The Role of the Accented-Vowel Onset in the Perception of German Early and Medial Peaks

Oliver Niebuhr

IPdS, Christian-Albrechts-Universiät, Kiel, Germany
on@ipds.uni-kiel.de

Abstract

Starting from a series of speech stimuli representing an F0 peak shift continuum from German early to medial peak, a series of non-speech stimuli is created. These non-speech stimuli show the F0 and intensity courses of the original speech stimuli, but with a constant formant structure. The results of a perception experiment reveal that the organisation of the peak shift continuum found for the identification of early and medial peaks in the speech stimuli can be replicated by the non-speech stimuli, indicating that early and medial peaks are signalled by an interplay of the F0 and intensity courses without reference to the spectral change at the accented-vowel onset.

1. Introduction

The phonology of the Kiel Intonation Model (KIM) for German distinguishes between two fundamental contour classes, peak contours and valley contours [1]. Founded on perception experiments (cf. [2]), further phonological distinctions within each class are set up with reference to the synchronisation of the contour unit to the accented-vowel onset. Within this framework, the present paper concentrates on two phonological categories called early and medial peak. They convey a diametrically opposed pragmatic meaning. With the early peak, a speaker judges a message as known or unalterable, whereas the medial peak is used to characterise a message as new and hence open for further discussion. F0 peaks having their maximum before the accented-vowel onset represent variants of the early peak category, while F0 peaks with a maximum (closely) after the accented-vowel onset consitute the medial peak category. The KIM is not the only model regarding the accented-vowel onset as the relevant point of reference for meaningful tonal units. In his holistic approach to melody in music and speech, Dombrowski [3] conceptualises melody in both domains as being composed of basic melodic accents which are further shaped by global processes and linked to critical points in time. As regards speech, these critical points are the accented-vowel onsets.

Starting from the phonology of KIM, Niebuhr (ongoing research, cf. [4]) performed perception experiments with different stimulus series based on a constant peak shift continuum from early to medial peak. Among others, he varied the internal timing (i.e. the global shape) of the F0 peak as well as the underlying intensity course. Subjects then indirectly identified the F0 peaks of the continuum as either early or medial by judging whether the stimulus utterances match pragmatically with a preceding context utterance. On the one hand, the results of all perception experiments showed a bipartite identification behaviour within the peak shift continuum. On the other hand, both peak shape and intensity course systematically influenced the perception of early and medial peaks within the peak shift continuum; e.g., the perceptual change from early to medial was accelerated or delayed.

On the basis of these findings, Niebuhr assumes that early and medial peaks are signalled by a different prominence pattern of the high and low portions within the rising-falling peak course, which is, among others, due to an interplay of F0 and intensity. In this perceptual strategy, the accented-vowel onset only plays a role in the form of the increase in intensity accompanying the consonant vowel transition. That is, the co-ordination of the F0 peak with the spectral change at the boundary between the accented vowel and the preceding consonant should not be relevant for perceptual differentiation between the two peak categories.

This paper presents a perception experiment investigating this assumption. It is hypothesised that the identification behaviour found for one of the stimulus series used in the perception experiment of [4] can be replicated by a new series, in which each stimulus shows the same F0 and intensity patterns as the corresponding original stimulus, while the formant structure is held constant.

2. Method

The stimuli of the present experiment were derived from one of the stimulus series used in the perception experiment of [4]. The series differed in the shape of the shifted F0 peak. Each series was created on the basis of the continuously voiced utterance *"Sie war mal Malerin"* ('She was once a painter') produced by a male speaker (the author) with *"Ma-"* ([ma:]) in *"Malerin"* ('painter') as the only accented syllable. The 11 stimuli of each series formed a peak shift continuum centered around the abrupt spectral change in the transition from the nasal [m] to the vowel [a:] of the accented syllable. That is, starting from an F0 peak coinciding with this spectral change (peak positions refer to the F0 maximum), the whole F0 peak was shifted in 5 steps of 20ms in both directions.

The stimulus series selected as the point of departure contained a symmetrical fast rising-falling F0 peak. The triangular shape was stylised at three contour points. Considering the natural F0 level of the speaker, the contour point constituting the peak maximum was placed at a frequency value of 150Hz. Extending over a frequency intervall of 45Hz or 6.2 semitones (st), the beginning of the rise started at 105Hz and covered 148ms, thus leading to a gradient of 41.9st/s for the rising movement. The last contour point was placed 133ms after the peak maximum at a frequency value of 110Hz. The frequency distance between this contour point and the preceding F0 maximum was 5.4st, resulting in a gradient of 40.6st/s for the falling movement. Both gradients fall below the physiological limits of F0 change as estimated by Xu and Sun [5]. The F0 peak was integrated into a stylised declination line spanning the utterance from 112Hz down to 90Hz.

To create the stimuli for the present experiment, each of the 11 stimuli of the selected series (referred to as the original stimuli) was subjected to an intensity and F0 analysis in *praat* [6] with the default analysis configurations. Using the

analysed F0 values of each original stimulus, a pulse train was generated and run through a sequence of second-order filters creating five formants at values around 580Hz, 1340Hz, 2370Hz, 3460Hz and 3770Hz ('hum' in *praat*). This procedure resulted in 11 synthesised signals representing an exact reproduction of the original F0 peak shift continuum within a constant hum sound determined by five steady formants.

In a following step, the intensity courses of the 11 synthesised signals were (value by value) multiplied by the intensity courses of the corresponding original stimuli. However, due to artifacts of this procedure, the resulting intensity courses of the synthesised signals matched the ones of the corresponding original stimuli only roughly. Therefore, the intensity courses of the synthesised signals had to be further manipulated. This manipulation was based on a visual comparison of the intensity courses of each synthesised signal and its original stimulus in *praat*. According to this comparison, small sections of the synthesised signal were multiplied by a factor higher or lower than 1 in *cool edit* [7]. The intensity course of the signal was then recalculated in *praat* and again compared with the intensity course of its original stimulus. This cyclic procedure was continued until the intensity courses in each pair of synthesised signal and original stimulus matched adequately. It has to be emphasized that this judgement is based on intensity analyses with a time window of 32ms effective length and a window shift of 10ms. For different settings, the deviations between the intensity courses of the synthesised signals and their corresponding original stimuli increase slightly.

The 11 synthesised signals were used as stimuli in the present perception experiment. Figure 1 shows a spectrogram with the five steady formants of the stimuli and the F0 peak shift continuum made up by the stimuli. Since the calculation of the intensity course is slightly influenced by F0, only the intensity course of the center stimulus 6 is given as an example in the middle section of Figure 1. As can be seen, the F0 peak was successively shifted into a section of increasing intensity, which was in the original stimuli brought about by the transition from the nasal [m] to the accented vowel [a:].

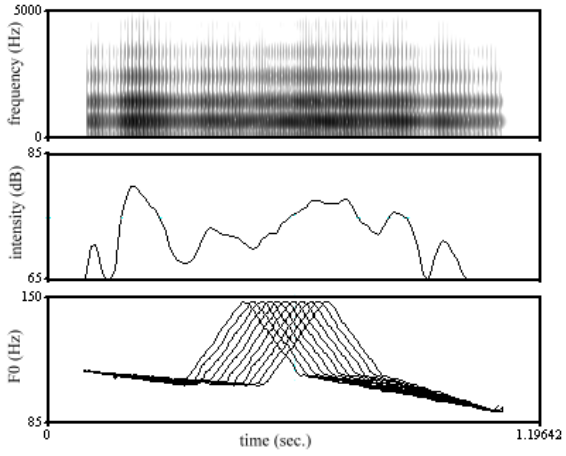

Figure 1: *Spectrogram (0-5kHz, top) and intensity course (65-85dB, middle) taken from the center stimulus 6 of 11 stimuli as well as the F0 courses (85-150Hz, buttom) of all 11 stimuli constituting the F0 peak shift continuum.*

In the perception experiments of [4], the identification of the F0 peaks contained in the stimulus utterances *"Sie war mal Malerin"* as either early or medial was guided by the constant preceding context utterance *"Jetzt versteh' ich das erst"* ('Now, I understand'), realisd with a medial peak on the accented syllable *"steh-"* of *"versteh"* ('understand'). Both context and stimulus came from the same male speaker (the author) and were produced in a comparable register with comparable voice quality and tempo. Hence, in the presentation, both could be interpreted by the subjects as part of one global utterance unit. This impression was further supported by an approriate pause duration between context and stimulus utterance. As regards their meaning relations, the two utterances were basically compatible on the lexical level. In this way, the final decision whether they matched or not was transferred to the intonational level. Here, the medial peak in the context utterance, together with the lexical meanings, expresses that the speaker is going to introduce new information in the following (stimulus) utterance. Consequently, an early peak characterising the information in the stimulus utterance as known is not compatible with the preceding context, whereas a medial peak is compatible. On this basis, subjects were asked in the experiment to judge whether context and stimulus utterances matched or not, in this way indirectly identifying the early and medial peak category in the stimulus utterances.

Since the present experiment consists of non-speech stimuli, the judgement of the F0 peaks in the continuum could not be guided by meaning. Moreover, it is assumed that in the experiments of Niebuhr [4] the categories to be identified were avaiable to the hearers as part of the knowledge of their language. Such kind of knowledge does not exist for the judgement of the non-speech stimuli in the present experiment.

Considering these differences, the stimuli of the present experiment were integrated into an AXB test. In this, the stimuli from both ends of the peak shift continuum, 1 and 11, served as A and B. For X, every stimulus of the continuum was used, including 1 and 11. Hence 11 AXB triplets were constructed. Two of them contained physically identical pairs of stimulus 1 or 11, respectively (referred to as AAB and ABB). The constant context frame of A and B provided the hearer with representatives of the melodic difference, in relation to which the X stimulus had to be judged. The context frame was thus meant as counterpart to the knowledge about the intonational categories the hearer could make use of in the speech experiments of Niebuhr [4]. Further, in the present non-speech experiment, the subjects were instructed to decide whether the melodic movement in the middle of the triplet had to be assigned to A or B, i.e. whether the triplet sounded like 'AAB' or 'ABB'. In order to identify pairs within the triplet, the subjects had to identify the X element as either A or B. With regard to the latter, this can be called an identification task. However, it is not a prototypical one, since an additional (bidirectional) discrimination component is introduced in the task by the constant context frame of A and B. Nevertheless, it is assumed that the AXB test produces data which can be compared to the data resulting from the identication test in Niebuhr [4]. To emphasize this comparability, the AXB test is in the following also regarded as an identification test. However, the reader should be aware of this simplification.

The AXB test consisted of 88 triplets, resulting from the 11 triplets copied eight times and arranged in a randomized order. Each of the triplets was introduced by a bleep and followed by a pause of three seconds in which the subjects had to judge the triplet. The elements of the triplets were separated by pauses of 500ms. Altogether, 19 native speakers of German (13 females, 6 males; age 22-34 years) participated in the experiment. Some of them had already participated in other perception experiments. Before the actual experiment, the sub-

jects were familiarised with their task by a short test consisting of the two triplets with physically identical elements, AAB and ABB, repeated five times in a randomized order. Subjects listened to the stimuli over headphones in a silent room. They responded to the stimulus triplets by putting ticks on a prepared sheet of paper.

3. Results

At first, it had to be considered that an identification task based on perceptual comparisons within triplets of non-speech stimuli would be much harder for subjects to fulfill than the identification task in the experiments of [4], in which subjects could rely on meaning relations between stimulus utterances and their preceding context. Therefore, the AAB and ABB triplets were used to filter out subjects who were not able to fulfill the task of the present experiment. Since these triplets contained physically identical pairs, the judgements of the subjects could objectively be evaluated as right or wrong. It was expected that subjects who managed the task would judge the AAB and ABB triplets in more than half of the 8 repetitions correctly. On this basis, four subjects (3 female, 1 male) were excluded from further analysis. A separate inspection of their data showed an identification rate around 50% for all 11 AXB triplets, hence supporting the assumption that they were not able to fulfill the task and judged the triplets by chance.

For the remaining 15 subjects, the black curve in Figure 2 shows for the 11 non-speech stimuli of the continuum, arranged with peak positions from left to right, the percentages of assignments 'to B'. Each value in this response profile represents 120 judgements. In addition to the response profile of the present experiment, the response profile received for the original speech stimuli in the experiment in [4] is given by the grey curve. It shows for 10 repetitions of all 11 context stimulus pairs the percentages of 'matching' judgements coming from 28 subjects. So, each value summarises 280 judgements.

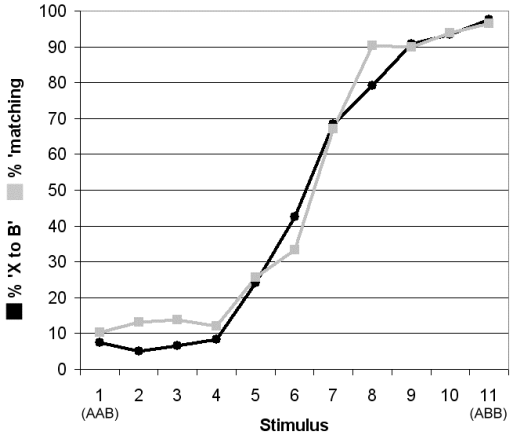

Figure 2: *Response profiles for the stimuli of the present experiment (black, percentage of X stimuli assigned to B, n=120) as well as for the original stimuli (grey, percentage of matching between stimulus and context utterance, n=280).*

The response profile of the present experiment illustrates that the melodic movements of stimuli 1-4, containing F0 peaks from the left edge of the continuum, were clearly assigned to the melodic movement of the preceding A stimulus. On the other hand, the melodic movements in stimuli 8-11 with F0 peaks from the right edge of the continuum were predominantly judged as belonging to the following B stimulus. Between these two sections of consistent assignment, there is a short transition phase concerning the three stimuli 5-7. In this, the change in the majority of assignments of X from A to B happens after the center stimulus 6, indicated by the response profile crossing the 50% line.

The properties described characterising the response profile of the present experiment can also be found in the response profile obtained for the original speech stimuli. That is, stimuli 8-11 matched with the preceding context utterance, whereas stimuli 1-4 did not fit into this context. A shift in the 'matching' judgements takes place between stimuli 5-7, with the change in the majority of judgements being located between stimuli 6 and 7. Deviations between the two response profiles are restricted to small differences in the sections of clear judgements and do not concern judgement behaviour per se. In this context, it has to be pointed out that for the deviation in stimulus 8, a closer look at the data suggests that its less clear identification as B is partially due to the presentation order of the stimulus triplets. This deviation will therefore be neglected in the following.

The visual comparison is supplemented by a statistical analysis, testing if the two response profiles differ significantly. For this, the response profile for each of the 43 subjects of both perception experiments was reduced to a single measurement, calculated by subtracting the relative frequency of 'X to B' or 'matching' judgements of stimulus 1 (number of judgements divided by number of stimulus repetitions) from the corresponding relative frequencies of stimulus 2 to 11 and by summing these differences. Finally, the relative frequency of 'X to B' or 'machting' judgements of stimulus 1 was added to this sum. A non-parametric *U* test revealed no significant differences between the 43 measurements of both experiments ($U_{(15,28;0.05)}$=203,5>144; p>0.05), in this way supporting the visual analysis of the response profiles.

4. Discussion and Conclusions

The results revealed that the identification behaviour of subjects for a series of speech stimuli constituting an F0 peak shift continuum from early to medial peak can be replicated by a series of non-speech stimuli only containing the F0 and intensity patterns of the original series. This outcome is in line with the hypothesis put forward in the introduction, i.e. the present results give further support to the assumption that the change in the spectral properties taking place at the boundary the consonant and the vowel of the accented syllable ([m] and [a:] in the present case) is no perceptual cue for the identification of early and medial peaks. Instead, considering that F0 and intensity were the only variable signal parameters in the stimulus series of the present experiment, the results point to an interplay of F0 and intensity as the basis for the perceptual differentiation of early and medial peaks.

To explain how this interplay might work, Figure 3 shows the relations between the F0 peak courses of stimuli 4 and 8 and the intensity course, representing the last clear A and the first clear B identification. It has to be noticed that Figure 3 only shows the intensity course of stimulus 8. Due to the influence of F0 on intensity, the intensity course of stimulus 4 is slightly deviating. However, this deviation does not touch the presented explanation. It can be seen from Figure 3 that the F0 maximum of the peak in stimulus 4 is located in an area of low intensity (originally belonging to [m]), immediately before a fast increase to a plateau of high intensity (originally belonging to [a:]). The latter plateau

spans the falling F0 movement as well as the low F0 at the end of the peak. Stimulus 8, on the other hand, contains the first F0 peak maximum lying within the high intensity plateau, while the low F0 values at the beginning and the end of the peak fall into a sections of decreasing or low intensity.

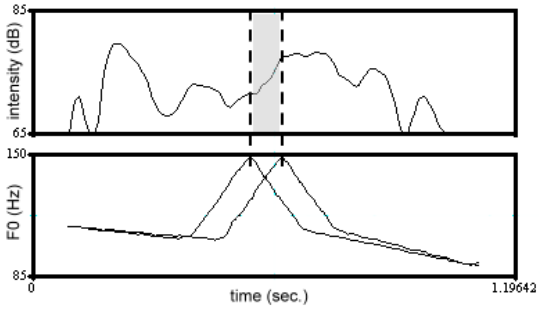

Figure 3: *The position of the F0 peaks in stimuli 4 and 8 in relation to the intensity course (of stimulus 8).*

These relations suggest that the distiction between early and medial peak, as represented by between stimulus 4 and 8, is based on specific prominence patterns of the high and low portions within the peak. In stimulus 4, the high portion of the peak around the F0 maximum is suppressed by its coincidence with a low intensity level, whereas the low portion at the end of the F0 peak is emphasised by the high intensity plateau. The opposite applies to the F0 peak in stimulus 8. Since for this stimulus, the high portion of the peak around the F0 maximum occurs within the high intensity plateau, it becomes more prominent than the surrounding low portions, which are located in sections with a low intensity level.

The interplays of F0 and intensity and the corresponding prominence patterns resulting for the stimuli 1-3 and 9-11 are comparable with the ones of stimuli 4 and 8. This accounts for the consistent identification of stimuli 1-4 and 8-11 as early or medial peak or as A and B, respectively. Furthermore, since the F0 peak maxima of stimuli 5-7 fall in the increasing intensity course preceding the high plateau, marked by a grey box in Figure 3, the high F0 portion of the peak is successively strengthened in perception. This explains the increase in the identification of medial peaks or B for this part of the peak shift continuum (cf. Fig. 2).

In this perspective, the perception of early and medial peaks for a constant peak shape is determined by the intensity course surrounding the F0 peak. In this connection, it has to be considered that the intensity course in an utterance consists of (at least) two fundamental components: one is fully determined by the articulatory properties of the underlying segmental string and the other is shaped by the intention of the speaker. With regard to the latter, the identification of early and medial peaks by a specific prominence pattern of the high and low portions in the rising-falling peak course is *in principle* independent from intrinsic intensity differences between single segments, e.g. the increase in intensity due to the consonant vowel transition within the accented syllable.

However, from an economic point of view, it seems likely that speakers make use of the instrinsically given intensity courses, instead of pursuing the extra effort of radically changing these courses. On this basis, the perception of early and medial peaks can indeed be regarded as guided by the coordination of the F0 peak with the increase in intensity due to the consonant vowel transition. This also means that the phonological differentation of intonational categories with reference to the accented-vowel onset, as it is done in KIM [1] and the model of Dombrowski [3], is basically an adequate approach. However, it needs to be refined. Moreover, the identification strategy outlined here not only concentrates on local events like the accented-vowel onset or the F0 maximum. It is contour based and integrates several parameters over a period of time.

In addition to the indications concerning the perceptual identification of early and medial, the research presented in this paper shows different intonational phenomena in a new light. One of these is the phenomenon of categorical perception. Based on a combination of discrimination and identification experiments, Kohler [2] postulates a categorical boundary between the early and medial peak. With regard to the outlined identification strategy, a categorical outcome would depend on the dynamics of the intensity increase within the peak shift continuum (cf. grey box in Fig 3) as well as on the slope of the rising and falling movement in the shifted F0 peak. As regards the latter, the steeper the slope of a peak shifted from left to right, the more abrupt the change in the perception from early to medial peak and the more pronounced the discrimination maximum between stimulus pairs will be. This indeed showed up in the perception experiments of [4]. When categorical perception is bound to certain segmental structures or peak shapes, it cannot be regarded as mirroring a phonological discreteness of the intonational categories involved, but as a by-product of the perceptual strategy. By contrasting the results of perception experiments for early and medial peaks and early and late valleys, this conclusion was also drawn by Niebuhr and Kohler [8].

Finally, a model of intonational categories which goes beyond F0 and integrates other parameters like intensity and which starts from perception might help to understand the phenomenon of segmental anchoring as well as alignment patterns observed for different segmental structures and speaking rate conditions. Considerations in this direction have already been made by Silverman and Pierrehumbert [9] and should be followed up.

5. References

[1] Kohler, K.J., 1991. Prosody in speech synthesis: the interplay between basic research and TTS application. *Journal of Phonetics* 19, 121-138.

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

[3] Dombrowski, E., 2003. Steps to a common description of melody in music and speech. *5th triannual ESCOM conference, Hannover*.

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

[5] Xu, Y. and X. Sun, 2002. Maximum speed of pitch change and how it may relate to speech. *JASA* 111(3), 1399-1413.

[6] Boersma, P. and D. Weenink. Praat : doing phonetics by computer. *http://www.fon.hum.uva.nl/praat/*

[7] For more information, see *http://www.cooledit.com*

[8] Niebuhr, O. and K.J. Kohler, 2004. Perception and cognitive processing of tonal alignment in German. *Proc. of the international symposium on tonal aspects of languages:emphasis on tone languages, Beijing*, 155-158.

[9] Silverman, K.E. and J.B. Pierrehumbert, 1990. The timing of prenuclear high accents in English. In *Papers in Laboratory Phonology I*, J. Kingston and M. Beckman (eds.). Cambridge: CUP, 72-106.