

# Categorical Perception in Intonation: a Matter of Signal Dynamics?

Oliver Niebuhr

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

on@ipds.uni-kiel.de

## Abstract

Results of recent perception experiments revealed that the signalling of rising-falling F0 peak categories in German intonation involves an interplay of F0 and intensity. Moreover, combining identification judgements and reaction times suggests that the abruptness of the perceptual change between the categories is determined by the signal dynamics in the sense of the *durations of the F0 peak movements and intensity transitions*. This undermines the use of categorical perception as an instrument to detect phonological intonation categories.

## 1. Introduction

Speech means codified communication. That is, information is converted into phonetic substance from which it is possible to recover the (same) information again. Within this process, it is beyond doubt that the information is composed by combinations of speech units. Since our cognitive representation of the outside world and hence our thinking and acting is organized in categories [1], it is logical to assume that these categories or categorization in general also mark the instruments that we develop, such as speech. That is, the units of the speech code are organized in a categorical structure [2]. They represent *communicative categories*. The centre of these categories constitutes the meaning component. It is completed by (replaceable) coding specifications, usually called the phonological form.

In the investigation of speech, a critical (but often implicit) assumption is made: The categorical structures of the code itself are also present in its transmission, i.e. in the speech chain. This assumption is, e.g., reflected in the search for discreteness of speech units resulting from *invariant properties* in the realms of production and/or acoustics. Within restricted laboratory conditions, such a search can be successful. However, since speech units (segmental and prosodic ones) are always coded by multiple factors in trading relations, and since the factors are additionally embedded in surround relations [3,1], it is clear that searching invariance is finally doomed to fail [4,5].

The next logical stage to search for categorical structures is the realm of perception, guided by the idea that the continuous acoustic input is split up into perceptually stable sections which are linked with the underlying (communicative) categories of the code [6]. This concept of *categorical perception* is even more problematic than the search for invariance, since it requires measuring the presence and quality of perceptual boundaries. Following the phoneme concept of the American structuralism, this is traditionally done by a combination of identification and discrimination tasks based on stimuli forming a continuum in a certain acoustic parameter. For categorical perception the results of the two tasks have to be related in the way that a (short) transition phase in the identification across the stimulus continuum coincides with a (clear local) maximum in the discrimination of the corresponding stimuli. Many objections were raised [7,8,9], e.g., conc-

erning the external validity of this experimental paradigm; and alternative approaches and statistical procedures have been used [10,11,15]. However, whether the acoustic continuum covered by the stimuli contains a categorical or a gradual change in perception seems to be itself a gradual decision. Nevertheless, categorical perception and the corresponding experimental paradigm have become a popular research tool to detect and to verify communicative categories of intonation and prosody, cf. e.g. [10,11,12,13,14,15]. Apart from the question whether this is a practical attempt, the present study raises doubts about the reliability of this tool by comparing the perceptual boundaries of German early, medial, and late peaks received for different stimulus conditions.

The three peak categories are part of the phonology of the Kiel Intonation Model (KIM) for German [16]. They represent rising-falling F0 peak contours, distinguished by their synchronization with the accented vowel. The phonological concept was founded on meaning-based perception experiments of [17], using an F0 peak shift continuum. It was found that peaks having the maximum before the accented-vowel onset are identified as early; and peaks with maxima after the vowel onset are identified as medial. For late peak identification, the peak must be shifted with its maximum beyond the vowel offset. Moreover, while the change from early to medial was categorical, the one from medial to late was gradual.

Further perception experiments of [18,19,20] revealed, e.g., that the transition from early to medial peak took place for those stimuli of the F0 peak shift continuum, in which the peak was shifted with its maximum across the increase in intensity from the low level of the consonant to the high level of the vowel in the accented syllable. The findings led to the idea that hearer do not identify early, medial, and late peaks by their synchronization, but by combinations of perceptual pitch and prominence patterns. In this refined concept, the pitch pattern arises from a decomposition of the rising-falling F0 peak into single pitch events, e.g. the low and high pitches represented by F0 sections around rise onset, maximum, and fall offset. The pitch patterns of early, medial, and late peaks are largely comparable, i.e. they are primarily differentiated by the prominence patterns. In this connection, the interplay of F0 and intensity plays an important role. For F0 peaks shifted into the vowel, e.g., the high pitch event around the peak maximum becomes successively more prominent by the increasing intensity at the CV boundary. Simultaneously, the low pitch around the F0 peak offset is shifted into the low intensity level of the postvocalic consonant and is hence step by step reduced in prominence. In this way, the perceptually outstanding pitch event changes from the low one at the peak offset to the high one at the peak maximum. This results in a change from early to medial peak identification. This signalling concept is explained in further details in [20,21].

It follows from this concept that, for a given peak shift, the prominence pattern and hence the identification of the peak categories change more rapidly, the faster the F0 and intensity movements are. The experimental results presented in the following meet this expectation. On this basis, it will

further be argued that the perceptual boundaries in the most dynamic stimulus conditions can count as categorical, where in the less dynamic conditions they are clearly gradual.

## 2. Method

The stimulus series (and their results) presented here are part of the comprehensive experiments of [18,19] and were derived from two types of F0 peak shift continua, both with local a rising-falling F0 peak contour stylized by points at rise onset, maximum, and fall offset. The peak contour was integrated into a global weak F0 declination across the stimulus utterances. The shifts and the resyntheses were done in *praat*.

One type of peak shift continua aims at the change from early to medial peak; i.e. the peak contour is shifted into the accented vowel. The other type is directed at the change from medial to late peak. Thus, the peak is shifted out of the accented vowel into the following syllable. In both types of continua, the shift was done in steps of 20ms. The underlying utterances were either “*Sie war mal Malerin*” or “*Sie’s mal Malerin gewesen*” (‘She was once a painter’). They differ in tense, but the domain of manipulation, the two words “*mal Malerin*” (‘once a painter’), was constant. In this, “*Ma-*” (produced as [ma:]) was the only accented syllable.

Considering the direction of the shift and the covered segments, the stimulus series resulting from the two types of peak shift continua are referred to as CV series or VC series, respectively. In [19], a total of 13 experimental conditions were created and integrated into the CV and VC series. From these, 3 conditions, i.e. 6 series, were selected for the present study. They differed in the peak shape and intensity levels (as well as durations) of the underlying syllables. The latter indirectly led to different durations of the intensity transitions between the segments. The intensity manipulations were done in *cool edit* (see <http://www.cooledit.com>).

The first condition was marked by a fast rising-falling F0 peak (*f/f*, slope durations 120-150ms, peak heights 5-7 semitones). In addition, also the intensity courses into the accented vowel and out of the accented syllable showed fast transitions from a low to a high or from a high to a low level, respectively. In the view of the present investigation, this is the basic – i.e. the most dynamic – condition. The corresponding two stimulus series are called CVff and VCff. The remaining two conditions were less dynamic, either with regard to the peak shape or to the intensity transitions. The two stimulus series of the second condition, CVss and VCss, showed fast intensity transitions comparable to ones in the basic condition, but a slow rising-falling F0 peak shape *s/s* with doubled durations of the F0 slopes. In the third condition, the intensity transitions into and out of the accented vowel are slower than in the basic condition, while the F0 peaks in both conditions have a comparable *f/f* shape. The resulting stimulus series are named CVint and VCint. Illustrations and audio examples for all stimulus series are provided in [21].

For each of the 6 stimulus series, one perception experiment was created. In this, the stimuli were preceded by a context utterance, constant for each experiment. Depending on the meaning of the F0 peak category perceived in the stimuli, the context stimulus pairs either match or not. So, by judging ‘matching’ or ‘not matching’ the subjects indirectly identified the early, medial, and late peaks in the stimuli. In each experiment, the context stimulus pairs were repeated several times in a randomized order. Groups of 18, 20, or 28 subjects took part in the experiments. They listened to the context stimulus pairs via loudspeakers in a sound-treated room and judged the pairs by pressing buttons. In addition to the buttons pressed,

the reaction times were collected. Following the arguments of [22,23] and considering the problems with the interpretation of the traditional discrimination test [7,8], the reaction time measurements were used as a substitute for the (AX or ABX) stimulus discrimination.

## 3. Results

The results of the six experiments are summarized in Figures 1a-c and 2a-c. Black lines represent the identification courses across the ascending stimulus numbers, which corresponds to a peak shift from left to right, i.e. into or out of the accented vowel [a:], respectively. The ‘matching’ and ‘not matching’ judgements given by the listeners were already translated into identifications of early, medial, and late peaks. For the CV series in Figures 1a-c, the percentages of medial peak identifications are shown. Figures 2a-c, concerning the VC series, show the percentages of late peak identifications. Depending on the stimulus repetitions and the number of subjects, each value represents 140, 180, or 280 judgements.

Grey lines show the corresponding reaction time values. In the analysis of these values, it was found that the stimulus series were judged with different reaction time levels and that higher reaction time levels are frequently accompanied by a larger range of values. To normalize for this phenomenon, percentages were calculated. Based on the mean values across all subjects, they show the increase in reaction time relative to the smallest mean value found for a stimulus in the series. In consequence, the latter value is 0%. Finally, since the reaction times of one/three subjects in the CVff/CVss conditions were not recorded, the corresponding percentages only represent 270/250 reaction times, respectively (cf. Fig. 1a,b).

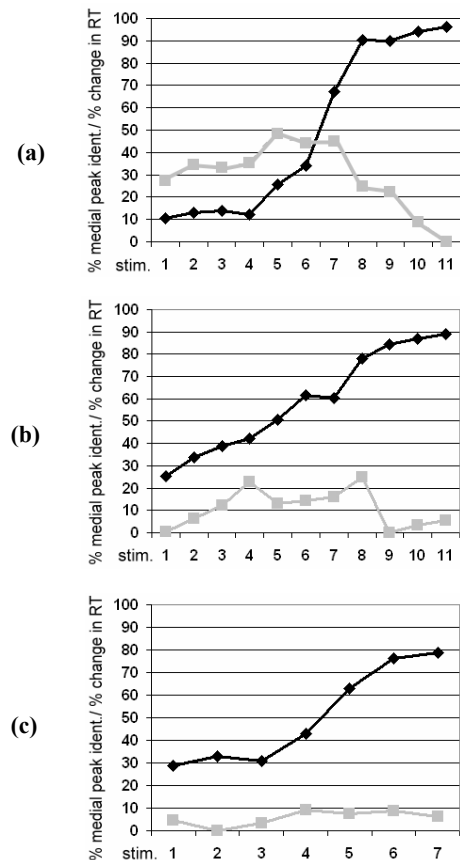


Figure 1: Identification courses (black) and reaction time courses (grey) for the stimulus series of conditions CVff (a,  $n=280/250$ ), CVss (b;  $n=280/270$ ), and CVint (c;  $n=140$ ).

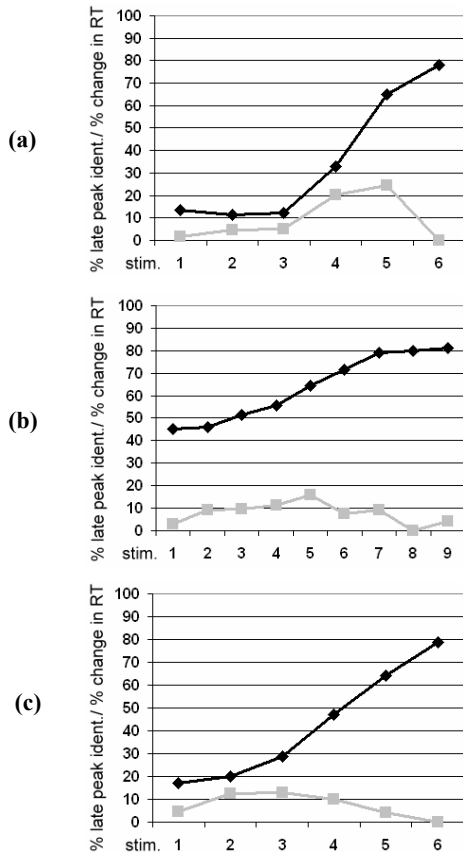


Figure 2: Identification courses (black) and reaction time courses (grey) for the stimulus series of conditions VCff (a,  $n=140$ ), VCss (b;  $n=180$ ), and VCint (c;  $n=140$ ).

As regards Figure 1a-c, the *f/f* peak shape and the underlying fast intensity transitions in the stimuli of the basic condition CVff (Fig. 1a) produce an abrupt transition from a clear identification of early to a clear identification of medial peaks. In addition, the short transition phase between stimuli 5 and 7 (corresponding to a peak shift of 60ms) coincides with a clear and pronounced increase in the reaction time. Compared with the smallest mean value of stimulus 11, which is clearly identified as containing a medial peak on “Malerin”, the reaction times at the maximum are almost 50% higher. The two stimulus series based on the less dynamic conditions, CVss and CVint (Fig. 1b-c), also produce a perceptual change from early to medial. However, the transition phase is broader, i.e. it spans more stimuli or peak shifts, respectively; in particular in case of CVss with the slow rising-falling peak shape. In the latter condition, the less abrupt change in identification is accompanied by a broad and bimodal reaction time plateau between stimuli 4 and 8. At the maxima of this plateau, the reaction time is only about 20% higher than the smallest reaction time found for stimulus 9. The reaction time course of condition CVint does not show a maximum at all. On the contrary, the reaction time course runs almost flat below 10%.

Comparable observations can be made in Figures 2a-c. The stimuli based on the most dynamic condition VCff yield a clear perceptual change from medial to late with a short transition phase in between (Fig. 2a). In this transition phase, a local and pronounced reaction time maximum develops, with a mean value almost 30% higher than the smallest one found for the final stimulus 6. Compared with this, the reaction time courses of the two less dynamic stimulus series VCss and VCint (Fig. 2b-c) do not show a clear maximum. On the contrary, most stimuli show similar reaction times of

about 10%. This goes well with the identification courses of VCss/int, which are marked by a rather slow perceptual change from medial to late across the peak shift continuum.

## 4. Discussion

The results of the present study in fact suggest that the dynamics of the F0 peak movements and the underlying intensity transitions are connected with the sharpness and quality of the perceptual boundary between the peak categories. Within an F0 peak shift continuum spanning the accented syllable and part of the following unaccented syllable, the range of peak positions which are unreliably identified as early, medial, and late shrinks, if the rising-falling peak shows fast movements, and if these movements additionally coincide with fast intensity transitions into and out of the accented vowel. This configuration, which characterizes the CVff and VCff series, also led to pronounced local reaction time maxima for the stimuli within the short transition phases of the identification courses.

If these properties of the identification and reaction time courses are jointly interpreted in the view of [22,23], they indicate that the perceptual boundaries between early and medial peak as well as between medial and late peak can count as categorical in the stimulus series CVff and VCff. On the other hand, the identification and reaction time courses for the stimulus series of the less dynamic conditions CV/VCss and CV/VCint, which were either marked by slower F0 peak movements or by slower intensity transitions, show clearly less abrupt and technically gradual changes in the perception from early to medial or from medial to late, respectively.

That is, the perceptual change from early to medial, which was regarded as categorical in [17,8] (cf. introduction), can be turned into a gradual one by decreasing the dynamics of the F0 and intensity movements. On the other hand, by increasing the dynamics of these movements, the change from medial to late, which was classified as gradual in [16], can become categorical. To support this interpretation also from the traditional point of view, additional discrimination experiments should be performed, since the terms ‘categorical’ and ‘gradual’ were originally based on statistical comparisons between the empirical and the theoretical stimulus discrimination derived from the identification course [6]. However, regarding [23] and own informal listening, it is expected that such additional discrimination courses will closely resemble the reaction time courses of the present study. Moreover, following the objections against the discrimination tests by [7,8,9,22], it is assumed that the present findings are sufficient to justify the given interpretation: categorical perception in intonation is a matter of F0 and intensity (and in this sense of signal) dynamics.

While the latter statement also supports the idea that the signalling of the German peak categories involves an interplay of F0 and intensity [19,20,21], it has to be pointed out that whole picture of findings cannot be explained with reference to the dynamic properties of F0 and intensity. So, e.g., [8] found a categorical change from early to medial for a rising-falling F0 peak shifted across the accented-vowel onset. Another stimulus series with an inverted falling-rising valley contour also yielded a comparably clear and abrupt change in the identification course covering the German early and late valley categories. However, no corresponding discrimination maximum appeared. Therefore, despite comparable dynamics of the relevant F0 and intensity movements in the peak and valley series, the perceptual change in the latter must be classified as gradual. Similarly, in their investigation of intonational emphasis in English based on a peak height

continuum, [14] found an abrupt change in identification but no corresponding maximum in discrimination. On the other hand, [12] found evidence for a categorical change within a peak height continuum (with a differently aligned peak maximum) aiming at the distinction between yes-no questions and wh-questions in Majorcan Catalan.

Apart from methodological aspects (e.g. step size, instruction and competence of the subjects), this inconsistent picture can be interpreted in two different ways: First, it is incorrect that the abruptness of the perceptual change is related to the dynamics of F0 and intensity movements. Even in this case, however, it remains a fact that German early, medial, and late peaks are not consistently separated by either categorical or gradual boundaries. The second and more likely interpretation is that the dynamics in the acoustic signal is *not the only* factor determining the quality of the perceptual boundary. Among others, a different signalling of the intonation categories has to be taken into account. For instance, if the signalling of the investigated categories involves structural differences like falling vs. rising movements or syntagmatic relations between local temporal events (as it is assumed for the German peak categories, cf. [8, 19]), stimulus continua can contain an artificial discontinuity (cf. [1]), which is then reflected in a special abruptness of the perceptual change (and frequently misinterpreted as the linguistic mode of “categorical perception”). Moreover, it can be assumed that the dynamics of acoustic parameters like F0 and intensity particularly affect perceptual changes between those categories that differ in syntagmatic (i.e. temporal) patterns. In these cases, finally, F0 and intensity may not be the only relevant acoustic parameters.

The discussion clearly shows that it is highly problematic to use categorical perception as a criterion to set up the intonation categories of a language. If at all, categorical perception tells us more about the signalling of intonation categories than about their presence. That is, once we really understand the signalling of the intonation categories within as well as across languages, we will in many cases be able to control whether a stimulus series yields a categorical or a gradual outcome. This could be a future perspective for categorical perception as a research tool. At present, categorical perception rather obstructs progress in intonation research. Instead of investigating the meanings and the signalling of intonation categories, the research efforts are directed towards their boundaries. A more promising way to establish communicative categories is an approach based on function/meaning, respectively ([8,19]).

Finally, the assumption that the categorical structure of the code itself also characterizes its transmission likely goes back to a too mechanistic concept of the communication process. The hearer is not a mere receiver of coded elements. Speech perception means active and continuously revised interpretation of (multimodal) information based on the represented categories and supported by powerful signal-external processes [1,5,19]. Reducing the incoming information at an early processing stage by means of categorical perception impedes the interpretation instead of supporting it, cf. [7].

## 5. References

- [1] Handel, S., *Listening – An introduction to the perception of auditory events*, MIT Press, Cambridge, 1986.
- [2] Rosch, E., *Principles of categorization*. In Rosch, E., Lloyd, B.B. (eds.), *Cognition and categorization*, Erlbaum, Hillsdale, 1978.
- [3] Repp, B.H., “Phonetic trading relations and context effects: new experimental evidence for a speech mode of perception”, *Psychological Bulletin*, 92, 1982, 81-110.
- [4] Nygaard, L.C., Pisoni, D.B., *Speech perception: New directions in research and theory*. In Miller, J.L., Eimas, P.D. (eds.), *Speech, language, and communication*, Academic Press, San Diego, 63-97, 1995.
- [5] Lindblom, B., *Explaining phonetic variation: A sketch of the H&H theory*. In Hardcastle, W.J., Marchal, A. (eds.), *Speech production and speech modelling*, Kluwer, Dordrecht, 403-439, 1990.
- [6] Liberman, A.M., Harris, K.S., Hoffman, H.S., Griffith, B.C., “The discrimination of speech sounds within and across phoneme boundaries”, *Journal of Experimental Psychology* 54, 1957, 358-368.
- [7] Massaro, D.W., “Categorical perception: Important phenomenon or lasting myth?”, *Proc. of the 5th ICSLP*, Sydney, Australia, 2275-2279, 1998.
- [8] Niebuhr, O., Kohler, K.J., “Perception and cognitive processing of tonal alignment in German”, *Proc. of TAL2004*, Beijing, China, 155-158, 2004.
- [9] Cummins, F., Doherty, C., Dille, L., “Phrase-final pitch discrimination in English”, *Proc. of the 3rd international conference of speech prosody*, Dresden, Germany, 467-470, 2006.
- [10] Pierrehumbert, J.B., Steele, S.A., “Categories of tonal alignment in English”, *Phonetica*, 46, 1989, 181-196.
- [11] Kleber, F., “Form and function of falling pitch contours in English”, *Proc. of the 3rd international conference of speech prosody*, Dresden, Germany, 61-64, 2006.
- [12] Vanrell, M. d. Mar, “A scaling contrast in Majorcan Catalan”, *Proc. of the 3rd international conference of speech prosody*, Dresden, Germany, 807-810, 2006.
- [13] Scheider, K., Lintfert, B., “Categorical perception of boundary tones”, *Proc. of the 15th ICPhS*, Barcelona, Spain, 631-634, 2003.
- [14] Ladd, D.R., Morton, R., “The perception of intonational emphasis: continuous or categorical?”, *Journal of Phonetics*, 25, 1997, 313-342.
- [15] Radtcke, T., Harrington, J., “Is there a distinction between H+!H\* and H+L\* in standard German? Evidence from acoustic and auditory analysis”, *Proc. of the 3rd international conference of speech prosody*, Dresden, Germany, 783-786, 2006.
- [16] Kohler, K.J., “Prosody in speech synthesis: The interplay between basic research and TTS application”, *Journal of Phonetics*, 19, 1991, 121-138.
- [17] Kohler, K.J., “Categorical pitch perception”, *Proc. of the 11th ICPhS*, Tallinn, Estonia, 331-333, 1987.
- [18] Niebuhr, O., “Perceptual study of timing variables in F0 peaks”, *Proc. of the 15th ICPhS*, Barcelona, Spain, 1225-1228, 2003.
- [19] Niebuhr, O., *Perzeption und kognitive Verarbeitung der Sprechmelodie. Theoretische Grundlagen und empirische Untersuchungen*, Ph.D. thesis, University Kiel, 2006.
- [20] Niebuhr, O., “The role of the accented-vowel onset in the perception of German early and medial peaks”, *Proc. of the 3rd international conference of speech prosody*, Dresden, Germany, 109-112, 2006.
- [21] Niebuhr, O., *Categorical perception in intonation and the dynamics of F0 and intensity movements*, [http://www.ipds.uni-kiel.de/on/on\\_CSPinto\\_07.html](http://www.ipds.uni-kiel.de/on/on_CSPinto_07.html), 2007.
- [22] Chen, A., “Reaction time as an indicator to discrete intonational contrasts in English”, *Proc. EUROSPEECH 2003*, Geneva, Switzerland, 97-100, 2003.
- [23] Pisoni, D.B., Tash, J., “Reaction times to comparisons within and across phonetic categories”, *Perception & Psychophysics*, 15, 1974, 285-290.