Reaction time and decision difficulty in the perception of intonation

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

[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]Department of Computational Linguistics and Phonetics, Saarland University, Germany
{katrin.schneider,grzegorz.dogil}@ims.uni-stuttgart.de, moebius@coli.uni-saarland.de

Abstract

An experiment was carried out to test the Categorical Perception as well as possible Perceptual Magnet Effects in the two boundary tone categories L% and H% in German, corresponding to *statement* vs. *question* interpretation, respectively. Additionally, reaction times (RT) were logged during all subtests to see if they support the results. Analyses revealed that RTs always increased with rising difficulty of the perceptual task, and decreased when the decision process was easy. Task-specific results showed that RT also correlated with the number of possible answers during a perceptual decision, i.e. more answer alternatives resulted in longer RT. Furthermore, female subjects generally reacted faster during all perceptual tasks, although this did not necessarily correlate with the accuracy of the results. Nevertheless, the results confirmed the usefulness of RT to support the analyses and the interpretation of perceptual data.

Index Terms: prosody perception, intonation, categorical perception, perceptual magnet effect, reaction time

1. Introduction

The use of reaction time (RT) in perception experiments is not new. In 1974, Pisoni and Tash [1] compared subjects' reaction times during a categorical perception test to investigate if RT licence conclusions about stimulus pairs, i.e. if pairs consisted of identical or of different stimuli. They used synthetic speech syllables ranging perceptually from /ba/ to /pa/, and adopted the reaction time matching paradigm developed by Posner and colleagues [2]. This design used RT measured for different types of discrimination or identification data to determine the level of analysis at which the comparisons were made. The time necessary to make the decision if two speech sounds are *identical* or *different* may reflect the level of the perceptual processing and therefore the type of information required for this decision. Pisoni and Tash assumed that judging two acoustically different speech sounds from the same category (*A–a pairs*) as being *identical*, "[...] involves a comparison of abstract phonetic features at a higher level of perceptual analysis than does classifying two acoustically identical stimuli as the *same*" ([1], p. 286). This additional stage of analysis might explain longer RTs when correctly discriminating *A–a pairs*, a finding not covered by the classical paradigm of Categorical Perception (CP) [3]. However, this process may fail if there is not enough time to complete it before making the decision, or if the stimuli are acoustically too similar, or if the listener has a high *response criterion* according to Signal Detection Theory (SDT) [4].

1.1. Reaction time, CP and PME

If different levels of processing are involved in the analysis and/or comparison of speech sounds, this should have an influence on RT in perception experiments testing for CP between two speech categories or for PME within a speech category. For both test designs a stimulus continuum is needed ranging from the first to the second category under examination. A CP test includes two subtasks: an identification and a discrimination test [3]. During identification, stimuli from the continuum have to be labeled as belonging to one of the proposed answer categories, and the identification function indicates where a category switch takes place. If the identification function corresponds to a steep curve, this is a first indication of the presence of CP. Accordingly, RT should be higher at the category switch than inside a category, because at the location of the switch identification is unclear and therefore processing should take longer. During discrimination, stimulus pairs have to be judged as consisting of identical or different stimuli. If the location of the discrimination maximum correlates with that of the category switch, CP is assumed to be present. According to [1] RT should be shorter when the acoustic difference between the stimuli is clearly perceptible, i.e. when stimuli from different categories are paired (*A–B pairs*). This is in line with CP, because CP hypothesizes that discrimination inside a category is difficult, and therefore RT should be longer, whereas discrimination between categories is easier, which should result in shorter RT.

A test for the Perceptual Magnet Effect (PME) evaluates stimulus differences within a category [5, 6]. Discrimination around a prototype (*P*), i.e. a stored item that best matches the main features defining the category, is harder than discrimination around a non-prototype (*NP*), i.e. a bad instance of the category. The perceptual space around *P* is warped: the category-internal acoustic differences between *P* and its neighbors are further reduced perceptually, which results in a particularly low discrimination performance around *P*. As neighbors of *P* are perceptually mapped onto *P*, RT for *A–a pairs* should be short, because the acoustic differences will be too small to be perceptible.

Batliner and Schiefer [7] reported that RT is longer in *A–a pairs* erroneously perceived as *identical* than in pairs of different stimuli that were perceived correctly. Moreover, RT was lower in *A–A pairs* correctly recognized as *identical* than when discrimination was incorrect. According to [1], listeners in such cases unnecessarily try to retrieve additional acoustic information which was then interpreted incorrectly.

Chen [8] found that during identification in a peak alignment and in a peak height continuum, participants' mean RT inside a category was shorter than mean RT at the category boundary. She concluded that these RT differences were essential properties of a real linguistical identification of the categories and that the RT peak at the identification boundary signaled difficulties arising in the decision process.

1.2. Reaction time and continuous perception

Massaro [9] suggested that the use of RT measures in conjunction with the CP paradigm might help decide whether speech perception effects are categorical or continuous. He argued that if an increase in RT corresponds to an increase of the ambiguity of the stimulus, perception is continuous rather than categorical. This implies that both the identification function and the corresponding RT function should vary accordingly, i.e. gradually from one end of the stimulus continuum to the other. However, if RT and the identification function show a significantly greater variation at a location in the continuum that corresponds to the category boundary, then CP is confirmed. This conclusion can be adopted in a discrimination task as well. If perception is continuous, then discrimination performance is comparably high at any point in the continuum, and the corresponding RT values should also be on a high level. However, if perception is discrete or categorical, RT should be on a comparable level inside each category, because there discrimination is difficult or impossible. At the category boundary, however, RT should decrease because here discrimination has to be performed between stimuli of different categories, which should be a much easier task.

Such experimental evidence and theoretical considerations suggest that RT might be an indicator of the complexity of a perceptual task. A simple task, such as identifying a stimulus that clearly belongs to one category or discriminating stimuli from different categories, leads to short RT, while a complex task, such as identifying a stimulus that lies perceptually between two categories or discriminating stimuli differing only in their acoustic features but not in their category membership, involves additional processes resulting in longer RT. However, even though these considerations are not new, RT measurements have rarely been included and analyzed in perception experiments in the domain of prosody [10, 8, 7].

Therefore, in the experiments testing for CP and PME in German boundary tones presented here, RTs were logged for all subtests and included in the statistical analyses. We hypothesize that identification within a category will lead to relatively short RT and identification at or near the category boundary, i.e. when the stimulus is ambiguous, will result in longer RT. During PME discrimination, RT should decrease when the acoustic difference between the stimuli is increased. However, as *P* warps the perceived distances, this RT increase should be faster in the *NP* compared to the *P* region. During CP discrimination, RT should be shorter when the paired stimuli are taken from different categories, and longer otherwise.

2. Method

A perception experiment was carried out to verify CP of the low (L%) and the high (H%) boundary tones in German, corresponding to an interpretation of *statement* vs. YES-NO-*question*, respectively. Additionally, it was examined if these two boundary tone categories have developed a PME [5, 6]. In all subtasks, RT was logged to investigate correlations between the perceptual data and RT.

2.1. Experimental design

To test for CP and/or PME in the two boundary tones in German, a stimulus continuum was created ranging from L% to H%. This was done by manipulating only one of the three main acoustic correlates of prosody, viz. the fundamental frequency (F_0) contour bearing the boundary tone. In German the position of the main verb in a sentence may bias the perception towards one of the possible interpretations (here *statement* vs. *question*). Therefore, the syntactically ambiguous phrase "nach Panama" (to Panama) spoken by a male native speaker of German was used as the test phrase. Nine stimulus steps between L% and H% were created using a step width of 0.35 ERB (the definition of the ERB scale can be found in Hermes and van Gestel [11]), and 4 steps below L% and 5 steps above H% were added to test for PME. These 20 stimuli were then resynthesized using PSOLA [12], representing an enlarged continuum from L% to H%.

Three subtests were carried out: identification, goodness rating, and discrimination. In the identification task the stimuli had to be labeled as tokens of one of three proposed answer alternatives: *statement*, *question*, or, if it was undecidable, *neither-nor*. In the goodness rating, the quality of each stimulus, i.e. how well it fitted in the proposed category, was rated on a scale from 1 ("very bad") to 9 ("very good"). This was done separately for each category. During discrimination, pairs of stimuli had to be evaluated as to whether they consisted of *identical* or *different* stimuli. 26 native German subjects without hearing deficits participated in all three subtests.

2.2. Measuring reaction time

Reaction time was logged for all participants in each subtest and the RT measurement always started immediately after the stimulus presentation. Although the instruction was to answer as fast as possible, some RTs were longer than 10 seconds. Participants reported that they were interrupted or needed a break at this point. As it is obviously impossible to correctly remember the presented stimulus after several seconds, all results were excluded whose RT values were more than two standard deviations away from the mean. As RT varied slightly between the subtests, this procedure was adopted separately to each subtest. After this elimination step approximately 95% of the data remained for further analyses.

3. Results

3.1. Reaction times in the identification task

RT differed significantly between the three response alternatives in the identification ($p < 0.0001$). A Waller-Duncan posthoc test revealed that the *neither-nor* answers had significantly higher RT values than the other two answer categories. As the *neither-nor* category did not reach an average identification of at least 50%, it did not appear to represent any boundary tone category; rather, it seemed to correspond to the crossover between the L% and H% categories. RT was significantly longer because the subjects were unsure about how to label the stimulus when its perceptual position was between L% and H%. This was observed in the individual results as well as for all subjects pooled (Figure 1, dashed-dotted line). Interestingly, female subjects reacted significantly faster than male subjects.

3.2. Reaction times in the goodness rating task

All stimuli that received an average identification higher than 60% were included in the goodness rating test of their category. Significant correlations were found between the average rating value of a stimulus and its RT ($p < 0.0001$), and Waller-Duncan post hoc tests revealed that for the L% and the H% category RT correlated with the goodness rating of the stimulus within a category: the higher the rating, the shorter RT, and vice

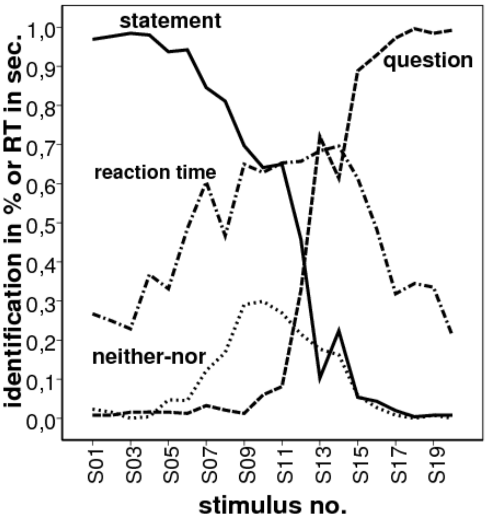

Figure 1: *Identification results and averaged RT values (in sec.).*

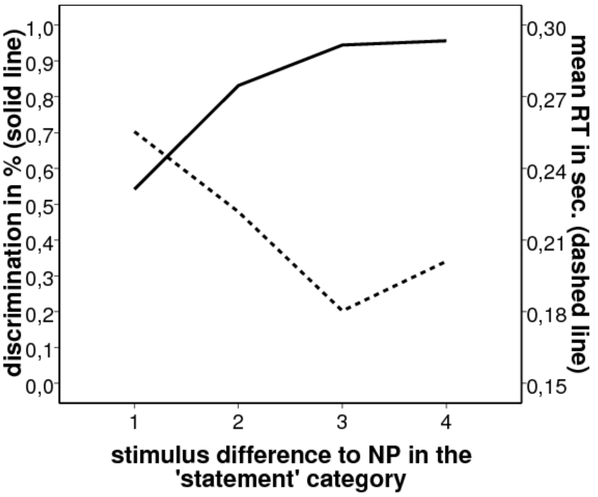

Figure 2: *Discrimination vs. RT in the NP region of the L% category.*

versa. Female subjects were faster in their responses than the male participants, although these differences did not reach significance. This finding was more pronounced for the L% than for the H% category.

3.3. Reaction times in the CP discrimination design

In the CP paradigm for the two boundary tones, the discrimination task involved pairs of immediately adjacent stimuli of the underlying stimulus continuum. These pairs had to be judged as to whether they consisted of *identical* or of *different* stimuli [3]. Discrimination performance was best between pairs 8–9 and 10–11 of the underlying stimulus continuum. RT was shorter in these pairs than in most of the other stimulus pairs, except for the pairs 2–3 and 3–4. The low RT values in these two pairs correlate with low discrimination performances, suggesting that these pairs were not perceived as consisting of identical rather than different stimuli. The RT values of pairs consisting of identical stimuli were as low as those for pairs consisting of different stimuli and with a maximum in the discrimination performance. However, a significant negative correlation ($p < 0.0001$) was found between RT and discrimination performance, i.e. RT was short when discrimination performance was high.

Moreover, RT was longest during discrimination inside the H% category, longer than around the discrimination maximum, but also longer than when discriminating inside the L% category. The results revealed that all subjects showed a comparably short RT at the discrimination peak, whereas inside each category females were generally faster in their decisions, even though they reached a lower total discrimination performance.

7 out of 26 participants had a discrimination performance below 50% pooled for all stimulus pairs. Therefore, we decided to separately analyze the results of the "high" performing group, i.e. subjects with a discrimination peak of at least 50%, and the "low" performing group, i.e. the 7 participants with the low discrimination results. Although the average RT did not differ between these groups, we found a negative correlation between the discrimination results and RT for the "high" performing group, but a positive correlation for the "low" performing group. In the "high" performing group, RT decreased at the discrimination maximum around pair 9–10, while in the "low" performing group RT was longest at the discrimination peak (only 31% discrimination performance).

3.4. Reaction times in the PME discrimination design

When testing for a perceptual magnet effect (PME) inside a category, discrimination around the prototype *P* and the non-prototype *NP* have to be compared [5, 6].

During discrimination in the L% category, RT differed significantly between the $P_{L\%}$ vs. the $NP_{L\%}$ region (first neighbor: $p = 0.028$, third neighbor: $p < 0.01$). As illustrated in Figure 2, in the $NP_{L\%}$ region, an increasing stimulus difference to $NP_{L\%}$ resulted in shorter RT. The slight increase in RT from stimulus difference 3 to 4 is non-significant and seems to result from RT variation within the standard deviation range. Contrary to this finding, RT in the $P_{L\%}$ region was shortest when $P_{L\%}$ was paired with its immediate neighbor or with its neighbor four steps away in the stimulus continuum, but longest when $P_{L\%}$ had to be discriminated from its second or third neighbor. Females discriminated significantly faster around $P_{L\%}$ ($p < 0.05$) than males. However, there was no significant difference in RT around $NP_{L\%}$. When separating the two groups of participants, the "high" performing group had the longest RT when discriminating $P_{L\%}$ from its second neighbor, while the "low" performing group had the longest RT when discriminating $P_{L\%}$ from its third neighbor. For the $NP_{L\%}$ region RT of both groups were similarly short for all four $NP_{L\%}$ neighbors.

When discriminating inside the H% category, there were no significant differences between both regions, neither in the discrimination results nor in RT. Increasing the acoustic difference between either $P_{H\%}$ or $NP_{H\%}$ and its paired neighbors improved discrimination performance and decreased RT. Females discriminated significantly faster than males in the $P_{H\%}$ region ($p < 0.05$), but not in the $NP_{H\%}$ region. Furthermore, RT in the $P_{H\%}$ and the $NP_{H\%}$ region was longer for the "high" performers than the "low" performers, but this difference was not significant. Interestingly, RT of the "high" performing group clearly decreased with rising acoustic difference of the stimulus to $P_{H\%}$ or $NP_{H\%}$, while RT of the "low" performing subjects increased minimally for the same pairs.

4. Discussion

During identification, a longer RT correlated significantly with the *neither-nor* answer. As this answer alternative had no clear linguistic definition, the subjects may have labeled all stimuli with an uncertain identification as *neither-nor*. This is evidence for the hypothesis that uncertainty leads to higher RT, as found in previous studies [1, 7, 8, 9, 10]. Interestingly, female subjects identified the stimuli produced by the (male) test voice significantly faster than male subjects. Maybe females generally make faster decisions in perceptual tasks. However, this finding may be restricted to speech perception. Testing such hypotheses is beyond the scope of this paper.

In the goodness rating tasks, higher ratings correlated with shorter RT, which further supports the hypothesis that RT reflects decision difficulty. Again, females reacted faster than males, and although the difference was not significant, this finding further supports the idea that females might be generally faster in speech perception. Furthermore, goodness rating appears to have been more difficult than identification, maybe because subjects had to select 1 out of 9 answers in the rating and only 1 out of 3 answers in the identification. Comparing average RT of all subtests we found that RT values were lowest in the discrimination task, followed by those in the identification task, and the longest average RT was observed in the two goodness rating tasks. We conclude that the number of answer alternatives may influence the speed of the decision process.

In discrimination, RT significantly correlated with the perceptual results. RT was shortest when discrimination performance was low (0–20%) or high (80–100%). Thus, irrespective of correctness, when subjects were certain they responded quickly, whereas the response was delayed when they were unsure. This correlation was more explicit for the female than for the male participants, which supports the idea that RT is also a gender-specific variable in speech perception. However, further tests are required to verify this finding.

Higher average RT in the *question* category suggests that this prosodic category may be less clearly defined than the *statement* category. This observation is additionally supported by the results of the PME discrimination task. In the *L%* category, RT clearly correlated with the discrimination performance, viz. RT decreased with rising performance. Thus, the RT results further support the existence of a PME in the *L%* category, because they varied with the discrimination performance around $P_{L\%}$ as compared to $NP_{L\%}$. In the *H%* category, there were no significant differences between the discrimination around $P_{H\%}$ that around $NP_{H\%}$, neither in the discrimination performance nor in RT. This finding was mainly caused by the opposite behavior of two groups of participants. For the "high" performers, better discrimination performance correlated with shorter RT, whereas for the "low" performers RT minimally increased with rising discrimination performance. These findings do not contradict the hypothesis that RT reflects the perceptual difficulty of the task. The discrimination performance as well as the RT values of the "low" performers were low throughout the test. The slight increase of RT with rising discrimination performance may be explained as follows: with rising stimulus difference the "low" performers received minimal perceptual cues that the stimuli from the same category are not identical. This lack of cues may have triggered an additional processing of acoustic features (cf. [1]). However, the results of these processes were mostly not sufficient to differentiate between the stimuli (cf. SDT [4]). Therefore, discrimination performance improved only slightly while RT increased as well.

5. Conclusions

The analysis of reaction time data obtained during our perception experiments confirm that reaction times can be useful as an indicator of the simplicity or difficulty of a perceptual decision. The easier a perceptual decision is, the shorter RT will be, and the longer RT, the more difficult the decision must have been. However, this correlation between RT and the perceptual results is not always statistically significant, as some of the subtests show. One reason for non-significance may be the pooling of data from all participants despite strong individual and group differences in perceptual performance and RT. Further experiments should consider a normalization of participants' RT, and they should examine gender differences more closely. Even so, the RT data in our experiments support the conclusion from our discrimination data that prosodic categories have an internal structure which may influence the speed or ease of perceptual processes. These results show that RT should be measured during all perceptual tasks, because RT can indicate where processing problems occur and which data are potentially unreliable.

6. Acknowledgements

We wish to thank Doug Whalen for his advise on how to eliminate experimental data with problematic RT values in a principled way.

7. References

[1] Pisoni, D.B. and Tash, J.,"Reaction times to comparisons within and across phonetic categories", Perception & Psychophysics, 15(2):285–290, 1974.

[2] Posner, M.I. and Mitchell, R.,"Cronometric analysis of classification", Psychological Review, 74:392–409, 1967.

[3] Repp, B., "Categorical perception: Issues, methods, and findings", in N.J. Lass [Ed], Speech and Language: Advances in Basic Research and Practice (10), 243–335, Ac. Press NY, 1984.

[4] Wickens, T.D., "Elementary Signal Detection Theory", OUP, Oxford, UK, 2002.

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

[6] Kuhl, P.K. and Iverson, P., "Linguistic experience and the 'Perceptual Magnet Effect' ", in W. Strange [Ed], Speech perception and linguistic experience: Issues in cross-language research, 121–154, MD: York Press, 1995.

[7] Batliner, A. and Schiefer, L., "Stimulus Category, Reaction Time, and Order Effect - An Experiment on Pitch Discrimination", Proc. ICPhS (3):46-49, 1987.

[8] Chen, A., "Reaction time as an indicator to discrete intonational contrasts in English", Proc. Eurospeech, 97–100, 2003.

[9] Massaro, D.W., "Speech perception by ear and eye: a paradigm for psychological inquiry", Hillsdale, NJ, London, 1987.

[10] Prieto, P., "Experimental methods and paradigms for prosodic analysis", in: A.C. Cohn, C. Fougeron, M.K. Huffman [Eds], Handbook in Laboratory Phonology, OUP, Oxford, in press.

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

[12] Moulines, E. and Charpentier, F., "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", Speech Communication (9), 453–467, 1990.