Chapter 4

Categorical perception of stop consonants and vowels

4.1 Introduction

In the preceding chapters we have shown that the degree of categorical perception of vowels varies as a function of the discrimination task. Perception of vowels can be very categorical, but it can also be completely non-categorical. We conclude from these results that an unvarying relationship between discrimination and classification of speech sounds does not exist.

However, supporters of the categorical perception model may find the observation that vowel discrimination varies as a function of the task not very convincing as an argument for rejecting their model. They may argue that it is not unusual for the degree of categorical perception to vary with vowels (e.g. Pisoni, 1975). Furthermore, they may claim that only consonant stimuli are perceived categorically, not vowels. Although the results with 2IFC discrimination (chapter 3) have clearly shown that there can be a relationship between observed and predicted vowel discrimination, it would be useful to substantiate our conclusion with data from a different phoneme contrast.

Categorical perception of stop consonants is often cited as a robust phenomenon, which has been found in many settings (see several reviews in Harnad, 1987). Even Massaro (1994), who is one of the prominent sceptics with respect to categorical perception, reluctantly admits that "there may be a few phoneme contrasts that qualify for a weak form of categorical perception." The few phoneme contrasts Massaro (1994) refers to are stop-consonant ones.

To explain the high degree of categorical perception for stop consonants, Liberman, Mattingly and Turvey (1972) suggested that the listener has no auditory image of the signal available, but only the output of a specialised processor that has stripped the signal of all normal sensory information and represents each phonetic segment categorically. This claim

has motivated researchers to prove the opposite, which is that listeners do hear differences between within-category stop consonant stimuli.

Firstly, reaction-time measurements have been used to show that listeners are sensitive to differences between within-stop-category stimuli (Massaro, 1987; Pisoni & Tash, 1974; Repp, 1981). According to the categorical perception model, listeners cannot discriminate differences within a category and therefore there is no basis for differences in reaction times to different stimuli. However, in Pisoni and Tash (1974) it was shown that the time it took subjects to respond "Same" to different within-category pairs of stimuli was significantly longer than the time to respond "Same" to pairs in which the stimuli were actually identical. This finding suggests that the auditory system did detect some differences between sounds. In Massaro (1987), reaction times were measured for classification of a stop consonant place continuum. Massaro's results revealed an increase in reaction times near the category boundary, which implies that stimuli within a stop-consonant category varied in ambiguity or the degree to which they represented the category. However, Samuel (1979) and Pisoni and Tash (1974) have failed to find different labelling reaction times for stop-consonant stimuli that were strictly within the same category.

Secondly, rating experiments have shown that listeners give different ratings to within-category stimuli. Apparently, they feel that these stimuli differ in the degree to which they represent the phoneme category (Massaro & Cohen, 1983).

Thirdly, it has been demonstrated that, although certain speech continua were perceived categorically by naive subjects, after discrimination training with the given continuum listeners were able to resolve very fine distinctions between within-category pairs (Carney, Widin & Viemeister, 1977; Pisoni et al., 1982; Samuel, 1977; Werker & Logan, 1985).

The rating, reaction-time and training studies demonstrate that, to some extent, listeners have access to auditory differences between within-stop-category stimuli. As for the training studies, however, there have also been several experiments in which extra training failed to improve within-stop-category discrimination (see review Strange & Jenkins, 1978). Furthermore, the extensive training and the use of highly practiced subjects is an important drawback of the training studies, since it does not show what listeners typically do in everyday speech perception. For instance, in Carney et al. (1977) three subjects, all with extensive experience in previous psychophysical experiments, were tested in more than a dozen sessions distributed over several weeks. It would be interesting to see if within-category discrimination could be improved without using highly experienced listeners and without numerous training sessions.

Only Pisoni and Lazarus (1974) have tried to demonstrate that within-category perception of stops can vary with tests that are used in the traditional categorical perception framework. They tested the perception of stimuli within a voice-onset continuum by comparing a classification task with ABX and 4IAX discrimination tasks. The results showed that both ABX and 4IAX discrimination were consistently better than predicted discrimination, although the shape of the observed and predicted functions was fairly similar. The level of within-category discrimination was slightly higher under the 4IAX test than under the ABX test, but only for two-step comparisons and not for one-step comparisons. At the peaks in the functions, performance in both discrimination tasks was comparable. According to Pisoni and Lazarus (1974), the two-step 4IAX results imply that listeners can discriminate between within-category stimuli in a non-phonetic mode. However, this study also included a condition in which the subjects were given extra training before performing the discrimination tests. This extra experience with the stimuli appears to have been the critical factor in the results, for Pisoni and Glanzman (1974) failed to find any difference between stop consonant

discrimination tested with the ABX or 4IAX tasks when no pretraining was provided.

The aim of the present study was, firstly, to see if degree of categorical perception of stops could be varied in a design similar to the one used by Pisoni and colleagues and, secondly, to compare stop-consonant perception with vowel perception. We compared classification performance with discrimination performance in two different discrimination tasks: 2IFC and 4IAX discrimination. By using 2IFC instead of ABX, we expected to find more differentiation between the discrimination results than Pisoni did. We used 4IAX instead of 4I-oddity since both tasks seem to be sensitive to the acoustic differences between stimuli, but 4IAX has the advantage of having a standard signal detection model with which d' can be calculated.

Our previous experiments have shown that, in 2IFC discrimination, listeners are strongly encouraged to use their internal phoneme representations to make a decision about the order of the stimuli, whereas 4IAX discrimination is mainly based on an auditory trace strategy (e.g. Pisoni, 1975; Schouten & Van Hessen, 1992). Thus, for 2IFC, observed discrimination is expected to be highly predictable from classification. For the 4IAX results we predict that observed discrimination will be less clearly related to the classification data.

Furthermore, we hypothesise not only that stop consonant discrimination will vary as a function of the discrimination task, just as was the case for vowels, but also that the stops will be perceived more categorically than the vowels, as has been shown by numerous previous studies (e.g. Fry et al., 1962; Pisoni, 1973; Repp, 1981; Stevens et al., 1969). Thus, although the relationship between observed and predicted discrimination may be different for the two discrimination tasks, there will always be a stronger relationship between the results for stop consonants than for vowels.

4.2 Method

4.2.1 Stimulus material

The speech material in this experiment consisted of seven stimuli in a stop consonant /pAk/-/tAk/ continuum and eight stimuli in a vowel /pup/-/pip/ continuum. The stop-consonant and vowel stimuli sounded like completely natural speech signals.

The /pAk/ - /tAk/ stop consonant continuum was similar to the one used in the experiments described by Van Hessen and Schouten (1999). The generation of the stop-consonant continuum was done with linear interpolation between the relative amplitudes of the spectral envelopes of /pAk/ and /tAk/ according to the method described in chapter 2. The original words were produced by a male native speaker of Dutch. Source spectra and filter of the original words were determined and separated frame-by-frame. The frames were 25.6 ms long and the frame shift was 6.4 ms, resulting in a great deal of frame-to frame overlap. Source spectra and spectral envelopes were described by means of the phases and amplitudes of 100 spectral components. The interpolation between the original stop consonants consisted of six steps and consequently resulted in seven stimuli.

The vowel stimuli were the ‘text vowels’ used in the experiments of chapter 2 (see section 2.2.1) and consisted of a 7-step continuum between /pup/ and /pip/. The original words had been read aloud in a written text at a fast speech rate by a male native speaker of Dutch.

4.2.2 Subjects

Twelve listeners took part in the experiments. In the vowel experiment, one subject's data were ignored because his responses turned out to consist largely of missing values, probably due to computer failure. Thus eventually there were 12 listeners for the stop consonant tests, but only 11 for the vowel tests. The subjects were all native speakers of Dutch who were paid a basic rate plus a reward depending on the number of correct responses they gave. The listeners in the current experiments were different from those in the previous experiments, since most subjects of the initial subject pool were no longer available by the time this test took place.

4.2.3 Design

All subjects took part in three different tasks on each of the two stimulus continua: fixed and roving 2IFC discrimination, fixed and roving 4IAX discrimination and classification. In one series of tests, the stop continuum between /pAk/ and /tAk/ was presented and in another series of tests the stimulus material consisted of the vowel continuum between /pup/ and /pip/. The order of presentation of the continua and the tests was balanced over subjects.

In 2IFC discrimination all trials consisted of two different stimuli, either AB or BA. The subjects responded by indicating the order in which the stimuli were presented. The inter-stimulus interval was 300 ms and response time was 1200 ms. In the fixed context each pair was presented repeatedly in one block of trials. For the vowel stimuli there were 7 blocks, and for the stops there were 6 blocks, one block for each pair. Each block contained 32 trials, 16 for each of the two possible combinations. In the roving discrimination experiment, the A and B stimuli to be discriminated were chosen randomly from the total range of stimuli and thus varied from trial to trial. In the roving 2IFC test for vowels, 7 (stimulus pairs) x 32 trials were presented, and in roving 2IFC for stop consonants, 6 (stimulus pairs) x 32 trials were presented.

In 4IAX discrimination, each trial consisted of two stimulus pairs, one of two identical stimuli and the other of two different stimuli; AB-AA, BA-AA, AA-BA, AA-AB, BA-BB, AB-BB, BB-AB, BB-BA. These 8 combinations were presented 4 times in random order. In the 4IAX test, subjects reported which pair of stimuli was the same, the first pair or the second pair. ISI was 300 ms, the interval between pairs was 600 ms and response time was 1200 ms. In the fixed context, all 32 trials of one, randomly chosen, stimulus pair were given before the next stimulus pair was presented. For the vowel stimuli there were 7 fixed blocks, and for the stops there were 6 blocks. In the roving context stimulus presentation was random.

Classification involved a forced choice between two alternatives, the vowels /u/ and /i/ or the stop consonants /p/ and /t/. There were separate tests for each of these two continua. All stimuli were presented to subjects 32 times in completely random order. Response time was 1200 ms.

4.2.4 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. Subjects were paid two cents for each correct response and had the same amount deducted for each

mistake (this meant that random respondents were paid exactly the basic hourly rate).

The subjects responded by pressing the mouse button on one of two response fields on a computer screen. In 2IFC, the response fields were labelled “ie-oe” or “oe-ie” (/ι-υ/ or /υ-ι/) for the experiment with the vowel stimuli. In the stop consonant continuum the 2IFC labels referred to the stop contrast. In 4IAX, the response labels were “pair 1” or “pair 2”. During discrimination listeners received visual feedback (500 ms) of the correct answer after each response. Discrimination training consisted of 56 trials for the stop consonant continuum and of 64 trials for the vowel continuum. Training was intended to familiarise subjects with their task.

In classification, one stimulus was played on each trial, and the subject had to identify it by mouse-clicking on one of two response fields (phoneme labels) on a computer screen. The only classification training consisted of 14 (stops) or 16 (vowels) trials. No feedback about correct responses was given.

4.3 Results

The results of the experiments with stop consonant and vowels are presented in sections 4.3.1 and 4.3.2. The figures display the listeners’ average observed and predicted discrimination performance. The results of the classification tests are presented as predicted discrimination scores. The 2IFC and 4IAX discrimination d' scores were calculated by subtracting z(FA) from z(H) and then transforming these scores into ‘true’ d'-values (Kaplan, Macmillan & Creelman, 1978; Macmillan & Creelman, 1991; Macmillan, Kaplan & Creelman, 1977). The values of p(H) and p(FA) were limited to the 0.99 to 0.01 range, which meant that the maximum d'-value that could be obtained was 3.29 for 2IFC. For 4IAX it was 5.14, and the maximum d'-value for classification was 4.65.

4.3.1 Stop-consonant results

The mean stop-consonant results are displayed in figure 4.1. The stop-consonant stimuli in pair 1 resemble /p/ and the stimuli in pair 6 sound like /t/. The bold line represents predicted discrimination/classification, the dotted lines represent the 2IFC discrimination results and the thin solid lines are the 4IAX results.

Figure 4.1 leads to the following observations: (1) there is no strict relationship between the results of either of the two discrimination tasks and the classification results; (2) fixed 2IFC contains a small peak but is very low, even lower than roving 2IFC. The roving 2IFC discrimination function appears to contain a discrimination peak which roughly coincides with the classification peak, and within-category discrimination is similar to classification performance; (3) the shapes of the 4IAX discrimination functions are unrelated to the classification function and within-category discrimination is much better than predicted. Fixed 4IAX seems to have two small discrimination peaks, the second of which coincides with the peak in 2IFC discrimination. However, it is very likely that these fixed 4IAX peaks are not significant.

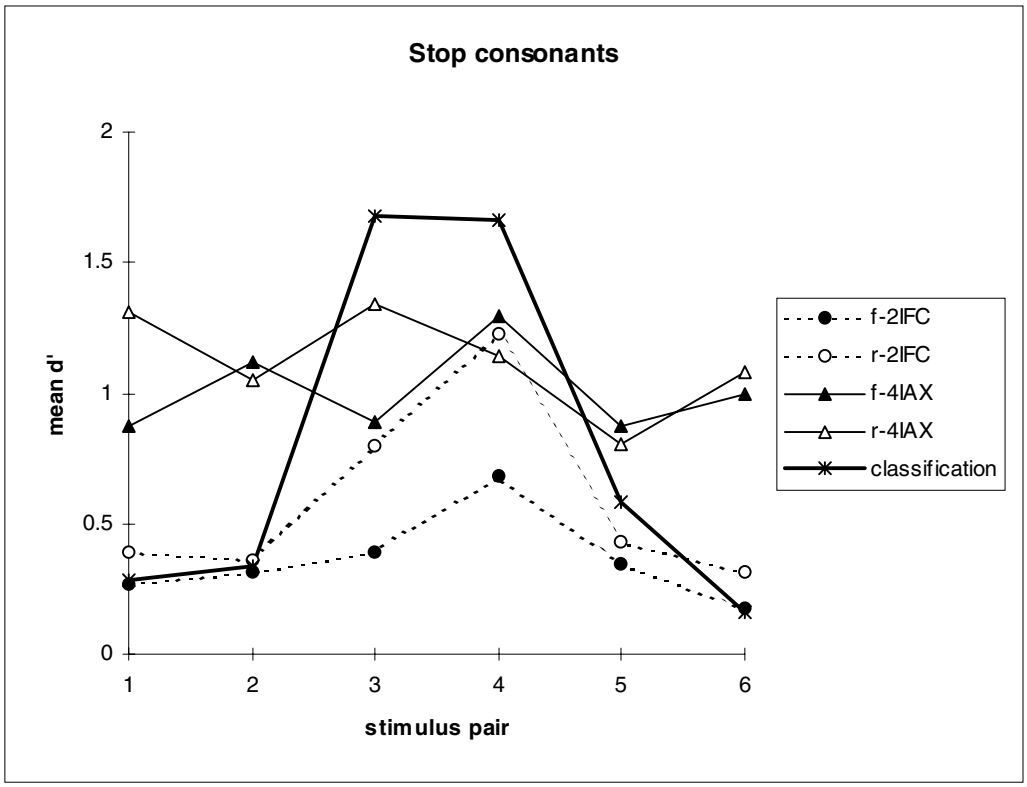

Figure 4.1. Classification and 2IFC and 4IAX discrimination results for the stop consonants.

Statistical analysis

A two-way analysis of variance was performed on the observed and predicted discrimination data. The two main effects under consideration were: Task (5 levels) and Stimulus (7 levels). Cell variance was over 12 subjects. The results of the analysis showed a significant difference between the tasks, F=32.87, p<.001. Discrimination performance was always better with the 4IAX tests than with the 2IFC tests or than predicted by classification. The main effect for Stimulus was also significant, F=20.97, p<.001. Furthermore, the Task by Stimulus interaction was significant, F=5.05, p<.001, indicating that the discrimination behaviour of the listeners varied as a function of the task at different points along the stimulus continuum.

To test for the significance of classification and discrimination peaks, separate one-way analyses and post-hoc Tukey HSD tests were conducted on the data of each task, with Stimulus as independent variable. There were no significant peaks in the fixed and roving 4IAX results. However, the analyses did reveal a significant peak at pair 4 for fixed and roving 2IFC discrimination (F=3.24, p=.011 and F=9.43, p<.001) and also at pairs 3 and 4 for classification (F=29.22, p<.001). The 2IFC results imply that listeners tried to use phoneme labels to discriminate the stimuli, but the results in figure 4.1 show that they were not completely successful in doing so, because at the discrimination peak 2IFC discrimination is lower than predicted by the labelling data. At pair 3, both fixed and roving 2IFC are significantly lower than predicted discrimination, F=14.27, p<.001. At pair 4 only fixed 2IFC is lower, F=3.319, p=.017. The data show that fixed 2IFC is worse than roving 2IFC discrimination. However, listeners performance is usually better in the fixed than in the roving context. We will return to this issue in the discussion of the stop-consonant results, section 4.3.1.2.

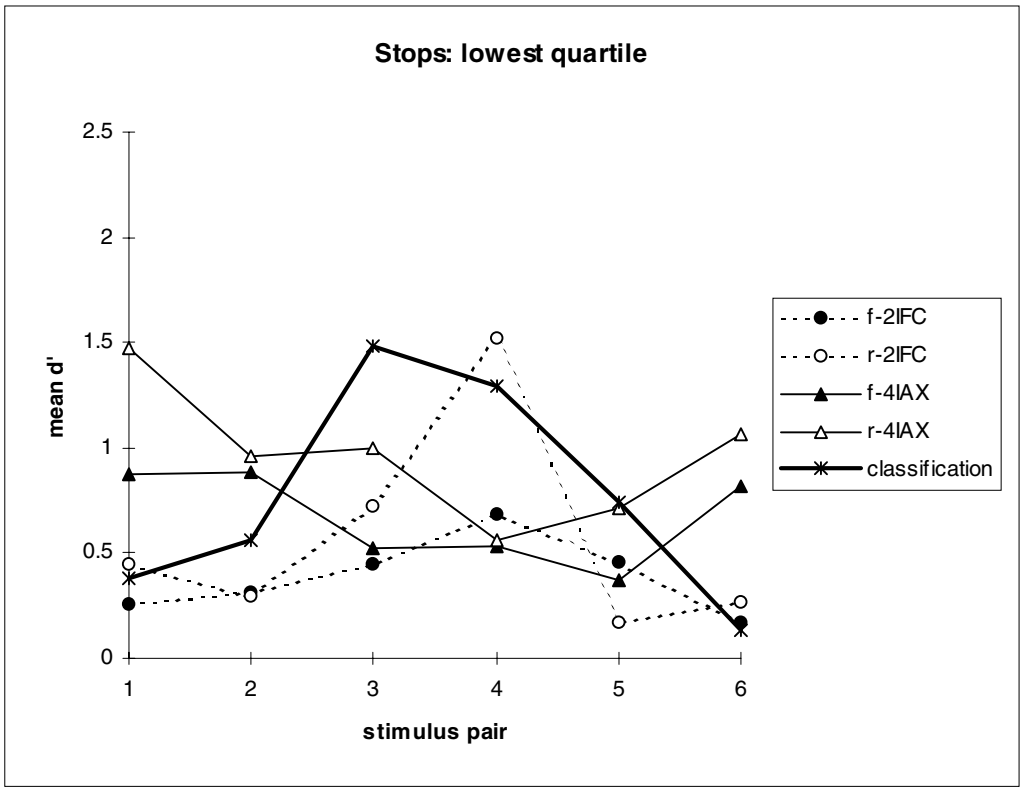

Figure 4.2. Stop-consonant classification and 2IFC and 4IAX discrimination results of the lowest quartile (3 subjects).

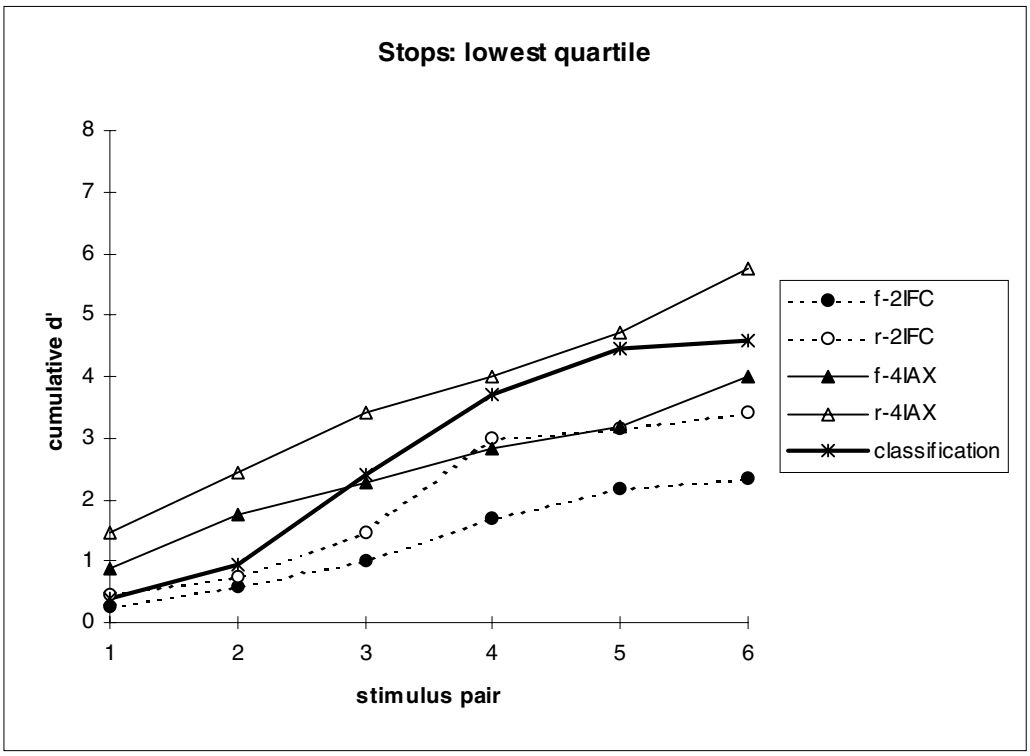

Figure 4.3. Cumulative stop-consonant classification and discrimination of the lowest quartile.

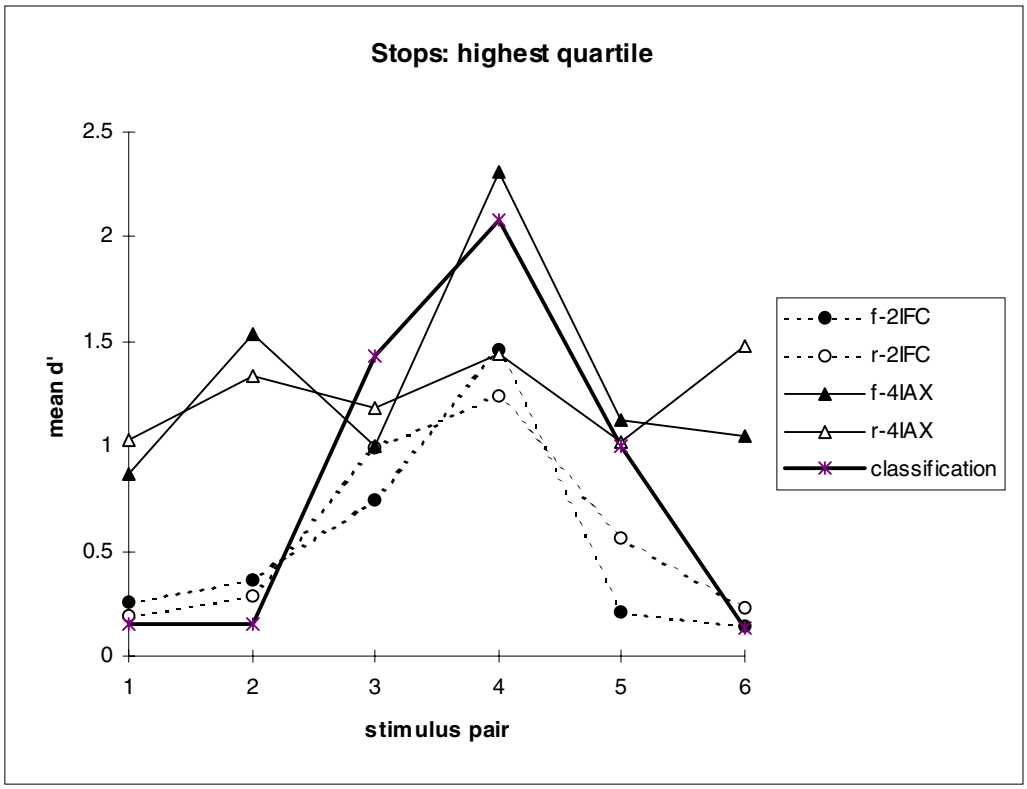

Figure 4.4. Stop-consonant classification and 2IFC and 4IAX discrimination results of the highest quartile (3 subjects).

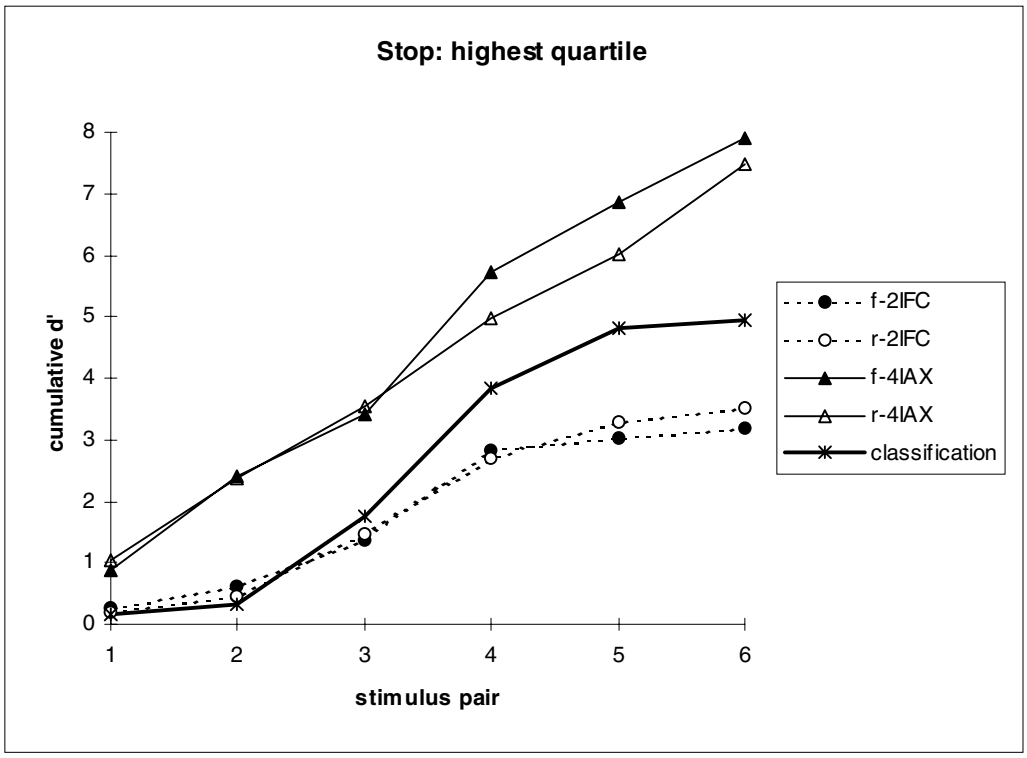

Figure 4.5. Cumulative stop-consonant classification and discrimination of the highest quartile.

Post hoc Tukey-HSD analysis of Task at each stimulus pair showed that fixed and roving 4IAX were significantly higher at all pairs except at pair 3, at which roving 4IAX and Classification were highest, and at pair 4, at which Classification was significantly higher than all other tasks. This indicates that within-phonetic-category comparisons benefited more from the 4IAX procedure than across-category comparisons, relative to the predicted discrimination performance.

4.3.1.1 Differences among subjects

In the stop-consonant experiment 54% of the total variance was explained by cell variance. As in the previous experiments, we took a closer look at listeners' performance by dividing the pool of subjects into quartiles. This time the criterion was based on fixed and roving 4IAX discrimination. Figures 4.2. and 4.4 display the mean results of the three subjects in the lowest quartile and the highest quartile, respectively. To get a better view of the differences between performance of the quartiles, the cumulative d' values are also presented (figures 4.3 and 4.5). The cumulative d' scores, which were calculated simply by adding up the d' values from left to right, give an indication of the overall level of performance as a function of task.

It has to be noted that there were only 3 subjects in each quartile, which is also the reason why we ran only one statistical analysis on this subset of data. The effect of Quartiles (4 levels) was tested on the data per task and was only significant for fixed 4IAX ($F=4.62$, $p<.005$). This is confirmed by the cumulative functions in figures 4.3 and 4.5 in which fixed 4IAX is generally worse than predicted in the lowest quartile, but much better than predicted in the highest quartile.

Figure 4.2 shows the averaged results of the three subjects in the lowest quartile. However, the results of each of the three subjects within this quartile differed considerably from this average pattern. One of the subjects showed a close relationship between fixed and roving 2IFC and classification. The second subject had results that are similar to the average pattern: roving 2IFC was very similar to classification. The discrimination results of the third subject were low across the board.

From figure 4.4, which displays the results of the highest quartile, a very different pattern emerges. However, the separate results of each of the three subjects within this quartile were also very different from the average pattern. One subject showed a much better than predicted peak in the fixed 4IAX results at pair 4, but had low overall results for the other three discrimination tests. The second subject showed a flat shape for 4IAX discrimination and a close relationship between both fixed and roving 2IFC and classification. The third subject had two marked peaks in fixed 4IAX (the first of which was at pair 2) and a peak in roving 4IAX, which roughly coincided with the classification peak. This subject's 2IFC was low.

It seems better, therefore, to focus on the differences and similarities between the lowest and highest quartiles, than to try to interpret the peaks and valleys in the separate figures. There was only a difference between quartiles in tasks in which the listeners were expected to be encouraged to use auditory trace coding. Performance in phoneme labelling tasks, i.e. classification and 2IFC discrimination, was not significantly different.

4.3.1.2 Discussion stop-consonant results

The question we addressed in this experiment was if the degree of categorical perception for stop consonants would vary as an effect of the discrimination task. The results show that this is indeed the case: the degree of categorical perception for stops varies as a function of 2IFC or 4IAX discrimination. However, there is no strict relationship between classification and either 2IFC or 4IAX discrimination. Nevertheless, it is evident that 2IFC discrimination is more closely related to classification than is 4IAX discrimination. Thus, as we expected, responses are more categorical for 2IFC, since this task presumably encourages listeners to use a phoneme labelling strategy during discrimination. On the other hand, perception is less categorical for 4IAX discrimination, because this task is more sensitive to the acoustic differences between the stimuli.

However, this interpretation of the 2IFC results only accounts for the roving data. In the fixed 2IFC results we see d' values that are worse than the values for roving 2IFC. Usually, fixed context results are better than roving context results, regardless of the discrimination task (e.g. Schouten & Van Hessen, 1992; Van Hessen & Schouten, 1999). In fixed discrimination, listeners hear the same stimuli over and over again and thus hear only a small subset from the total range of stimuli in the continuum (see also chapter 2, section 2.1.5). It was expected that, as a consequence of the small range, stimulus uncertainty would be lower in fixed than in roving discrimination. This leads to a higher degree of labelling in the fixed condition. According to Durlach and Braida (1961) listeners will construct a temporary internal representation of the range of stimuli presented in a particular experiment, which in the case of speech sounds may coincide with phoneme categories in long-term memory. If the temporary labels represent phoneme categories, the results will represent the effect of phoneme labelling processes.

The present fixed 2IFC results are not in accordance with our expectations. It appears that listeners were unable to use any labelling at all. We can only speculate about the causes of this unusual behaviour, particularly since a labelling strategy *was* followed in roving 2IFC discrimination. Apparently, applying the two phoneme labels was easier when the full stimulus range was available than when only two very similar stimuli were presented from trial to trial so that listeners did not regularly hear clear exemplars from both categories.

However, it has to be noted that there was a large variation within the fixed 2IFC results due to individual differences. For 6 of the 12 subjects, fixed 2IFC was worse than roving 2IFC. For 4 subjects, fixed and roving 2IFC performance was similar and two subjects had better fixed than roving 2IFC. We have to conclude, therefore, that the use of labelling is not only dependent on the discrimination task itself, but also on the stimulus context and on the listener.

One could also argue that the interpretation of roving 2IFC is incorrect, because the scores are too low at the phoneme boundary (worse than predicted). A hypothetical explanation is given by Schouten and Van Hessen (1992), who found similar results with almost identical stimuli. The low 2IFC results motivated them to propose a new model for the calculation of d' for 2IFC discrimination. If we had used the model by Schouten and Van Hessen (1992), the present 2IFC results would have been $\sqrt{2}$ times as high, and this would have resulted in a very close relationship between observed and predicted discrimination and a high degree of categorical perception. In fact, the results would have been identical to those found for stop consonants by Schouten and Van Hessen (1992). However, we used the standard model by Macmillan et al. (1977). We will elaborate on the two models to explain their differences and our choice of the standard model.

In 2IFC discrimination, two different stimuli are presented and the task of the listener

is to determine the order of the stimuli (AB or BA). According to Macmillan et al. (1977), the optimal strategy for the listener is to subtract the first stimulus from the second. Each of the two stimuli has, by definition, a normal distribution with a variance of 1. The subtraction of the two stimuli results in two new normal distributions with means that are equal to the differences (A-B and B-A) of the two original means, and variances that are equal to the sums of the two original variances. On the basis of the two response categories (AB and BA), the listener has to create an internal decision criterion. If the difference between the auditory traces surpasses this criterion, the listener decides to choose the order BA, if not, the order is AB.

According to Schouten and Van Hessen (1992), 2IFC discrimination of speech stimuli involves different decision processes than those proposed in the standard model, which is based on 2IFC discrimination of meaningless psychoacoustic stimuli like pure tones. In their view, a listener cannot decide about the order of speech stimuli, without assigning a 'quality code' to the stimuli first. This quality code is necessary, because differences between speech sounds are, unlike differences between pitch, duration and intensity, dependent on the phoneme categories they belong to. The quality code contains the phoneme label of the stimulus, and an estimate of the distance between this stimulus and its phoneme category. If the second stimulus receives the same phoneme label as the first one, a second quality code is created. After this coding, the order of the stimuli is determined in the way described for the standard model. This means that an extra step is added to the listeners' decision strategy in the new model. As a consequence of this extra step, the comparison of stimulus and phoneme category, an extra source of variance is introduced. The variance of the phoneme category is estimated to be identical to the variance of the stimulus. As a result, a different calculation of d' for 2IFC discrimination is proposed, resulting in a 'true' d' that is $\sqrt{2}$ times as high as d' calculated according to the standard model. However, an important factor of dispute is the estimated variance of the phoneme category. If phoneme category representations in long term memory are perfect, variance equals 0. Because it is likely that the representations are less than perfect, the variance is expected to be higher than 0. However, it is unclear how much higher it really is. Assuming that the variance is 1 is a rather arbitrary choice; this is the reason why we preferred the standard model instead of risking that, in absolute terms, our 2IFC values would be too high.

The fact remains that, calculated either according to the standard or to the new model, our stop-consonant 2IFC results are almost identical to those found by Schouten and Van Hessen (1992).

As for the fixed and roving 4IAX data, there appeared to be large between-subject variation around the phoneme boundary location. In some of the results, fixed and roving 4IAX contained a marked discrimination peak, which coincided with the peak in classification. This peak was present in the fixed condition for 4 of the 12 subjects and in the roving condition for 2 of the 12 subjects. Overall, our 4IAX results are less categorical than the results of Pisoni and Lazarus (1974), which may be due to differences between the stimulus material. In their experiment, synthetic /bA/ and /pA/ syllables were used, which varied only in voice-onset-time. In the present study, a stimulus continuum spanning a stop place contrast was used in the meaningful words /pAk/ and /tAk/. The continuum was generated by interpolating between the spectral envelopes of the original words instead of interpolating between one specific parameter like VOT or onset formant frequency.

The next question of interest was whether these stop-consonant results would be different from vowel results in degree of categorical perception. We will return to this question in sections 4.3.2.2 and 4.4.

4.3.2 Vowel results

The discrimination and classification results for the vowel stimuli are displayed in figure 4.4. Vowel stimuli in pair 1 resemble /u/ and stimuli in pair 7 sound like /i/.

The general vowel patterns are very similar to those in the stop-consonant results. There is no strict relationship between observed and predicted discrimination performance for either of the two discrimination tasks. The fixed and roving 2IFC discrimination functions contain a discrimination peak, which roughly coincides with the classification peak, although this peak is very small in the fixed 2IFC function. Observed fixed 4IAX discrimination contains two peaks, the second of which seems to be related to the classification peak and to the roving 2IFC discrimination peak. Within-category 4IAX discrimination is much higher than predicted.

A minor difference between the stop and vowel data is that within-category 2IFC discrimination is slightly higher for vowels than for stops, although this is probably not significant.

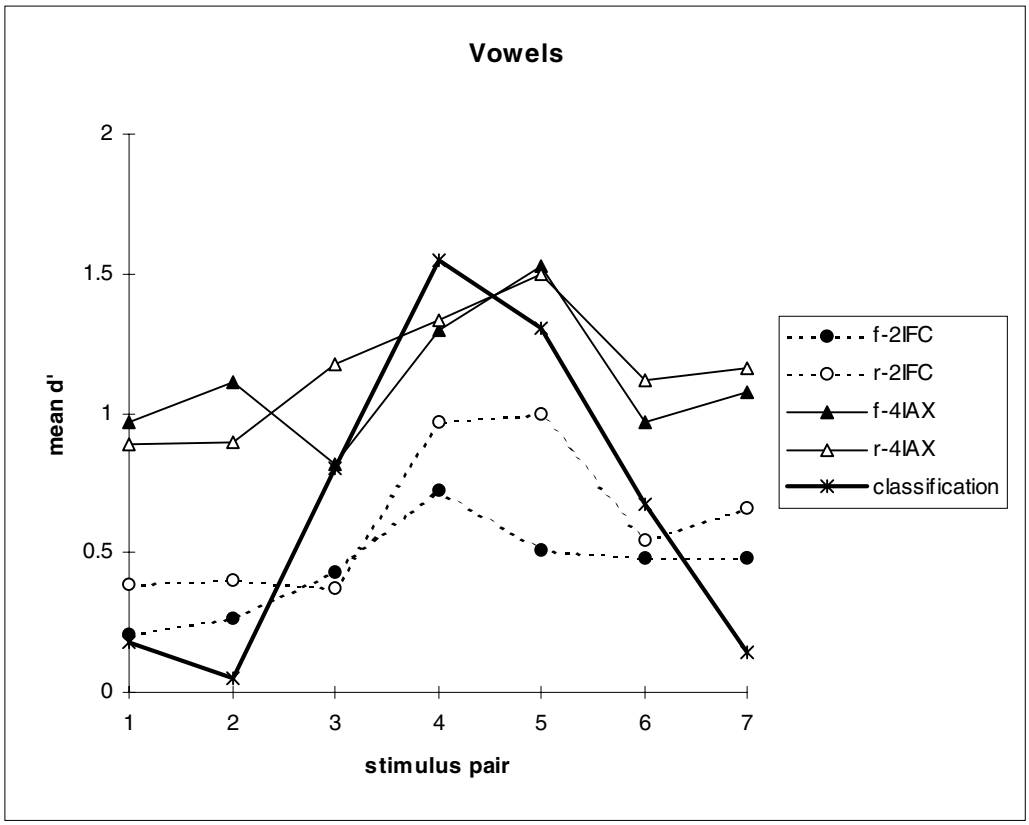

Figure 4.6. Classification and 2IFC and 4IAX discrimination results for the text vowels.

Statistical analysis

A two-way analysis of variance on the vowel data with factors Task (5 levels) and Stimulus (7 levels) demonstrated main effects of Task and Stimulus and a Task by Stimulus interaction effect (respectively $F=24.34$, $p<.001$; $F=12.63$, $p<.001$ and $F=1.70$, $p=.023$). Cell variance was over 11 subjects. A post hoc Tukey HSD test revealed that the main effect of Task implied that the mean d'-values for fixed and roving 4IAX were significantly higher than those for the

other tasks. The post hoc analysis on the factor Stimulus demonstrated that the d'-scores at stimulus pairs 4 and 5 were significantly higher than the scores at the other stimulus pairs, which means that there was an overall discrimination peak at what appears to be the phoneme boundary location.

One-way analyses of variance per task were conducted, with Stimulus as independent variable, to reveal significant peaks in the discrimination results. Significant peaks were found for roving 2IFC and for classification, both at pairs 4 and 5 (roving 2IFC $F=4.36$, $p=.001$; Classification $F=11.54$, $p<.001$). There were no significant peaks in fixed 2IFC and fixed and roving 4IAX discrimination. Only in roving 2IFC was there a significant peak that coincided with the peak in the classification results. Thus roving 2IFC discrimination is more closely related to classification (which implies a higher degree of categorical perception) than are fixed 2IFC and fixed and roving 4IAX discrimination. However, even the roving 2IFC peak is lower than the classification peak. Similar results were found for the stop consonants. Apparently listeners use their phoneme representations in long term memory to discriminate stimuli, but they are less successful in using this strategy during 2IFC discrimination than they are in classification.

An analysis of the difference between the d'-scores for each stimulus pair revealed that fixed 2IFC was always significantly lower than the other tasks and that fixed and roving 4IAX were always higher except at pairs 4 and 5. At stimulus pair 4, Classification was highest and at pair 5 there was no significant difference between the d'-scores as a function of task. The analysis of the 4IAX data confirms what is shown in figure 4.6: within-category 4IAX discrimination was significantly better than predicted by classification. The within-category results in 2IFC discrimination are not significantly better than predicted by classification. These results support the previous conclusion: 2IFC discrimination is more closely related to classification than is 4IAX discrimination.

4.3.2.1 Differences among subjects

In the vowel experiment 71% of the total variance was explained by cell variance. Once again we divided the total subject group into four subgroups based on their 4IAX performance. Figures 4.7 to 4.10 show discrimination and classification performance in the highest (three subjects) and lowest quartiles (three subjects). It has to be noted that the subjects in these vowel quartiles are not identical to the subjects in the stop-consonant quartiles. Only in the lowest quartile was there an overlap of 2 of the 3 subjects. In the highest quartiles there was no overlap, which means that subjects differed in their performance as an effect of the stimuli or the phoneme contrast under investigation.

Separate analyses of variance on the vowel data per task with factor Quartiles (4 levels) as the independent variable revealed only a significant effect on roving 4IAX discrimination ($F=4.46$, $p=.006$). There were no significant differences in fixed 4IAX between quartiles, probably because of the large variation within the quartiles. Thus, like in the case of the stop-consonant results, there was no significant difference between the quartiles for the phoneme labelling tasks, classification and 2IFC.

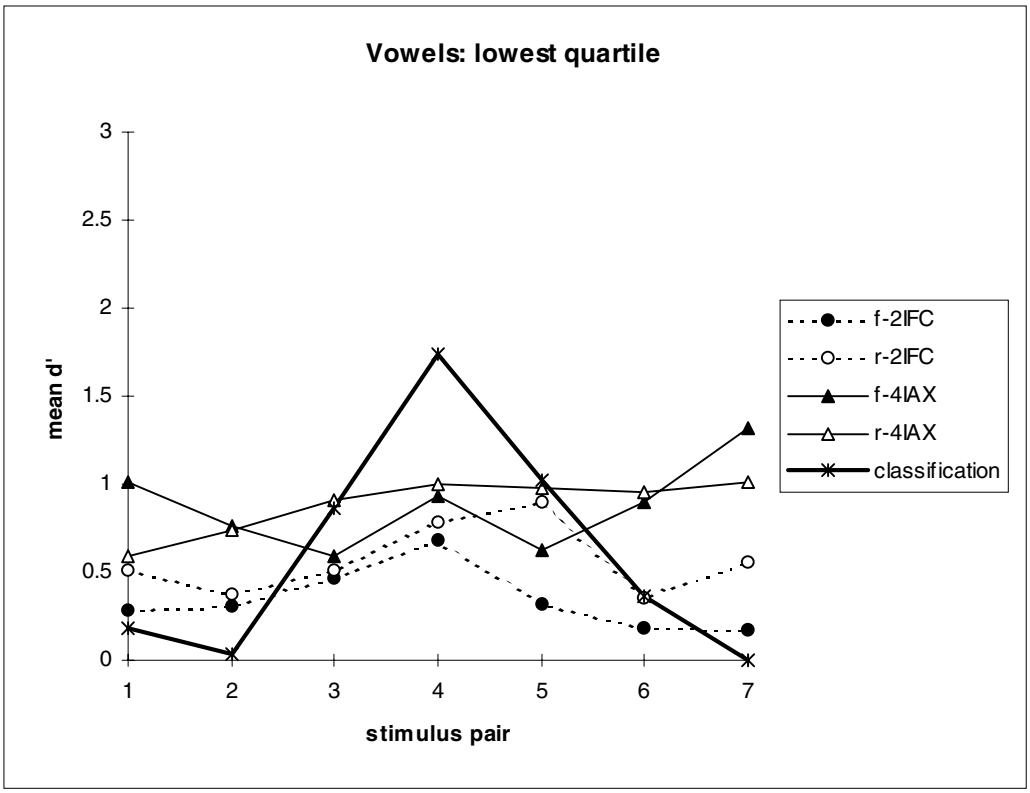

Figure 4.7 Vowel classification and 2IFC and 4IAX discrimination results of the lowest quartile (3 subjects).

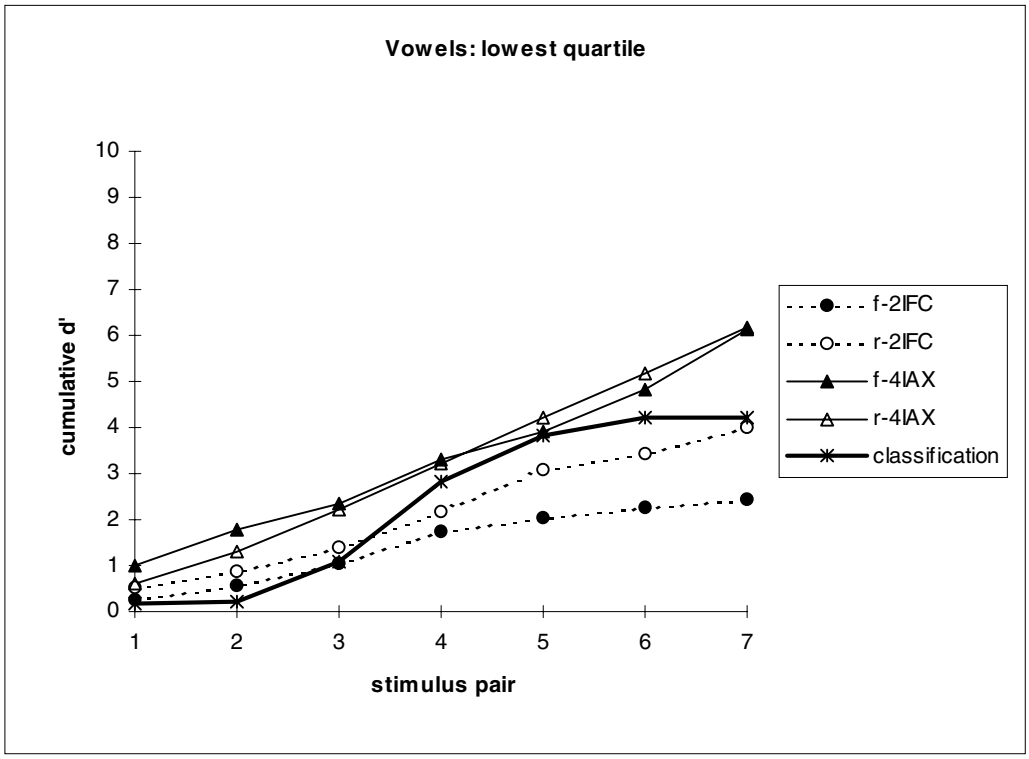

Figure 4.8 Cumulative vowel classification and discrimination results of the lowest quartile.

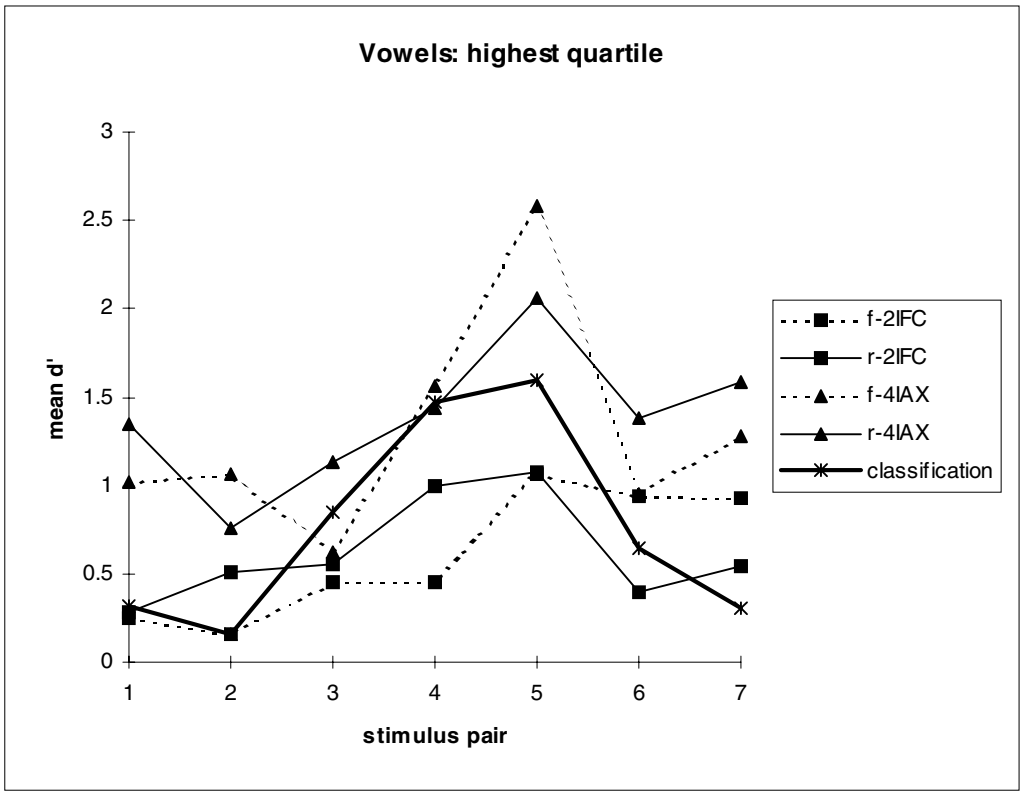

Figure 4.9 Vowel classification and 2IFC and 4IAX discrimination results of the highest quartile.

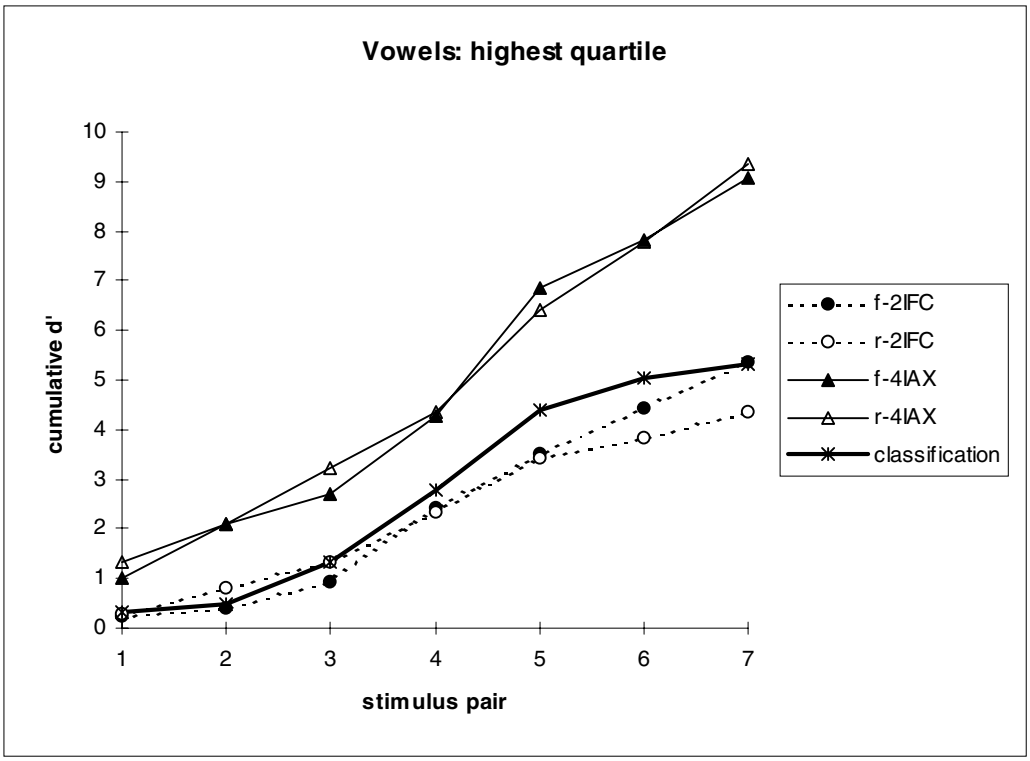

Figure 4.10 Cumulative vowel classification and discrimination results of the highest quartile (3 subjects).

Figure 4.7 displays the mean results of the lowest quartile. The individual data of these three subjects show that discrimination was always unrelated to classification, which is also the pattern in the average results. The individual results in the highest quartile (figure 4.9) were quite different from the average pattern. The most important individual effects are, for the first subject, a high peak in fixed 4IAX, which was at its maximum d' value (5.14), and for the second subject, a close relationship between roving 2IFC and classification. The results of the third subject showed a close relationship between fixed 2IFC and classification.

4.3.2.2 Discussion vowel results

The results clearly indicate that there is a difference in the degree of categorical perception for text vowels as a function of the discrimination task. 2IFC discrimination is more categorical (more related to classification) than 4IAX discrimination. The results are very similar to the stop-consonant results. Again we see that fixed 2IFC discrimination was lower overall than roving 2IFC discrimination. However, if we look at the individual vowel data we see that this was only true for 5 of the 11 subjects. Four subjects demonstrated similar fixed and roving 2IFC, and for two subjects fixed 2IFC was higher than roving 2IFC discrimination. It will be clear that there is a large variation within the subject group.

Another type of between-subject variation was found in the fixed 4IAX and roving 4IAX data. In some of the results fixed and roving 4IAX contained a marked discrimination peak, which coincided with the peak in classification. This peak was present in both the fixed and roving conditions for 4 of the 11 subjects. We found similar roving 4IAX results for some subjects with the word vowels in chapter 3, where, as an effect of a relatively long inter stimulus interval, some listeners tended to use more phoneme labelling, which caused scores to peak around the phoneme boundary.

As in the stop consonant findings, we see that roving 2IFC for vowels is worse than predicted by the labelling data. Again these results are very similar to those found by Schouten and Van Hessen (1992). Interestingly, the present vowel results imply that the text vowels used here are perceived differently from the word vowels. Even though we did not find any differences in 4I-oddity discrimination as an effect of vowel condition, we now see that 2IFC and 4IAX discrimination do reveal perceptual differences.

If we had used the model by Schouten and Van Hessen (1992), the roving 2IFC results for the text vowels would have shown a very close relationship between classification (similar to that found for the stop consonants). However, the recalculated 2IFC data for the word vowels would be better than predicted, which is exactly what Schouten and Van Hessen report for their ‘word’ vowel stimuli. This indicates that with the model by Schouten and van Hessen, the degree of categorical perception would be higher for the text vowels than for the word vowels.

In addition, 4IAX discrimination for the word vowels did not show any effect of phoneme labelling: there was no discrimination peak in the 4IAX word-vowel results. In the 4IAX results for text vowels we did find a discrimination peak for some subjects. We think this difference is caused by differences in complexity. Listeners will use their labelling strategy to discriminate stimuli especially when the spectral coding of the stimuli is more complex.

If we return to the question whether word vowels are perceived differently from text vowels, the answer is a qualified no. In the Schouten and Van Hessen model, text vowels are perceived more similarly to stop consonants than word vowels are. However, this interpretation is not really open to us since we chose to use the standard model by Macmillan

et al. (1977). Therefore, the 'true' d' scores in the present study show that there is closer relationship between 2IFC and classification performance for word vowels than for text vowels (and also for stop consonants). For the 4IAX results the average functions are unrelated to classification in both vowel conditions, although there are some listeners that show a peak in 4IAX for the text vowels.

4.4 Discussion and conclusion

The two aims of this study were to see if the degree of categorical perception of stop consonants varies as a function of the discrimination task and whether the degree of categorical perception for stop consonants is higher than the degree of categorical perception for vowels.

Firstly, we found that the degree of categorical perception for stop consonants does vary as an effect of the discrimination task. If we follow the strict definition of Liberman et al. (1957), we have to conclude that the results show no categorical perception for stop consonants, because both 2IFC discrimination and 4IAX discrimination deviate from classification. Still, as we predicted, 2IFC performance is more closely related to classification performance than 4IAX discrimination. These results go against the general explanation of the high degree of categorical perception for stops proposed by Liberman et al. (1972). According to this explanation, the listener has no auditory image of the signal available, but only the output of a specialised processor that has stripped the signal of all sensory information and represents each phonetic segment categorically. The present findings show that listeners do have access to the auditory image of stop consonants. However, it is clear that the relative amount of auditory trace information that the listeners use during discrimination varies as a function of the task, the experimental context, and the listener.

Secondly, the present experiment has revealed that, as an effect of the use of two different discrimination tasks, the degree of categorical perception varies for vowels as well as for stop consonants. The effects of the different tasks are identical for both types of phoneme, thus stop consonant perception is not more categorical than vowel perception. The vowel results were not strictly categorical either, but again 2IFC performance was more closely related to classification performance than 4IAX discrimination. The main difference between the vowel and stop-consonant results was that within subject variance was 17 % higher in the vowel experiment. The resemblance between the stop and vowel results goes against the general view that these two types of phoneme are perceived in different perceptual modes, phonetic and nonphonetic (e.g. Ades, 1977; Fujisaki & Kawashima, 1970, 1971; Liberman, et al., 1972; Pisoni, 1973).

The results are not as clear as we might have hoped. From the earlier vowel experiments we draw the conclusion that, during discrimination of speech sounds, listeners use one of two possible strategies: they are either in a psychoacoustic mode and deal with the stimuli as essentially meaningless auditory events, or they are in a phonetic mode in which stimuli are interpreted as members of two possible phoneme classes (and auditory information has presumably been lost in the process and is unavailable to the discriminator).

The distinction between the two modes of perception was very clear in the previous experiments. All the evidence indicated that in 4I-oddity, listeners were not influenced by phoneme labels at all; they often failed to discriminate stimuli that they assigned to different phoneme categories during classification. There were good and poor discriminators, but neither group showed any relationship between discrimination and phoneme classification. In

2IFC, on the other hand, performance was largely, although not completely, determined by phoneme classification.

The present experiments with vowels and stop consonants reveal a different pattern. In the average data, 4IAX discrimination is quite similar to the earlier 4I-odity discrimination, and roving 2IFC is again determined largely by phoneme classification. Fixed 2IFC is different, however: performance is much lower than in classification and roving 2IFC (although, only for a subgroup of subjects), but there is a small peak at the phoneme boundary. In fixed 2IFC, where listeners know that they will be hearing the same two stimuli a large number of times, this knowledge does not seem to induce phoneme labelling at all: on the contrary, in roving 2IFC they are very much in the labelling mode, but in fixed 2IFC they are much more in the auditory mode. The labelling mode would lead to much more success, so there must be something about this particular series of experiments which causes listeners to abandon this strategy. However, we do not know what this might have been.

There is another difference with the earlier experiments: the highest quartile differs from the lowest quartile mainly around the phoneme boundary; this means that the difference between good and poor discriminators is not that the former are psychoacoustically better than the latter, but that they are better because they are capable of using labelling information when it is useful to them. Consequently, for some listeners, fixed and roving 4IAX discrimination were very similar to classification. This is true both of the vowel data and of the stop consonant data, which indicates that this result cannot be attributed to a difference in perception between these two types of phonemes.

Again, but this time with stop consonants, we see that the degree of categorical perception varies as a function of the discrimination task. However, the results of the experiments in the present chapter demonstrate that it is not just the task that determines the degree of categorical perception, but also the experimental context, the individual subjects, and perhaps the way subjects were motivated.