Perception & Psychophysics
2004, 66 (3), 363-376

Categorical perception depends on the discrimination task

E. GERRITS and M. E. H. SCHOUTEN
Utrecht University, Utrecht, The Netherlands

Speech sounds are said to be perceived *categorically*. This notion is usually operationalized as the extent to which discrimination of stimuli is predictable from phoneme classification of the same stimuli. In this article, vowel continua were presented to listeners in a four-interval discrimination task (2IFC with flankers, or 4I2AFC) and a classification task. The results showed that there was no indication of categorical perception at all, since observed discrimination was found not to be predictable from the classification data. Variation in design, such as different step sizes or longer interstimulus intervals, did not affect this outcome, but a 2IFC experiment (without flankers, or 2I2AFC) involving the same stimuli elicited the traditional categorical results. These results indicate that the four-interval task made it difficult for listeners to use phonetic information and, hence, that categorical perception may be a function of the type of task used for discrimination.

Traditionally, research into the mental representation of phonetic categories has focused on the relationship between discrimination and classification of speech sounds on a stimulus continuum. The first such experiment was performed by Liberman, Harris, Hoffman, and Griffith (1957). Their hypothesis was that discrimination of certain speech sounds would be limited by classification; two different stimuli would be discriminated only to the extent that they were classified differently (this was later referred to as *categorical perception*; Eimas, 1963). Liberman et al. concluded that their results did not agree with their own "extreme assumption": Discrimination results were better than predicted from the classification results. The difference presumably represented the listener's ability to distinguish the speech sounds not solely on the basis of the phonemic labels, but also on the basis of acoustic differences. Despite this conclusion, however, this first study is often cited as typically demonstrating categorical perception (for a review, see Repp, 1984). And even though a clear relationship between discrimination and classification has rarely been demonstrated in subsequent research, the results are often interpreted as indicating absolutely categorical perception (Macmillan, 1987). In other words, there is no explicit criterion for the maximum difference between discrimination and classification results that would still be compatible with categorical perception. In this article, we will use the original definition by Liberman et al.: Perception is fully categorical only if there is no significant difference between phoneme categorization (i.e., predicted discrimination) and actually measured discrimination. In all other cases, one can talk only about various degrees of categorical perception. A great deal depends, of course, on the model used to derive the prediction; more about this will be said below.

This article is about the effects of different discrimination tasks on categorical perception. But that is not what we set out to do. Our original intention was to investigate the perception of vowels and to try to find an answer to the question of why vowels are generally perceived much less categorically than, for example, stop consonants (see, e.g., Fry, Abramson, Eimas, & Liberman, 1962; Repp, 1981). The results of Experiment 1, however, did not provide an answer to this question, but did raise questions about the various aspects of the experimental task that had been used, thus changing the focus of the research and of this article. Since the design of Experiment 1 was, of course, determined completely by the original hypothesis, this hypothesis will be briefly introduced, before we shift attention to the discrimination task.

The Original Hypothesis: Categorical Perception of Vowels?

There is a clear difference in degree of categorical perception between stop consonants and vowels. An explanation proposed by Pisoni (1973, 1975) and Tartter (1981) is that this may be due to differences in cue duration. The essential acoustic cues for stop consonants are rapidly changing formant transitions and a brief noise burst (Liberman et al., 1957; Tartter, 1982; see also Sawusch, Nusbaum, & Schwab, 1980, Experiment 3). In contrast, vowels are assumed to remain uniform over a much longer duration (Delattre, Liberman, Cooper, & Gerstman, 1952). This difference in cue duration has an effect on the availability of auditory memory for these two classes of speech sounds.

We thank Sieb Nooteboom for his thoughtful comments and suggestions. We are grateful to Theo Veenker for programming help. We also express our appreciation to Aad Houtsma, Cecile Kuijpers, and Frank Wijnen. Correspondence concerning this article should be addressed to M. E. H. Schouten, Utrecht Institute of Linguistics OTS, Utrecht University, Trans 10, 3512 JK Utrecht, The Netherlands (e-mail: bert.schouten@let.uu.nl).

Note—This article was accepted by the previous editorial team, headed by Neil Macmillan.

According to Pisoni (1973) and Fujisaki and Kawashima (1971), discrimination may be performed in an auditory mode or in a labeling mode (our terms). The auditory (or psychoacoustic) mode is supposed to use only *bottom-up* stimulus information, whereas in the labeling (or phonetic) mode a subject first categorizes the stimuli and then bases the discrimination decision on the categories to which he has assigned the stimuli. The short cue duration for consonants is responsible for the inferior performance of the auditory mode. Presumably, the decay of rapidly changing acoustic information is too fast to make an auditory comparison of consonantal stimuli possible, with the result that discrimination is performed in a phonetic mode. This is not the case for vowels, which are, consequently, perceived much less categorically. Discussing the results of Pisoni (1973), Schouten and Van Hessen (1992) suggested that this low degree of categorical perception of vowels could be due to the nature of the stimulus material. Up to then, the vowels used as stimulus material had been modeled on productions in isolated words. When produced in isolated words, vowels are lengthened (hyperarticulation). We hypothesized that, in running speech, temporal reduction and more complex spectral coding of the vowel would make vowel perception more categorical. To test this hypothesis, we intended to study the difference in perception between vowels spoken in isolated words and in a text read at a fast rate. However, as will be shown in the Results section, we obtained no categorical perception in either condition, so the original hypothesis could not be confirmed. There was no visible relationship between observed and predicted discrimination, and observed discrimination was actually worse than would be predicted from classification. The focus of our interest then shifted to the question of why we had failed to find even a hint of categorical perception.

The New Hypothesis: The Effect of the Discrimination Task on Categorical Perception

Great care had been taken over the choice of a discrimination task that would test the original hypothesis. If at all possible, we wanted to select a task that would leave a subject free to use both auditory and phonetic information and would not encourage the use of one type of information at the expense of the other—that is, would not "push" the subjects into a particular mode. We therefore examined the available tasks, speculating about subject behavior in each of them. In Experiment 1, only one task was used, so only one of these speculations about subject behavior could be tested.

The ABX and AXB Tasks

The prototypical discrimination test used for assessing categorical perception is the ABX task, in which each trial consists of three intervals and a subject has to decide whether stimulus X in the third interval is the same as A in the first interval or B in the second interval (Liberman et al., 1957). In view of the relatively short time span of auditory memory (200–300 msec), however, the rather high degree of categorical perception often found with the ABX task may, according to Massaro and Cohen (1983), reflect the exclusive use of phonetic memory. Subjects may try to remember both the auditory memory traces and the labels assigned to the A and B sounds. By the time X is presented, these auditory traces may have faded away. If they have, the subjects must rely on the labels they have assigned to A and B. This strategy may produce results indicating a high degree of categorical perception. Moreover, as B. Schouten, Gerrits, and Van Hessen (2003), using a signal detection analysis, have shown, the ABX task is subject to a very strong bias toward the response "$B = X$." In theory, this need not worry us, since a signal detection analysis should provide us with a clear separation between sensitivity and bias, but in practice, the greater the bias, the less likely the conditions for such an analysis are to be met.

The use of a variant of the ABX task, AXB, in which the second stimulus is identical to either the first or the third one and is close in time to both, has yielded contradictory results. Van Hessen and Schouten (1999) reported that their subjects often ignored the third stimulus, thus annulling the expected advantage over ABX. Gerrits (2001, pp. 42–49), however, found considerable differences between AXB and AX discrimination.

The 2I2AFC Task

Similar problems hold for another paradigm: two-interval two-alternative forced-choice (2I2AFC) discrimination. In the 2I2AFC paradigm, the stimuli are always different, and the subject has to determine the order in which they are presented (AB or BA). This makes it necessary to explain to the subjects what the term *order* means and even to mention the phoneme categories in the instructions, at the risk of encouraging labeling behavior (Schouten & Van Hessen, 1992). Response bias is much smaller than that in the ABX task; it favors a response that says "the first stimulus is closer to the phoneme prototype than the second" (B. Schouten et al., 2003).

The AX Task

To avoid strategies that rely exclusively on labeling, we need a task that reduces the load on auditory memory. Such a task could be AX (same–different) discrimination. In an AX discrimination experiment, the subject has to determine whether the two stimuli in a trial are the same or different. A disadvantage of this paradigm is that subjects may decide to respond "different" only if they are very sure of their decision. This means that AX may be strongly biased: A subject's response could be completely dominated by a subjective phoneme-based criterion, very close to one end of a scale between *same* and *different*.

The 4IAX and 4I2AFC Tasks

A discrimination test that is regarded as sensitive to auditory differences between speech stimuli is the 4IAX task (Pisoni, 1975). In the 4IAX test, two pairs of stimuli are presented on every trial; one pair is the same, and one pair is different (e.g., AB–AA, AA–BA, or BA–BB), and the

subject has to decide which pair contains the odd one out. The 4IAX task is assumed to be more sensitive to purely auditory cues, since a correct decision can be largely based on bottom-up auditory information and is thought not to be subject to strong top-down skewing by subjective criteria, such as phoneme boundaries. However, we decided not to use a 4IAX task, but a task in which important aspects of the 2I2AFC and the 4IAX tasks are combined: the 4I2AFC task (see, e.g., Heller & Trahiotis, 1995; Trahiotis & Bernstein 1990). In this task, a subject is expected to be free to use both auditory processing and phoneme labeling. The *A* and *B* stimuli are presented randomly in the two possible orders AABA or ABAA, with a 50% a priori probability. Stimulus *A* at the beginning and end of each quadruplet functions as a reference. The subject has to decide whether the odd one out occurred in the second or in the third interval. The flanking stimuli are there to make direct auditory comparisons of the stimuli easier, and they may also make a low-bias 4IAX type of strategy possible, in which two differences are compared. On the other hand, the 2I2AFC aspect (order detection: AB or BA, leaving the two flanking A stimuli out of consideration) may encourage labeling. This task, therefore, should be a useful diagnostic instrument for determining whether and to what extent categorical perception occurs and was the focus of interest in Experiment I. The aim of this article is to find out whether the 4I2AFC task really does allow both auditory and phonetic processing.

EXPERIMENT 1

Method

Stimuli. The stimulus material consisted of two continua of eight vowel stimuli ranging from /u/ to /i/ in a /pVp/ context. The vowels /u/ and /i/ were selected because it was expected that the differences between the two speech conditions would be greater with these two vowels than with most other vowels, due to the relatively long articulatory trajectories required to reach them. The first step in stimulus generation was recording the vowels /u/ and /i/ in the meaningful words /pup/ and /pip/ produced both in isolation at a rather slow rate and in a text that was read aloud at a rapid speech rate by a male native speaker of Dutch. The speaker was instructed to read the text three times at an increasing speech rate. The third recording was selected, since it was the most rapidly read version, measured as the total amount of reading time for the whole text.

In each of the two conditions, there were nine repetitions by the speaker of the /pup/ and /pip/ words. Five phonetically untrained listeners identified 30-msec segments from these vowels in an open-set identification task (12 monophthongs). The vowels spoken in isolated words, or *word vowels*, were significantly more often identified as /u/ and /i/ than were the vowels from the fast text, or *text vowels* (65% vs. 32% for /u/ and 42% vs. 7% for /i/). Other frequently used response categories were /o/ for /u/ and /y/ for /i/. All the words were rated on a 7-point acceptability scale by a listening panel that consisted of five phoneticians. The word pairs that were used as endpoints in the two stimulus continua were selected on the basis of acceptability and of matching vowel duration within a pair.

The acoustic differences between the word and the text vowels were determined with analyses of duration and formant frequency. The text vowels were temporally reduced, as compared with the more carefully articulated word vowels. The duration of the word /u/ was 90 msec; its steady-state component was 60 msec. The duration of the text /u/ was 70 msec, with a steady-state component of 30 msec, a reduction of 22% and 50%, respectively. The duration of the isolated word /i/ was 90 msec; its steady-state component was 50 msec. And the duration of the text /i/ was 70 msec, with a steady-state component of 15 msec, a reduction of 22% and 70%, respectively. This temporal reduction is comparable to the reduction reported in the Dutch studies by Schouten and Pols (1979) and Van Son and Pols (1990). The temporal vowel reduction found by Schouten and Pols was 28%. The reduction of steady-state segment duration was, on average, 38%. Van Son and Pols found a temporal reduction of 15% between vowels in a text that was read at a normal speech rate and those in a text read as fast as possible.

An analysis of the formant frequencies of the vowels in the two speech conditions was also attempted. Since formant extraction failed with the text vowels (no second formant for /u/ could be found, nor a third formant for /i/), it was impossible to quantify the degree of spectral reduction. The absence of these formants, however, suggests some loss of spectral detail in the text vowels.

The stimuli in the continua between the original utterances were obtained by interpolation between the relative amplitudes of the spectral envelopes of the vowels. The interpolation method had been used successfully in studies on categorical perception by Schouten and Van Hessen (1992) and Van Hessen and Schouten (1992). This method was preferred to working in the formant domain: Since no second formant could be defined for text /u/, interpolating between the formants of the text vowels was impossible. Moreover, in this way, we avoided the risk that, after having listened for a while to stimuli in which only one or two parameters were varied, some subjects would learn to attend selectively to those parameters. The experimental design was intended to motivate the listeners to focus on the speech signal as a whole. (Van Hessen & Schouten, 1999, have shown that there is an increase in categorical perception as synthesis quality improves from a simple synthesis by rule, via linear predictive coding (LPC) synthesis, to the more complex method used in the present study.) The importance of stimuli in which more than one parameter is varied was also mentioned by Liberman (1996), who predicted that, with proper synthesis, when the acoustic signal changes in all relevant aspects and not just one cue is varied, the discrimination functions will come much closer to being perfectly categorical.

The first step in the interpolation method was an analysis of the spectral envelopes of the original vowels in terms of the phases and amplitudes of up to 70 spectral components between 80 and 5000 Hz, depending on spectral density. Before interpolation, the signal was split into a source spectrum and a filter spectrum by means of cepstral deconvolution. The spectral envelopes of the eight stimuli, obtained by means of seven linear interpolation steps between each of the 70 pairs of spectral components, were then reconvolved with the original source spectrum of the /u/. The interpolation was always done in overlapping 25.6-msec time frames over the full length of the vowel (frame shift was 6.4 msec). Parameters such as *F*0, duration, and voice quality remained constant. For more details of this procedure, see Schouten and Van Hessen (1992) or Van Hessen (1992).

Stimulus generation resulted in two continua of eight stimuli that sounded completely natural and convincingly like utterances from the original speaker. In each continuum, the initial /p/ and the final /p/ of the stimuli were copied from the original word /pup/. In a pilot experiment, the stimuli of the two continua were identified (open set) by a listening panel that consisted of five phoneticians, all well-trained listeners. The listeners' identification responses were always /u/ or /i/. In none of the cases were the stimuli identified as the Dutch vowel /y/, which might have been expected, because the *F*2 of this central vowel lies between those of /u/ and /i/. The absence of an intermediate /y/ is a result of the interpolation method, since it moves from one vowel to another by progressively lowering and raising spectral peaks. Formant peaks that do not occur in either endpoint cannot occur in any of the interpolated stimuli.

The fundamental frequency and duration of the stimuli were the same as those of the original /pup/. In the word vowel continuum, *F*0 was 120 Hz, and stimulus duration was 215 msec (vowel, 90 msec; steady state, 60 msec). The duration of the stimuli in the text vowel continuum was 187 msec (vowel, 70 msec, steady state, 30 msec), and *F*0 was 125 Hz.

Interstimulus interval. Since auditory memory is time dependent, it was important to make a considered decision about the interstimulus interval (ISI). If this interval exceeds the life span of auditory memory, all that will be left of the first stimulus will be a representation coding its relationship to the other stimuli in the experiment, to preestablished categories, or to both (Pisoni, 1973). Massaro (1972a, 1972b, 1974) tried to determine the time a sound pattern is held in some unanalyzed form. His results indicated that a processing time of approximately 250 msec is sufficient for recognition of a speech signal. These results were in agreement with those of Dorman, Kewley-Port, Brady and Turvey (1977) and Plomp (1964).

Cowan and Morse (1986), Pisoni (1973), and Van Hessen and Schouten (1992) tested the effect of varying the ISI on discrimination performance for speech sounds. On the basis of Massaro's (1974) results, within-category discrimination should decrease with an increasing interval, reflecting the fading of the memory trace. This is in agreement with the results of Van Hessen and Schouten (1992), who found a decrease of within-category stop-consonant discrimination with increases in ISI from 100 to 300 msec. The between-category results of the discrimination studies by Cowan and Morse (1986), Pisoni (1973), and Van Hessen and Schouten (1992) confirmed the notion that processing of the auditory signal is not terminated after 100–200 msec. Their results indicated that when listeners use a labeling strategy to compare stimuli, discrimination improves as an effect of increasing ISI: Discrimination performance increases rapidly between 100 and 500 msec, reaches a maximum between 500 and 1,000 msec and falls gradually as the ISI increases further.

On the basis of these results, we assumed that, after an ISI of more than 100–200 msec, labeling processes would take over from direct auditory comparison. We did not want to use an interval shorter than 200 msec, since this might have increased the chance of mutual masking among the stimuli. In line with Massaro (1974), Pisoni (1975), and the within-category results of Van Hessen and Schouten (1992), we therefore decided to use 200-msec intervals, hoping that the required information would be available for direct auditory comparison of successive stimuli in a trial (200 msec was, of course, a nominal interval, indicating only the silences between successive stimuli; the vowels were additionally separated by the durations of two instances of /p/, which added a total of 117 and 125 msec to the intervals between the text and the word vowels, respectively).

Subjects. The subjects were 19 students at the Faculty of Arts at Utrecht University. They had no known hearing deficits and were all native speakers of Dutch. They were paid a fixed hourly rate.

Design. Experiment 1 consisted of six tests, three for each of two vowel continua, involving the same subjects. The tests were fixed discrimination, roving discrimination, and classification. The subjects took the tests in a fixed order, counterbalancing fixed and roving discrimination and the word and text vowel continua across subjects, but the classification tests were always performed after all the discrimination tests.

The discrimination used in the experiment was the 4I2AFC task (AABA/ABAA; the subjects had to indicate whether Stimulus *B* occurred in the second or the third interval). The stimuli in the second and third intervals always differed by one step along the continuum; the number of comparisons was, therefore, seven. The intertrial interval was determined by the response time. The ISI within a trial was 200 msec.

In the fixed-discrimination experiment, only one stimulus pair (*A* and *B*) was presented during a block of trials. The fixed-discrimination test consisted of seven blocks, one for each of the stimulus pairs, which were clearly separated from each other. Each block contained 64 trials, 32 for each of the two possible combinations, AABA and ABAA. The order of blocks was randomized for the fixed discrimination experiment. In the roving-discrimination experiment, the A and B stimuli to be discriminated were drawn randomly from the total range of stimuli and, thus, varied from trial to trial. In the roving-discrimination test, 7×64 trials were presented.

In the classification test, all eight stimuli had to be identified 64 times in a random order. Classification involved a forced choice between two alternatives, the vowels /u/ and /i/. There was no response time limit.

Procedure. The stimuli were presented to the subjects over headphones in a sound-treated booth. In the discrimination tests, it was stressed that differences between the stimuli would be small and, in most cases, could be detected only by listening carefully to all details of a stimulus. No phoneme labels were mentioned in the instructions, but the subjects were told that three of the four stimuli were going to be identical and that the oddball was either the second or the third one. They responded by mouse clicking on one of two response fields (labeled "2" and "3") on a computer screen. After the response had been made, visual feedback of the correct answer was given, so that the subject was able to judge and possibly improve his or her performance. Discrimination training consisted of 128 trials (responses to which were not stored) and was intended to familiarize the subjects with their task. In the fixed-discrimination context, the first 10 trials of every block were considered practice and were not included in the data analysis.

In classification, one stimulus was played on each trial, and the subject had to identify it by mouse clicking on a response field labeled "oe" or "ie" (/u/ or /i/). The only training consisted of 16 trials, two repetitions of the eight stimuli in the continuum, presented randomly.

Results

The results of Experiment 1 are displayed in Figures 1–3 and in Table 1. Figures 1 and 2 show the classification and discrimination data for the word vowels and the text vowels, respectively. The data displayed in the figures represent the averages of 19 subjects' individual d' scores (the individual subjects' data can be found on the World-Wide Web: www.let.uu.nl/~bert.schouten/personal/gerrits.htm). The numbers (n) along the abscissa refer to stimulus pairs, consisting of the (n) and ($n + 1$) stimuli; n is, therefore, a number between 1 and 7. The d' score at Stimulus Pair 6, for example, represents the discrimination of Stimulus 6 and Stimulus 7. The stimuli in Pair 1 resemble /u/, and the stimuli in Pair 7 sound like /i/.

The discrimination d' scores were calculated by subtracting z(FA) from z(H), with $\sigma = \sqrt{2}$ (Macmillan & Creelman, 1991, p. 121). The results of the classification test are presented as predicted discrimination scores. The transformation of the classification data into predicted discrimination was done as follows. For each pair of stimuli A and B and response alternatives /u/ and /i/, the proportion of /u/-responses to stimulus A (position n) was regarded as an estimate of p(H), and the proportion of /u/-responses to stimulus B (position $n + 1$) was taken as an estimate of p(FA). The classification d' values were determined by subtracting z(FA) from z(H). The values of p(H) and p(FA) were limited to the .99–.01 range, which

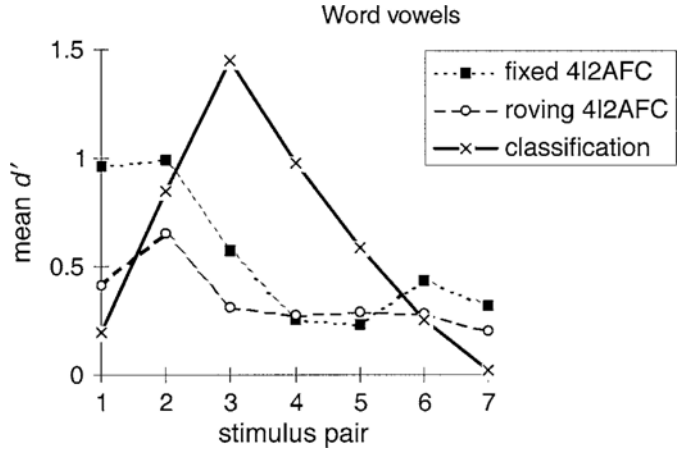

Figure 1. Predicted discrimination (classification) and actually measured discrimination (fixed and roving) for the word vowels. 4I2AFC, four-interval two-alternative forced choice.

meant that the maximum d' values that could be obtained were $4.65 * 0.5\sqrt{2} = 3.29$ for discrimination and 4.65 for classification.

Figures 1–3 show that there was not much difference between the two vowel conditions. We expected the data in Figure 1 to be less categorical than those in Figure 2 and, hence, the d' difference scores in Figure 3 to be much higher for word vowels than for text vowels. This clearly was not the case: The word vowels were not perceived less categorically than the text vowels. Neither figure shows anything like the expected relationship between observed and predicted discrimination, so we can conclude that there is no indication of categorical perception for either of the two vowel conditions.

The results of an analysis of variance (ANOVA) confirmed what is shown by the figures. Fixed independent variables were task (three levels), vowel condition (two levels), and stimulus (seven levels, nested under vowel condition). Cell variance was over 19 subjects. Performance was not significantly affected by vowel condition [$F(1,41) = 0.12, p = .731$]. The effect of task was significant [$F(2,41) = 14.62, p < .01$]. There was also a significant effect of stimulus and a significant interaction between task and stimulus [$F(6,41) = 5.58, p < .01$, and $F(12,41) = 4.56, p < .01$, respectively]. A Newman–Keuls test on the task factor revealed a significant difference between the means for roving discrimination and classification [$F(795,2) = 12.45, p < .01$]. A Newman–Keuls test on the stimulus factor (word vowel continuum) showed a significant peak for fixed discrimination at Stimulus Pairs 1 and 2 and a significant peak for classification at Pair 3. A similar test on the data from the text vowel continuum revealed a significant peak in the classification curve at Stimulus Pairs 3, 4, and 5.

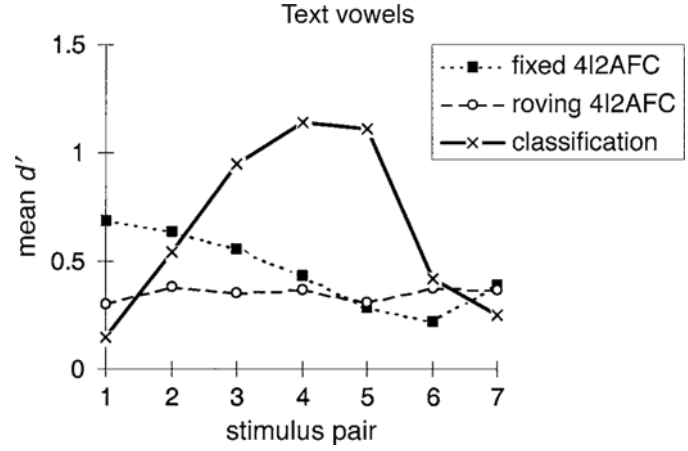

Figure 2. Predicted discrimination (classification) and actually measured discrimination (fixed and roving) for the text vowels. 4I2AFC, four-interval two-alternative forced choice.

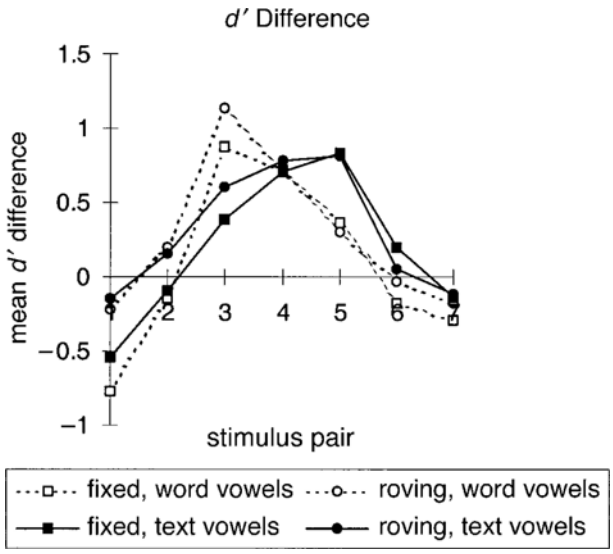

Figure 3. Difference scores between the obtained and the predicted data for each stimulus pair in Figures 1 and 2. The dashed lines represent the difference scores for the word vowels; the solid lines represent the difference scores for the text vowels.

The two vowel conditions were compared in two different analyses: a calculation of difference scores and a calculation of a categorical perception (CP) index, a measure of the degree of categorical perception. In Figure 3, the two vowel conditions are compared by calculating the difference scores between obtained discrimination and discrimination predicted by the classification data for each stimulus pair (as in Pisoni, 1975, but here in terms of averaged individual d' differences). If text vowels are perceived more categorically than word vowels, we should expect a smaller difference between the obtained and the predicted functions for the former condition than for the latter (if perception is fully categorical, the difference scores should be zero). A series of paired-samples sign tests on the d' difference scores between classification and discrimination in Figure 3 revealed that there were only two stimulus pairs with a significant effect of vowel condition: Stimulus Pairs 3 and 5. At Stimulus Pair 3, the difference between obtained and predicted discrimination of word vowels was higher than it was for text vowels, and at Stimulus Pair 5 the opposite occurred: The difference between obtained and predicted discrimination of text vowels was higher than that for the word vowels.

Table 1
Mean Categorical Perception-Indices and Standard Errors of the Mean for Four-Interval Two-Alternative Forced-Choice (4I2AFC) Discrimination of Word Vowels and Text Vowels

	Vowels			
	Fixed		Roving	
Task	*M*	*SE*	*M*	*SE*
	Word Vowels			
4I2AFC	−7	8	1	9
	Text Vowels			
4I2AFC	6	9	1	10

Difference scores estimate the difference in degree of categorical perception between word and text vowels. However, in an ANOVA or sign test, only means are compared, and not the shapes of the various functions. Therefore, another criterion for degree of categorical perception has been proposed, the so-called CP index introduced by Van Hessen and Schouten (1999). Van Hessen and Schouten (1999) have shown that the CP index can be used to estimate differences in categorical perception between various stimulus synthesis modes. They calculated the CP index as follows:

$$\mathrm{CP} = \frac{r}{1 + 0.2 \times \mathrm{abs}\left[d'(\mathrm{class}) - d'(\mathrm{discr})\right]} \times 100. \qquad (1)$$

In this equation, *CP* is the degree of categorical perception ranging from 0 to 100 (or to −100, in the case of negative correlations), r is the coefficient of correlation between classification and discrimination, and the denominator contains a term determined by the averaged absolute differences between the data points of the classification and the discrimination functions, multiplied by a constant chosen in such a way that the full range can be used. This equation expresses the degree of categorical perception as a function of the resemblance (numerator) and proximity (denominator) of the two functions. In Table 1, the CP indices, averaged over the 19 subjects, are shown for the word vowels versus the text vowels.

Our hypothesis that word vowels are perceived less categorically than text vowels would predict that the CP index of the word vowels should be lower than that of the text vowels. This was not really the case: All CP indices were effectively zero. Across the board, the CP indices presented the same picture as the difference scores: There was no indication of categorical perception for either of the two vowel conditions.

Discussion

The lack of any clear correlation between the discrimination and the classification results is counterintuitive to anyone who has ever spent any time comparing classification and discrimination within a categorical perception context and makes it impossible to evaluate the original hypothesis about categorical perception of vowels. Moreover, not only was there no evidence of the expected relationship, but also the subjects did not even manage to detect any differences between stimuli to which, during classification, they gave, fairly consistently, different phonetic labels. There appears to be something about the 4I2AFC discrimination task that prevents subjects from using phonetic information. If this is true, it could point to two different perceptual strategies: an auditory comparison of stimulus information during discrimination and phonetic labeling during classification. In other words, during discrimination, listeners are in an *auditory mode* and, during classification, in a *phonetic mode*. Such a conclusion would be almost the exact opposite of our predic-

tion that the 4I2AFC task would make both auditory and phonetic information equally available. Before allowing ourselves to draw such a far-reaching conclusion, however, we decided to take a closer look at two aspects of the discrimination task that could have affected the subjects' behavior decisively—namely, physical distance between the stimuli in a trial and ISI. Both were very small, and this could have made the task unusually difficult, so in Experiment 2, we ran two control conditions: one with a larger distance and one with a longer interval. In another condition in Experiment 2, we replicated the discrimination test with another task, 2I2AFC discrimination (i.e., one without flanking stimuli). Finally, we returned to the results of Experiment 1 in order to examine differences between subjects.

EXPERIMENT 2

Task

The reason for choosing the 4I2AFC task was to avoid an exclusive reliance on phoneme-labeling processes during discrimination. The results seem to indicate that we succeeded too well: Discrimination deviated from classification in a way that suggested an exclusive reliance on auditory coding. One way of testing this interpretation would be to compare the present discrimination results with discrimination of the same stimuli in a task in which phonetic-labeling processes are known to play an important role, such as the 2I2AFC task. In this task, the stimuli are always different, and the subject has to determine the order in which they are presented. This makes it necessary to explain to subjects what the term *order* means by mentioning the phoneme categories in the instructions (e.g., "ba–da" or "da–ba"). If our interpretation is correct, we should now obtain the "regular" relationship between classification and discrimination, in which discrimination can be predicted to some extent by, and does not fall below, classification.

Interstimulus interval

In Experiment 1, we had, on the basis of Massaro's (1974) study of auditory storage time, decided to use an interval of 200 msec, in which stimulus information should be available for direct auditory comparison. As a result, the listeners may have been discouraged from using a labeling strategy during discrimination, because there may not have been enough time between the successive stimuli to synthesize the acoustic information into phonetic percepts. This interpretation is suggested by data from Cowan and Morse (1986), Pisoni (1973), and Van Hessen and Schouten (1992), showing that discrimination performance based on labeling reaches a maximum after between 500 and 1,000 msec; if the interpretation is correct, it could explain the lack of a clear relationship between classification and discrimination in our data.

In addition, although the auditory image of a stimulus is usually held to fade away within 200–300 msec, this has only ever been tested in two-interval presentations. When auditory information has to be extracted from four stimuli presented in rapid succession, listeners might need more processing time than they do in the masking task Massaro (1974) used in his experiments. In other words, auditory analysis and comparison of four successive speech sounds may require more time than 200 msec each, and the arrival of new stimuli may disrupt processing of the previous stimuli.

In Experiment 2, we expanded the interval from 200 to 500 msec. If this is enough time to allow listeners to use auditory information for discrimination, as a result of reduced interference between the four stimuli within a trial, performance should increase independently of classification. In addition, the listener should have more time to assign labels to the stimuli (Van Hessen & Schouten, 1992), and discrimination performance should increase differentially as a function of classification (predicted discrimination), which means that stimuli with different labels should now be easier to discriminate than stimuli within one label category and, thus, perception should be more categorical.

Physical distance between the stimuli

The physical distance between the stimuli in a trial was only one step along the stimulus continuum, because we expected discrimination across two steps to be too easy and, hence, to result in a ceiling effect. According to Trahiotis (Heller & Trahiotis, 1995; Trahiotis & Bernstein, 1990), the flanking stimuli in a 4I2AFC task should have a facilitating effect; in his experiments, d' was much higher in 4I2AFC than in 2I2AFC tasks. However, our discrimination results did not seem to show any facilitation, being as low as they were (although a comparison with the 2I2AFC task was not carried out until our Experiment 2; see below). It could be that facilitation occurs only when stimuli are reasonably discriminable. This was tested with a step size of two between the stimuli.

Method

Stimuli. The stimuli used in Experiment 2 were those from the word vowel continuum in Experiment 1. The choice of word vowels rather than text vowels was an arbitrary one, since Experiment 1 had failed to show a significant difference between these two conditions.

Subjects. The subjects were 14 students from the pool of 19 subjects who had participated in Experiment 1. Five students were no longer available for testing.

Design. Each listener took part in six tests, three roving 2I2AFC discrimination tests, two roving 4I2AFC discrimination tests (with flanking stimuli), and a classification test. The 2I2AFC discrimination was tested under three conditions: (1) *baseline*, with one step between the stimuli in a trial and a short ISI; (2) *ISI*, with one step between the stimuli and a long ISI; and (3) *step size*, with two steps between the stimuli and a short ISI. The 4I2AFC discrimination was tested only in Conditions 2 and 3. For the 4I2AFC baseline condition, the data in Experiment 1 were used (roving 4I2AFC discrimination, word vowels) minus the results of the 5 subjects that did not participate in Experiment 2. The order of the discrimination tests was randomized across subjects, but classification was always performed last.

In 2I2AFC discrimination, all the trials consisted of two different stimuli, either AB or BA. These combinations were presented 32 times in a random order. The subjects responded by indicating the

order in which the stimuli had been presented. The short ISI was 200 msec, and the long one was 500 msec. In the one-step discrimination test, 7 (stimulus pairs) × 64 trials were presented. In the two-step discrimination test, 6 (stimulus pairs) × 64 trials were presented.

In 4I2AFC discrimination, the trials consisted of AABA and ABAA combinations. These combinations were presented 32 times in a random order. The subjects had to indicate whether the "odd-ball" was in the second or the third interval. The short ISI was 200 msec and the long one was 500 msec. In the one-step discrimination test, 7 (stimulus pairs) × 64 trials were presented. In the two-step discrimination test, there were 6 (stimulus pairs) × 64 trials.

The classification test was a replication of the test in Experiment 1, using 14 of the 19 original subjects.

Procedure. The procedures for discrimination were identical to the procedure described in Experiment 1. The only difference was the notation in the response fields. In the 2I2AFC test, the subjects responded by mouse clicking on a response field with "oe–ie" or with "ie–oe," which represented the order of the stimuli. The total experiment (six tests) lasted approximately 3 h.

Results

The discrimination and classification results are displayed in Figures 4–6. The data in the figures represent the averages of the 14 subjects' individual d' scores. The numbers (n) along the abscissa refer to stimulus pairs, consisting of the (n) and ($n + 1$) stimuli for one-step discrimination and the (n) and ($n + 2$) stimuli for two-step discrimination. The stimuli in Pair 1 resemble /u/, and the stimuli in Pair 7 sound like /i/. In order to compare classification and discrimination, the classification scores were transformed into predicted discrimination scores. The degree of categorical perception of the various functions in the three test conditions was estimated with the CP index.

Task. In Figure 4, the results of the baseline condition are presented. The relationship between 4I2AFC discrimination and classification in Figure 4 is the same as the one shown in Figure 1: no correlation between observed and predicted discrimination, so no indication of categorical perception. However, 2I2AFC discrimination is more closely related to classification. It is higher, except at the first and last stimulus pairs. (The 4I2AFC functions in Figures 1 and 4 are slightly different, because only the data from the 14 listeners who participated in both experiments are plotted.)

As Table 2 shows, the CP index for 4I2AFC discrimination was −2, whereas for 2I2AFC discrimination, it was 40. These CP indices confirm what is shown in Figure 4. The degree of categorical perception is very low in the 4I2AFC task, but 2I2AFC discrimination is more closely related to classification. A two-way ANOVA was performed on the d' data, with task (three levels) and stimulus (seven levels) as factors. There were significant effects of the task and stimulus factors [$F(2,20) = 6.33, p < .005$, and $F(6,20) = 6.39, p < .001$], and there was a significant interaction between task and stimulus [$F(12,20) = 1.96$, $p < .05$]. A Newman–Keuls test revealed that 4I2AFC discrimination was significantly different from 2I2AFC discrimination and from classification [$F(2,291) = 5.50$, $p < .01$] but that there was no significant difference between two-interval discrimination and classification. A Newman–Keuls post hoc analysis was carried out on each separate task, with stimuli as an independent variable. This was done mainly to test the significance of the classification and discrimination peaks. There turned out to be significant peaks at Pair 4 for 2I2AFC discrimination and at Pair 3 for classification [$F(6,91) = 2.68, p < .05$, and $F(6,91) = 14.39, p < .001$].

Interstimulus interval. Figure 5 shows the results for vowel classification and discrimination with a longer ISI. A comparison with the discrimination results in Figures 1

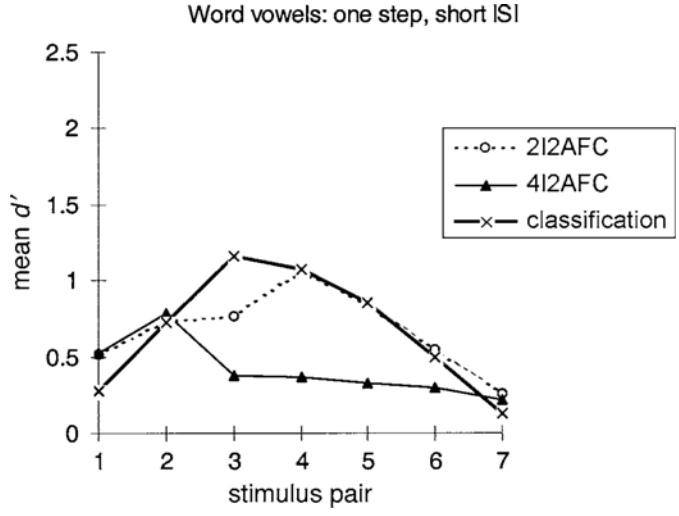

Figure 4. Predicted one-step discrimination (classification) and actually measured roving discrimination (two-interval two-alternative forced choice [2I2AFC] and 4I2AFC) from the baseline condition for short interstimulus interval (ISI), one-step discrimination.

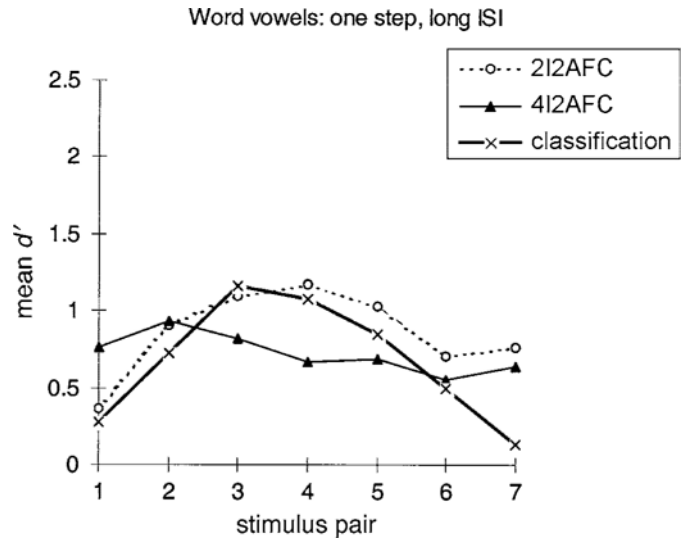

Figure 5. Predicted one-step discrimination (classification) and actually measured roving discrimination (two-interval two-alternative forced choice [2I2AFC] and 4I2AFC) from the interstimulus interval (ISI) condition for long-ISI one-step discrimination.

and 4 indicates that the scores in this test were higher, which means that the longer ISI facilitated discrimination. There is still no apparent correlation between observed 4I2AFC discrimination and discrimination as predicted by classification: For the stimuli around the phoneme boundary, discrimination remained worse than classification. This can mean only that comparison of auditory traces benefits from a longer ISI and that the optimum interval of 200–300 msec usually found in two-interval tasks should not be generalized to other situations. The scores in 2I2AFC discrimination also increased, but the difference from the baseline condition was smaller than in the

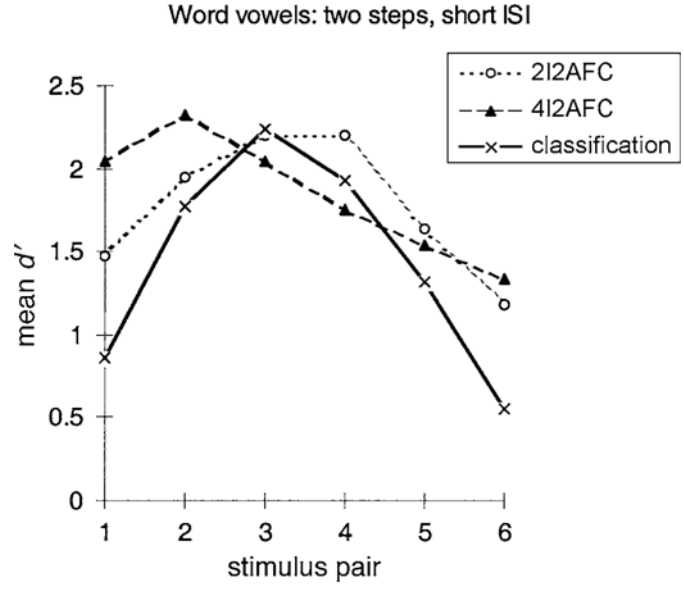

Figure 6. Predicted two-step discrimination (classification) and actually measured roving discrimination (two-interval two-alternative forced choice [2I2AFC] and 4I2AFC) from the step size condition for short interstimulus interval (ISI) two-step discrimination.

Table 2
Mean Categorical Perception Indices and Standard Errors for Four-Interval Two-Alternative Forced-Choice (4I2AFC) and 2I2AFC Discrimination With Interstimulus Intervals (ISIs) of 200 or 500 msec, and With a Physical Distance of One or Two Steps Between Stimuli in a Trial

		4I2AFC		2I2AFC	
ISI (msec)	Distance	*M*	*SE*	*M*	*SE*
200	one step	−2	12	40	9
500	one step	18	10	57	7
200	two steps	32	11	65	6

case of four-interval discrimination. The 2I2AFC discrimination results still agreed well with those for classification, indicating that perception was categorical. It has to be concluded that there was an effect of ISI on discrimination but that this effect was rather small and affected only overall performance. It is unclear why having more time between stimuli did not induce the subjects to use phoneme labels.

This conclusion is modified to some extent by the CP indices in Table 2. The CP index for the 500-msec condition was a little higher, which indicates that degree of categorical perception depended to some extent on ISI.

The effect of ISI was tested in a three-way ANOVA of the discrimination data from the present experiment, combined with the data from the baseline condition (in Figure 4). The independent variables were task (two levels, nested under ISI), stimulus (seven levels) and ISI (two levels). The classification data were excluded from this statistical analysis, since the effect of ISI does not play a role in classification. Cell variance was over 14 subjects. The analysis revealed significant main effects of task [$F(1,27) = 9.8, p = .002$], stimulus [$F(6,27) = 2.89, p = .009$], and ISI [$F(1,27) = 14.85, p < .001$] and a significant interaction between task and stimulus [$F(6,27) = 2.58, p = .018$]. The main effect of ISI means that the long-ISI d' values were significantly higher than the short-ISI ones, which indicates that the longer ISI facilitated discrimination. Since there was no significant interaction between task and ISI, this facilitation effect seems to be equally large for 4I2AFC and 2I2AFC discrimination. In a second statistical analysis, the long-ISI discrimination data were compared with the classification results. A two-way ANOVA with task (three levels) and stimulus (seven levels) as factors showed only a significant main effect of stimulus [$F(6,20) = 4.61, p < .001$]. There were no significant peaks or valleys in the 4I2AFC and 2I2AFC functions, as was shown by a separate Newman– Keuls post hoc analysis.

Physical distance between the stimuli. The results of two-step discrimination are presented in Figure 6. Of course, two-step classification and discrimination generated higher scores than did their one-step counterparts, shown in Figures 4 and 5. In addition, the 4I2AFC scores are no longer worse than would be predicted by the classification data. But again, there is little similarity between discrimination and classification and, thus, hardly any indication of categorical perception. On the other hand, the similarity between 2I2AFC and classification does point to categorical perception.

As can be seen in Table 2, the CP index for 4I2AFC discrimination in the two-step condition remained lower than the one for 2I2AFC discrimination, although it increased as an effect of step size, due to a decreased difference (more proximity) in the denominator of the CP equation. The 2I2AFC CP index also increased as an effect of step size, indicating a relatively high degree of categoricalness of 2I2AFC discrimination.

In a two-way ANOVA, the task (three levels) and stimulus (six levels) factors turned out to be significant [$F(2,17) = 3.49, p < .05$, and $F(5,17) = 8.25, p < .001$]. Newman–Keuls tests revealed significant peaks in both 2I2AFC discrimination and classification at Stimulus Pairs 3 and 4 [$F(5,78) = 2.94, p < .05$, and $F(5,78) = 5.82, p < .001$].

Discussion

The results of Experiment 2 show that neither physical distance nor ISI has a large effect on degree of categorical perception (as shown with the CP index) and indicate that listeners use different processes during the two discrimination tasks. This confirms the suspicion we formulated in the Discussion section in Experiment 1 (and which falsifies our hypothesis about the nature of the 4I2AFC task). The 4I2AFC task relies on auditory processing, whereas the 2I2AFC task elicits phoneme labeling. These findings suggest that our first interpretation of the discrimination results in Experiment 1 was correct: The subjects were in an auditory mode and outside their normal phonetic mode.

In summarizing the test conditions in Experiment 2, it can be concluded that the counterintuitive discrimination results in Experiment 1 were caused not so much by too short an ISI or too small a step size between the stimuli, but mainly by the auditory nature of the task, which seems to preclude phoneme labeling.

DIFFERENCES AMONG SUBJECTS IN EXPERIMENT 1

Now that Experiment 2 has shown that our failure to find categorical perception with the 4I2AFC task was probably due to the task itself, we will return to Experiment 1 in order to explore subject differences. A large proportion of the total variance in Experiment 1 (Figures 1–3) can be explained by cell variance (78.5%). This means that there must have been considerable differences, probably in performance level, between the 19 subjects. If our interpretation of the overall results from Experiment 1 is correct—that is, if 4I2AFC discrimination (with flankers) is an auditory task that, for whatever reason, precludes labeling—all subjects should have one thing in common: Regardless of their level of performance, none of them should show a strong relationship between classification and discrimination. A corollary prediction with respect to classification is that, if we grade subjects according to their

discrimination performance, this will tell us nothing about classification performance, which should be roughly the same over discrimination quartiles.

In order to test these predictions, the subjects were divided into quartiles on the basis of their roving-discrimination d' scores. Discrimination performance decreased gradually over the quartiles; classification, however, was at roughly the same level in all the quartiles, as was predicted. These results were confirmed by an ANOVA: The effect of the quartiles factor on discrimination performance was significant [fixed, $F(3,129) = 9.09$, $p < .01$; roving, $F(3,129) = 27.60$, $p < .01$], whereas there was no significant effect of quartiles on classification performance. There were considerable differences in the position of the phoneme boundary between individuals (as is often the case with vowel stimuli) and, hence, between quartiles.

Figures 7 and 8 show the discrimination and classification results for the word vowels (the results were the same for the two vowel conditions), obtained from the highest and the lowest quartiles (4 and 5 subjects, respectively). In Table 3, the CP indices are presented.

There was a marked difference between the discrimination performances of the two subject groups (they were, after all, selected on that basis). In Figure 7 (lowest quartile), the classification results show a peak at the phoneme boundary, with d' scores decreasing to chance level at the extremes of the continuum. The discrimination scores are at chance level for all the stimuli, indicating that the listeners could not detect differences, not even at the phoneme boundary. In Figure 8, which represents the performance of the subjects in the highest quartile, discrimination is, in general, as high as, or higher than, classification. In neither case does discrimination performance appear to have much in common with classification performance, which confirms our prediction. The CP indices, shown in Table 3, tell a slightly different story, confirming the lack of categorical perception in the lower quartile (the negative CP index is due to negative correlations), but indicating at least a small amount of categorical perception for the higher quartile. These findings indicate that, during discrimination, hardly any phonetic information was used: The listeners were in an auditory mode. In classification, however, all the listeners had to operate in the phonetic mode and showed the same performance.

GENERAL DISCUSSION

In a number of experiments, we have investigated categorical perception of a series of vowel stimuli as a function of two 2AFC tasks: one consisting of two intervals, the other of four intervals. In the four-interval task, two intervals served as facilitating *flankers* but did not add any extra information. It was expected that this task would permit the use of both bottom-up auditory and top-down phonetic information, mainly because the subjects were expected to have a choice between (combinations of) two strategies: a 2IFC strategy, in which the two flanking stimuli are ignored, and a 4IAX strategy, in which auditory differences within two pairs are compared. The first strategy was expected to lead to the use of phoneme labels, whereas the second strategy should have made auditory information available. However, our results did not provide any evidence of the first strategy, mainly because there was no apparent relationship between observed and predicted discrimination and, thus, no indication of categorical perception in the 4I2AFC task. For the 2I2AFC task, however, the results turned out to be fairly categorical, as they

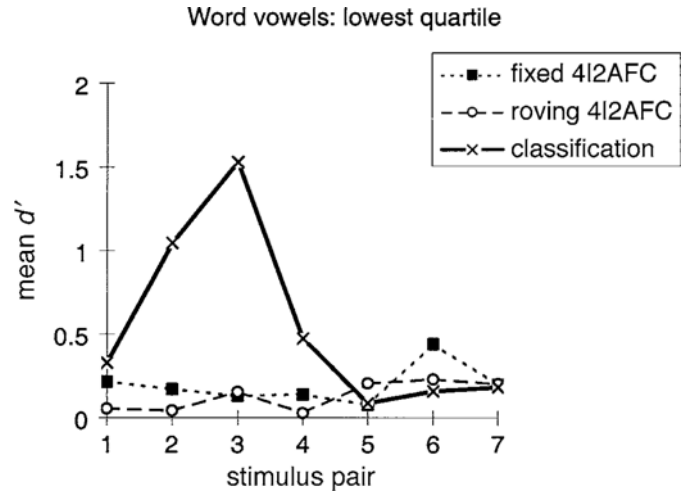

Figure 7. Predicted discrimination (classification) and actually measured discrimination (fixed and roving) of the word vowels in the lowest quartile. 4I2AFC, four-interval two-alternative forced choice.

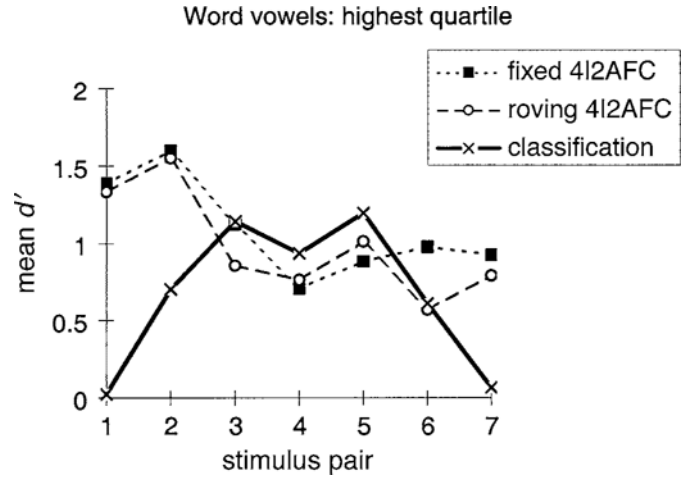

Figure 8. Predicted discrimination (classification) and actually measured discrimination (fixed and roving) of the word vowels in the highest quartile. 4I2AFC, four-interval two-alternative forced choice.

usually are for this task, which makes explicit use of phoneme labels.

Discrimination performance in the 4I2AFC task was poor and fell generally below that predicted on the basis of classification. This paradoxical finding (paradoxical because discrimination performance has nearly always been better than would be predicted from classification; see, e.g., Macmillan, Goldberg, & Braida, 1988; Pastore, 1987; Repp, 1984; Schouten & Van Hessen, 1992) could be explained partly as being due to the difficulty of the task: A greater physical spacing between stimuli in a discrimination trial improved performance but did not really make it much more categorical. It improved discrimination more than it did classification, so that the former rose above the latter on over half the stimuli. Increasing the ISI had a similar effect; it is unclear why having more time between stimuli did not induce the subjects to use phoneme labels more. There were considerable differences in discrimination performance levels among the subjects: All of them performed at about the same level in classification, but they could easily be divided into good and poor discriminators, and the performances of both groups were largely uncategorical.

Table 3
Mean Categorical Perception Indices and Standard Errors for Fixed and Roving Four-Interval Two-Alternative Forced-Choice (4I2AFC) Discrimination (Word Vowels) of the Lowest and Highest Quartiles of Subjects

	Vowels			
	Fixed		Roving	
Task	*M*	*SE*	*M*	*SE*
	Lowest Quartile			
4I2AFC	−25	7	−32	21
	Highest Quartile			
4I2AFC	15	13	22	13

The conclusion seems to be almost inescapable: When listening to speech, all subjects perform at roughly the same level, as they would be expected to do in everyday speech situations; when listening in the auditory mode, however, they show great differences in performance, sorting out themselves into good and poor (and intermediate) listeners. This is a common psychoacoustic pattern; several studies have indicated that subjects listening to speech-like stimuli may have widely different scores if they are operating in an auditory mode (Best, Morrongiello, & Robson, 1981; Dorman et al., 1977; Foard & Kemler-Nelson, 1984; Repp, 1981; Rosen & Howell, 1981). The fact remains that all our listeners, good or poor, behaved during 4I2AFC discrimination as if our stimuli had nothing to do with speech at all. This finding is, as far as we know, without precedent in the categorical perception literature, but then, so is the 4I2AFC discrimination task. The question is, what is it that makes this task so different from all other discrimination tasks ever used in categorical perception, such as the regular 2I2AFC task in Experiment 2? And what do our results tell us about the categorical perception of speech?

A first attempt to account for the traditional finding (discrimination better than would be predicted by the classification data) was made with the dual-process model for the discrimination of speech stimuli by Fujisaki and Kawashima (1971). This model explicitly distinguished between categorical phonemic judgments and judgments based on auditory memory for acoustic stimulus attributes. The authors proposed that two perceptual modes are active simultaneously (or in rapid sequence). One of them is strictly categorical and represents phonetic classification and the associated verbal short-term memory. The other mode is not categorical and represents processes common to all auditory perception. The results of any particular speech discrimination experiment are assumed to reflect a mixture of both components. The part of perfor-

mance that can be predicted from labeling probabilities is attributed to categorical judgments, whereas the remainder (the deviation from ideal categorical perception) is assigned to comparison of acoustic stimulus properties (Repp, 1984).

In our selection of the 4I2AFC task as a discrimination paradigm, we were led by the idea that a discrimination task in which auditory coding would not be excluded in advance was needed to assess categorical perception. If a discrimination task is used that *prevents* a direct comparison between successive stimuli, listeners are *forced* to use a phonetic-labeling strategy for discrimination, and results will *inevitably* be highly categorical. We were even more successful than we expected in stimulating an auditory-coding strategy: The results show that there was no categorical perception at all. Moreover, discrimination performance was lower than would be predicted from classification. This was especially true of the lowest quartile, whose classification results were quite normal, but whose discrimination results were at chance level for all stimuli, indicating that the listeners could not detect differences, not even at the phoneme boundary. The discrimination performance of the subjects in the highest quartile, however, was, in general, as high as, or higher than, classification. In neither case did discrimination performance show any clear relationship with classification performance. This means that, during discrimination, no phonetic information was used: The listeners were in the auditory mode. In classification, however, all the listeners had to operate in a phonetic mode and showed the same behavior. Why is it that the 4I2AFC task puts subjects into an auditory mode, whereas all other speech discrimination tasks that have been used up to now, including the 2I2AFC task in the present Experiment 2, produce at least a mixture of the two modes, so that results always turn out categorical to some extent?

We suspect that this is due to the nature of the four-interval task, which is less close to the traditional 2I2AFC task than we had expected, and closer to the 4IAX task. The crucial difference here could be whether or not a subject's decision is dominated by a criterion, or bias, that is external to the stimuli. In the traditional 2I2AFC task, in which a subject has to indicate the order of two stimuli, this can, in the case of speech stimuli, be done only with reference to *top-down* labeling criteria that are external to the stimuli, such as boundaries between phonemes or between categories like *high* versus *low* in psychoacoustic stimuli. However, the 4I2AFC results were more like those that may be expected from the 4IAX task, in which subjects are presented with two pairs of stimuli and have to decide which pair contains the odd one out. The 4IAX task does not refer subjects to criteria that are external to the stimuli: Subjects hear two differences and have to decide which is the greater one, a decision that can be taken on the basis of stimulus information and does not depend very much on the position of a criterion along a scale, apart from a preference for one response type over another, which is unlikely to be strong. This is why the 4IAX task is generally regarded as relatively bias free in psychophysics.

Our subjects' comments and our own experience when performing the 4I2AFC task confirmed the similarity between it and the 4IAX task, at least with vowel stimuli: It soon became apparent to the subjects that attempts to label the stimuli did not produce correct responses, so they claimed to have given up the attempt and to have applied a strategy in which differences between two pairs of stimuli were used in order to reach a decision. As a result, we ended up with a discrimination experiment in which phoneme categories did not play a part. This was unexpected: Even if we had foreseen that the subjects would treat the 4I2AFC task as a variant of the 4IAX task, we still would not have foreseen the complete absence of phonetic information in the discrimination response. The 4IAX task may be relatively bias free, but we had expected phoneme labels to be inevitable: A human subject should not be able to treat speech as if it does not consist of phonemes or words. We were wrong: It *is* possible to get subjects completely into the auditory mode when they have to discriminate speech sounds, although we cannot really claim to know why they were incapable of using available phonetic information that would have improved their performance.

In normal everyday speech perception, we perceive categorically. It has always been assumed that this mental process can be investigated by looking at the relationship between two psychophysical tasks: classification and discrimination. It now appears that this assumption rests crucially on the discrimination task that is chosen. When a four-interval discrimination task is used, no resemblance between phoneme classification and discrimination is found. However, an inability to discriminate sounds does not prevent listeners from being able to assign different labels to these same sounds.

REFERENCES

Best, C. T., Morrongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception & Psychophysics*, **29**, 191-211.

Cowan, N., & Morse, P. A. (1986). The use of auditory and phonetic memory in vowel discrimination. *Journal of the Acoustical Society of America*, **79**, 500-507.

Delattre, P. C., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel colour: Observations on one- and two-formant vowels synthesised from spectrographic patterns. *Word*, **8**, 195-210.

Dorman, M. F., Kewley-Port, D., Brady, S., & Turvey, M. T. (1977). Vowel recognition: Inferences from studies of forward and backward masking. *Quarterly Journal of Experimental Psychology*, **29**, 483-497.

Eimas, P. D. (1963). The relation between identification and discrimination along speech and nonspeech continua. *Language & Speech*, **6**, 206-217.

Foard, C. F., & Kemler-Nelson, D. G. (1984). Holistic and analytic modes of processing: The multiple determinants of perceptual analysis. *Journal of Experimental Psychology: General*, **113**, 94-111.

Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language & Speech*, **5**, 171-189.

Fujisaki, H., & Kawashima, T. (1971). A model of the mechanisms for speech perception: Quantitative analyses of categorical effects in discrimination. *Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo*, pp. 59-68.

Gerrits, E. (2001). *The categorisation of speech sounds by adults and children*. Unpublished doctoral dissertation, Utrecht University.

Heller, L. M., & Trahiotis, C. (1995). The discrimination of samples of noise in monotonic, diotic and dichotic conditions. *Journal of the Acoustical Society of America*, **97**, 3775-3781.

Liberman, A. M. (1996), *Speech: A special code.* Cambridge, MA: MIT Press.

Liberman, A. M., Harris, K., Hoffman, H. S., & Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, **54**, 358-368.

Macmillan, N. A. (1987). Beyond the categorical/continuous distinction: A psychophysical approach to processing modes. In S. Harnad (Ed.), *Categorical perception* (pp. 53-85). New York: Cambridge University Press.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.

Macmillan, N. A., Goldberg, R. F., & Braida, L. D. (1988). Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua. *Journal of the Acoustical Society of America*, **84**, 1262-1280.

Massaro, D. W. (1972a). Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review*, **79**, 124-145.

Massaro, D. W. (1972b). Stimulus information vs processing time in auditory pattern recognition. *Perception & Psychophysics*, **12**, 50-56.

Massaro, D. W. (1974). Perceptual units in speech recognition. *Journal of Experimental Psychology*, **102**, 199-208.

Massaro, D. W., & Cohen, M. M. (1983). Categorical or continuous speech perception: A new test. *Speech Communication*, **2**, 15-35.

Pastore, R. E. (1987) Categorical perception: Some psychophysical models. In S. Harnad (Ed.), *Categorical Perception* (pp. 29-52). New York: Cambridge University Press.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, **13**, 253-260.

Pisoni, D. B. (1975). Auditory short-term memory and vowel perception. *Memory & Cognition*, **3**, 7-18.

Plomp, R. (1964). Decay of auditory sensation. *Journal of the Acoustical Society of America*, **36**, 277-282.

Repp, B. H. (1981). Two strategies in fricative discrimination. *Perception & Psychophysics*, **30**, 217-227.

Repp, B. H. (1984). Categorical perception: Issues, methods, findings. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 10, pp. 244-322). Orlando, FL: Academic Press.

Rosen, S. M., & Howell P. (1981). Plucks and bows are not categorically perceived. *Perception & Psychophysics*, **30**, 156-168.

Sawusch, J. R., Nusbaum, H. C., & Schwab, E. C. (1980). Contextual effects in vowel perception II: Evidence for two processing mechanisms. *Perception & Psychophysics*, **27**, 421-434.

Schouten, B., Gerrits, E., & Van Hessen, A. (2003). The end of categorical perception as we know it. *Speech Communication*, **41**, 71-80.

Schouten, M. E. H., & Pols, L. C. W. (1979). Vowel segments in consonantal contexts: A spectral study of coarticulation–part I. *Journal of Phonetics*, **7**, 1-23.

Schouten, M. E. H., & Van Hessen, A. J. (1992). Modelling phoneme perception: I. Categorical perception. *Journal of the Acoustical Society of America*, **92**, 1841-1855.

Tartter, V. C. (1981). A comparison of the identification and discrimination of synthetic vowel and stop-consonant stimuli with various acoustic properties. *Journal of Phonetics*, **9**, 477-486.

Tartter, V. C. (1982). Vowel and consonant manipulations and the dual-coding model of auditory storage: A re-evaluation. *Journal of Phonetics*, **10**, 217-223.

Trahiotis, C., & Bernstein, L. R. (1990). Detectability of interaural delays over select spectral regions: Effects of flanking noise. *Journal of the Acoustical Society of America*, **87**, 810-813.

Van Hessen, A. J. (1992). *Discrimination of familiar and unfamiliar speech sounds.* Unpublished doctoral dissertation, Utrecht University.

Van Hessen, A. J., & Schouten, M. E. H. (1992). Modelling phoneme perception: II. A model of stop consonant discrimination. *Journal of the Acoustical Society of America*, **92**, 1856-1868.

Van Hessen, A. J., & Schouten, M. E. H. (1999). Categorical perception as a function of stimulus quality. *Phonetica*, **56**, 56-72.

Van Son, R. J. J. H., & Pols, L. C. W. (1990). Formant frequencies of Dutch vowels in a text, read at normal and fast rate. *Journal of the Acoustical Society of America*, **88**, 1683-1693.

(Manuscript received June 19, 2001;
revision accepted for publication August 6, 2003.)