) and fast rising, slowly falling F0 peaks (f/s) to be synchronised before the accented-vowel onset (Figure 1c–d). Shifting them into the vowel was not necessary. So, unlike postulated by the KIM, peaks with the same synchronisation can belong to different intonational categories. Moreover, a peak synchronisation after the accented-vowel onset is not an absolute prerequisite and
hence not an invariant phonological feature of the medial-peak intonation category. That shape differences mess up the KIM’s synchronisation concept becomes even clearer in a later experiment, in which Niebuhr (2011a) compared rightward peak shifts with leftward extensions of a high F0 plateau in between rise and fall. He found that a plateau contour, which extended over 100 ms away from the accented vowel, was functionally and perceptually equivalent to a pointed peak, which was shifted in the opposite direction 30 ms into the accented vowel (Figure 1e). In both cases, the percept was that of a medial peak, although the rise offsets were 130 ms apart and on different sides of the accented-vowel onset.

In a following step, Niebuhr et al. (2011) demonstrated in a large production study with context-based elicitations of early and medial peaks that speakers actually make use of the perceptual interplay of synchronisation and shape. Comparing pairs of early and medial peak realisations across 35 speakers revealed a highly significant negative correlation ($r=-0.67$, $df=523$, $p<0.001$) between the shape difference (rise duration divided by fall duration) and the synchronisation difference in a pair. That is, the more ambiguously early and medial peaks were realised along the synchronisation dimension the clearer they were separated in terms of peak shape (Figure 2). About 15% of the speakers created the early vs. medial contrast only by means of shape differences while synchronising both peak categories at about the vowel onset. These speakers were called “shapers”. Speakers with the opposite strategy, i.e. “aligners”, were more frequent, but the majority of speakers used both synchronisation and shape to different degrees for signalling early and medial peaks.

**Figure 1:** Differently shaped and synchronised F0 contours that turned out to be functionally and perceptually equivalent, clear triggers of the medial peak category.
The acoustic complexity of intonation

As regards the medial vs. late contrast, the results of the perception experiments of Niebuhr (2007b) also yielded a clear effect of shape in addition to that of synchronisation. The shifted F0 peak with the pointed shape (f/f) caused a perceptual change from medial to late as soon as the peak maximum occurred after the accented-vowel offset. In contrast, the same peak synchronisation continuum with the blunt shape (s/s) was completely unable to trigger the medial-late transition, irrespective of how far the peak was shifted beyond the accented vowel. The effects of the two asymmetrically shaped peaks f/s and s/f were similar to those of f/f and s/s, respectively. However, the findings for f/s and s/f must no longer be interpreted in terms of the medial vs. late contrast. Due to recent phonological refinements of the KIM, the shape difference between f/s and s/f is (for synchronisations around the vowel offset) now associated with separate intonational categories of the ‘late’ type (cf. Introduction).

As regards further parameters of contour shape, there are Anglo-Frisian languages – most prominently Standard British English – in which the extension of peak maxima (ranging from pointed peaks to peaks with a more or less long plateau in between rise and fall) is communicatively relevant (Knight, 2008). Another shape parameter with communicative relevance is the F0 curvature along a convex-concave continuum. In contrast to the peak-plateau parameter, it seems that the convex-concave parameter is less susceptible to contextual variation and more widely exploited across languages, including German (Dombrowski & Niebuhr, 2010), Estonian (Asu, 2006) and Neapolitan Italian (Cangemi, 2009).

3. Duration and intensity patterns

Unlike postulated by the KIM, the coding of intonational categories goes beyond F0 and includes at least two further prosodic dimensions: duration and intensity. Based on Niebuhr (2007a), Niebuhr & Pfitzinger (2010) used read-speech dialogues to elicit the short utterance “Eine Malerin” (a painter) with early, medial and late peak categories on “Ma-”. They found that the two syllables around the
accented one – i.e. “-ne” and “-le-” – differed in their durations and intensity levels, in this way creating the impression of intonation-category specific micro-rhythms across the triplet of pre-accented, accented, and post-accented syllable. The pre-accented syllable of the early peak was fairly long and had a high intensity similar to that of “Ma-”, whereas the post-accented syllable was very short and of low intensity (Figure 3a). The opposite was true in combination with late peaks (Figure 3c). Here, it was the post-accented syllable whose duration and intensity levels were of a similar order of magnitude as in the accented syllable, whereas the pre-accented syllable had a clearly shorter duration and a much lower intensity. For the medial peak (Figure 3b), both the pre-accented and the post-accented syllables were similarly reduced in terms of duration and intensity relative to the accented syllable.

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 the original productions of “Eine Malerin” and (2) PSOLA resyntheses with interchanged F0 contours. The interchanged contours had the same proportional synchronisations relative to the vowel boundaries. Naturally produced shape differences 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 type (1) were judged in a separate block after those of type (2). The results show firstly that 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 patterns. They biased judgments towards the semantic profiles of the peak categories with which they naturally co-occurred. Thus, the findings suggest that listeners are sensitive to the duration and intensity patterns (i.e. to the specific micro-rhythms) that underlie the peak contours, and that duration and intensity make a separate contribution to identifying or conveying the communicative functions of intonation categories.

Figure 3: Waveforms, spectrograms and F0 contours of “Eine Malerin”, produced with (a) early, (b) medial, and (c) late peaks on the accented syllable “Ma-“.
The signalling of peak categories is also affected by the intensity course and its durational characteristics *within* the accented syllable. Niebuhr (2006) performed perception experiments based on two stimulus series. The first series resulted from shifting a pointed rising-falling F0 peak (f/f) in 11 steps from an early to a medial position into the accented vowel of “Ma-” in “Sie war mal Malerin”. The second series was derived from the first series such that each stimulus showed 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 in separate experiments. Both experiments presented the stimuli ten times in an overall randomised order. 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 contextualisation in the indirect-identification text, A and B also provided a constant context frame against which the ‘hum’ stimuli (X) were judged to sound either like A or like B.

![Figure 4](image_url)

*Figure 4:* 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.
The first stimulus series yielded an abrupt change from the early to the medial peak category. The second series (‘hum’) was able to replicate this perceptual change from early-peak intonation to medial-peak intonation solely on the basis of the F0 and intensity patterns of the first series (Figure 4a–b). Moreover, the change from early to medial peak intonation co-occurred with the intensity increase from the low level of the consonant to the high level of the vowel of the accented syllable (cf. dotted lines in Figure 4). Motivated by this fact, Niebuhr (2007c) repeated the experiments 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 Adobe Audition (cf. grey lines in Figure 4). The results for this additional pair of perception experiments showed that the less dynamic intensity increase was clearly paralleled by a less dynamic perceptual transition from early to medial peak intonation across both the original speech and the ‘hum’ stimulus series.

Niebuhr (2007c) also applied the same experimental procedure to peak-shift continua 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 stimuli. In addition, the perceptual change from medial to late was more gradual for a more gradual intensity decrease after the accented-vowel offset.

The findings of Niebuhr (2006, 2007c) provide a simple explanation for the well-known fact that different consonant and vowel qualities in the accented syllable affect the produced synchronisation of intonation categories. For example, the F0 maximum of an intonation category like the medial peak occurs later in open or low-vowel syllables than in closed or high-vowel syllables; and it occurs earlier for approximants as opposed to other voiced consonants in the syllable onset (e.g. Kohler, 1991; Gartenberg & Panzlaff-Reuter, 1991). Parallel synchronisation differences were also found for other intonation categories and/or in other languages than German. These parallels made Ladd (1999: 1544) assume “that similar principles govern the alignment of F0 movements with the segmental string in all languages”. A major governing principle could simply be the intensity course across the syllable: depending on the syllable structure and the intrinsic sonorities of the combined consonants and vowels (cf. Möbius, 2003), the intensity changes into and out of the accented syllable are more or less large and dynamic. For example, approximants in the syllable onset and/or high-vowels in the syllable nucleus result in a smaller and less steep intensity increase across the CV sequence. This relatively flat and high intensity in the CV sequence allows an intonation category like the medial peak to be synchronized earlier in the syllable without changing its perceptual identification from medial to early (cf. Figure 4, black vs. grey lines). In a nutshell, it is assumed here that many (if not all) cross-linguistically similar peak synchronisation dif-
The acoustic complexity of intonation

4. Segmental intonation

Like many other intonation models, including the autosegmental-metrical model (Pierrehumbert, 1980), the KIM is focussed on F0. Speech sounds serve as docking points for intonational categories on an allegedly separate layer of the speech signal. But, apart from this contribution, speech sounds are considered to be troublemakers whose micro-prosodic perturbations must be filtered out by listeners as well as by researchers, in order to identify the phonologically relevant (macro-prosodic) intonation patterns. This perspective severely underestimates the role of speech sounds in intonation, as it ignores that speech sounds themselves can also induce pitch impressions. For example, the articulatory movements of vowels directly raise or lower F0 (cf. intrinsic F0, Fowler & Brown, 1997), and voiceless fricatives (including aspiration noises) trigger aperiodic pitch impressions that vary with the amount of noise energy and its distribution across the spectrum (cf. Traunmüller, 1987).

Figure 5: Examples of segmental intonation: cochleagrams of two vocoids (a) and spectrograms of two /ʃ/ (b), produced at the end of the utterance-final words “lecker” (a, delicious) and “Fisch” (b, fish) after falling and rising F0 movements.

A series of production experiments by Niebuhr (2009), Niebuhr et al. (2011) and Niebuhr (2012) revealed that the sound qualities of vowels and fricatives in German differ systematically between high and low F0 contexts. Formant measurements suggest in accord with auditory analyses that the vowel qualities of [ə], and [ɐ] (the vocalised <-er> suffix in Standard German) are closer and further front after high-rising than after terminal-falling utterance-final F0 movements.
For example, F1 and F2 of [ɐ] are further apart after a final rise so that [ɐ] changes towards [ɛ̃] (Figure 5). Similarly, when produced in combination with a final F0 rise, the glides of the diphthongs [au] and [ɔʏ] set in earlier and approximate the formant frequencies of isolated German [iː]. Voiceless fricatives like [f], [s], [ʃ], and [x] all have greater intensity and higher spectral centre-of-gravity levels after utterance-final F0 rises than after utterance-final F0 falls (Figure 5).

Differences like these also occur utterance-medially in combination with high F0 peaks and low F0 valleys. The peak vs. valley contrast can even influence the degree of /s/-to-[ʃ] assimilation in utterance-medial /ʃʃ/ sequences. That is, the assimilation of the light sibilant noise of [s] towards the dark sibilant noise of [ʃ] is typically stronger in the context of a low F0 valley than in the context of a high F0 peak (Niebuhr et al., 2011).

It is already implied in the influence of F0 on the degree of light-to-dark sibilant assimilation that the sound-quality differences summarised above are all suitable to support the F0-based signalling of intonation categories. The higher aperiodic pitch impressions of fricatives with more energy at higher frequencies can perceptually complement a high F0 level; and vowels which are produced closer and further front can facilitate reaching a high F0 level due to the contribution of intrinsic F0. Niebuhr (2009) introduced the term “segmental intonation” to refer to such correspondences between the pitch characteristics of a sound segment and its F0 context. A growing body of evidence suggests that listeners in fact use segmental intonations to identify (the functions of) intonational categories and to perceptually compensate for truncated F0 movements (Niebuhr, 2008, 2011b; Kohler, 2011). So far, pitch-driven systematic differences in the pronunciation of speech sounds were considered to occur only for some types of sounds and limited to whispered speech (cf. Whalen & Xu, 1992). However, in fact they seem to be the rule rather than the exception.

5. Summary and conclusion
Differences in F0 shape are either directly or indirectly phonologically distinctive in German. In the indirect case, they interact with other F0 parameters like synchronisation (or scaling, cf. Niebuhr, 2007b) in distinguishing intonational categories. Examples for directly phonologically distinctive shape difference are the /ʃ/ vs. /ʃʃ/ contrast among intonation categories of the ‘late’ type and the convex vs. concave contrast among utterance-final rises (Dombrowski & Niebuhr, 2010). The relevance of shape supports the key postulation of the KIM that the basic elements of intonation are contours (or tonal configurations) rather than
separate tonal events. So far, however, the phonological relevance of contour features in the KIM is restricted to the peak vs. valley distinction. A refined KIM must reflect that contour features also play a role within the peak and valley classes.

Additionally, sections 3–4 emphasise – in opposition to the KIM’s first key postulation – that it is essential for intonation models to go beyond F0 and include other prosodic and even segmental aspects of the speech signal. That intonation cannot be equated with F0 may be considered a turning point in intonational modelling, but there is no way around accepting this challenge.

As for intensity, the findings corroborate the key role of the accented-vowel boundaries for the synchronisation of accent contours. However, unlike postulated by the KIM, it is not the vowel segment itself that is important, but the intensity transitions on both sides of the vowel. This conclusion is actually not that surprising, since contrasts between accent contours can also be conveyed in interjections that only consist of single vowel or consonant sounds like [aː] or [mː]. These vowels and consonants then show different intensity patterns for different accent contours (Niebuhr, 2007b). In addition, the fact that the abruptness of the perceptual change between intonation categories depends on the steepness of the underlying intensity transition clearly argues against the suitability of categorical-perception experiments for separating phonological contrasts from phonetic variation (cf. Niebuhr & Kohler, 2004).

Segmental intonation is a novel idea. However, the observations underlying this idea are actually not new. Jones (1950) already identified systematic, melody-related changes in speech sounds and therefore included voice pitch in those factors that cause allophonic variation of phonemes. However, due to the strict separation of sounds and prosodies the implications of this variation for pitch perception never received much attention. It is still an open question whether segmental intonations are just an epiphenomenon of the phonation process, or whether they are intentionally added by speakers in order to support the coding of intonational forms or functions.

Although there are many open questions like the one above, the hardly manageable complexity of intonation categories in the acoustic domain suggests that the future of intonation models lies in the perceptual domain. The major advantage of such a reorientation would be that a high number of acoustic parameters merge into a small number of perceptual parameters. One promising approach in this direction is the tonal centre-of-gravity concept of Barnes et al. (2010, 2012). It integrates acoustic variation in intensity, F0-peak synchronisation and shape by translating them into timing differences of the perceived tonal targets.

Another elaborated approach is the contrast theory of Niebuhr (2007b). This theory capitalises on the fact that perception follows the same basic principles
across modalities. Therefore, the perception of intonational categories is analysed in analogy to the relatively well understood perception of visual objects. A basic assumption of the contrast theory is that intonational categories consist of two Gestalt like patterns: a pitch Gestalt and a prominence Gestalt (prominence in the sense of perceptual salience, evoked by differences in traditional prominence-cueing parameters). The pitch Gestalt is formed by a combination of tones and movements that result from segmental intonations and spectral constraints in the perceptual implementation of F0 movements (cf. House, 1990). The crucial point is that each tonal element of the pitch Gestalt is contrasted with the preceding element (within the same Gestalt unit) in terms of its pitch, length, and loudness properties. These multi-dimensional backward contrasts – and the concomitant contrast enhancements – create a specific prominence level for each tonal element and hence a specific prominence pattern across all elements.

The contrast theory assumes that the early, medial, and late peak categories of the KIM are all represented by the same pitch Gestalt, i.e. a rise-fall. The phonological differences are in the prominence Gestalt, which is waxing for the early peak, waning for the late peak, and waning-waxing for the medial peak. F0 shape as well as syllable duration and intensity come into play via their effects on the prominence Gestalt. Variation in F0 shape has a twofold effect. Firstly, it changes the temporal distance between tonal elements and hence the degree to which contrast enhancement occurs. Secondly, it directly changes the length of tonal elements and hence their second most important cue to prominence. Synchronisation, the phonological key feature of the KIM, is not considered to be phonologically relevant. Rather, the contrast theory assumes that synchronisation is merely the most effective means to create a prominence pattern for a given pitch pattern by exploiting the durations and intrinsic intensity levels of the segmental string, which is there anyway. However, speakers are free to go for other means, such as F0 shape, in order to create the same perceptual result.

It will be a task of follow-up studies to continue elaborating perception-based intonational phonologies like those outlined above. Apart from addressing the many open questions concerning the acoustic complexity of intonation, this will require interdisciplinary approaches and methodological innovations.

References


Next contribution