Chapter 2

Categorical perception of word and text vowels[1]

2.1 Introduction

Several studies have shown that there is a difference in the degree of categorical perception between stop consonants and vowels. Stop consonants are said to be more categorically perceived than vowels (e.g. Fry et al., 1962; Pisoni, 1973; Repp, 1981; Stevens et al., 1969).

This difference in perception has received various explanations. Early explanations, based on the Motor Theory of Speech Perception, posit that it is the result of a difference in the underlying articulatory gestures (Liberman et al., 1967). Perception is categorical when the articulatory space is relatively discontinuous. Stop consonants show marked acoustic variability, but speakers seem to use discrete places of production for them. Vowel articulation is much less discrete: there is a fair amount of articulatory overlap between adjacent vowel categories (Liberman et al., 1967). However, the mechanisms by which perceptual processes might refer to articulation have always remained obscure, and this has led many researchers to dismiss the motor theory (see review Repp, 1984).

A second explanation for the processing differences between vowels and stop consonants is given by Ades (1977), Rosen and Howell (1987) and Sachs (1969). They speculate that the differences arise from differences in the perceptual distances between stop consonants and between vowels: two adjacent vowels span a wider perceptual range in terms of JND's than two adjacent consonants. Thus there should be a larger range of ambiguity along a vowel continuum than along a consonant continuum. One would therefore expect a smaller discrepancy between classification and discrimination (in other words, the results

[1] This chapter is an adapted version of Gerrits and Schouten (1998).

would be more categorical) for consonants than for vowels. However, the results of Ades (1977) and Macmillan, Goldberg and Braida (1988) revealed no difference in perceptual range in their stimulus sets: the total number of JND's between /ba/ and /pa/ was about the same as between /i/ and /I/.

A third explanation is proposed by Pisoni (1973, 1975), Sachs (1969), and Tartter (1981, 1982). They explain the difference in perception between stop consonants and vowels by the difference in cue duration. The essential acoustic cues for stop consonants are rapidly changing F1 and F2 transitions and a brief noise burst (Liberman et al., 1954; Tartter, 1981, 1982). By contrast, vowels are assumed to remain uniform over a much longer duration (Delattre et al., 1952). This difference in duration between vowels and stop consonants has an effect on the availability of auditory memory for these two classes of speech sounds. According to Pisoni (1973), the outcome of a stimulus comparison during discrimination depends on two types of memory components: trace memory and (phonetic) label memory. This is in line with the dual-process model of Fujisaki and Kawashima (1971), who propose that discrimination may be performed in a trace comparison or in a labelling mode. They explain the difference in perception between stop consonants and vowels by differences in the duration of the critical information in these speech signals. The short cue duration in consonants is responsible for the inferior performance of trace memory. Presumably, the decay of rapidly changing acoustic information is too fast to make an acoustic comparison of consonantal stimuli possible, with the result that discrimination is performed in a phonetic mode. This is not the case for vowels produced in isolated words (citation form).

In line with this explanation, Schouten and Van Hessen (1992) propose that the lack of categorical perception of vowels may be due to the nature of the stimulus material that has been used. Up to now, the vowels used as stimulus material have been modeled on productions in isolated words. When produced in isolated words, vowels will be lengthened (overarticulation). The duration of a plosive will be less affected by overarticulation than that of a vowel: the articulation of the burst can hardly be lengthened and the length of the formant transitions is also restricted since too much lengthening would lead to a change in phoneme identity. It is hypothesised that in running speech temporal reduction and more complex spectral coding of the vowel will make vowel perception more categorical (see also Pisoni, 1973; Repp, 1984; Sawusch, Nusbaum & Schwab, 1980; Schouten & Van Hessen, 1992; Stevens, 1968; Studdert-Kennedy, 1976; Tartter, 1981). The shorter duration and the more complex coding of vowels is expected to force listeners to make a quick decision about the phoneme category, especially when stimuli are difficult to discriminate. This will strengthen the relationship between discrimination and classification of the same stimuli. To test this hypothesis, the difference in perception was studied between vowels spoken in isolated words and vowels in a text read at a fast rate. The question of interest was whether text vowels would be perceived more categorically than word vowels.

2.2 Method

2.2.1 Stimulus material

The first step in stimulus generation was recording the vowels /u/ and /i/ in the meaningful words /pup/ and /pip/ produced both in isolation and in a text that was read aloud at a rapid speech rate by a male native speaker of Dutch. The vowels produced in the isolated words will be called 'word vowels' and the vowels produced in the words in the text will be referred to as 'text vowels'. Since the perception of naturally produced vowels was of interest in this study,

one could argue that the ideal stimulus material should be based on vowels from spontaneous speech. It was decided to elicit vowels in a rapidly read text in order to control the stimulus material, on the assumption that fast reading produces the same type of underspecification as spontaneous speech. 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.

There were nine repetitions of /pup/ and /pip/ in the list of words and in the text. Five phonetically untrained listeners identified 30 ms segments of these vowels in an open-set identification task. The vowels spoken in isolation were significantly more often identified as /u/ and /i/ than the vowels from the fast read text (65% versus 32% for /u/ and 42% versus 7% for /i/). Frequently used other response categories were /o/ for /u/ and /y/ for /i/. The word pairs that were used as endpoints in the two stimulus continua were selected by a listening panel that consisted of five phoneticians. They rated the word pairs on a 7-point acceptability scale. The most acceptable ones were selected for the experiment.

Description of the original vowels

To see if there were any differences between the vowels spoken in isolated words and fast text, the durations and formant frequencies were measured. The text vowels were temporally reduced compared to the word vowels. Vowel duration was defined as the duration of the voiced portion between the two /p/'s. The steady-state part of a vowel was defined as that part during which the frequency of F2 did not differ by more than 50 Hz in successive analysis frames, which were 16 ms apart and 32 ms wide. The rest of the vowel consisted of formant transitions.

The duration of the word vowel /u/ was 90 ms, the steady-state component was 60 ms. The duration of the text vowel /u/ was 70 ms, with a steady-state component of 30 ms, a reduction of 22% and 50%, respectively. The duration of the word vowel /i/ was 90 ms, the steady-state component was 50 ms. And the duration of the text vowel /i/ was 70 ms, with a steady-state component of 15 ms, a reduction of, respectively, 22% and 70%. This temporal reduction is comparable to the reduction reported in studies by Schouten and Pols (1979) and Van Son and Pols (1990). The temporal vowel reduction found by Schouten and Pols (1979) was 28%. The reduction of steady-state segment duration was on average 38%. Van Son and Pols (1990) found a temporal reduction of 15% between vowels in a text that was read at a normal speech rate and as fast as possible.

In addition to the duration measurements, an analysis of the formant frequencies of the vowels in the two speech conditions was performed. Since automatic formant extraction failed with the text vowels, formant frequencies in all vowels were estimated by eye. Figure 2.1 shows the spectral envelopes of the word vowels /u/ and /i/, figure 2.2 does the same for text vowels /u/ and /i/. The FFT analysing window was 25 ms and contained the highest amplitude in the vowel. The spectrum was cepstrally smoothed with a cepstral lifter of 5 ms in the quefrency domain.

For the word vowels there were no difficulties in locating the formants in the frequency spectrum. The curves in figure 2.1 show resonance peaks that can be interpreted as formants. The estimated values of the first three formants are 370 Hz, 683 Hz, and 2086 Hz for /u/ and 284 Hz, 2065 Hz, and 2741 Hz for /i/. This is in accordance with the formant frequencies of Dutch /u/ and /i/ measured by Pols, Tromp, and Plomp (1973) and Van Son and Pols (1990).

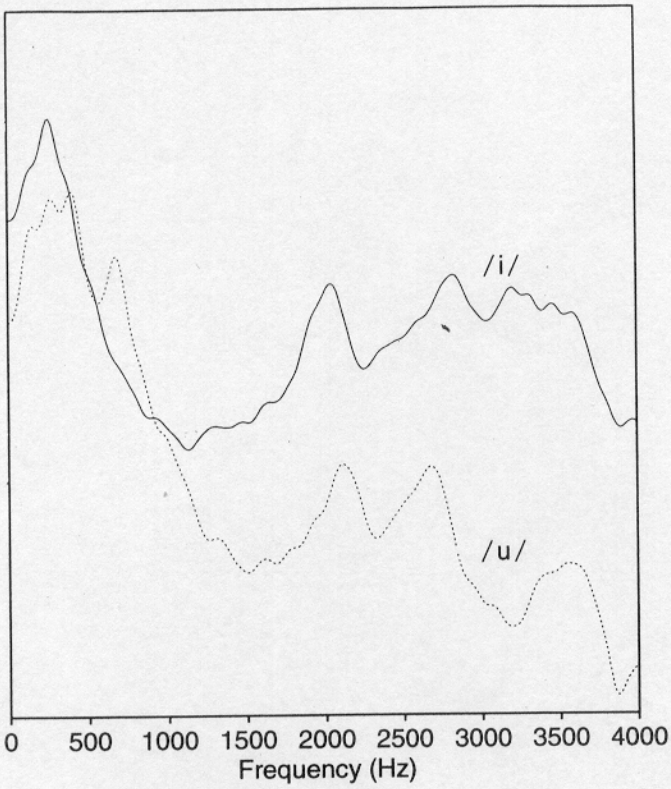

Figure 2.1 Spectral envelopes for the word vowels /u/ and /i/.

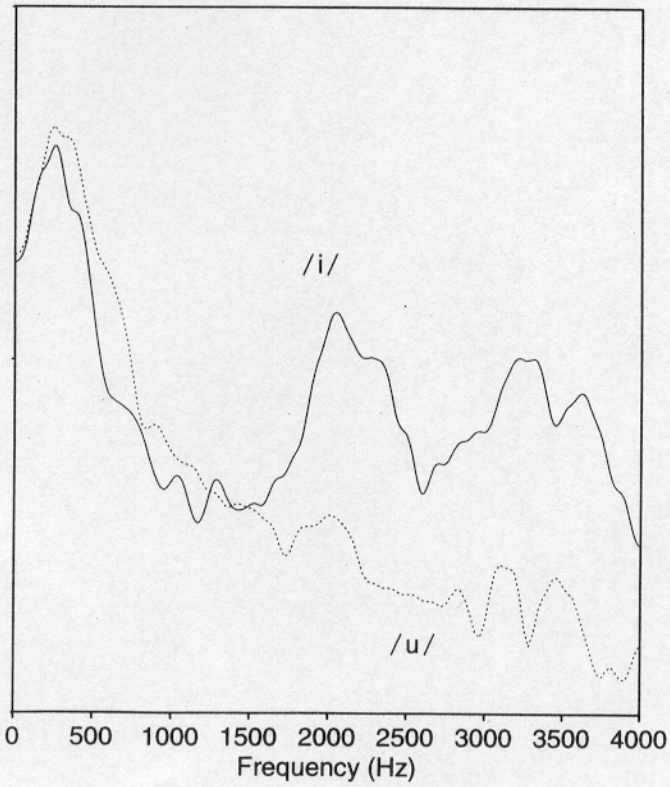

Figure 2.2 Spectral envelopes for the text vowels /u/ and /i/.

In figure 2.2, the spectral envelopes of *both* vowels show amplitude peaks around 250 Hz, 2100 Hz, and 3200 Hz, although at different relative amplitudes. If there had been no a priori knowledge of where the formants should be located, these peaks might have been interpreted as representing the first three formants. Despite the spectral similarity between the two curves in figure 2.2, the vowel that is represented by the dashed line clearly sounds like an /u/, and the vowel of the solid line is without doubt an /i/. Ladefoged (1967) noted that it is often impossible to locate the second formant when it is close to the first formant and lower in intensity. Either the lowest harmonics constitute two formants, or there is no second formant with sufficient intensity for its frequency location to be measured. Ladefoged (1967) added that when F2 is low in intensity, F3 is often impossible to locate for the same reason. Both Ladefoged (1967) and Pols et al. (1973) write that without knowing what vowel phoneme the speaker intended, and without previous knowledge of where the formants of that vowel should be, it is impossible to make valid measurements of the formant frequencies. The spectral data of our text vowels seem to be a nice illustration of their observations.

In short, the duration and formant frequency measurements show that there is a temporal reduction of the text vowels and that there are spectral differences, notably a clustering of formants for the text vowels.

Synthesis of the vowel continua

The interpolation method used had been developed for the study of categorical perception by Van Hessen (1992). Stimuli in a continuum between two natural utterances are obtained by interpolation between the relative amplitudes of the spectral envelopes of these utterances. Parameters such as fundamental frequency, duration, and voice quality, remain constant. The first step in this interpolation method is a separation between the source (vocal folds) and the filter (vocal tract) of the original speech signals. Only the filter characteristics are used for interpolation. The second step is an analysis of the spectral envelopes of the original words in terms of the phases and amplitudes of a large number of spectral components between 80 and 8000 Hz, depending on spectral density. By means of linear interpolation of the amplitudes and interpolation of the phases, spectral envelopes are constructed for the stimuli in between the original words. In the last step, each new spectral envelope (filter) is multiplied with the spectrum of the source of one of the original speech signals. The reconstructed endpoint signals are nearly indistinguishable from the original ones. More details of this procedure are described in Schouten and Van Hessen (1992), or Van Hessen (1992). For the interpolation between the vowels in the present study, the spectral envelopes were analysed in phases and amplitudes of 70 spectral components. 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-ms time frames over the full length of the vowel (frame shift was 6.4 ms).

A disadvantage of the interpolation method is that it generates a speech continuum which has no correlate in speech production. Usually, a speech stimulus continuum is generated with an acoustic parameter that does have articulatory meaning, such as a formant frequency. Nevertheless, for our vowel study the amplitude-interpolation method was preferred to working in the formant domain. In this way the risk was avoided 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 and Schouten, 1999, have shown that there is an increase in categorical perception as synthesis quality

improves from a simple synthesis by rule, via LPC synthesis, to the more complex method used in the present study.) The importance of stimuli in which more than one parameter is varied is also mentioned by Repp (1984) and Liberman (1996), who predict 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. Moreover, since no second formant could be defined, interpolating the formants of the text vowels was impossible.

Stimulus continua

Stimulus generation resulted in two continua of 8 stimuli that sounded completely natural and convincingly like utterances from the original speaker. The continua were named as follows: 1) Word Vowel Continuum; 2) Text Vowel Continuum. In each continuum, the initial and 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 possible because F2 of this central vowel is in between F2 of /u/ and /i/. The absence of an intermediate /y/ is an effect of the interpolation method.

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

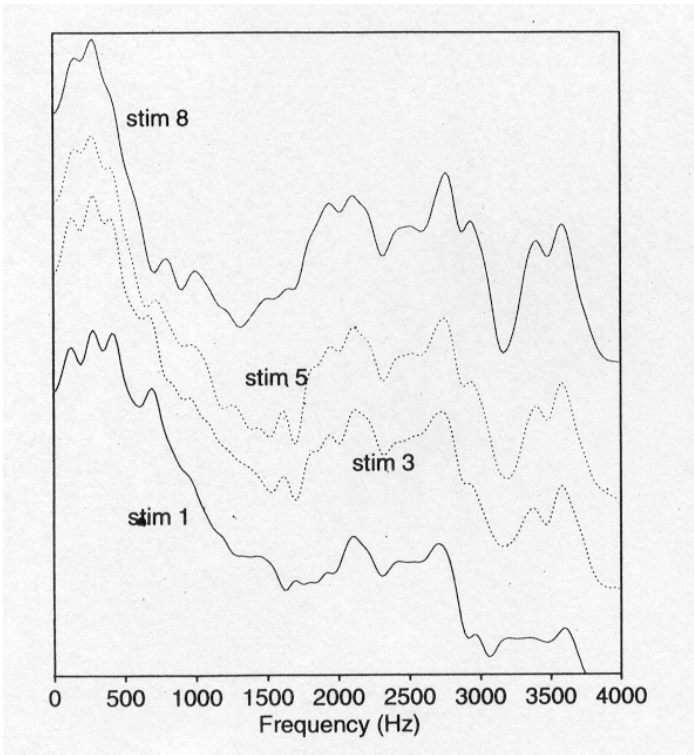

Figure 2.3 Spectral envelopes for stimuli 1, 3, 5, and 8 of the word vowel continuum.

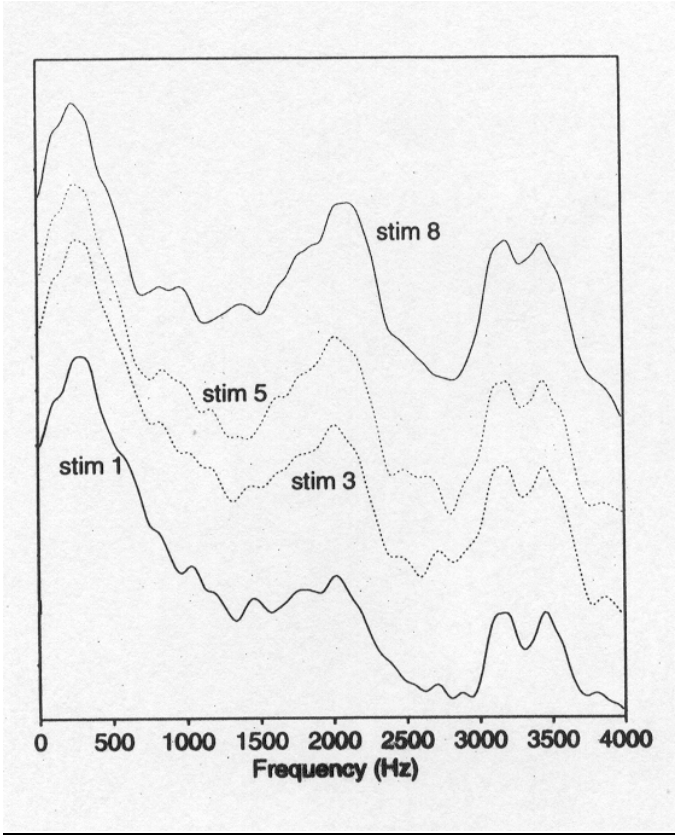

Figure 2.4 Spectral envelopes for stimuli 1, 3, 5, and 8 of the text vowel continuum.

The spectral differences between the two stimulus continua are illustrated in figures 2.3 and 2.4. These figures show the spectral envelopes of stimuli 1, 3, 5, and 8 of both the word and the text vowel continuum. The various curves have been shifted along the vertical axis for maximum clarity. In both figures stimulus 1 sounded like /u/ and stimulus 8 sounded like /i/. The FFT analysing window was 25 ms and contained the highest amplitude in the vowel. The spectrum was cepstrally smoothed with a cepstral lifter of 5 ms in the quefrency domain.

In figure 2.3 it can be seen that the spectral envelopes of stimuli 1 and 8 (the endpoints) are identical to those of the original vowels in figure 2.1. It is clear that the differences between the spectra of the endpoint stimuli and the intermediate stimuli (only 3 and 5 are plotted) consist of shifts in the relative amplitudes of the spectral envelope of the vowel in stimulus 1 towards the relative amplitudes of the spectral envelope of the vowel in stimulus 8.

In figure 2.4 it is shown that the spectral envelopes of the endpoint stimuli have the same shape as the envelopes of the original vowel in figure 2.2. The spectral envelope of stimulus 1 has no significant amplitude peaks that indicate the presence of a second formant for /u/, and in the spectrum of the vowel in stimulus 8 there is no amplitude peak that can be unambiguously classified as the third formant for /i/. Nevertheless, stimulus 1 did sound like /u/ and stimulus 8 sounded like /i/. Figure 2.4 also shows that the stimuli between the endpoints differed in the relative amplitudes of the spectral envelopes.

2.2.2 Subjects

The subjects were 19 students of 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.

2.2.3 Inter-stimulus interval (ISI)

Since it was desirable to encourage a comparison of the acoustic cues in stimuli during discrimination and since the auditory trace of speech sounds is time-dependent, it was important to make a considered decision about the inter-stimulus interval. If the inter-stimulus interval exceeds the duration of the auditory trace of the vowel stimuli, all that is left of the first stimulus is a representation coding the relationship of the stimulus to the other stimuli in the experiment, or to pre-established categories, or to both (Pisoni, 1973). Therefore, an interval was preferred that did not exceed the auditory trace of the stimulus. Massaro (1972a, 1972b, 1974) tried to determine the time a sound pattern is held in some preperceptual form. His results indicated that the auditory trace of an acoustic signal has faded after approximately 250 ms, which is in agreement with the findings by Dorman et al. (1977) and Plomp (1964).

Cowan and Morse (1986), Pisoni (1973), and Van Hessen and Schouten (1992) tested the effect of a varying inter-stimulus interval on discrimination performance. On the basis of Massaro's results, *within-category* discrimination should decrease with increasing interval, reflecting the fading of the 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 an increase in ISI from 100 to 300 ms. The *across-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 to 200 ms. Their results indicated that if listeners use a labeling strategy to compare the stimuli, discrimination improves as an effect of increasing ISI: discrimination performance increases rapidly between 100 and 500 ms, reaches a maximum between 500 and 1000 ms and falls gradually as ISI increases further.

On the basis of these results it was assumed that, after an ISI of more than 100 - 200 ms, labeling processes would take over from trace coding. An interval shorter than 200 ms was not used, since this might increase the chance of mutual masking among the stimuli. In line with Massaro (1972a, 1972b, 1974), Pisoni (1975) and the within-category results of Van Hessen and Schouten (1992), it was therefore decided to use 200 ms intervals to make sure that trace information was available for direct auditory comparison of successive stimuli in a trial.

2.2.4 The discrimination task

The prediction test is the most widely used formal criterion of categorical perception (the Haskins model). It requires a close correspondence between the actual discrimination of speech stimuli along a continuum and discrimination performance predicted from classification results. The procedure of the prediction test has been criticised by Massaro and Cohen (1983). The prototypical discrimination test used to assess categorical perception is the ABX task (Liberman et al., 1957). In view of the relatively short time span of trace memory, however, the observed phenomenon of categorical perception may, according to Massaro and

Cohen (1983), reflect the exclusive use of phonetic memory. Subjects in the ABX task may try to remember both the auditory traces and the labels assigned to the A and B sounds. When X is presented, they try to match the sound of X with the auditory traces of A and B, but by this time they may have "forgotten" what A and B sounded like. If they have, the subjects must rely on the labels they have assigned to A and B and choose the one that matches the label they have assigned to X. This strategy will produce the results usually attributed to categorical perception. Pisoni (1975) and Schouten (1987) also note that the ABX task prevents a direct comparison between successive stimuli and forces the listener to use an encoded categorisation in discrimination. Therefore, the ABX task is not the most appropriate task to use when assessing categorical perception.

The same problems as for ABX discrimination may hold for another paradigm: 2IFC discrimination. In this paradigm, two different stimuli are presented, AB or BA. The subject has to determine the order in which the stimuli are presented. This task has a long tradition in research on intensity resolution (e.g. Macmillan et al., 1977). In intensity experiments, the listener can refer to the order as "soft-loud" or "loud-soft". In the case of speech stimuli, it is necessary to explain to the subjects what the term "order" means and to mention the phoneme categories in the instructions, e.g. /u-i/ and /i-u/. This is at the risk of encouraging labeling behaviour.

To avoid strategies that rely exclusively on labeling, it was necessary to opt for a task that reduced the load on auditory memory and encouraged a direct auditory comparison between the stimuli in a trial (Massaro & Cohen, 1983). Such a task is AX (Same-Different) discrimination, in which AA, AB, BB, and BA combinations are presented. A disadvantage of this paradigm is that, when subjects are asked if two stimuli are different or equal, they may decide to respond "different" only if they are very sure of their decision. This means that AX is not bias free: a subject's response is determined by a subjective criterion (i.e. a boundary between two categories) on a scale between "same" and "different".

A discrimination test that has been shown to be more sensitive to acoustic differences between speech stimuli is the 4IAX task (Pisoni, 1975). The trials in this task consist of eight combinations: ABAA, BAAA, AAAB, AABA, and BABB, ABBB, BBBA, BBAB. The temporal interval between stimuli 2 and 3 is longer than between the other stimuli, so that listeners hear two pairs within one trial. They have to decide which pair contained identical stimuli, pair 1 or pair 2. It is assumed that listeners first determine the differences between the stimuli within the pairs and, in a second step, determine which of the two differences is the smaller one. The correct decision is based on trace information and does not involve subjective criteria.

Yet, it was decided not to use a 4IAX task, but 4I-oddity. In 4I-oddity, the A and B stimuli are presented randomly in the two orders AABA or ABAA, with a 50% a priori probability. Stimulus A at the beginning and end of each quadruplet functions as a reference. Listeners have to respond by indicating if the 'oddball' (stimulus B) is in the second or third stimulus interval. In principle this task is as bias free as 4IAX, and it has a much shorter experimental duration. However, although it is a four-interval task, the optimal decision rules defined by Macmillan and Creelman (1991) predict that the ideal observer will ignore the reference stimuli (interval 1 and interval 4) and perform the 4I-oddity task as if it is a standard 2IFC task (see also Heller & Trahiotis, 1995; Trahiotis & Bernstein, 1990; and Verbunt, 1996). The advantage over 2IFC is that listeners *can* decide about the oddball without needing to refer to any internal criterion. In short, it is expected that 4I-oddity combines some of the important aspects of 2IFC and 4IAX and that listeners will have the choice between two perceptual strategies: a 2IFC-like phoneme-labelling strategy and a 4IAX-like trace-coding strategy. This is also the conclusion drawn by Heller and Trahiotis (1995): their 4I-oddity

results indicate that listeners do not always use the optimal listening strategy but also use the reference stimuli to reach a decision about the oddball. In line with these results it is predicted that the use of the reference stimuli is optional and therefore that 4I-oddity will be, hopefully, neutral with respect to auditory processing and phoneme labelling. This neutral position should make 4I-oddity a proper diagnostic instrument for determining whether categorical perception occurs.

2.2.5 Fixed and roving discrimination

The stimuli were presented in a 'fixed discrimination' and a 'roving discrimination' test (Macmillan, 1988). In the fixed context, listeners have to respond to the same stimulus pair repeatedly during a block of trials before the next pair (which is chosen at random) is presented in another block. This means that within one block the stimulus range is very small, which reduces stimulus uncertainty. Listeners know which stimulus pair will be presented within the block and are consequently able to build up a temporary representation of the two stimuli (Durlach & Braida, 1961). Discrimination performance in the fixed context is expected to be higher than in the roving context.

In the roving context stimuli were drawn randomly from the total stimulus continuum. For instance, if there are 8 stimuli, the first trial may use 4 and 5, the second 1 and 2, etc. The roving context is the one almost invariably used in speech discrimination experiments. The stimulus range in this context is the whole continuum and therefore stimulus uncertainty is much higher than in the fixed context. As a consequence of high stimulus uncertainty it is expected that listeners are unable to form temporary labels for each of the speech stimuli and therefore that labelling effects in the roving data reflect the use of category labels stored in long-term memory.

2.2.6 Design

The experiment consisted of six tests, three for each of two vowel continua, involving the same subjects. The tests were fixed and roving 4I-oddity discrimination and classification. The subjects took the tests in a fixed order: the classification experiment was always performed after the discrimination tests.

The discrimination experiment used was the 4I-oddity task (AABA/ABAA). In this test, the subjects' task was to indicate whether stimulus B (the oddball) 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 inter-trial interval was determined by the response time. The inter-stimulus interval within a trial was 200 ms.

The fixed test consisted of 7 blocks, one for each comparison, which were clearly separated from each other. Each block contained 64 trials, 32 for each of the two possible combinations, AABA and ABAA. In the roving discrimination test, 7 x 64 trials were presented randomly.

In the classification test, each stimulus 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.

2.2.7 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 only be detected by listening carefully to all details of a stimulus. The subjects responded by mouse-clicking on one of two response fields (labelled “2” and “3”) on a computer screen. After the response had been made, visual feedback (250 ms) 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 126 trials, and was intended to familiarise subjects with their task. In the fixed discrimination context the first ten 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 labelled “oe” or “ie” (/u/ and /i/). The only training consisted of 16 trials.

2.3 Results

The discrimination and classification results were computed in terms of d' (e.g. Kaplan, Macmillan & Creelman, 1978; Macmillan & Creelman, 1991). First, the proportions of hits (H) and false alarms (FA) of the adjacent stimulus pairs on the continuum were transformed in z-scores and, second, the z-scores were converted into d' according to the decision model appropriate for the task. The 4I-oddity discrimination d' scores were calculated by subtracting z(FA) from z(H) with a standard deviation $\sigma = \sqrt{2}$ (Macmillan & Creelman, p. 121, 1991). 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 AB 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 forcibly limited to the 0.99 to 0.01 range, which meant that the maximum d'-value that could be obtained was $4.65*0.5\sqrt{2}= 3.29$ for oddity discrimination and 4.65 for classification.

The results are displayed in figures 2.5 to 2.7. Figures 2.5 and 2.6 show the classification and discrimination data for the word vowels and the text vowels, respectively. In figure 2.7, the results are recalculated to get a better indication of the difference between the two vowel conditions. The data in the figures represent the averages of 19 subjects’ individual d'-scores. The numbers (n) along the abscissa refer to stimulus pairs, consisting of stimuli (n) and (n+1); n is therefore a number between 1 and 7. The d' result at stimulus pair 6, for example, represents the discrimination of stimulus 6 and stimulus 7. Stimuli in pair 1 resemble /u/ and stimuli in pair 7 sound like /i/.

In figure 2.7, the two vowel conditions (displayed in figures 2.5 and 2.6) 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). If text vowels are perceived more categorically than word vowels, a smaller difference is expected between the obtained and predicted functions for the former condition than for the latter.

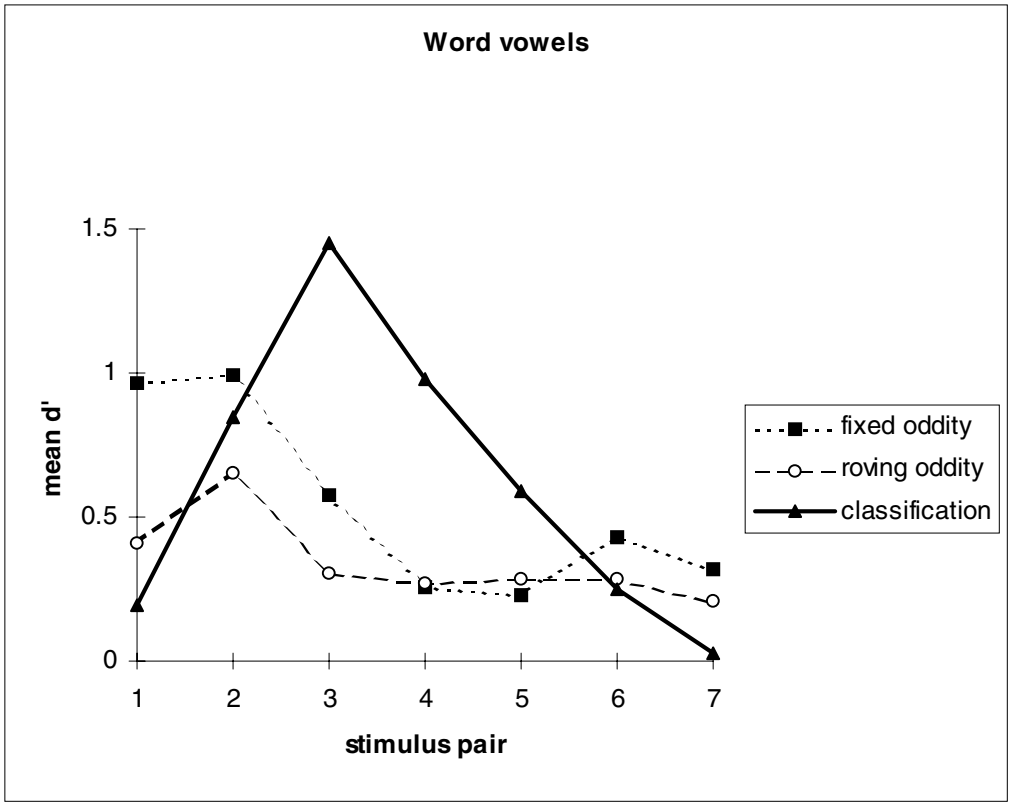

Figure 2.5 Classification and discrimination results for the word vowels.

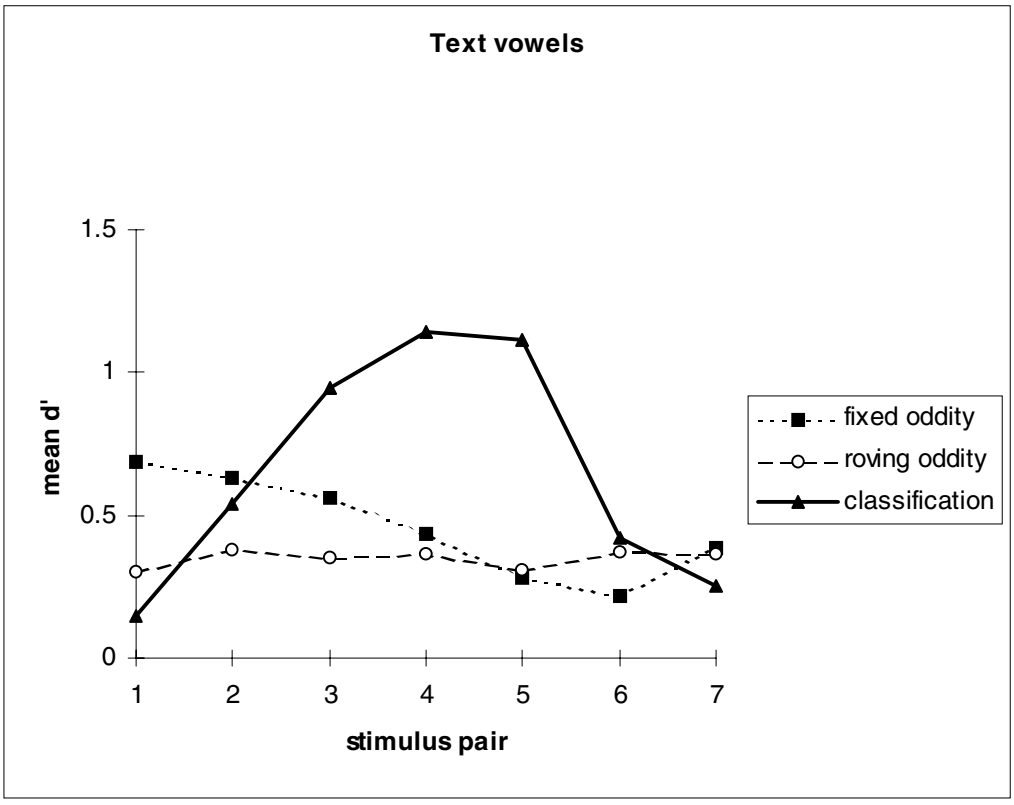

Figure 2.6 Classification and discrimination results for the text vowels.

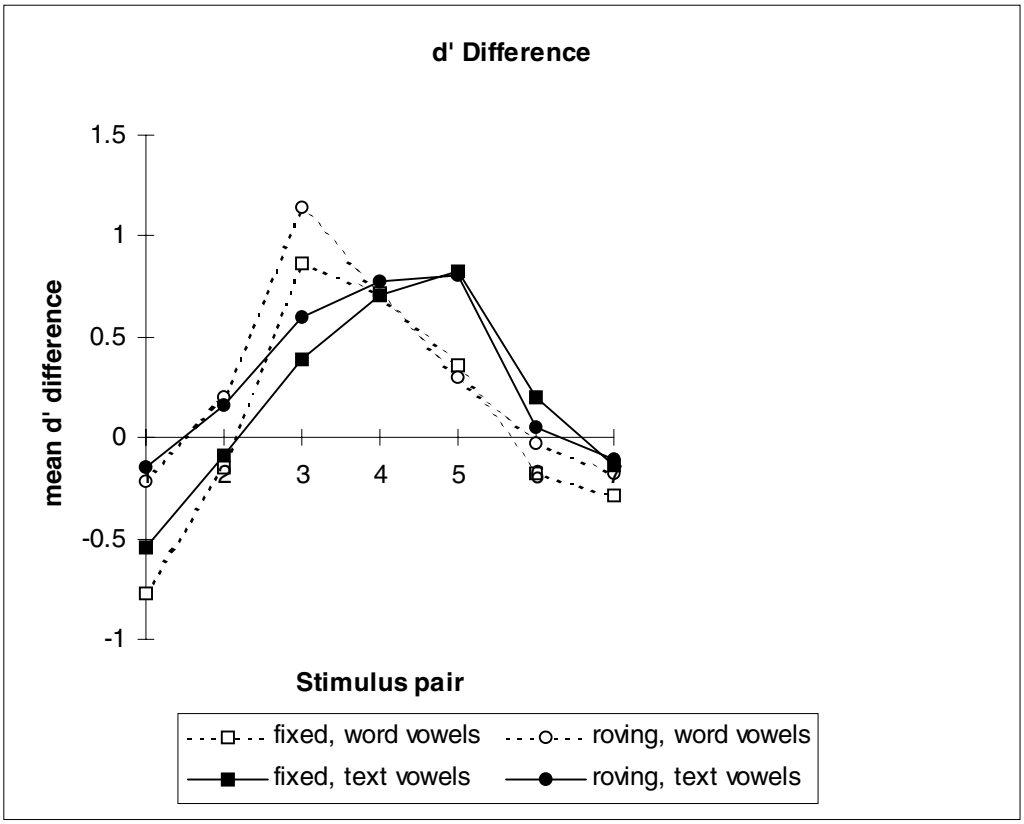

Figure 2.7 Difference scores between the obtained and predicted data for each stimulus pair in figures 2.5 and 2.6. The dashed lines represent the difference scores for the word vowels, the solid lines represent the difference scores for the text vowels.

A series of paired samples sign tests on the d'-difference scores between classification and discrimination in figure 2.7 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, but at stimulus pair 5 the opposite occurred: there it was higher for the text vowels. The d'-difference scores confirm that there is no difference in the relationship between discrimination and classification as an effect of the naturalness of the vowel stimuli. It could be that there was not enough difference between the two vowel continua to induce differences in perception, in the sense that the vowels from the rapidly read text were perhaps not sufficiently naturally.

Figures 2.5 and 2.6 show that there is not much difference in the results of the two vowel conditions. The data in figure 2.5 were expected to be less categorical than in figure 2.6 and hence the d'-difference scores in figure 2.7 to be much higher for word vowels than for text vowels. This is clearly not the case: the word vowels are not perceived less categorically than the text vowels. Neither figure shows any relationship between observed and predicted discrimination, so it can be concluded that there is no indication of categorical perception for either of the two vowel conditions. Unexpectedly, discrimination performance seems to be worse than predicted by the labelling data. The presumed ordering of the results was: fixed discrimination best, followed by roving discrimination, and then classification (Macmillan et al., 1988; Schouten & Van Hessen, 1992).

The results of a multivariate analysis of variance confirm this interpretation of the results. Fixed independent variables were Task (5 levels), Vowel Condition (2 levels), and Stimuli (7 levels: nested under Vowel Condition). Cell variance was over 19 subjects. There were significant main effects of Task and Stimulus within Vowel Condition (Task F=14.62,

p<.001; Stimulus F=5.58, p<.001). Furthermore, there was a significant interaction between Task and Stimulus within Vowel Condition (F=4.56, p<.001). Discrimination and classification performance were not significantly affected by Vowel Condition (F=0.12, p=0.731).

Separate one-way analyses per task with factor Stimulus as independent variable and post-hoc Tukey HSD tests were conducted to test for the significance of classification and discrimination peaks. There were no significant peaks in the discrimination results except for fixed discrimination, word vowel continuum, in which the d'-values of pairs 1 and 2 were significantly higher than those of the other pairs (F=3.811, p=.002). In the classification results there were significant peaks at pair 3 for the word vowels (F=14.39, p<.001), and at pairs 4 and 5 for the text vowels (F=14.24, p<.001). This confirms that there is a peak in the classification results indicating the phoneme boundary, whereas such a peak is completely absent in the discrimination results. Furthermore, discrimination scores are significantly lower than the classification scores at the phoneme boundary location (word vowels F=13.93, p<.001; text vowels F=5.07, p=.003 and F=6.17, p=.001).

In summary, these findings show that classification performance is better than discrimination, which is rather awkward since it indicates that listeners did hear differences between stimuli when classifying them, but could not hear differences between the *same* stimuli during discrimination. It seems as if listeners have applied different perceptual strategies during classification and discrimination. Listeners used a phoneme labelling strategy during classification, but apparently could not assign labels to the stimuli during discrimination. Furthermore, without labelling listeners were incapable of discriminating the stimuli, possibly because the acoustic differences between the stimuli were too small to profit from auditory trace processes. In the next sections, it will be shown that differences among subjects and results from a control experiment confirm that 4I-oddity is a purely psychoacoustic task.

2.3.1 Differences among subjects

A large proportion of the total variance was explained by cell variance (78.5%); this means that there must have been considerable differences, probably in performance level, among the 19 subjects. If our interpretation of the overall results is correct, i.e. if 4I-oddity discrimination is a purely psychoacoustic task that precludes labelling, all subjects should have one thing in common: regardless of their level of performance, none of them should show any relationship between classification and discrimination. A corollary prediction with respect to classification is that, if subjects are graded 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 see whether these predictions were correct, the subjects were divided into quartiles on the basis of the roving 4I-oddity d'-scores.

Figures 2.8 and 2.9 show the discrimination and classification results for the word vowels (the results were the same for the two vowel conditions), obtained from the highest (four subjects) and lowest quartiles (five subjects). There is a marked difference between the discrimination performance of the two subject groups (they were selected on that basis). In figure 2.8 (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.

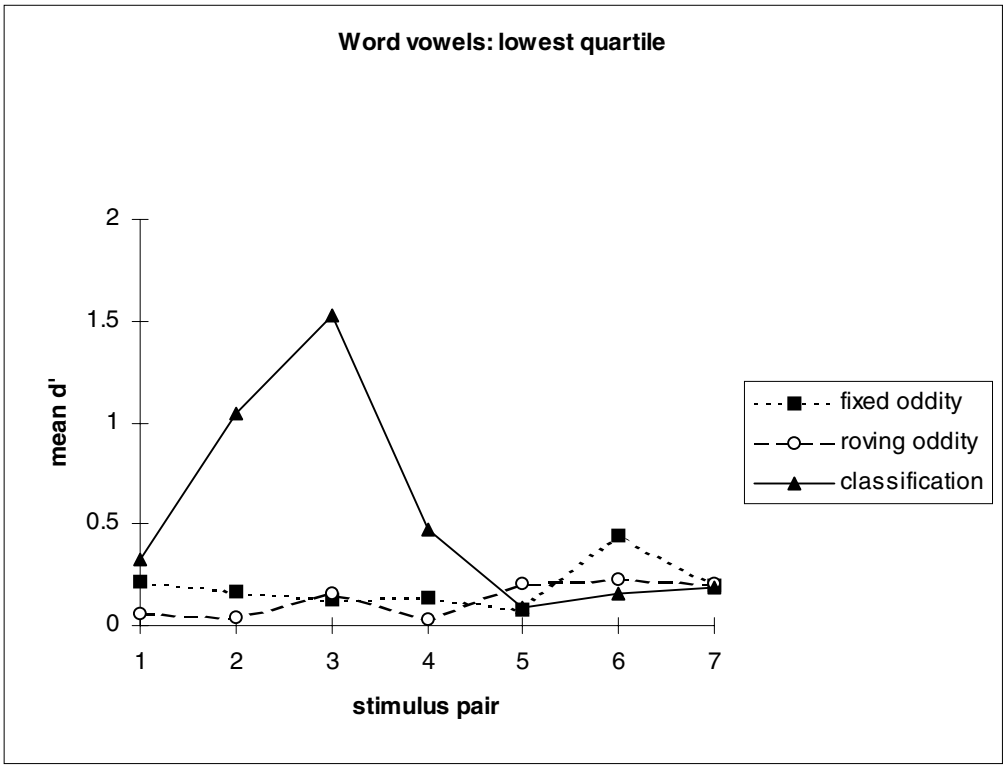

Figure 2.8. Word vowel classification and discrimination of the lowest quartile (5 subjects).

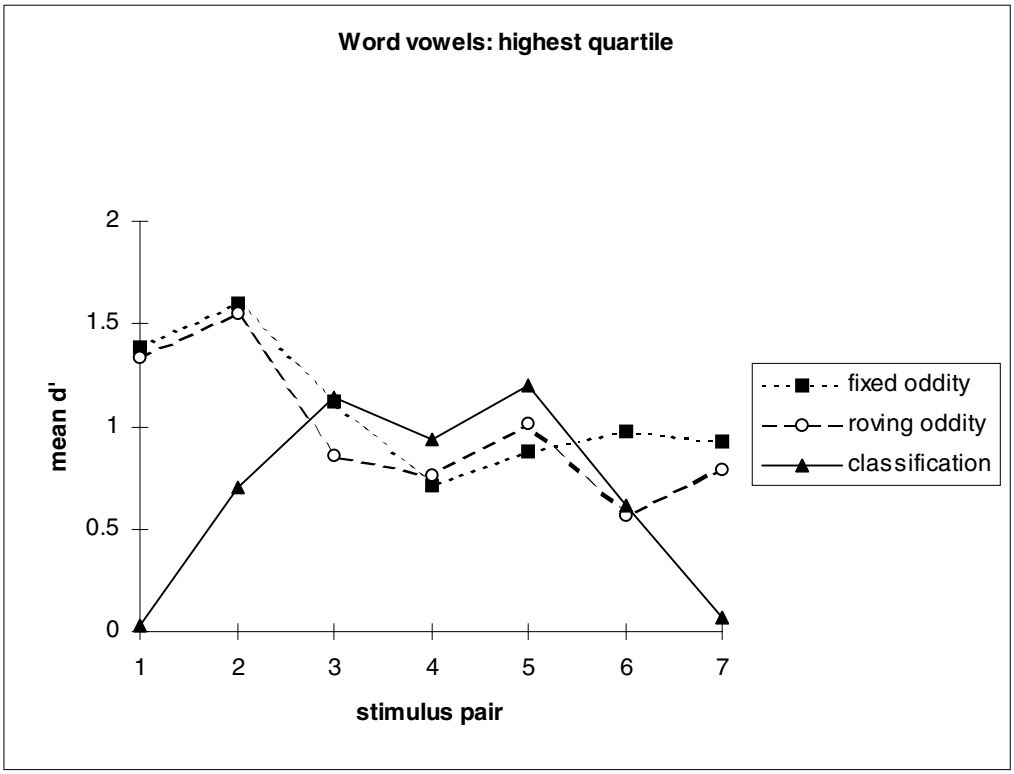

Figure 2.9. Word vowel classification and discrimination of the highest quartile (4 subjects).

The discrimination results are at chance level for all stimuli, indicating that listeners could not detect differences, not even at the phoneme boundary. In figure 2.9, which represents the performance of the subjects in the highest quartile, discrimination is in general as high as, or higher than, classification. The slightly lower d' score at pair 4 is caused by considerable differences in the position of the phoneme boundary (as is often the case with vowel stimuli).

The results for the lowest and highest quartiles show that in neither case does discrimination performance show any relationship with classification performance, which confirms our prediction. This means that, during discrimination, no phonetic information was used: listeners were in the psychoacoustic, or trace, mode. In classification, however, all listeners had to operate in the speech mode and showed the same performance.

These results were confirmed by separate analyses of variance on the data of each task: the effect of the factor Quartiles (4 levels) on discrimination performance was significant (fixed 4I-oddity: $F=9.09$, $p<.001$; roving 4I-oddity: $F=27.60$, $p<.001$), whereas there was no significant effect of Quartiles on classification performance.

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 psychoacoustic mode, however, they show great differences in performance, sorting 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 different scores if they are operating in a trace mode: some subjects are more ‘analytical’ than others (Best, 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 discrimination as if our stimuli had nothing to do with speech at all.

2.3.2 Control experiment

In the previous section it was shown that discrimination performance for the highest quartile was equal to or better than predicted by the labelling data. However, the highest quartile only contained 4 of the 19 subjects. The discrimination results for the subjects in the other quartiles were worse. This indicates that for these subjects the discrimination task might have been too difficult and therefore that their 4I-oddity results were an artefact of the experiment. To test this hypothesis a control experiment was conducted. In this control experiment the physical distance between the stimuli within trials was changed from one step to two steps along the continuum.

The stimuli used in the control experiment were those from the ‘word vowel continuum’. The subjects were 14 students from the pool of 19 subjects who had participated in the one-step experiment. Unfortunately, 5 students were no longer available for testing. The design consisted of only one task: roving two-step oddity discrimination. In this task 6 stimulus pairs were presented 64 times, following the same procedure as in the previous experiment.

We hypothesised that the larger physical distance between the stimuli would induce a general increase of the discrimination results of all listeners, thus including the subjects in the lowest quartile. Results were not expected to change in the degree of categoricalness, since the results of the highest quartile predict that even if subjects can discriminate between the stimuli, there is no relationship between discrimination and classification performance. The results of two-step discrimination and classification are presented in figure 2.10. The

classification results are taken from the previous experiment (minus the data of the five subjects that did not participate in the two-step discrimination task), and recalculated to predict a two-step comparison.

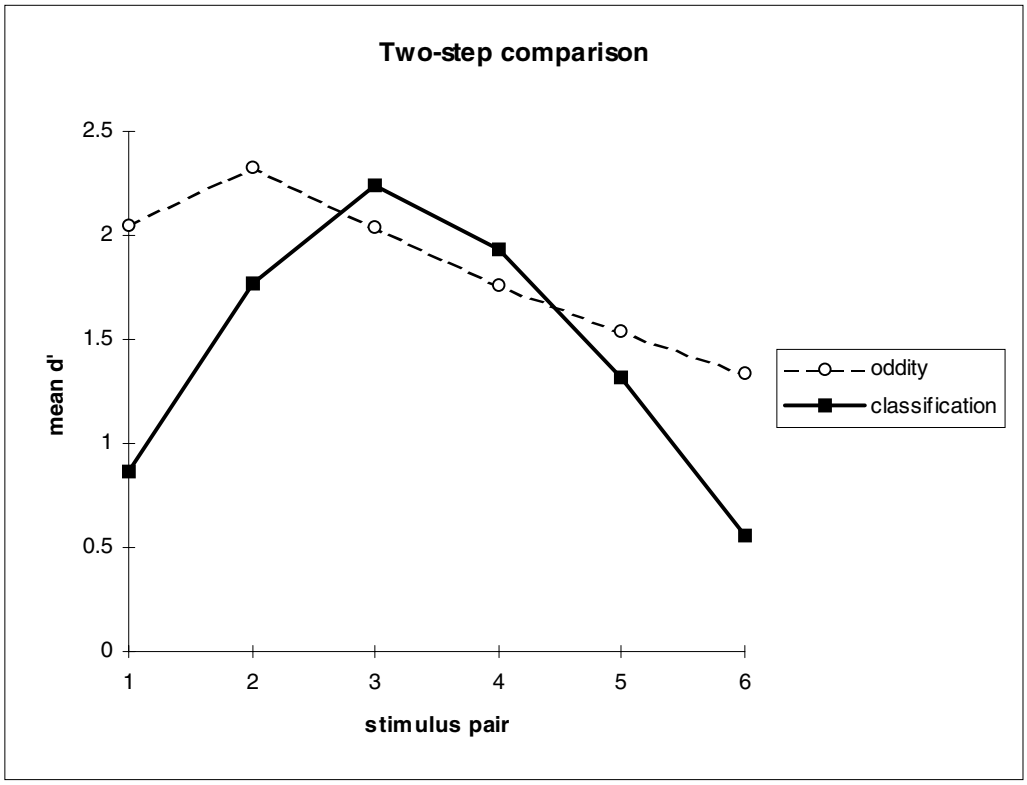

Figure 2.10 Classification and two-step roving discrimination results for the word vowels.

As expected, two-step discrimination generated higher scores than one-step discrimination (see figure 2.5), and the 4I-oddity scores are no longer worse than predicted by the classification data. But again, there is no relation between 4I-oddity discrimination and classification and thus no indication of categorical perception.

The effect of the physical distance between the stimuli was tested with a two-way analysis of variance on the two-step data, combined with the data from the one-step experiment (see figure 2.5, but again the data of the five subjects that did not participate in the control experiment were excluded). The independent variables were Step-size (2 levels) and Task within Step-size (2 levels). Both effects turned out to be significant (Step-size $F=142.2$, $p<.001$; Task within Step-size $F=6.41$, $p=.002$). A one-way analysis of variance per task with Stimulus as independent variable and post hoc Tukey HSD tests revealed a significant peak in two-step Classification at stimulus pair 3, at the phoneme boundary ($F=2.94$, $p=.02$ and $F=5.82$, $p<.001$). There was no significant peak in the two-step discrimination results.

Summarising the control condition, it can be concluded that the counterintuitive 4I-oddity discrimination results are not caused by too small a step size between the stimuli. Even with a larger step size, discrimination results show no relationship with the classification data. Again the interpretation of the results must be that different perceptual processes were used for these tasks: an auditory trace strategy during discrimination and phonetic labelling during classification.

2.4 Discussion and conclusion

In a series of experiments the perception of word vowels was compared with that of text vowels. It was expected that the temporal reduction and spectral complexity of the latter would result in a greater dependence on labelling and hence in more categorical perception. However, our results did not confirm this expectation, mainly because there was no relationship between observed and predicted discrimination, and thus no indication of categorical perception for either of the two vowel conditions.

Discrimination performance 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 predicted - see, e.g., Macmillan et al., 1988; Pastore, 1987; Repp, 1984; Schouten & Van Hessen, 1992) could not be explained away as being due to the unprecedented difficulty of the task: a greater physical spacing between stimuli in a discrimination trial did not make any significant difference.

A first attempt to account for the traditional difference (discrimination better than predicted by the classification data) was made in the dual-process model for the discrimination of speech stimuli by Fujisaki and Kawashima (1971). This model explicitly distinguishes between categorical phonemic judgements and judgements based on auditory memory for acoustic stimulus attributes. The authors propose 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 performance that can be predicted from labelling probabilities is attributed to categorical judgements, whereas the remainder (the deviation from ideal categorical perception) is assigned to comparison of acoustic stimulus properties (Repp, 1984).

The selection of a discrimination paradigm was led by the idea that a discrimination task in which trace 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 labelling strategy for discrimination and results will *inevitably* be highly categorical. The task was even more successful than expected in stimulating a trace coding strategy: the results show that there was no categorical perception at all. Moreover, discrimination performance was lower than predicted by 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 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 relationship with classification performance. Similar results were found for all subjects in the control experiment with a larger physical distance, which made discrimination easier but did not lead to a stronger relationship between discrimination and classification results. This means that, during discrimination, no phonetic information was used: listeners were in the psychoacoustic, or trace, mode. In classification, however, all listeners had to operate in the phonetic mode and showed the same behaviour. Why is it that 4I-oddity puts subjects into the psychoacoustic mode, whereas all other speech discrimination tasks that have been used up to now produce at least a mixture of the two modes, so that results always turn out categorical to some extent?

It is suspected that this is due to the nature of the oddity task, which is less close to the traditional 2IFC task than was expected, and closer to 4IAX. The crucial difference here is whether a subject's decision is influenced by a criterion, or bias, that is external to the stimuli. In 2IFC, in which a subject has to indicate the order of two stimuli, this can only be done with reference to criteria that are external to the stimuli, such as phoneme boundaries or categories such as "high" vs. "low" in psychoacoustic experiments. However, the oddity results were more like those that may be expected from 4IAX, 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 one is the greater one, a decision that can only be taken on the basis of stimulus information and not on the basis of a criterion along a scale. This is why 4IAX is generally regarded as bias-free in psychophysics.

Subjects' comments, and our own experience when performing the oddity task, confirm the similarity between 4I-oddity and 4IAX, at least with vowel stimuli: it soon became apparent to subjects that attempts to label the stimuli did not work, so they gave up the attempt and apparently applied a strategy in which differences between two pairs of stimuli were used in order to reach a decision. As a result, the discrimination experiment turned out to be one in which phoneme categories did not play any role whatsoever. This was unexpected: even though Heller and Trahiotis (1995) have suggested that subjects might treat 4I-oddity as a variant of 4IAX, it still was not foreseen that there would be a complete absence of phonetic information in the discrimination response. 4IAX may be bias-free, but phoneme labels were expected to be inevitable: a human subject should not be able to treat speech as if it does not consist of phonemes or words. This is wrong: it *is* possible to get subjects completely into the psychoacoustic trace mode when they have to discriminate speech sounds.

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. The present results suggest that this assumption rests crucially on the use of biased discrimination tasks, in which listeners are compelled to use subjective decision criteria. When a bias-free discrimination task is used, no resemblance between phoneme classification and discrimination is found at all. This indicates that categorical perception is highly task dependent. In the next chapter, the hypothesis that the degree of categorical perception varies as an effect of the discrimination task and of the duration of the inter-stimulus interval was explicitly tested.