Perception and Cognitive Processing of Tonal Alignment in German

Oliver Niebuhr & Klaus J. Kohler

Institute of Phonetics and Digital Speech Processing (IPDS)
Christian-Albrechts-University, Kiel, Germany

Abstract

Results are presented of two parallel sets of discrimination and identification experiments for peak and valley shift in German.

1. Introduction

Kohler [1,2] used a shift paradigm, in which a complete peak contour, defined by three f0 points (left base, peak maximum, right base), was moved successively in equal steps of 30ms over a stretch of vocal tract articulation from the syllable preceding to the one following the accented syllable in a natural production of the German sentence *sie hat ja ge<u>lo</u>gen* ("she has been lying"; male speaker, accented syllable underlined). Listeners had to assess, as 'same/different', pairs of resynthesized stimuli separated along this left-to-right shift series by one or two steps, respectively. The experiment was repeated with similar sentences providing different syllabic contexts. The results were the same in every case: The discrimination function showed a maximum for the pairing in which the f0 peak maximum was shifted across the consonant – vowel boundary, thus changing from a falling to a rising pattern in the transition and thus focussing on a lower vs a higher pitch level in the accented vowel.

In a second set of experiments, the sentence *sie hat ja ge<u>lo</u>gen* was contextualized according to the semantic categories *'knowing'* vs *'observing'* vs *'observing in contrast to one's expectation'*, by providing appropriate introductory phrases. Listeners then had to judge whether the introductory phrase and the stimulus sentence carrying different f0 peak positions matched, thus providing a semantic identification of the test stimuli. Among others, the series of f0 peaks from the leftmost position to the accented vowel centre were put in the context *jetzt versteh ich das erst* ("now I understand"), simulating the semantic constellation of *'observing'*. The results are very clear: there is a change in the identification function across the consonant-vowel transition, the point where the discrimination function shows a maximum. So this seems to be an obvious case of categorical perception in the synchronization of f0 with articulation.

The peak shift paradigm was elaborated by Niebuhr [7,8], who, in a series of discrimination and identification tests with the sentence *sie war mal <u>Ma</u>lerin* ("she was once a painter"), showed that over and above the synchronization of f0 peak contours with articulation, their rise and fall speeds are also perceptually relevant and influence the categorical change-over points in the discrimination and identification functions.

On the basis of these perception data, the Kiel Intonation Model (KIM) [3,4] sets up three phonological peak categories for German: early, medial, late. It also postulates early and late valleys according to the position of the f0 minimum either before or well into the accented vowel. The f0 contour shift paradigm was also applied to valley patterns, in three ways:

(a) In the naturally produced utterance *sie hat ge<u>lo</u>gen* (male speaker) a valley pattern was defined by three f0 points: 105Hz sentence-initially, 85Hz at voice-onset of *ge<u>lo</u>gen*, and 160/180/200/220 Hz at the end. The central low point was shifted in 30ms steps [2].

(b) The basis for the shift series was the same as for (a), but the low f0 point stayed in position throughout; in the first right-shift a second low f0 point of 83Hz was set at 30ms after the original one, and was successively moved in 30ms steps [2].

(c) In the naturally produced utterance *haben Sie die Ro<u>man</u>e ge<u>le</u>sen?* ("have you read the novels?"; female speaker, potential accent syllables underlined) a valley pattern was defined by three f0 points – 210Hz at the boundary between /r/ and /o:/ of *Ro<u>man</u>e*, 165Hz at the boundary between /o:/ and /m/, and 330Hz at the end of the sentence. The first two points were then shifted in 30ms steps until the low point was positioned at the end of /e:/ of *ge<u>le</u>sen*. This series created two sets, one with valley patterns at a single accent *Ro<u>man</u>e*, a second one with valley patterns at a single accent *ge<u>le</u>sen* [6].

In the case of (a) and (b), informal listening to the series produced evidence of a perceptual continuum with clear differences between the end points, but with no abrupt discriminatory changes along the scale, contrary to what was the case in the peak shift series. Therefore, formal discrimination tests were not considered necessary. To explain this finding it was hypothesized that in both procedures the shift of the low valley point from the prenucleus to the nucleus syllable did not produce a rise-to-fall change across the consonant – vowel boundary, comparable to, i.e. as extensive and as fast as, the fall-to-rise change at this point in the peak series. With the f0 onset being at 105Hz, the descent to 85Hz over a progressively longer period produces a very small drop as the valley point enters the vowel, and the general pattern remains rising, with very similar, low-level f0 precursors, before successively later rise points. So these valley shifts create f0 pattern continua that are not divided with reference to distinct f0 trajectories across local acoustic landmarks determined by vocal tract timing. The acoustic continuum is mapped onto a perceptual continuum, and discrimination within stimulus pairs remains similar all the way through the series. But stimuli from different ends of the scale are saliently different and can be used to signal different valley categories in an intonational phonology for semantic coding. The differences of meaning relate to the speaker's feeling and concern for the listener: the *early valley* expresses casualness, routine politeness and lack of interest, whereas the *late valley* brings out the speaker's involvement and concern.

In view of these results, the descent into the valley was made sharper in (c), at both accent positions. Altough the absolute drop was very similar – 3.6st in (a), (b) vs 4.1st in (c) – the rate of change was very different: 11,25 st/s vs 26.62

st/s. But in spite of the greater similarity with the peak series in the dynamics of f0 change across the syllable landmark, the discrimination functions are completely different: the valley series shows no abrupt change. So the perception of differently synchronized valley contours seems to be scalar as against categorical perception in the case of peak synchronizations. The same conclusion was reached by Redi [9] for English within a different experimental paradigm.

2. Further experiments

2.1. The research question

Although quite indicative, the tests carried out with the valley shift paradigm are still not conclusive. The question remains as to whether the different discrimination results for the peak and valley series are due to the different types of pitch contour, or whether the conditions for the generation of the two series were not congruent. If the answer is in favour of the former alternative an explanation must be found in relation to different perceptual and/or cognitive processing. To decide on these alternatives a new set of experiments was devised, which was to evaluate the perceptual discrimination and identification of valley shift against peak shift in the same experimental design. It had to take the following conditions into account:

(1) The sentence for the test stimulus generation is to be identical for both series.

(2) The intial f0 point for the sentence is to be kept the same, and the f0 points defining the peak and valley patterns are to be chosen such that pitch movements from the initial f0 point through the utterance are exact up or down reversals in semitones.

(3) The analysis of *late valleys* in natural corpus data has shown that they have a low f0 precursor in the accented vowel before the rise, in addition to a later synchronization. From this angle, shift paradigm (b) is the most appropriate one of the 3 versions used previously. Its low pecursor is to be implemented in the new series.

(4) The internal timing of the peak and valley contours is to be identical.

(5) The starting position for the peak and valley shifts is to be the alignment of the peak maximum/valley minimum with the consonant – vowel boundary of the accented syllable. The defined peak/valley contour is then to be shifted in equal steps to the left and to the right.

(6) The context precursor for the identification tests is to set the semantic frames of *'opening an argument'* in case of peak stimuli and of *'friendly concern for the listener'* in case of valley stimuli. The occurrence of *medial* or *early peaks* in the former is then expected to yield 'matching'/'not matching' responses, and correspondingly for *late/early valleys*.

2.2. Hypotheses

(I) Pairings of identical stimuli are less frequently judged different than those of non-identical ones, in both series.

(II) Previous peak shift results are replicated: Discrimination shows a significant maximum for the pair of peak stimuli whose f0 peak points span the accented vowel onset.

(III) There is no such maximum in the discrimination of valley stimuli.

(IV) Previous peak shift results are replicated: The identification function shows a category change in the peak series, which is interpretable as a change from the *early* to the *medial peak* in the intonational phonology of German. Together with (II), this constitutes a case of categorical perception.

(V) The identification function shows a category change in the valley series, which is interpretable as a change from the *early* to the *late valley* in the intonational phonology of German. Together with (III), this does not support categorical perception.

2.3. Method

The stimuli for the listening test were prepared as follows:

- *und wie ist dein Name?* ("and what's your name?") was produced with a *medial peak* and with a *late valley* on *Name* by a male speaker (the 2nd author). The former was preceded by the context *aha* ("OK"), also with a *medial peak*, the latter by *aha* with a *late valley*. The communicative particle *aha* with the appropriate intonation provides the required setting for +/-matching judgements of the peak/valley series in the test stimuli.

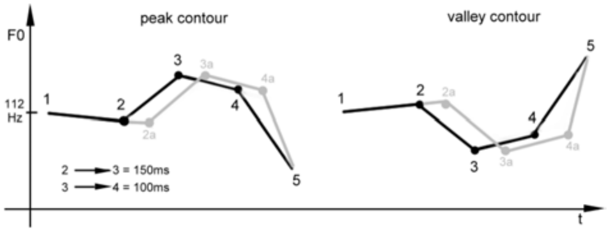

Figure 1: *Schematic illustration of the peak and the valley contour and of the alignment shifts.*

- f0 manipulation and resynthesis were carried out in praat, with the following f0 and duration values (see Figure 1): (1) 112Hz initially, (2) 1.1st down/up (105/120Hz), (3) 6.2st up/down (150/84Hz), (4) 1.2st down/up (140/90), (5) down to 70Hz or up to 220Hz, for peak and valley, respectively. (2) and (3) were separated by 150ms, (3) and (4) by 100ms, (5) was placed at the end of voicing. The exact reversal of down and up movements for (5) was not possible, because an octave rise was not sufficient to create a convincing friendly question intonation.
- As a point of departure, f0 (3) was located at the boundary between /n/ and /a:/ of the accented syllable, and then shifted in 4 equal steps of 25ms to the left as well as in 5 steps to the right. The two original stimuli were then resynthesized creating two alignment continua with 10 peak or 10 valley contours, respectively.

Following the classical paradigm of categorical perception, one discrimination and one identification test were created for the stimuli of each alignment continuum. For the discrimination experiments, each stimulus in each continuum was paired with the stimulus two removed in an ascending order, resulting in 8 unequal pairs (1/3 up to 8/10). In addition, the odd-numbered stimuli were combined to 5 equal pairs (1/1 to 9/9). Each of the total of 13 pairs per continuum was preceded by a signal tone and followed by a silent intervall of 4 sec for subjects to make their judgements. In each pair, the two stimuli were separated by a silent interval of 1.5 sec. In both tests, the 13 pairs were copied 5 times and randomized. Subjects were instructed to concentrate on the speech melody of the stimuli and to judge whether they perceived both stimuli in a pair as ‘same’ or ‘different’.

As regards the two identification tests, each of the ten stimuli was preceded by the constant context phrase *a<u>ha</u>*, either with a medial peak for the peak series or with a late valley for the valley series. A silent interval of 250ms between context phrase and stimulus was suitable to perceive both phrases as parts of one utterance. Every combination of context and stimulus was preceded by a signal tone and followed by a silent intervall of 4 sec for judgement. The 10 combinations in each test were copied 5 times and randomized. Subjects had to judge whether context and stimulus were a semantic match. It is assumed that subjects base these judgements on the semantic function of the coupled contours, indirectly identifying the corresponding intonational categories. The medial peak in *a<u>ha</u>* expresses pleasant surprise about new information received from the hearer. For a match, the features 'pleasant surprise' and 'new' must be continued or followed up in the next phrase *und wie ist dein <u>Na</u>me?* This can only be done by a further medial peak. The early peak, which may be paraphrased as "so let's have your name and be finished with it!", is not compatible with this context. Analogously, the late valley in *a<u>ha</u>* expresses a friendly concern for the hearer, which sounds odd when it is followed by an early valley, expressing casualness. So, a late valley in the stimulus is necessary for judging both phrases as 'matching'.

18 native speakers of German, 6 male and 12 female, average age 23, participated in the experiments. They were grouped into three sessions. In all sessions, the experiments proceeded from the peak to the valley continua, with identification following discrimination tests in each case. Each sub-session was preceded by a short practice run with stimuli from either end and from the centre of the particular series. The experiments were presented over loud-speaker in a sound-attenuated room, and subjects gave their responses by pressing one of two buttons ('same' vs 'different' or 'matching' vs 'not matching') of a computerized system registering reactions to speech stimuli.

3. Results

Figure 2 illustrates the results of the discrimination experiments as the percentage of 'different' judgements for each pair pooled over all 18 subjects (each data point = 90 judgements: 18 sbs. x 5 repetitions). Unequal pairs yielded more 'different' judgements than the equal ones in either series, and the unequal pairings of the peak stimuli more 'different' judgements than the unequal valley pairs, with a clear discrimination maximum (stimuli 5/7), which is absent from the discrimination function of unequal valley pairs, as well as from both equal pair functions. The results of the identification experiments for peak and valley stimuli are presented in Figure 3 (each data point = 90 judgements). There is a clear response change for both contour types in the shift from left to right. While for the stimuli in the left half of the continuum the majority of judgements are 'not matching', the opposite is true for the stimuli in the right half. This change takes place between stimuli 4 and 7, and in the case of the peak series, coincides approximately with the peak stimulus spanned by the discrimination maximum as shown in Figure 2. It can also be seen from Figure 3 that peak and valley functions have similar ranges and slopes. The two curves are nearly congruent, except for the deviation around stimulus 5.

Because the data from the perception experiments do not meet the distributional requirements essential for parametric testing, non-parametric tests were used. In view of hypotheses (I) and (III), the first test was to examine whether there are significant deviations in the frequency of 'different' judgements between the equal and unequal stimulus pairs of both alignment continua. The corresponding test design consisted of 4 conditions (k) of the independent variable, 'peak (un)equal', 'valley (un)equal'. To obtain samples of equal size containing comparable stimuli, the equal pairs were contrasted with those unequal pairs that have the same initial odd-numbered stimulus. Due to the asymmetricality of stimulus combinations in the series, 9/9 had to be linked to 8/10. The selected data were reduced to only one measurement per person and independent variable by summing the 'different' responses. A Wilcoxon-Wilcox multiple comparisons test was then performed. The results are given in Table 1. It shows no significant deviation in the frequencies of 'different' judgements between the equal pairs of peaks and valleys. On the other hand, the unequal pairs containing peak stimuli were judged as 'different' significantly more often ($p<0.05$) than any other combination.

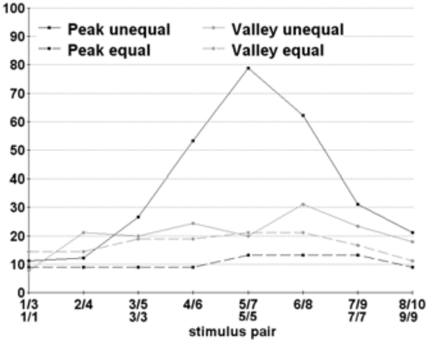

Figure 2: *Discrimination functions of the peak and valley stimuli for (un)equal pairs, 18 sbs, n=90*

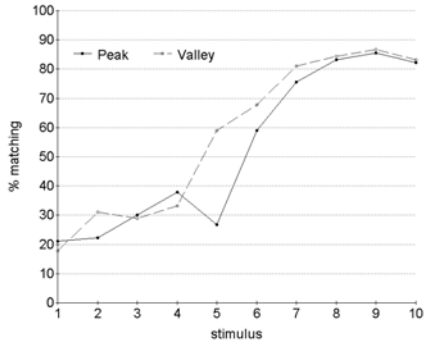

Figure 3: *Identification curves of the peak and valley stimuli, 18 sbs., n=90*

Table 1: *Ranking sums (R) and their calculated differences for the Wilcoxon-Wilcox multiple comparisons test (two-tailed).*

Variable		Valley equal	Valley unequal	Peak unequal
	R	38,5	43	67
Peak equal	31,5	7	11,5	35,5*
Valley equal	38,5		4,5	38,5*
Valley unequal	43			24*

With regard to hypotheses (IV) and (V), a second statistical analysis was performed, testing whether there are significant differences between the two 'matching' functions of the peak and valley stimuli (see Figure 3). The 50 responses per subject (10 stimuli x 5 repetitions) were reduced to only one measurement representing the properties of the whole response profile. This was done by subtracting, for each subject, the frequency of 'matching' judgements for stimulus 1 from the corresponding frequencies for stimuli 2 to 10 (varying between 0 and 5 for each stimulus) and by summing these differences. The resulting value is positive for rising curve slopes (as in Figure 3), negative for falling slopes and 0 for flat ones. Further, the earlier the function rises/falls between stimuli 1 and 10 the larger is the positive/negative value. A Wilcoxon matched pairs signed rank test was performed, using the calculated values as the dependent variable and the two contour types, 'peak' and 'valley' (k), as well as the subjects, as the independent variables. This analysis yielded no significant difference between the identification functions of the peak and valley stimuli (Rp=79, Rn=74; p>0.05 for k=2 and n=18).

4. Discussion and Conclusions

All five hypotheses have been confirmed by the data and their statistical evaluation. Listeners were able to separate pairs of physically identical stimuli from non-identical ones. They did this most clearly in the middle range of the peak continuum, i.e. in the region where the identification function shows a transition between two categories. So, their responses were systematic, not random. This is an exact replication of all previous experiments with the peak shift paradigm, and seems to point to the classical concept of categorical perception. The abrupt perceptual transition takes place between stimuli 5 and 6. This is slightly earlier than was found by Kohler [2]. This difference may be connected with the peak shape used in this study. The internal timing of the rising and falling f0 movements (Fig.1) is similar to the fast rising, slow falling peak shape, which was one of 4 shapes used by Niebuhr [7,8] in his experiments, and for which he found the same earlier location of the perceptual transition in relation to the accented vowel onset. Thus, these findings are also replicated by this study.

Although physical differences are perceived as different more often than physical identity in the valley continuum as well, the perceptual differentiation is very much smaller, stays the same across the continuum, and does not reach significance. On the other hand, the identification functions for the peak and the valley continuum are not significantly different. This means that the classical concept of categorical perception is not applicable to the valley continuum. Since the experimental design ensured an absolute parallelism in the perceptual testing of the two stimulus series, it can now be concluded that the listener processes valley and peak contours differently.

In a first step towards explaining this finding, it is necessary to define what *identification* means in the context of this experimental design. Listeners had to project their perception of pitch patterns onto cognitive categories of semantic interpretation provided by a context – test sentence frame. So the task is not simply perceptual, but also involves an association with meaning. Through this reference to meaning, the investigator can deduce an association with the *early/ medial peak* or *early/late valley* categories of the intonational phonology of German, the categories to be identified in the perception experiment. As the identification functions for both series show, the listeners were successful with this perception – meaning association, and the metalinguistic deduction is justified. In performing such a perceptuo-cognitive task, listeners will refer to prototypical prosodic realisations of the semantic constellations, viz. they select stimuli from outside the central region of the continuum to associate with either 'matching' or 'not matching', and show greater uncertainty as to the classification inside this region. So the identification function represents clear categorizations at either end of the scale and a gradual transition between them.

The *discrimination* of pitch patterns within each series is more closely related to perception. Although the results of Niebuhr [7,8] demontrate that the categorical boundary between early and medial peaks can be shifted to both sides of the accented vowel onset by the holistic influence of the steepness of the rising and falling f0 slopes, the range of possible boundary shifts is limited to an area closely around the accented vowel onset. This may indicate that the f0 turning point from rising to falling and its synchronization with the spectral turning point from consonant to vowel are the primary acoustic cues for perceiving early or medial peaks. This prosodic – segmental link seems to be responsible for pairs whose peak positions span the transition region to be especially salient perceptually. In the case of the valley series, however, a sharp perceptual decision between early and late valleys is not possible for three reasons: (1) The perception of early or late valleys depends on more than one salient cue (synchronization as well as position and duration of a low precursor). (2) Some of these cues are not points in time but periods of time. (3) The decisive differentiator between *early* and *late* is linked to *inside* the vowel, which lacks local landmarks. So discrimination is not supported by a prosodic – segmental link at a certain position in the acoustic continuum, and is consequently similar, and low, all the way through the series: an acoustic continuum is mapped onto a perceptual contiuum. The same applies to continua of medial-to-late peak shift [2], peak height [5], or final rise point. This more differentiated and cognition-orientated treatment of categorical perception challenges the traditional view of phonology as mediator between acoustic continua and categorical perception: scalar perceptual changes do not preclude the existence of underlying phonological categories and vice versa. Thus the categorical perception concept needs revision.

5. References

[1] Kohler, K. J., 1987. Categorical pitch perception. *Proc. 11th ICPhS, Tallinn, vol. 5*, 331-333.

[2] Kohler, K. J., 1991a. Terminal intonation patterns in single-accent utterances of German: Phonetics, phonology, semantics. In K. J. Kohler (ed.), *Studies in German Intonation*. AIPUK 25, 295-368.

[3] Kohler, K. J., 1991b. A model of German intonation. In K. J. Kohler (ed.), *Studies in German Intonation*. AIPUK 25, 295-368.

[4] Kohler, K. J., 1997. Modelling prosody in spontaneous speech. In Y. Sagisaka, N. Campbell, and N. Higuchi (eds.), *Computing Prosody*, 187-210. Springer: Berlin/ Heidelberg/New York/Tokyo.

[5] Ladd, D. R., Morton, R., 1997. The perception of intonational emphasis: continuous or categorical? *JPhon* 25, 313-342.

[6] Landgraf, K., 2003. *Steigende Intonationskonturen im Deutschen. Experimentalphonetische Untersuchungen zur auditiven Kategorisierung.* MA Diss., Kiel.

[7] Niebuhr, O., 2003a. *Perzeptorische Untersuchungen zu Zeitvariablen in Grundfrequenzgipfeln.* MA Diss., Kiel.

[8] Niebuhr, O., 2003b. Perceptual study of timing variables in F0 peaks. *Proc. 15th ICPhS, Barcelona*, 1225-1228.

[9] Redi, L. C.,2003. Categorical effects in production of pitch contours in English. *Proc. 15th ICPhS, Barcelona*, 2921-2924.