Perceptual magnet effect in German boundary tones

Katrin Schneider, Bernd Möbius

Institute of Natural Language Processing, Experimental Phonetics Group
University of Stuttgart, Germany
{katrin.schneider,bernd.moebius}@ims.uni-stuttgart.de

Abstract

The experiment described in this paper tests for the perceptual magnet effect within the categories of high and low boundary tones in German, referring to question and statement, respectively. The experiment is based on previous work in which the categorical status of the two German boundary tones had been evaluated. The results found there showed that there was a discrimination ability within categories which could not be explained by the classical definition of categorical perception. The results reported in the present paper show that a perceptual magnet exists in the statement category but not in the question category.

1. Introduction

Three essential phonetic cues are used for coding prosodic information: duration, intensity and fundamental frequency (F_0). Pitch, i.e. perceived F_0, can express many functions such as tone and accent, intonational meaning and discourse structure [1]. But how can one single cue give rise to so many interpretations? How can speakers and listeners distinguish the different possibilities of interpretation?

This presentation focuses on a small part of the German intonation system: the perception of boundary tones. In a preceding experiment [2] categorical perception (CP) of low and high boundary tones in German, presented by "statement" and "question" respectively, was found, but some kind of discrimination ability within each category was also shown which is not in line with the strict definition of categorical perception. This result could be better explained by the concept of the perceptual magnet effect (PME) introduced by Kuhl [3]. The experiment presented in this paper was designed with its main focus on PME in the two categories examined, i.e. the category of the low (L%) and the high (H%) boundary tone. To test for PME we chose a sentence that was syntactically ambiguous between statement and question, the contrast being encoded by the phrase-final intonation. We created a series of stimuli which differed in the phrase-final F_0 contour. The original intensity contour and the original timing were retained. For manipulation we used the ERB (Equivalent Rectangular Bandwidth) scale, because this frequency scale is considered to be the most satisfactory psychophysical transformation of pitch intervals in human speech [4].

2. Methods

2.1. Stimuli

In this experiment the same test sentence as in [2] was used: "Steht alles im Kochbuch" ("It's all in the cookbook"). The sentence had been selected from a recording of several dialogs with natural intonation. A professional male speaker of German read these dialogs in the anechoic chamber at the Institute of Natural Language Processing (IMS). The selected sentence satisfied the following conditions for manipulation: it was syntactically ambiguous between statement and question; and the sentence final syllable was unaccented, did not contain a schwa, and its F_0 formed a plateau-like contour ending on medium level (128 Hz) within the speaker's pitch range.

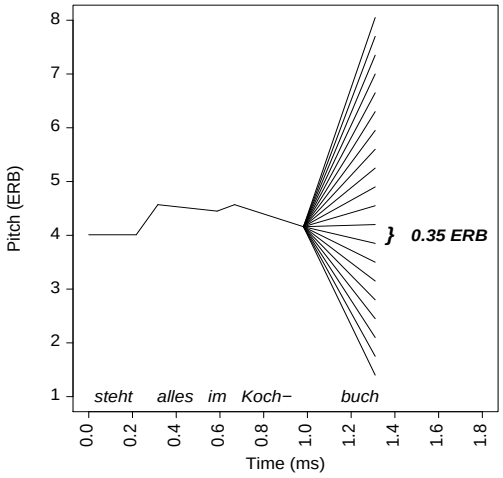

Figure 1: *Set of 20 stimuli used in the PME tests, differing in the phrase-final F_0 contour.*

Averaged over all sentences produced by this speaker, F_0 rises by 100 Hz to reach a typical H%, and falls by 50 Hz to reach a typical L%, which resulted in 228 Hz for H% and 78 Hz for L% in our test sentence. The 11-step continuum along the ERB scale with a step size of 0.35 ERB created for the CP experiment [2] was expanded to a 20-step continuum by producing stimuli with the same step size below and above the typical L% and H%, respectively. Further stimuli were produced as long as they sounded natural, which was evaluated by several listeners. In the new stimulus continuum the lowest boundary tone has an F_0 value of 35.3 Hz and the highest boundary tone has an F_0 value of 337.5 Hz. With this extension of the stimulus set we ensured that the presumed perceptual magnets of each boundary tone category were included. The F_0 contours were resynthesized by means of PSOLA [5] and the stimuli were numbered from 1 (lowest boundary tone) to 20 (highest boundary tone). Figure 1 shows a schematic illustration of the manipulation.

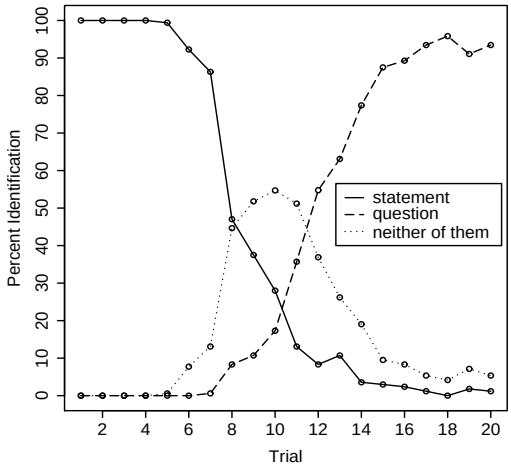

Figure 2: *Percent identification of all stimuli as a function of stimulus step number, averaged over listeners.*

2.2. Experimental procedures

Three subtests were created according to Kuhl [3] to test for the perceptual magnet effect, namely an identification test, a goodness rating task and a discrimination test. The tests had to be performed in this order because the results of the identification render the input stimuli for the goodness rating, and the results of the goodness rating render the input stimuli for the discrimination. Because of the high number of stimuli and because of the need to test both categories separately during the second and the third subtests, the subjects had to complete seven subtests in total (1 identification, 2 goodness ratings and 4 discriminations) with a break of at least one hour between any two subtests.

21 native German listeners, 10 males and 11 females, who were students or staff at the IMS, participated in the experiment on a voluntary basis. No subject reported any perceptual deficits. The experiment was performed individually. The subjects listened to the stimuli via headphones. The volume of the stimuli was set to a comfortable level at the beginning of each subtest. After listening to a stimulus the subjects had to choose one of the response alternatives before listening to the next one.

3. Experimental results

3.1. Identification

During identification listeners were asked to classify the stimuli as either *question* or *statement* or *neither question nor statement*. The third category should be used when the subject was not sure to which of the two main categories the stimulus belonged. There were 20 stimuli occurring 8 times each in the test in randomized order, i.e. subjects had to identify 160 stimuli in total.

The individual results diverge slightly. Most subjects used all three given categories and demonstrated that there are in fact cases in which they are unsure about the classification of the stimulus. Two subjects did not use the category *neither question nor statement*. They reported that they had always been sure to which category the stimulus belonged. As shown in Figure 2 the *statement* category was almost correctly identified and the *question* category was above 90%, whereas the category *neither question nor statement* had an identification rate of at most 55%. This seems to be counterevidence for the existence of a third category between the low and the high boundary tone in German, which was proposed in our previous paper [2]. Another observation is that the category *statement* has more constant identification rates within the category itself than the category *question*. This is in accordance with the fact that 4 subjects reported that they classified the stimuli 18 to 20 as belonging to the category *neither question nor statement* because these stimuli sounded unnatural, i.e. too high, for them.

Based on the results of the identification the two categories *statement* and *question* had to be isolated before starting the goodness rating task for each category. To ensure that only those stimuli are included in each of the two categories that are indeed members of the pertinent category, a stimulus was accepted for the goodness rating of its category if it was identified as a member of this category by more than 75% of the subjects. Two proportion tests verified that the stimuli 1 through 7 as well as the stimuli 14 through 20 did not differ significantly in their identification rates. According to these results the stimuli 1 through 7 and 14 through 20 were included in the goodness rating tasks for the categories *statement* and *question*, respectively.

3.2. Goodness rating

There were two goodness ratings, one for each category. The listeners were asked to label the quality of each stimulus within its category on a scale from 1 (*very bad exemplar of this category*) to 7 (*very good exemplar of this category*). Prior to both ratings there was a training session to acquaint the listeners with the range of the stimuli. There were 7 different stimuli within each category, labelled from A1 to A7 for *statement* and from F1 to F7 for *question*. They were repeated 10 times and presented in randomized order. This resulted in 70 stimuli for each goodness rating task.

For the *statement* category there were only slight individual differences in the rating of the stimuli. To get the prototype and the nonprototype for the *statement* category the stimulus with the highest and the stimulus with the lowest rating were calculated for each subject, resulting in two sets, one set of individual prototypes and the other of individual nonprototypes. The median of the first set corresponded to stimulus 2, i.e. this was the prototype of the *statement* category (P_S), and the median of the second set corresponded to stimulus 7, i.e. this was the nonprototype of the *statement* category (NP_S).

For the category *question* there were greater individual differences in the ratings of the stimuli than in the *statement* category. It was more difficult for all subjects to decide between good and not so good exemplars of this category than it was for the *statement* category. Four female subjects evaluated the highest stimuli as bad ones – they reported them to be unnaturally high – whereas all the other subjects rated these stimuli as the best ones. But as subjects had been asked to label the stimuli according to their individual impression, these four subjects were not excluded from further evaluation. The procedure to calculate the *question* prototype (P_Q) and the *question* nonprototype (NP_Q) was the same as for the *statement* category. For the category *question* the prototype corresponded to stimulus 6 and the nonprototype corresponded to stimulus 1. The difficulties in obtaining a clear prototype in the *question* category may be viewed as a first cue that there is no perceptual magnet in this category. This issue will be addressed in section 4 below.

3.3. Discrimination

In the discrimination test subjects listened to stimulus pairs that were either identical (AA) or different. In the latter case, the second stimulus was either higher (AB) or lower (BA) than the first one. One stimulus of each sentence pair was always either the prototype (P) or the nonprototype (NP), the other stimulus in the pair was one of the immediate neighbors of P or NP, respectively. The interstimulus interval within each pair was 500 ms [6] and each pair was repeated 7 times in the test. Discrimination was tested separately for each category. In addition there were two different test designs: first, the random design, in which the pairs including the prototype and those including the nonprototype were randomly mixed; second, the block design in which the pairs including P were tested separately from those including NP. Moreover, two subdesigns arose from the possibility that the order of presentation of these two blocks might affect the discrimination results, namely the P-first and the NP-first design. Every design included 126 different and 126 identical stimulus pairs for each category. Each subject participated in only one subdesign; the assignment of subjects was decided randomly. Because the number of stimuli was too large for one test session for each category, discrimination was split up into two tests for the *statement* and two tests for the *question* category. Again, a training session before each subtest served for the subjects to become acquainted with the stimuli and the differences within the pairs. During training listeners received feedback for their anwers, but not during the test.

3.3.1. Signal Detection Theory

Although the discrimination task involved different test designs, no order of presentation effect was found in the results. However, results of the individual subjects differed not only in their hit rates, i.e. how many pairs they correctly recognized as including different stimuli, but also in their false alarm rates, i.e. how many pairs they wrongly recognized as including different stimuli. This is in line with Signal Detection Theory (SDT) [7, 8]. 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. On any trial, the answer of the observer is YES if the evidence for the signal is larger than some value known as the *response criterion*, and NO when it is smaller than this value. Therefore the number of hits and false alarms depends on this criterion, known as λ_{Center}. For the description of the results we therefore have to take into account, first, the hit rate (*h*), i.e. the number of hits against the total number of really different stimulus pairs, second, the false-alarm rate (*f*), i.e. the number of false alarms against the total number of really identical stimulus pairs, and third, λ_{Center} with $\lambda_{Center} = (-f) - 0.5 * (h - f)$ [7].

3.3.2. Statement category

For the category *statement* the hit rates of the stimuli in the immediate vicinity of the prototype P_S differ significantly in their mean values from those in the immediate vicinity of the nonprototype NP_S: there are significantly more hits in the NP_S environment. Concerning λ_{Center} there is also a significant difference between the P_S and the NP_S vicinities. As shown in Figure 3, the values for λ_{Center} are significantly higher around P_S than around NP_S ($p \ll 0.0001$, $r = 0.763$), especially for the two immediate neighbors of P_S, i.e. P.min1 and P.plus1. Post-hoc tests evaluated this finding and confirmed that the immediate

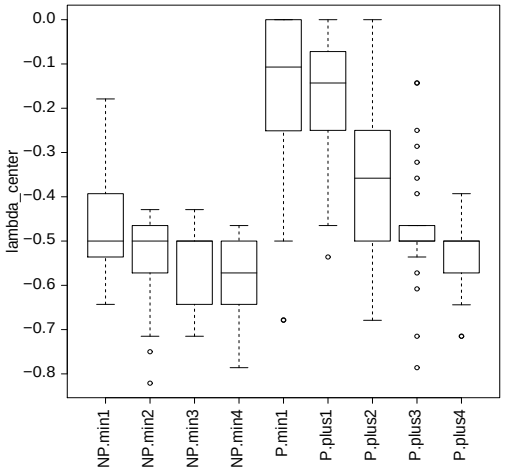

Figure 3: λ_{Center} *for the statement category differs significantly between NP and P trials.*

neighbors of P_S differ significantly in their values for λ_{Center} from all other stimuli in the surrounding of P_S or NP_S. Another correlation ($p \ll 0.0001$, $r = 0.425$) was found for the response criterion and the results of the goodness rating task for the P_S environment, with the response criterion increasing with better goodness rating values, i.e. the higher the goodness rating value the lower the hit rate for the environment of the prototype. So there is reduced discriminability in the surrounding of P_S but not in the surrounding of NP_S. This is exactly what PME assumes: perception is warped around P but not around NP. Thus, we found strong evidence for PME for the *statement* category in German.

3.3.3. Question category

For the *question* category there are no significant differences between the hit rates of the immediate environments of P_Q and NP_Q (Figure 4). The same holds for the response criterion λ_{Center}: both surroundings do not differ significantly in their λ_{Center} values. But as in the category *statement*, there is a correlation between λ_{Center} and the trials ($p \ll 0.0001$, $r = 0.721$). There are significant differences in the mean values for hit rates and λ_{Center} ($p = 0.005$) between the first and the second immediate neighbors of NP_Q, i.e between NP.plus1 and NP.plus2; between the two immediate P_Q neighbors (P.min1 and P.plus1) and the second P_Q neighbor (P.min2); and between the second (P.min2) and the third (P.min3) P_Q neighbors. The result indicates that the discrimination ability around the supposed prototype of this category is worse than around the nonprototype, but this difference is found only with the second neighbor of P_Q and NP_Q, respectively, and it is far from reaching significance. We conclude that there is no evidence for a magnet effect in the *question* category in German.

4. Discussion

The results of our experiments show that a perceptual magnet exists for the *statement* category in German. There were only slight differences between subjects in the results of the identification, the goodness rating and the discrimination for this cat-

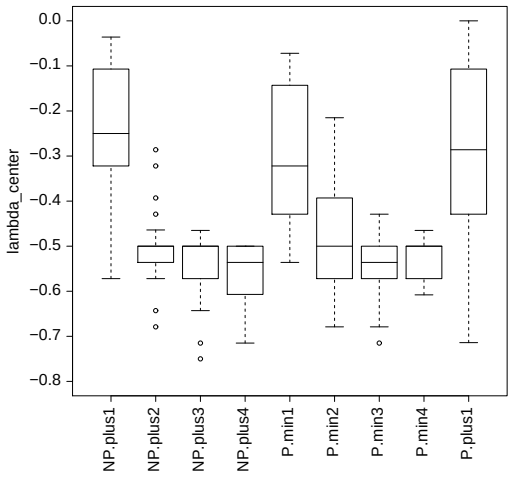

Figure 4: λ_{Center} *for the category question shows minor differences between NP and P trials.*

egory. We therefore conclude that the *statement* category with its low boundary tone is well established in German and has clear restrictions and contours. We successfully determined a prototype of this category which shows a very low F_0. As this prototype was almost the lowest one in the stimulus continuum we created and still was accepted as sounding naturally, we suppose that the most important cue for the perception of *statement* is its low boundary tone.

For the category *question* there are no such clear results. There was no evidence for a perceptual magnet in this category. This finding may be explained in different ways. First, there might be a third boundary tone called continuation (%). This boundary tone differs from L% in the height of F_0 and from H% only in being non-terminal. A high but non-terminal boundary tone signals that the speaker will continue with his turn, whereas a high and terminal boundary tone (H%) signals that the listener may take the turn. Therefore it might be the case that the better discrimination ability in the *question* category is a consequence of discriminating within a perceptual space where two boundary tones are present, possibly comprising two perceptual magnets. This assumption might explain why the discrimination values for the immediate NP_Q neighbor are as bad as for the immediate neighbors of P_Q: P_Q corresponds to the *question* prototype and NP_Q corresponds to the continuation prototype (and not to the nonprototype of the *question* category as we supposed). However, we did not find clear evidence for the existence of a third boundary tone category in German in our identification task.

Second, according to the fixed step size in ERB during stimulus creation the differences in Hz between the *question* stimuli are larger than those between the *statement* stimuli. In our earlier experiment [2] we had mentioned that the better discrimination ability within the *question* category might be a test artefact because the step size of 0.35 ERB might be too large for this category. To check this possibility one would have to create a new stimulus set with a smaller step size than we used here.

The third explanation uses arguments from exemplar theory. If each category emerges from all the exemplars the listener perceived, then the definition of a prototype of a category states that this is the place in the exemplar cloud with the highest exemplar density. Discrimination is almost impossible there because the exemplars are so close to each other. Precisely at this point goodness ratings are expected to be maximal. The goodness ratings for the *question* category demonstrate that several subjects do not accept the highest stimuli as good exemplars of this category. Maybe these subjects have developed different exemplar clouds for the category *question*. If so, they may not accept the highest stimuli as questions and use other, possibly nonlinguistic, criteria during discrimination.

5. Conclusions

In this experiment we demonstrated that the concept of the perceptual magnet effect can be applied to the perception of intonation. There is clear evidence for the existence of one perceptual magnet in the *statement* category in German, as perception is warped around the prototype of this category, which results in a low discrimination rate around the prototype of the *statement* category. Perception is not warped around the nonprototypes in this category, i.e. discrimination of different stimuli of this category is quite good. This result could not be repeated for the German *question* category: discrimination in this category is quite good in the entire stimulus set and perception is not warped around one specific point in the exemplar cloud. Therefore, we conclude that in German no prototype exists for the *question* category, but a clear prototype does exist for the *statement* category.

6. Acknowledgements

This experiment was carried out as part of the project *Target Oriented Production of Prosody* funded by the German Research Foundation (DFG, Grant DO 536/4–1).

7. References

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

[2] Schneider, K. and Lintfert, B., “Categorical perception of boundary tones in German”, Proc. 15th ICPhS, Barcelona, pp. 631–634, 2003.

[3] Kuhl, P. K., “Human adults and human infants show a ‘perceptual magnet effect’ for the prototypes of speech categories, monkeys do not”, Perception & Psychophysics, vol. 50 (2), pp. 93–107, 1991.

[4] Hermes, D. J. and van Gestel, J. C., “The frequency scale of speech intonation”, J. Acoust. Soc. Am., vol. 90, pp. 97–102, 1991.

[5] Moulines, E. and Charpentier, F., “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones”, Speech Communication, vol. 9, pp. 453–467, 1990.

[6] Batliner, A., “Wieviele Halbtöne braucht die Frage? Merkmale, Dimensionen, Kategorien”, in: Altmann, H., Batliner, A. and Oppenrieder, W. (eds.), Zur Intonation von Modus und Fokus im Deutschen, Tübingen: Niemeyer, pp. 111–162, 1989.

[7] Wickens, T. D., “Elementary Signal Detection Theory”, Oxford University Press, 2002.

[8] Heeger, D., “Signal Detection Theory”, http://www.cns.nyu.edu/˜david/sdt/sdt.html, 2003, last accessed 03/2005.