Detecting categorical perception in continuous discrimination data

Paul Boersma, Kateřina Chládková

Amsterdam Center for Language and Communication, University of Amsterdam, The Netherlands
paul.boersma@uva.nl, k.chladkova@uva.nl

Abstract

We present a method for assessing categorical perception from continuous discrimination data. Until recently, categorical perception of speech has exclusively been measured by discrimination and identification experiments with a small number of repeatedly presented stimuli. Experiments by Rogers and Davis [1] have shown that using non-repeating stimuli along a densely-sampled phonetic continuum yields a more reliable measure of categorization. However, no analysis method has been proposed that would preserve the continuous nature of the obtained discrimination data. In the present study, we describe a method of analysis that can be applied to continuous discrimination data without having to discretize the raw data at any time during the analysis.

Index Terms: categorical perception, continuous stimuli, discrimination

1. Introduction

In speech perception research, categorical perception of vowels and consonants has been assessed by experiments that involve identification and discrimination tasks. In an identification task, sounds that belong to the same category receive the same label. Identification has mostly been tested by means of a multiple-forced choice experiment in which listeners label each stimulus as one of the phonemes of their native (or second) language.

In categorical perception, discrimination of sounds across a category boundary is easier than discrimination of sounds within a category. To test the discrimination of speech sounds, various laboratory tasks have been designed and utilized; among these are the AX ("same"–"different") task, in which listeners indicate whether the sounds of a pair are the same or different, the AXB task, in which listeners identify the second sound of a stimulus triplet either with the first sound or with the last sound, or the 4IAX (four-interval "same"–"different"; or ABAA) task, in which listeners have to indicate whether the first or the second pair of a stimulus quadruplet contained a deviant sound; see [2] for a review.

Discrimination experiments reported in the vast majority of previous studies have used a relatively small number of stimuli that were repeated multiple times within a single experiment; e.g. 7 tokens in [3], 15 tokens in [4], or 14 in [5]. However, a recent study has shown that "stimulus repetition reduces discrimination of within-category differences, and enhances between-category discrimination" [1, p.379]. In their study, Rogers and Davis [1] compared the results of a discrimination task with 2 stimulus pairs repeated 416 times to the results of a discrimination task with 96 stimulus pairs repeated 16 times, and found that numerous repetitions of a small number of stimuli increase the 'categorical bias'. To obtain a reliable measure of listeners' categorical perception, discrimination experiments should thus be designed with a large number of non-repeating stimuli created along a densely sampled phonetic continuum. In such a 'continuous' design, then, a plausible method of analysis should conform to the continuity of the obtained data. This was, however, not the case in Rogers and Davis' study. In the data analysis they report, the continuous discrimination data are collapsed into several bins (namely, three) and the continuum thus becomes sparsely sampled. Analogously to their paper, we use a continuous experimental design, creating a discrimination experiment with 260 different stimuli sampled along a single phonetic continuum. Improving on their paper, however, we introduce a method that preserves the continuous nature of the obtained discrimination data throughout the analysis.

2. The experiment

In this section we report an actual perception experiment, which addresses discrimination within the continuum between [i] and [ɛ]. In section 3 we try to infer categorical perception along this continuum on the sole basis of the discrimination data obtained here. We did not elicit identification data, because the experiment was a part of a larger experiment that included continua on which the participants' language had no categories. That larger experiment, which has a research question on feature generalization, will be reported elsewhere; the subject of the present paper is only the analysis method.

2.1. Stimuli

The stimuli were vowels along an F1 continuum synthesized with the Klatt synthesizer built into the program Praat [6]. The vowels all had the same F2 value, namely 2700 Hz, and the same F3 value of 3300 Hz. Along the F1 continuum, which ranged from 280 Hz (6.93 erb) to 725 Hz (12.86 erb), we synthesized 260 vowel tokens, that is, 130 stimulus pairs. The F1 distance between the two vowels *within* a stimulus pair was 0.9 erb, and the F1 distance *between* two neighboring stimulus pairs was much smaller, namely 0.039 erb, thus maximizing the degree of continuity of the stimulus set. Both the within-pair F1 distance and the between-pairs F1 distance were kept the same for all the 130 stimulus pairs along the continuum.

Figure 1: *The 130 stimulus pairs. Each pair consists of two points along the horizontal axis, connected here by an arc. The distance between the members of a pair is constant, i.e.* $s_{12} - s_{11} = s_{22} - s_{21}$

2.2. Procedure

Vowel discrimination was tested by means of a traditional AX task. The inter-stimulus interval (i.e. the time interval between the two members of a pair) was 500 ms, and the trial-initial silence (i.e. the time interval between the participant's mouse click and the first member of the next pair) was 600 ms. Each of the 130 stimulus pairs occurred twice, that is, in one trial the pair member with the lower F1 was played first, while in the other trial with the same pair the member with the higher F1 was played first; this was to factor out any stimulus-order

effects that have been reported in previous vowel discrimination experiments [7]. The complete set of 260 pairs of stimuli was presented in random order.

As described above, the two members of a stimulus pair were never identical, and in fact the auditory distance between the two members of a pair was the same for every trial. Despite the fact that the two sounds were always different, we asked the listeners to indicate whether the sounds were different or the same. The F1 difference between the sounds was as small as 0.9 erb, i.e., about the size of a just noticeable difference for F1 [8]; in a pilot experiment, this turned out to be just small enough to generally make the number of "same" judgments of the same order of magnitude as the number of "different" judgments.

In line with the definition of categorical speech perception, our listeners (whose language has at least two segmental phonemes along the presented vowel continuum) were expected to perceive stimulus pairs in some regions of the F1 continuum as different (i.e., stimuli across a category boundary) and stimulus pairs in other regions of the F1 continuum as identical (i.e., stimuli that lie within one category). We can find categorical perception if our listeners have more "different" responses for stimulus pairs in some regions along the vowel continuum than for stimulus pairs in other regions. The location of the category boundary will lie between the sounds that elicit the largest number of "different" responses.

2.3. Participants

The subjects in this experiment were 62 monolingual Czech speakers. They were university or high-school students between 18 and 29 years of age. They were paid a fixed hourly rate for their participation. In the present paper, which only addresses the analysis method, we discuss only three of these participants; we choose these three people because they seem to reflect the three most common strategies found among the 62 listeners.

3. Analyzing a listener

The analysis of the data of a single listener runs as follows. The listener is confronted with N (here: 130) different stimulus pairs. The nth stimulus pair (n = 1..N) is repeated K_n (here: always 2) times. Of these K_n replications, the listener judges a pair as "same" s_n times, and as "different" d_n times, with $s_n + d_n = K_n$.

Figure 2 shows the raw data of three listeners, all of whom gave at least 0 and at most 2 "different" responses for every stimulus pair. Since the visualization of the raw data by poles is not very informative with respect to where the discrimination peaks lie, the Figure also shows smoothed versions of the data, obtained by convolving the raw data with a unit-area Gaussian with a standard deviation of 10; an edge correction is obtained by dividing the resulting curve by the convolution of that same Gaussian with data consisting of all ones [9]. The smoothed curves suggest that participant 1 has a constant probability of judging "different", that participant 2 has a single discrimination peak around stimulus pair 49, and that participant 3 could have discrimination peaks around stimulus pairs 53 and 113. Whether these visual suggestions are correct, e.g. whether the small right-hand bump of listener 2 is indeed irrelevant and the taller right-hand bump of listener 3 is not caused by random variation, remains to be seen. The following three subsections therefore submit these data to several maximum-likelihood analyses, each of which corresponds to a different model of what the listener is doing.

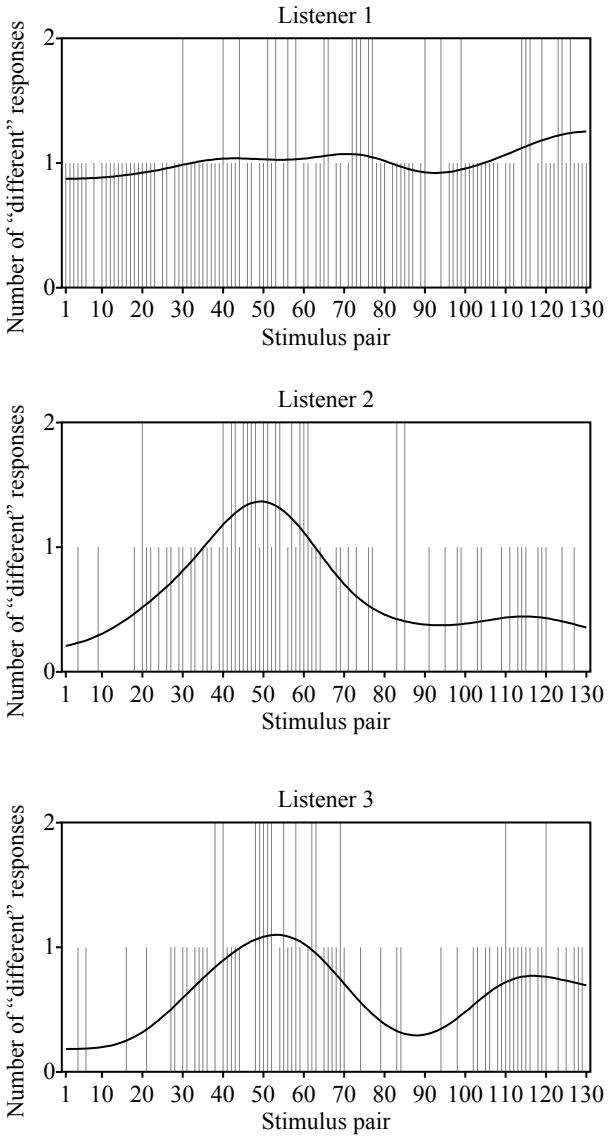

Figure 2: *Raw data (grey poles) and smoothed data (solid curves) of three participants with apparently zero, one, and two discrimination peaks, respectively*

3.1. First model: no discrimination peaks

Our first, simplest, model assumes that the listener has no categorical perception along the continuum but instead only has an acoustic discrimination strategy. Since we used constant distances along the auditorily uniform erb scale, an ideal acoustic listener has a constant probability p_{const} of judging any stimulus pair as "different". In other words, the probability p_n that the nth stimulus pair is judged as "different" is simply

$$p_n = p_{const} \quad (1)$$

Although an estimate of the parameter p_{const} could simply be computed by dividing the total number of "different" judgments by the total number of trials (260), we here provide a more general method for estimating p_{const}, which can also be used for more complicated formulas for p_n, as we do in sections 3.2 and 3.3.

The likelihood of the data, given the values of p_n, d_n and s_n from the experiment, is

$$L = \prod_{n=1}^{N} p_n^{\,d_n} \left(1 - p_n\right)^{s_n} \tag{2}$$

The logarithm of this is the "log-likelihood"

$$LL = \sum_{n=1}^{N} \left(d_n \ln p_n + s_n \ln\left(1 - p_n\right) \right) \tag{3}$$

We now want to find the value of p_n that maximizes *LL*. We initially assign to the parameter p_{const} a random value between 0 and 1 and subsequently add small positive or negative values to it, always checking whether *LL* improves (becomes less negative) according to formulas (1) and (3). Whenever *LL* improves, we keep the changed p_{const} as our new best value of p_{const}, and we subsequently start again from this new value. After many iterations, in which the changes get exponentially smaller, we arrive at *the* best value of p_{const}. For listener 1 it is 0.508, for listener 2 it is 0.319, and for listener 3 it is 0.304. The top row of Figure 3 shows these values, together with the best *LL* values obtained. The optimized p_{const} values are indeed identical to the overall fraction of "different" responses. The Figure suggests that the fit is good for listener 1 but not for listeners 2 and 3.

3.2. Second model: one discrimination peak

Our second model assumes that the listener mixes an acoustic discrimination strategy with a categorical perception strategy based on the existence of two categories along the continuum. We assume, therefore, that the probability of a "different" judgment shows one peak somewhere along the continuum:

$$p_n = p_- + \left(p_+ - p_-\right) e^{-\frac{(n-\mu)^2}{2\sigma^2}} \tag{4}$$

under the condition that p_+ is always greater than p_-. We again start with random values of the four parameters p_-, p_+, μ, and σ, and randomly change these parameters so as to increase the value of *LL* according to (4) and (3). Since this procedure can arrive in a local optimum, it is repeated 100 times in order to find the "best best *LL*". The results are shown in the middle row of Figure 3. The Figure shows both the fitted p_n itself and its smoothed version, which ought to be

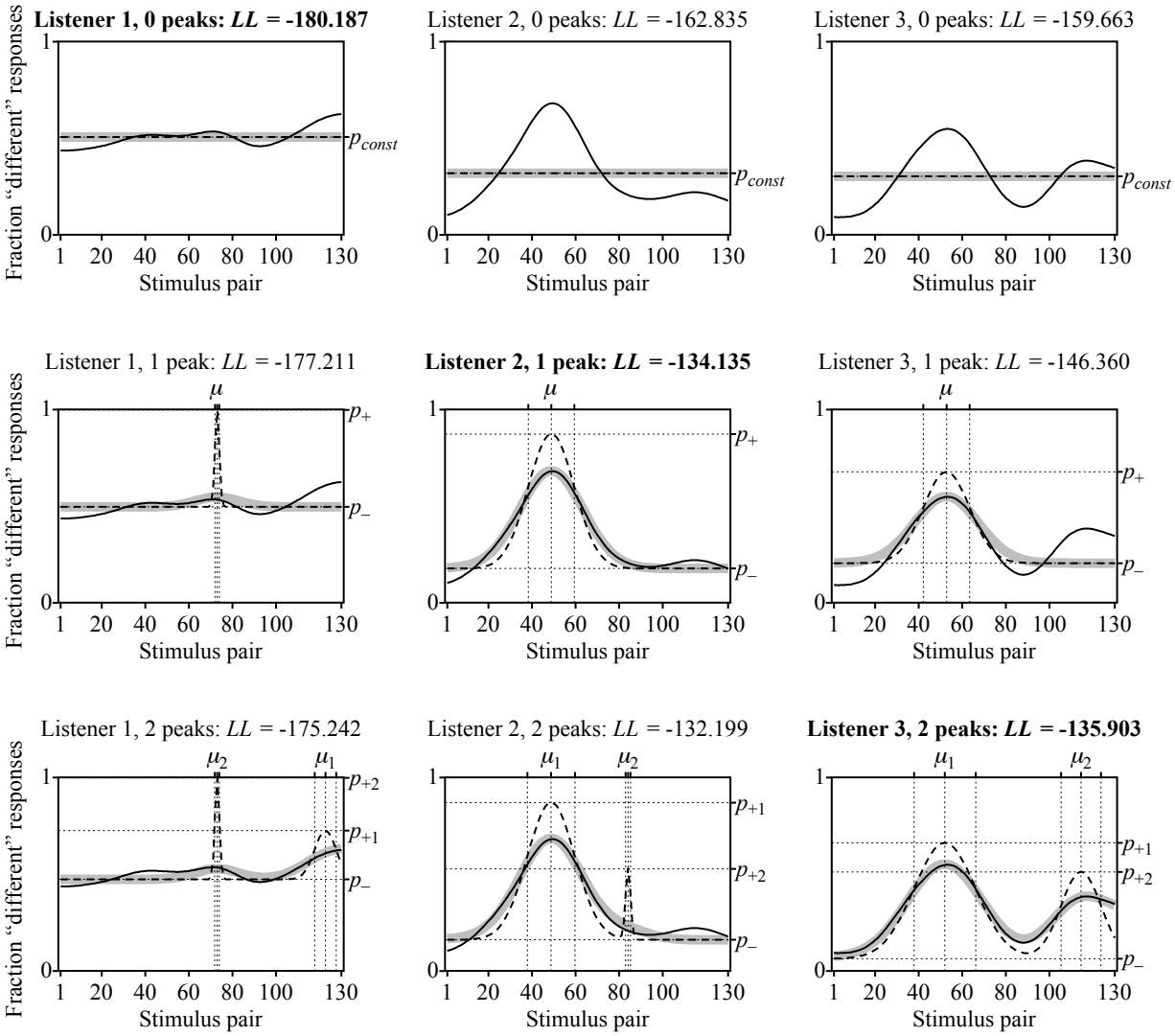

Figure 3: *Maximum-likelihood fitting of three listeners, each with zero, one, and two peaks. Solid curves: smoothed data (copied from Figure 2). Dashed curve: fit (unlabelled vertical dotted lines: $\mu \pm \sigma$). Thick grey curve: smoothed fit.*

close to the smoothed data. We see that visually, the smoothed fit for listener 2 is indeed very close to her smoothed data.

3.3. Third model: two discrimination peaks

Our third model assumes that the listener has three categories along the continuum, and therefore two discrimination peaks:

$$p_n = p_- + \left(p_{+1} - p_-\right) e^{-\frac{\left(n-\mu_1\right)^2}{2\sigma_1^2}} + \left(p_{+2} - p_-\right) e^{-\frac{\left(n-\mu_2\right)^2}{2\sigma_2^2}} \quad (5)$$

When we optimize the seven parameters in such a way that *LL* is maximized, we obtain the bottom row of Figure 3. For each listener, the smoothed fit is now close to her smoothed data.

3.4. Comparison of the three models

Instead of judging visually how an increase in the number of model parameters improves the fit or not, we should ask the question: does the likelihood rise significantly with each addition of parameters? The table below summarizes the values of *LL* for the three listeners, together with ΔLL, the increase in *LL* from the next simpler model. In accordance with what is common practice in the use of logistic regression, the *p* values in the table are derived from performing a χ^2 test on $-2\Delta LL$ (with 3 degrees of freedom, which is the number of parameters added to the model with each peak).

Model	Listener 1	Listener 2	Listener 3
No peaks	**−180.187**	**−162.835**	**−159.663**
One peak	−177.210	**−134.135**	**−146.360**
improvement	+2.977	**+28.700**	**+13.303**
p	0.11	**$2.1 \cdot 10^{-12}$**	**$7.1 \cdot 10^{-6}$**
Two peaks	−175.242	−132.199	**−135.903**
improvement	+1.968	+1.936	**+10.457**
p	0.27	0.28	**0.00011**
Three peaks	−174.798	−131.987	−135.671
improvement	+0.444	+0.212	+0.232
p	0.83	0.94	0.93

Table 1: *Development of log-likelihood as a function of the number of modelled distribution peaks. Bold = statistically significant improvement.*

We see that the data of listener 1 show no evidence for any discrimination peak, i.e. that they are consistent with the idea that she listens acoustically (with a probability p_{const} of hearing the difference) or that she has only one category (with a bias p_{const} toward responding "different"); a mix of these two strategies is also possible. The data of listener 2 prove that she has at least one discrimination peak, i.e. at least two categories (again, p_- reflects the success of acoustic listening and/or a bias toward responding "different"); there is no evidence for more categories than two. The data of listener 3 prove that she has at least two discrimination peaks, i.e. at least three categories; there is no evidence for more.

4. Discussion

Rogers and Davis [1] have found that discrimination tasks with a large number of different non-repeating stimuli are a more reliable measure of categorical perception than tasks with a small number of repeating stimuli. Along with that finding comes the need for a method of analysis suitable for such continuous discrimination data; devising such a method was our aim in this study. We presented a method that, unlike Rogers and Davis' own analysis method, preserves the continuity (i.e. dense sampling) of the raw data throughout the analysis, and that thereby contributes to the reliability of any claims about categorical perception made on the basis of continuous data. We illustrated how the method works on continuous discrimination data of three real listeners.

The present method fits the obtained discrimination function with several models that assume different numbers of discrimination peaks. Given that a peak in the discrimination function corresponds to a category boundary [5], this method determines a plausible (or at least minimum) number of categories along the stimulus continuum. The method also determines the locations and crispnesses of the boundaries. Of course one cannot divide the 62 listeners into three groups solely on the basis of the *p* values in Table 1 (one cannot prove that a listener does not have more peaks). Such a division may require adding latent variables to the model.

The method of data analysis that we described is based on the search for peaks. Note, however, that this method of assessing categorical perception differs from the methods used previously not only in its continuous nature. In the literature, categorical perception has been assessed by comparing the obtained and the expected discrimination that was computed from identification results [5]; such a method is not feasible once listeners have no phonemic representations (i.e. no labels) for the stimuli. Thus, next to the main purpose of the present study, which was to provide a method for *analyzing continuous discrimination data*, we therefore also showed how categorical perception can be assessed solely on the basis of obtained continuous discrimination data, i.e. without reference to identification results.

5. References

[1] Rogers, J. C. and Davis, M. H. (2009). Categorical perception of speech without stimulus repetition. In *Proceedings of Interspeech 2009*, 376–379.

[2] Gerrits, E. and Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception and Psychophysics*, 66(3):363–376.

[3] Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, 13(2):253–260.

[4] Schouten, M. E. H. and van Hessen, A. J. (1992). Modelling phoneme perception. I: Categorical perception. *JASA*, 92(4):1841–1855.

[5] Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1995). The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychology: HPP*, 54(3):358–368.

[6] Boersma, P. and Weenink, D. (1992–2010). Praat: doing phonetics by computer [Computer program]. Version 5.1.30, retrieved 1 April 2010 from http://www.praat.org/.

[7] Polka, L. and Bohn, O.-S. (2003). Asymmetries in vowel perception. *Speech Communication*, 41:221–231.

[8] Kewley-Port, D. (1995). Thresholds for formant frequency discrimination of vowels in consonantal context. *JASA*, 97(5):485–496.

[9] Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, University of Amsterdam, 17:97–110.