INTERSPEECH 2009 BRIGHTON

German Boundary tones show Categorical Perception and a Perceptual Magnet Effect when presented in different contexts

Katrin Schneider[1], *Grzegorz Dogil*[1], *Bernd Möbius*[1,2]

[1]Institut of Natural Language Processing, University of Stuttgart, Germany
[2]Institute of Communication Sciences, University of Bonn, Germany
katrin.schneider,grzegorz.dogil,bernd.moebius@ims.uni-stuttgart.de

10.21437/Interspeech.2009-664

Abstract

The experiment presented in this paper examines categorical perception as well as the perceptual magnet effect in German boundary tones, taking also context information into account. The test phrase is preceded by different context sentences that are assumed to affect the location of the category boundary in the stimulus continuum between the low and the high boundary tone. Results provide evidence for the existence of a low and a high boundary tone in German, corresponding to *statement* versus *question* interpretation, respectively. Furthermore, in contrast to previous findings, a prototype was found not only in the category of the low but also in the category of the high boundary tone, supporting the hypothesis that context might have been taken into account to solve a possible ambiguity between H% and a previously hypothesized non-low and non-terminal boundary tone.

Index Terms: speech prosody, categorical perception, perceptual magnet effect

1. Introduction

Prosody has an integrating function in the organization and production of spoken language by embedding information on different levels of the linguistic structure in a consistent reference frame [1]. Prosody can even change the semantic content: "...when there is a discrepancy between the prosody of the utterance and its overt semantic content we usually trust the prosody rather than the semantics." [2]

Three phonetic cues are essential for coding prosodic information: duration, intensity, and fundamental frequency (F_0). Pitch (perceived F_0) can express many functions such as tone, accent, intonational meaning and discourse structure [1], and speaker and listener need to distinguish between different possible interpretations to transfer or extract the information coded in the speech signal. This leads to the hypothesis that, besides segmental categories, prosodic categories might exist as well.

1.1. Testing categorical perception (CP)

The categorical perception (CP) paradigm, as described by Repp in [3], was successfully adapted to the prosodic research area, e.g. in [4, 5]. To test for CP, a stimulus continuum has to be created that covers the perceptual space between the hypothesized categories. During an identification test, subjects have to assign each stimulus to one of the proposed categories. In a discrimination test, pairs of stimuli have to be evaluated as consisting of either identical or different stimuli. If a pair consists of different stimuli, the stimuli are only one step apart from each other in the stimulus continuum (e.g. S7-S8). Results of experiments in different languages, e.g. for Dutch [4], for hummed stimuli in European Portuguese [6], and for German [5], provide evidence for the existence of at least two distinct boundary tone categories, a low (L%) and a high boundary tone (H%), corresponding to *statement* and some kind of *question* interpretation, respectively. In all these studies out-of-the-blue sentences were used, i.e. the stimuli were presented to the listeners without any context information. Whereas there was always a clear s-shaped curve obtained in identification, the discrimination results differed across studies. In [6] no discrimination peak was found, and in [4] and [5] some kind of discrimination plateau occurred around the category crossover, and a correlation between intra-subject crossovers and discrimination peaks was observed. Therefore, the authors concluded that categorical perception was found for the boundary tones under investigation.

1.2. Testing the perceptual magnet effect (PME)

Discrimination differences also inside the obtained categories were reported in [5]. These results appear to be more compatible with the concept of a perceptual magnet effect (PME) introduced by Kuhl [7] rather than with CP. According to [7], language acquisition warps the perceived distances between different instances of a given category. During speech perception, perceived instances of a category are stored in the perceptual space of the listener. Each instance has a quality value that describes its goodness of fit to the respective category; the best instance of the category is called the prototype (P). Discrimination ability is significantly reduced around P, i.e., it is very hard to distinguish P perceptually from its neighbors. Discrimination is not reduced around a non-prototype (NP).

We have previously adopted this design to investigate the boundary tone categories in German [8]. The results demonstrated a reduced discrimination sensitivity around P inside the *statement* category (L%), but in the *question* category (H%) discrimination sensitivity was reduced around P as well as around NP. Therefore we hypothesized a third category to be present in the perceptual space of H%, represented by a high, but non-terminal, boundary tone with a *continuation* interpretation.

This paper aims to verify if the boundary tones L% and H% in German are categorically perceived when context information is presented. Furthermore, the effect of context information on possible perceptual magnets in both categories is tested.

2. Methods

2.1. Stimuli

To remove syntactic biases, a verbless prepositional phrase (PP) *nach Panama* 'to Panama' was selected as the test phrase for

the perception experiment. The PP is ambiguous between a *statement* or *question* interpretation in German. Furthermore, *Panama* is a polysyllabic noun carrying lexical stress and a pitch accent on a non-final syllable, which helps disentangle the intonational effects of pitch accent and boundary tone. A male native speaker of German produced a set of tokens of the PP, and a token that sounded maximally ambiguous between *statement* or *question* was selected as the basis for creating an acoustic stimulus continuum.

The speaker produces an average rise by 90 Hz to reach an H% and an average fall by 50 Hz to reach an L%, as measured in a larger speech corpus. These values were taken as the range for the boundary tone height, which was then divided into 11 equidistant steps of 0.338 ERB width. The ERB scale was used because it is considered to be the most satisfactory psychophysical transformation of pitch intervals in human speech [9]. Then, F_0 was interpolated between the end of the first syllable of *Panama* and the new boundary tone. The PSOLA technique available in Praat was used to resynthesize the stimuli. To test around a possible P, additional stimuli were added above H% and below L% as long as they sounded natural. This procedure resulted in a set of 20 stimuli, and each stimulus was presented in 3 possible contexts:

1. *L% condition*: The PP is preceded by an unambiguous statement ending in a low boundary tone:
   Er will verreisen. Nach Panama./?
   'He wants to make a journey. To Panama./?'
2. *H% condition*: The PP is preceded by the same sentence as in 1., but ended in a high boundary tone and was therefore interpreted as an unambiguous question:
   Er will verreisen? Nach Panama./?
   'He wants to make a journey? To Panama./?'
3. *Wh_L% condition*: The PP is preceded by a syntactic question containing a question particle and ending in a low boundary tone:
   Was liegt da? Ein Ticket nach Panama./?
   'What is this? A ticket to Panama./?'

We hypothesized, first, that CP will be observed for the two boundary tones, L% and H%, and that PME will be found at least in the *statement* category (as in [8]); and second, that the presence of context information will affect the location of the category boundary in the stimulus continuum. Thus, if a question (marked either by boundary tone height or by a question word) is presented preceding the PP, then the boundary tone of the PP will have to be higher to be interpreted as a question than when a statement is presented as the preceding context.

2.2. Experimental procedure

To test the same subjects for CP as well as for PME, and because the tests for CP and for PME are identical in their identification subtest, we decided to combine both experimental designs. During all the subtests, reaction time (RT) was recorded. The subjects were encouraged to answer as quickly as possible.

The first subtest was an identification test. Subjects were asked to decide whether the PP presented after the context sentence corresponds to a *statement* or a *question*.

The second subtest was a goodness rating task, carried out separately for each context and each boundary tone category. Here the subjects were asked to decide how well a given stimulus fitted into its assigned category. All stimuli identified at better than 60% as an instance of the pertinent category were included into the rating task. Subsequently, a prototype (P) and a non-prototype (NP) were determined for each category and each context, with P corresponding to the stimulus with the highest, and NP corresponding to the stimulus with the lowest, average rating of the category.

The third subtest was a CP discrimination test. Subjects listened to pairs of stimuli and were asked to decide whether or not the two stimuli in a pair were identical with respect to boundary tone height. As each stimulus actually consisted of 2 phrases (context and target), participants had to listen to 4 stimuli in each trial and to compare the second and fourth phrases to each other. The test was carried out separately for each context, and the difference between the stimuli within one pair, if there was a difference, was one manipulation step.

The fourth subtest was a PME discrimination test in which task of the subjects was the same as in the third subtest. However, the acoustic difference between the stimuli in each pair was larger than in the CP discrimination, viz. between 2 and 4 manipulation steps. Furthermore, one of the stimuli in each pair was always P, or NP, paired with one of its neighbors. Each context and each boundary tone category was tested separately.

2.3. Participants

The CP experiment was completed by 36 participants (23 females, 13 males), whereas the PME part of the experiment, including the goodness rating test, was completed only by 29 participants (17 females, 12 males). All participants were students without any specific knowledge of prosody and they were paid for their attendance. The lower completion rate for the PME part was evidently due to the demand imposed by the length of the experiment. All session combined lasted about 11 hours, and in order to minimize possible learning effects sessions were separated by weeks without testing.

3. Experimental results

An initial look at the reaction times (RTs) revealed some extreme values. Therefore, all outliers, i.e. all results with an RT longer than $2 * sdev(RTs) + mean(RTs)$, were excluded from further analyses in each subtest. Hence, all results with a RT below 3 seconds were analysed.

An interesting global result with respect to RT was that in all experimental subtests the female participants were significantly faster than the male ones. At this time, the reasons for this finding are up for interpretation.

3.1. Identification test

There were 3 different context sentences, 20 manipulation steps and 10 repetitions of each stimulus, yielding a total of 600 stimuli for the identification task. Stimuli were presented in randomized order in 3 sessions to keep session durations acceptable.

The results pooled for all contexts (Figure 1) as well as in each single context condition (not displayed) showed clear s-shaped curves. Significant differences between the three contexts were only found in the vicinity of, but not directly at, the category crossover. As expected, the location of the category boundary shifts towards the *question* interpretation, if a question is presented as the preceding context. Interestingly, no differential effect was observed for the context question being marked by a high boundary tone vs. by a question particle.

Female participants showed a slightly, but not significantly, earlier category crossover than the male subjects. This might be due to the use of a male test voice in the experiment.

Furthermore, we found evidence that RT values can be used

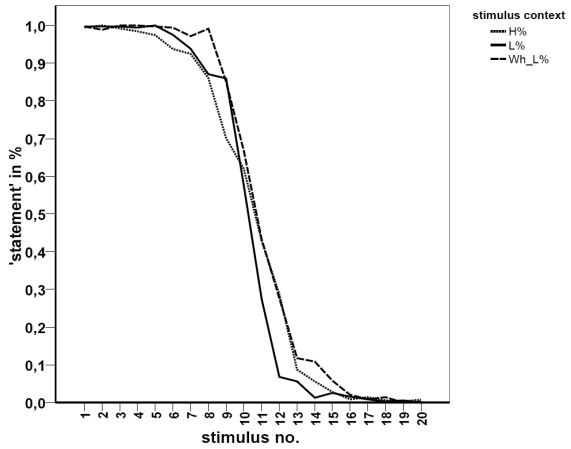

Figure 1: *Identification function for each context separately.*

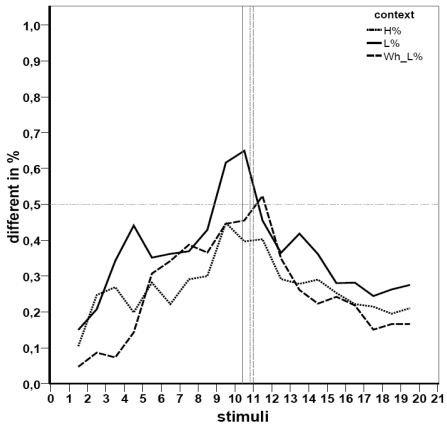

Figure 2: *CP discrimination pooled over all subjects. Vertical lines correspond to crossovers in different contexts.*

as an index for the location of a category boundary between *statement* and *question* interpretation in German as there was a significant correlation between the highest RTs and the location of the category crossover for each context and for both genders.

3.2. Goodness rating test

With respect to the 60% identification criterion, stimuli 1 to 10 were included in the rating for the *statement* category and stimuli 11 to 20 had to be rated as instances of the *question* category. All stimuli were repeated 10 times and presented in randomized order, split in 6 sessions with 100 stimuli each. The rating scale ranged from 1 (*very bad*) to 9 (*very good*), i.e., a higher rating corresponded to a better fit into the assigned category.

Average ratings were computed for each stimulus in each context to determine P and NP for each category and all contexts. Although there were slight context dependent differences in the location of P and NP for the *statement* category, these rating differences were small, yielding a constant correspondence of stimulus 1 as P, and stimulus 10 as NP, for this category. For the *question* category, there were no rating differences between contexts at all. Therefore, stimulus 20 corresponded to P, and stimulus 11 to the NP, of the *question* category.

3.3. Discrimination test: CP design

In this task, 19 AB pairs (B stimulus has a higher boundary tone than A), 19 BA pairs, and 38 AA pairs (identical stimuli) were tested. Each pair was repeated 5 times, resulting in a total of 380 stimulus pairs in each of the contexts (2 subtests).

An order of presentation effect was found in the direction observed in our previous experiment [8], viz. discrimination is better in AB pairs than in BA pairs.

Discrimination sensitivity is relatively low and differs between contexts (Figure 2). When a statement (L% condition) precedes the PP, discrimination is better than with a preceding question (H% or Wh_L% condition). Therefore, a question context, especially when the question ends in an H%, seems to impede the discrimination decision. Nevertheless, the low number of false alarms suggests that for most participants the discrimination task was feasible. In general, female subjects were worse than male ones in this task, which might again be due to the use of a male test voice.

There was a fair correlation between the crossover location and the discrimination peak for each context pooled over all participants, and a non-significant correlation between the individual peaks and crossover locations for two of the three contexts. This finding supports the CP hypothesis for the boundary tones L% and H% in German when context information is included.

3.4. Discrimination test: PME design

For both boundary tone categories, the discrimination results revealed that discrimination ability rises with increased boundary tone height difference in the stimulus pairs. However, there were subject-specific differences in the number of correctly discriminated pairs (hits) and the number of pairs that were wrongly classified as consisting of different stimuli (false alarms). Therefore, Signal Detection Theory (SDT) was applied to the analysis. According to SDT, listeners who share the same perceptual pre-condition (identical auditory threshold) can produce different results in a perception test because they use *response criteria* of different sizes [10]. On any trial, the answer of the observer is *yes* if the evidence for the signal is larger than the response criterion λ_{Center}, and *no* otherwise. Therefore, our data were transformed into λ_{Center} values, which take the hit rate (h) and the false alarm rate (f) into account: $\lambda_{Center} = -0.5 * (Z(f) + Z(h))$.

The results for the *statement* category show that λ_{Center} values around P were significantly higher than around NP (Figure 3, black lines), indicating a significantly worse discrimination ability around P than around NP. This can be interpreted as evidence for PME in the *statement* category.

In the *question* category, only the two nearest neighbors of P and NP differ significantly in their λ_{Center} values from each other (Figure 3, grey lines). Although this effect is not as strong as in the *statement* category, it can be seen as evidencing PME.

4. Discussion

Our experimental results confirm a categorical difference between the boundary tones L% and H% in German and support the existence of a perceptual magnet in both categories.

In identification, although there were slight gender-specific as well as context-specific differences in the location of the category boundary, the clear s-shaped curves are the first indications of the existence of two clearly distinct boundary tone categories in German, i.e. a low one (L%) corresponding to *statement* and

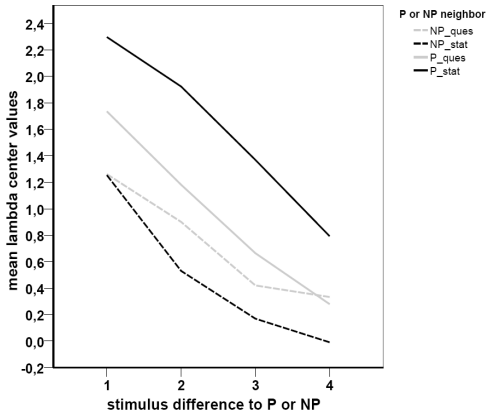

Figure 3: *Discrimination performance (given as lambda center values) around P and NP in both categories.*

a high one (H%) corresponding to *question* interpretation.

The CP discrimination test supported the finding of two categories of German boundary tones, because it showed a correlation between the general category crossover location and the discrimination peak, for each context. This is in line with the definition of CP, although discrimination performance was relatively low even at the category crossover. There might be different reasons for this finding. First, our stimuli were much longer than simple out-of-the-blue phrases. Subjects had to compare the last 2 syllables of the second phrase with the same syllables of the fourth phrase. The temporal distance might have caused discrimination difficulties due to short-term memory capacity limitations of different listeners. This problem might be solved by boosting the stimulus step difference during manipulation when using stimuli of this length. Furthermore, repetition of the stimuli during discrimination could be allowed. Because of the stimulus length, repetition should cause no auditory discrimination effect. Second, the low discrimination performance might result from the fact that there were more female than male participants, and women were worse in their discrimination than men, possibly because of the male test voice. Conceivably, listeners are better at interpreting the boundary tone height of speakers of their own gender. To test this hypothesis, an experiment using a female test voice and the same experimental design is currently being carried out.

In the goodness rating, a clear prototype as well as a clear non-prototype were determined for both boundary tone categories. Furthermore, PME discrimination results demonstrate a warping of the perceptual space towards P but not towards NP. The slightly lower discrimination performance around NP might be caused by the generally lower discrimination performance when there is only one manipulation step between NP and its neighbor, and not because there is a warping around NP too. Although the warping around P is much stronger in the *statement* than in the *question* category, the results support PME in the *question* category as well. This is in contrast to previous results in out-of-the-blue sentences [8], where PME was only found in the *statement* category. Maybe the use of context information limits the interpretation alternatives in the category of the high boundary tone to *question* only, thereby ruling out a *continuation* interpretation, which was available in the out-of-the-blue sentences.

Our data show an order of presentation effect in that AB pairs, i.e. when the second stimulus has the higher boundary tone, are better discriminated than BA pairs. This might be because AB pairs behave against the expected F_0 declination. Declination occurs in natural speech, i.e. F_0 is lowered during the course of the utterance, such that pitch accents at the end of a phrase are lower than at the beginning of a phrase, everything else (such as relative prominence) being equal. If this behavior is projected to the perception of boundary tones, then, when producing two subsequent phrases, the second one should have a slightly lower boundary tone than the first one, due to global F_0 declination. Therefore, an AB pair behaves contrary to expectation and might thus be better discriminated than a BA pair with the same stimuli, because in BA pairs declination "masks" the boundary tone difference.

The next analysis steps will be to take the RTs measured during all subtests into account in the interpretation of the results of the presented experiment. Furthermore, the results of the continuing experiment using a female voice will be compared to the results reported here.

5. Conclusions

In a series of experiments, the perception of a stimulus continuum of German boundary tones was investigated, ranging from an unambiguous *statement* to an unambiguous *question* interpretation, preceded by three different contexts. Although discrimination performance was not very good in general, experimental evidence for a categorical perception of two intended boundary tones was obtained. In contrast to previous experiments, a perceptual magnet effect was found in both boundary tone categories, leading to the interpretation that context has an influence on the strength of a possible magnet effect, in particular when it resticts competing interpretations of the stimulus.

6. Acknowledgements

This work was supported by a grant from the German Research Foundation (DFG) in the framework of Priority Programme SPP-1234 (grant MO 597/2).

7. References

[1] Dogil, G.,"Understanding Prosody", in Rickheit, G., Hermann, T. and Deutsch, W. (eds.), Psycholinguistics. An International Handbook, Berlin: de Gruyter, 544-565, 2003.

[2] Hirst, D.,"The Phonology and Phonetics of Speech Prosody: Between Acoustics and Interpretation", in Proc. Speech Prosody, Nara, 163-170, 2004.

[3] Repp, B.H.,"Categorical Perception: Issues, Methods, Findings", in Lass, N.J. (ed), Speech and Language: Advances in Basic Research and Practice (10), New York: Acad. Press, 243-335, 1984.

[4] Remijsen, B. and van Heuven, V.J.,"Gradient and Categorical Pitch Dimensions in Dutch: Diagnostic Tests", in Proc. 14th ICPhS, 1865-1868, 1999.

[5] Schneider, K. and Lintfert, B.,"Categorical perception of boundary tones in German", in Proc. 15th ICPhS, 631-634, 2003.

[6] Falé, I. and Hub Faria, I.,"Categorical Perception of intonational contrasts in European Portuguese", in Proc. Speech Prosody, Dresden, 171-174, 2006.

[7] Kuhl, P.K.,"Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not", in Perception & Psychophysics vol.50(2), 93-107, 1991.

[8] Schneider, K. and Möbius, B.,"Perceptual Magnet Effect in German boundary tones", in Proc. Interspeech, Lisbon, 41-44, 2005.

[9] Hermes, D.J. and van Gestel, J.C.,"The frequency scale of speech intonation", JASA 90(1), 97-102, 1991.

[10] Wickens, T.D.,"Elementary Signal Detection Theory",OUP,2002.