Response distributions in intensity resolution and speech discrimination

M. E. H. Schouten and A. J. van Hessen
Research Institute for Language and Speech, University of Utrecht, Trans 10, 3512 JK Utrecht, The Netherlands

(Received 25 November 1996; revised 18 June 1998; accepted 2 July 1998)

In this paper the assumption of an equal, Gaussian distribution of the response to each stimulus in an experiment, an assumption which has to be met if d' is to be estimated by calculating the difference between $z(\mathrm{H})$ and $z(\mathrm{FA})$, is tested for two different sets of stimuli: 1000-Hz tones differing in level only, and a continuum of stop consonants, obtained by full spectral interpolation between /p/, /t/, and /k/. Response distributions were measured directly by means of a form of non-numerical magnitude estimation, in which subjects had to indicate the position of each stimulus on a quasi-continuous rating scale. It could be shown that, in general, all distributions were sufficiently unimodal, but that their variances differed. The consequences for the calculation of d' are unlikely to be serious. © *1998 Acoustical Society of America.* [S0001-4966(98)00211-2]

PACS numbers: 43.66.Ba, 43.66.Fe, 43.66.Lj, 43.71.An [RHD]

INTRODUCTION

In an ideal world, every estimate of stimulus resolution expressed in terms of d' would be obtained by fitting a ROC (receiver operating characteristic) curve through a large number of paired hit (H) and false-alarm (FA) proportions. Each pair of proportions would be the result of a separate experiment involving the same stimuli and the same conditions, but with a different decision criterion for the observer in each experiment. Criteria may be influenced by manipulating the *a priori* probability of the stimuli or by changing the reward for a hit or the penalty for a false alarm.

Signal detection theory states that successive presentations of a stimulus give rise to a range of sensations distributed around a mean. These sensation distributions are usually assumed to be Gaussian and all distributions in an experiment are usually assumed to have equal variance. If these conditions are met, and if we plot the H and FA pairs along normal coordinates [i.e., $z(\mathrm{FA})$ vs $z(\mathrm{H})$], then the ROC fitted through them is a straight line with unit slope, and d' equals the intercept along the $z(\mathrm{H})$ axis. If the distributions are not Gaussian, a straight line will not provide a good fit; if the variances are unequal, slope will not be unity (see Swets *et al.*, 1961). This only makes sense, incidentally, if the variances of the distributions are expressed in terms of physical units or of perceptual units psychoacoustically derived from physical units.

It is fairly easy to check the Gaussian and equal variance assumptions, but it is also rather costly, since it requires a number of stimulus presentations that should be large enough for reliable estimates of hit and false-alarm probabilities over a wide range of criterion positions along the rating scale. As a consequence, experimenters usually restrict themselves to one criterion, yielding single estimates of $z(\mathrm{H})$ and $z(\mathrm{FA})$ which are then subtracted to yield a d' estimate, under the assumption that they define a straight line with a slope of unity. However, in an experiment with visual stimuli presented by Swets *et al.* (1961), a doubling of the mean stimulus value resulted in a 25% increase in variance, although the distributions appeared to be Gaussian. On the other hand, most of the available evidence indicates that, for simple auditory stimuli, the assumption is a tenable one. For example, Braida and Durlach (1972) provide magnitude-estimation and absolute-identification data obtained with 1000-Hz tones differing only in intensity; the various sets of $z(\mathrm{H})$ vs $z(\mathrm{FA})$ data points for the stimulus pairs are quite well described by straight lines of unit slope. The authors are careful to point out, however, that theirs is not really a rigorous test of the degree to which the data do or do not violate the assumption of Gaussian, equal-variance distributions.

This assumption may constitute a good approximation to the probability-density functions which underlie the rating distributions that are usually associated with simple auditory stimuli. However, the situation is likely to be quite different for more complicated stimuli, especially if these are associated with well-learned categories, such as speech sounds. The hypothetical decision axis is, in such cases, not just a combination of sensation axes related to the various stimulus parameters, but it may be greatly affected by higher-order concepts, such as category boundaries or prototypes. An increasing number of speech perception researchers have come to apply a signal-detection analysis to their data, e.g., Pisoni (1973), Macmillan *et al.* (1977), Rosner (1984), Cowan and Morse (1986), Samuel (1987), Macmillan *et al.* (1987, 1988), Uchanski *et al.* (1992), Schouten and van Hessen (1992), and van Hessen and Schouten (1992). In each of these studies, d' estimates are based on single $z(\mathrm{H})-z(\mathrm{FA})$ pairs and therefore on the assumption that all members of a series of stimuli cause equal, Gaussian variances.

The present authors are engaged in a series of experiments concerning the categorical perception of speech sounds. In our previous papers, cited in the last paragraph, we have found that stop consonants are perceived categorically, and we have attempted to model the discrimination of stop consonants as a function of time. All of this has, however, been done on the rather shaky foundation of an assumption which may be incorrect. Before proceeding with

our experiments, we therefore wanted to put this assumption to the test in a rating-scale experiment, not to obtain a large number of $z(\mathrm{H})-z(\mathrm{FA})$ pairs, but to get a more direct picture of perceptual variance, in a way that will be described briefly in the next paragraph. Since our method involves a type of magnitude estimation, it was decided to first try and replicate the relevant experiment of Braida and Durlach (1972) in experiment I and only then to apply it to speech sounds in experiment II.

We reasoned that it should be possible to obtain a good picture of a subject's perceptual variance by presenting each stimulus often enough and requiring the subject to give a non-numerical estimation of its magnitude, the advantage of a non-numerical estimate being that it does not anchor subjects to, e.g., whole numbers or multiples of 10, but encourages them to use the resolution they are capable of using. In all other respects, this task is equivalent to the one employed by Braida and Durlach (1972), in which subjects were instructed to assign the number 100 to the loudness of a stimulus just below the middle of the range (which was presented ten times before each group of 100 trials). They had to use a ratio scale to rate the loudness of all subsequent stimuli (50 meant half as loud, 200 twice as loud). In our experiments, subjects performed their task by placing a mouse pointer at the appropriate spot of a horizontal bar which spanned the width of their monitor screen, and which represented the full stimulus range used in the experiment. Three reference points were regularly reinforced: the end points of the scale, which represented two stimuli just outside the experimental range (by exactly one stimulus step), and the exact middle of the range. The method made it undesirable to give any form of feedback, since this would have provided subjects with the information that the number of different stimuli was very limited. It was therefore decided to give subjects extensive training with the same stimuli, but using a different task—absolute identification—in such a way that they would not realize that they were being trained and that the same limited number of stimuli was used in both experiments.

Our expectations were that the intensity differences in experiment I would lead to distributions which would be unimodal and have approximately the same variance, especially after training. The timbre differences in experiment II, however, were expected to have a relatively low associated variance near the best representatives of a speech category (phoneme), but a much higher variance for stimuli near a phoneme boundary. The reason for this expectation is to be found in the notion of categorical perception, which says that stimuli that belong to the same category are perceived as identical and will therefore be given the same rating on our rating scale; moreover, the variance associated with this rating will be small, since subjects do not have access to variations in sensation that are due to sensory noise. On the other hand, this same sensory variance will cause a stimulus near a category boundary to be classified both ways, resulting either in a bimodal rating distribution, or in a much wider distribution if the stimulus is actually perceived as being not readily classifiable. We therefore preferred a direct picture of the rating distributions, particularly those evoked by the speech stimuli, over ROC curves, which could easily be constructed from the same rating data by shifting the criterion along the rating scale.

I. EXPERIMENT I: 1000-Hz TONES

Experiment I consisted of two parts: experiment Ia, in which extensive identification training was given, but which will be only selectively reported, since the data showed that the response range had been too narrow, and experiment Ib, which was carried out a year later using a much wider response range, but without training and with a smaller number of subjects. In effect, experiment Ia served as a pilot study for experiment Ib, the main experiment.

A. Method

1. Stimuli

Since experiment I was in many ways intended as a replication of the magnitude-estimation experiment by Braida and Durlach (1972), the stimuli were as similar to theirs as circumstances allowed. There were ten basic stimuli, consisting of 1000-Hz tones of 500-ms duration and with 25-ms cosine-shaped onset and offset windows. Stimuli 1–10 had levels of 50, 54, 58, 62, 66, 70, 74, 78, 82, and 86 dB SPL. In addition, there were three reference stimuli: stimulus 0 had a level of 46 dB, and stimulus 11 one of 90 dB; a third reference stimulus, one of 68 dB in the middle of the range, will not be used in the presentation of the results and was therefore not given a number.

Sampling frequency was 20 kHz, and resolution was 16 bits.

2. Subjects

Six subjects took part in experiment Ia—five female students and one male student of Utrecht University, all in their early twenties. Four of them returned a year later for experiment Ib. They received a basic hourly rate, apart from bonuses and penalties for correct and incorrect responses in absolute identification.

3. General procedure

Experiment Ia consisted of seven tests, taken on consecutive weekdays:

(i) ME-1: magnitude estimation without feedback,
(ii) AI-1 to AI-5: absolute identification with feedback,
(iii) ME-2: magnitude estimation without feedback.

AI-1 to AI-5 were primarily intended as a form of training for ME-2.

Experiment Ib consisted of a single magnitude-estimation test, but with a wider response range.

Magnitude estimation trials involved only stimuli 1–10 (50–86 dB SPL); for absolute identification stimuli 0 and 11 (46 and 90 dB) were added, so the number of identification categories was twelve.

All tests were carried out with subjects seated in one of two sound-treated, but not completely insulated booths—SPL of the least intense stimulus (the one of 46 dB) could

not be measured directly, due to low-frequency ambient noise interference, so to check the level of this particular stimulus an A-weighting had to be used.

Stimuli were presented binaurally over Beyerdynamic DT 770 PRO headphones, which were chosen in preference over the standard Beyer DT 49 headphones, since the latter would have been unsuitable for the speech stimuli in experiment II, in view of their poor frequency response above 4000 Hz. Moreover, the DT 770 are much more comfortable to wear over long periods. Calibration was carried out by means of an artificial ear.

4. Procedure for magnitude estimation

Subjects were seated in front of a monitor screen on which an undivided horizontal bar was displayed; the left end of this bar was marked with the word ''soft,'' the right end with the word ''loud.''

Prior to each block of 100 trials, the two (Ib) or three (Ia) reference stimuli were presented five times in a fixed order: 46 dB (stimulus 0), 68 dB, and 90 dB (stimulus 11), with an interstimulus interval of 2.5 s. In experiment Ia a marker was visible during this interval at the extreme left, in the exact middle, or at the extreme right of the bar. In experiment Ib, the 68-dB reference in the middle of the range was left out, and the positions of reference stimuli 0 and 11 were pulled toward the center of the response bar by two stimulus steps. As a result, the extreme ends of the response bar were much further away from the actual stimulus range than they were in experiment Ia, producing enough latitude to accommodate both tails of each response distribution.

The test itself consisted of 400 presentations of each of the stimuli 1–10, in a completely random order (interrupted by the reference stimuli after every 100 presentations). Subjects responded by moving the mouse pointer to the appropriate position along the horizontal bar and then pressing a mouse button. As soon as they had done this, the next stimulus was presented; if they did not press within 2.5 s, a nonresponse was recorded and the next stimulus was presented.

Maximum net duration of the test was 3 h and 50 min for a subject who needed 2.5 s for every decision. Breaks could be taken at any time at the end of a series of 100 trials; nearly all subjects took breaks after every 1000 trials (the number of remaining trials was displayed at the bottom of the screen).

Since there could not be any ''correct'' responses, no feedback was given. In experiment Ia, subjects were not rewarded or punished in any way but in experiment Ib they were told that they could raise their earnings if their ratings were to show the lowest average variance of all four subjects.

5. Procedure for absolute identification (experiment Ia only)

Subjects were seated in front of a monitor screen which displayed a horizontal bar, divided into 12 segments, marked with the numbers 1 to 12, corresponding to stimuli 0–11 (46–90 dB).

Each of the twelve stimuli was presented 150 times in each of the five tests. Order was completely random. Presentation was self-paced, since there was no maximum response time. After the stimulus had been presented, the subject moved the mouse pointer to the chosen segment and recorded her or his response by pressing a mouse button. Feedback was given immediately: in case of a correct response, the word ''OK'' was displayed for one second in the ''correct'' segment; otherwise, a cross was shown for one second in the ''correct'' segment.

Breaks could be taken between any two stimulus presentations; most subjects took one break exactly halfway through the experiment (the number of remaining stimuli was displayed at the bottom of the screen).

An increasingly severe system of rewards and penalties was enforced over the five tests. A correct response was always rewarded with 5 cents, but incorrect responses were punished increasingly severely. Subjects knew at all times how much they had gained or lost on the response they had just given.

B. Results and discussion

1. Experiment Ia

Since all data from experiment Ia are flawed in the same way, only the final identification and magnitude-estimation tests will be presented and discussed, mainly to indicate what lessons can be learned from them and how particular aspects of the results from experiment Ib may be interpreted.

Figure 1 presents the results from the fifth (and last) absolute identification session. Please note that stimuli 0–11 are represented along the vertical axis, whereas the response categories 0–11 are presented horizontally. The data points show each subject's mean identification rating of each stimulus; the thin horizontal bars around them indicate standard deviations. The barely visible thick line connects the average ratings calculated over the six subjects.

The picture presented by Fig. 1 is simple. Feedback was, by itself, enough to produce accurate identification ratings, since the results for the four preceding sessions were almost exactly the same. The only effect of training was that variance decreased between the first and the last session. This can be seen in Fig. 2, where the diamonds represent standard deviations in the first session (AI-1) and the squares standard deviations in the last session (AI-5). A three-way analysis of variance, with subjects (6) as a random independent variable and tests (5) and stimuli (12) as fixed independent variables, showed a significant effect of the tests factor, and no interaction between tests and stimuli.

The mean magnitude-estimation results for ME-2 (after training) are presented in the top panel of Fig. 3. In this figure stimuli are represented along the ordinate and responses along the abscissa, just as in Fig. 1. What is plotted in the top panel is the number of responses of the type indicated along the abscissa; since there were ten different stimuli, there are ten such plots. The bottom panel shows the mean ratings for each subject.

In Fig. 3 we see that most subjects, after five days of identification training, have learned to correctly identify the stimuli. Truncations occur in the extreme stimuli, especially in stimulus 1. In addition, there is what seems to be a bimo-

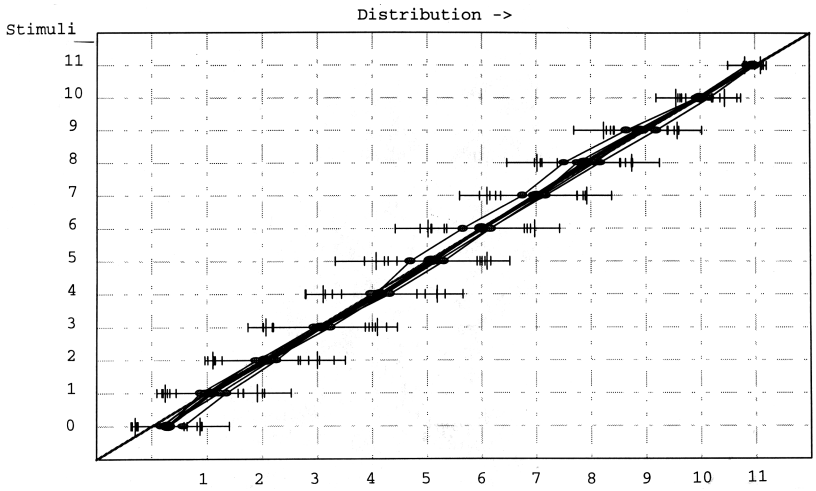

FIG. 1. Experiment Ia, 1000-Hz tones. Distribution means from AI-5 (the last absolute identification session). The stimuli are represented along the ordinate; stimuli 1–10 are the ten test stimuli, while stimuli 0 and 11 are the reference stimuli used in the magnitude-estimation tasks. The response bar, which is represented along the abscissa, contained 12 possible responses. Stimuli range in 4-dB steps from 46 (stimulus 0) to 90 (stimulus 11) dB SPL.

dal distribution in the response to stimuli 4–7, which is shared by half the subjects. These stimuli occupy the middle of the perceptual range; sometimes they are classed with the ''softer'' stimuli (5), sometimes with the ''louder'' ones (7). Some subjects reported that they had three anchors: not only the extreme stimuli, but also the position halfway between these; this was, of course, encouraged by the use of a reference stimulus in the middle of the range.

Figure 4 shows the standard-deviation estimates from ME-1 (before training, diamonds) and those from ME-2 (after training, squares). A certain amount of accuracy has clearly been lost with respect to the lower-level stimuli: standard deviations, expressed in number of stimulus steps, are up after identification training, whereas there is a considerable fall for the other stimuli. A three-way analysis of variance, with subjects (6), tests (2), and stimuli (10) as independent variables, showed that the tests factor did not have a significant effect; it did interact significantly, however, with the stimulus factor.

What is surprising, however, is the difference in shape between the functions marked by square data points in Figs. 2 and 4. Why doesn't the overall improvement in identification (AI-5, Fig. 2) carry over into magnitude estimation (ME-2, Fig. 4)? Average identification σ's fall by 0.28 stimulus steps as a result of training, whereas average estimation σ's decrease by only 0.15 steps. Moreover, there is no across-the-board improvement in magnitude estimation: σ's actually rise for stimuli 1–3. The main effect of training on magnitude estimation of the low-level stimuli seems to be to pull them apart (see the discussion of Fig. 5 below), with no concurrent increase in accuracy, even though Fig. 2

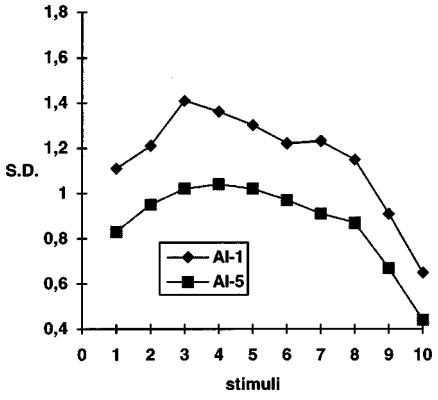

FIG. 2. Experiment Ia, 1000-Hz tones. Standard deviations in absolute identification from the first (AI-1, diamonds) and the last (AI-5, squares) of five identification sessions. Stimuli 1–10 range in 4-dB steps from 50 to 86 dB SPL.

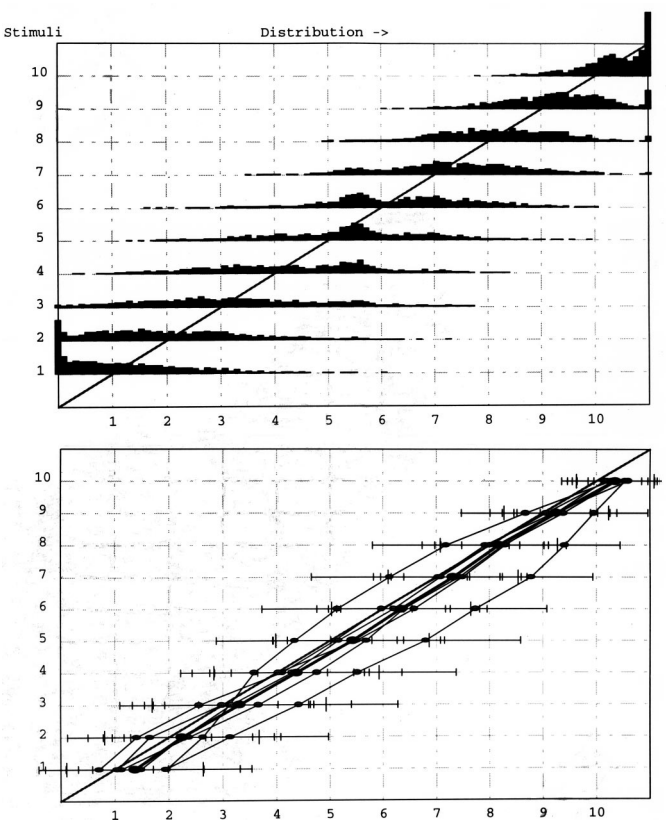

FIG. 3. Experiment Ia, 1000-Hz tones. Mean response distributions (top) and distribution means (bottom) for ME-2 (magnitude estimation after training). The abscissae represent responses along a continuous response bar, and the ordinates represent the ten stimuli. What is plotted is the number of responses of the type indicated along the abscissa; with ten different stimuli, there are ten such plots. In the top panel, the abscissa is divided into 100 segments, corresponding to the accuracy provided by the response bar. In the bottom panel, thin lines connect the distribution means of the separate subjects, and the thick line connects the average distribution means over the six subjects.

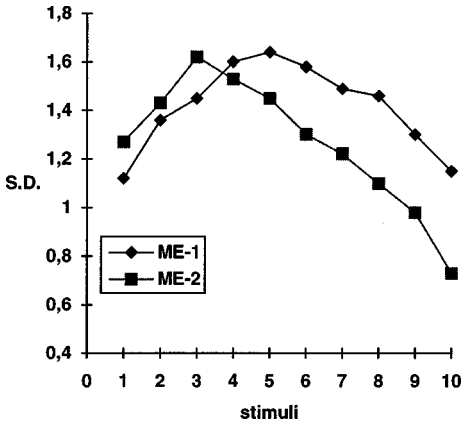

FIG. 4. Experiment Ia, 1000-Hz tones. Standard deviations in magnitude estimation before (ME-1, diamonds) and after (ME-2, squares) identification training. Stimuli range in 4-dB steps from 50 to 86 dB SPL.

shows a clear increase in accuracy for absolute identification. Apparently, accuracy cannot be maintained in the absence of feedback on stimuli that are hard to discriminate.

Braida and Durlach (1972) present d' estimates for their magnitude-estimation experiments, which are based on the discrete numerical responses they obtained from their subjects. We could have ''binned'' our responses to get the same effect, but we decided to use all the available information and therefore to divide the differences between two response means by their averaged standard deviations (separately for each subject), to obtain the d' estimates (averaged over the individual subjects' d' values) shown in Fig. 5. The data points in this figure (stimulus pairs along the abscissa) have been placed halfway between the stimuli that are compared (ticks along the abscissa). Figure 3 has shown that the calculated standard deviations for stimuli 1, 2, and 10 in Fig. 4 are probably too low; this means that the d' values for these stimuli in Fig. 5 are probably too high. If so, the d' values for ME-1 (lower graph) lie on a line that is practically straight. This does not apply to ME-2 (upper graph): as a result of training, all d' values have been lifted, but not all to the same extent. In fact, subjects seem to have learned that

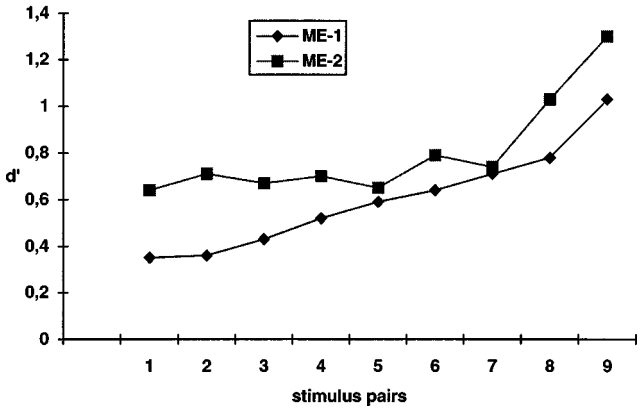

FIG. 5. Experiment Ia, 1000-Hz tones. Magnitude-estimation d' values before (ME-1, diamonds) and after (ME-2, squares) identification training, calculated by dividing the difference in two distribution means by their averaged standard deviations. Stimuli range from 50 to 86 dB SPL in 4-dB steps.

stimuli 1–8 are all at equal distances from each other, namely at the same distance as stimuli 7 and 8.

Cumulative Δ', obtained by simply adding the d' values from left to right, was 5.41 in ME-1 and 7.23 in ME-2, which compares quite well with the value of 6.5 found by Braida and Durlach (1972). The fact that these values are rather low can probably be attributed entirely to the stimulus range, which was the same in both studies (36 dB). As Braida and Durlach (1972) show in their Fig. 3c, which depicts absolute identification results, a stimulus range of 54 dB leads to a (cumulative) Δ' of around 13, whereas for a range of 2 dB (half the stimulus distance in the present experiment), Δ' equals almost 2, even at the low end of the dB range.

Non-numerical magnitude estimation appears to give a good estimate of perceptual variance for 1000-Hz stimuli differing only in level. Variance is a function of stimulus magnitude (see Fig. 4): it seems to be high in the middle of the range and seems to fall towards its edges; this should be taken into account when d' is calculated. However, part of this difference is probably due to the fact that variance is underestimated at the edges of the range, as a result of truncation, whereas it is overestimated in the middle of the range, as a result of the bimodality of the distributions there.

Experiment Ib was set up to remedy these deficiencies. The reference stimulus in the middle of the range was omitted, and the response range was widened by two stimulus steps on each side. Only four of the original six subjects were still available a year after having taken part in experiment Ia.

2. Experiment Ib

The results of the magnitude-estimation session involving 1000-Hz tones are shown separately for the four subjects in Fig. 6. The effect of earlier experience seems to have largely worn off: the ratings are quite similar to what they were before training (ME-1 in experiment Ia, not shown). The main difference is that the end-point stimuli are now accommodated fairly comfortably within the enlarged response range.

Some stimuli still invoke bimodal distributions: 5, 6, 7, and 8 for subject 2, and 6, 7, and 8 for subject 3. As in experiment Ia (Fig. 3), one of the modes coincides with the exact middle of the response range, where there had been an anchor 12 months before. However, it is unlikely that these bimodal distributions are a carryover from experiment Ia: subject 3 was the only subject with bimodal distributions on both occasions, whereas in experiment Ia subject 2 did not exhibit any peaks at or near 5.5 along the abscissa, despite the regularly reinforced anchor in the middle of the range. Subject 2 had apparently changed her strategy between experiments.

Figure 7 presents the standard deviations, averaged over the individual subjects' standard deviations per stimulus. The d' values in Fig. 8 are based directly on these standard deviations. If we compare them to the values from experiment Ia in Fig. 5, we see that they are slightly higher than they were before training for stimulus pairs 1–7, and considerably higher for pairs 8 and 9. The trend in the data is predicted by

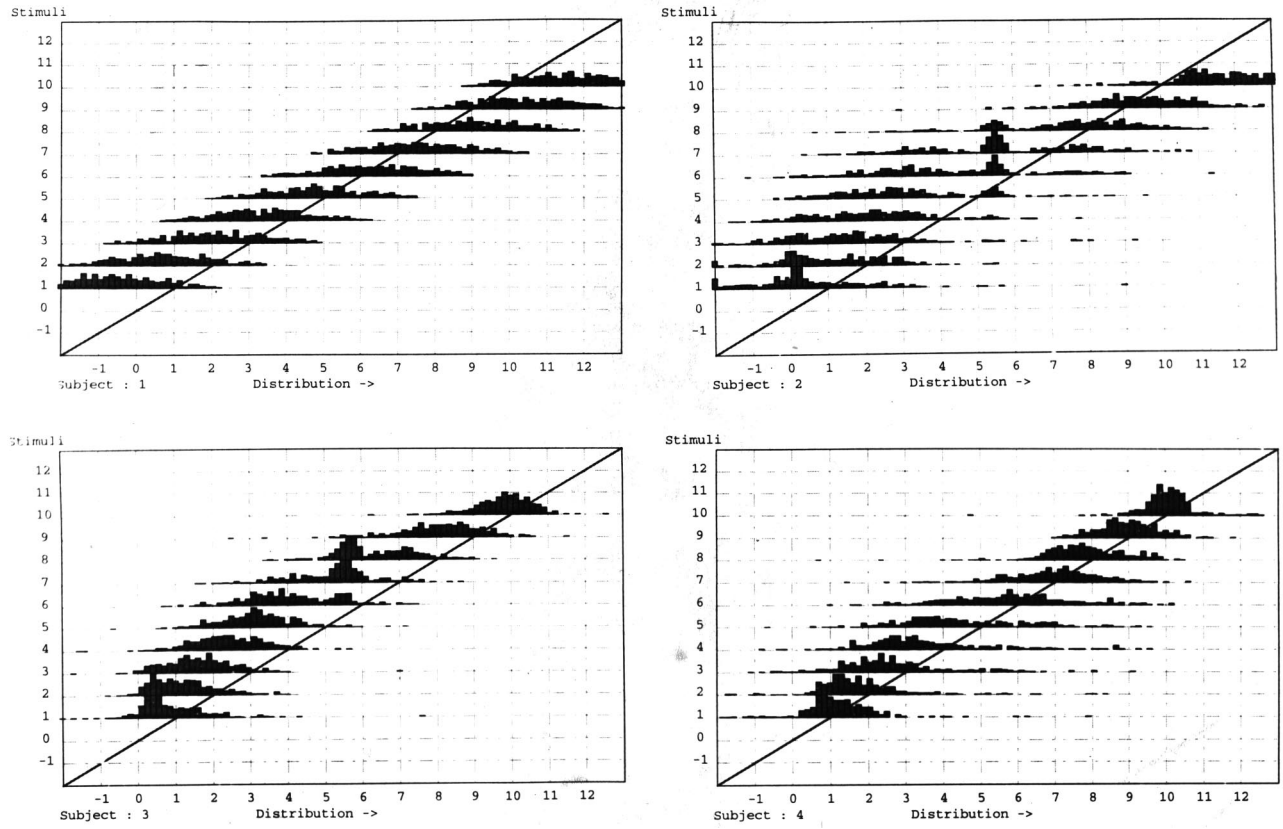

FIG. 6. Experiment Ib, 1000-Hz tones, extended rating scale. Individual magnitude-estimation response distributions. The abscissae represent responses along a continuous response bar divided into 14 positions (from −1 to 12) in the figure, while the ordinates represent the ten stimuli. Stimuli range from 50 to 86 dB SPL in 4-dB steps.

the near miss to Weber's law: resolution is positively correlated with stimulus level.

II. EXPERIMENT II: STOP CONSONANTS

A. Method

1. Stimuli

Just as in experiment I, which was run parallel to experiment II, there were ten stimuli, along with three reference stimuli. The reference stimuli were produced first, as direct resyntheses of the Dutch syllables /pak/, /tak/, and /kak/, pronounced by a male native speaker, using the same source signal for all three syllables (cepstral deconvolution, followed by convolution with one of the source signals). The test stimuli were then calculated by spectral interpolation of the stimuli 1–5 between references /tak/ and /pak/ on the one hand, and of the stimuli 7–11 between references /tak/ and /kak/ on the other [see Schouten and van Hessen (1992) for more details of the method]. As a result of this procedure, there was a gap between stimuli 5 and 7 of the test continuum in experiment IIa, since stimulus 6, the ''original''

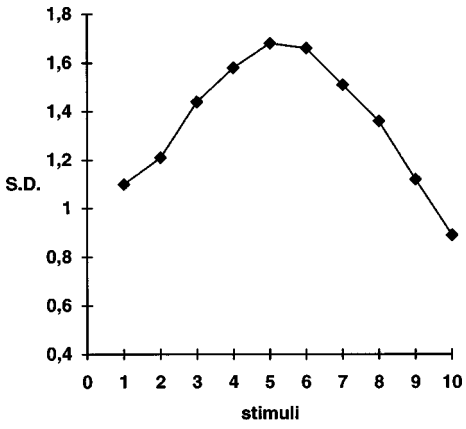

FIG. 7. Experiment Ib, 1000-Hz tones, extended rating scale. Standard deviations for magnitude estimation. Stimuli range in 4-dB steps from 50 to 86 dB SPL.

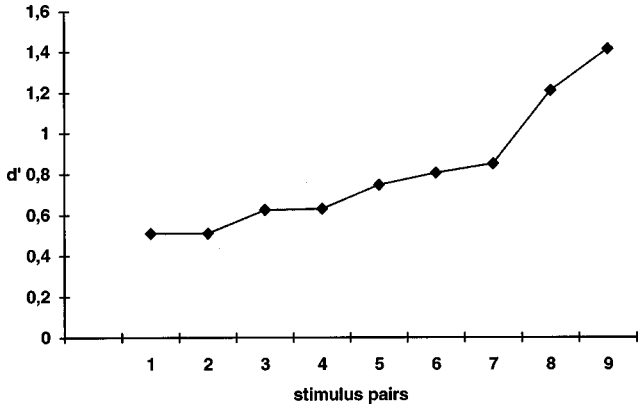

FIG. 8. Experiment Ib, 1000-Hz tones, extended rating scale. Magnitude-estimation d' values, calculated by dividing the difference in two distribution means by their averaged standard deviations. Stimuli range from 50 to 86 dB SPL in 4-dB steps.

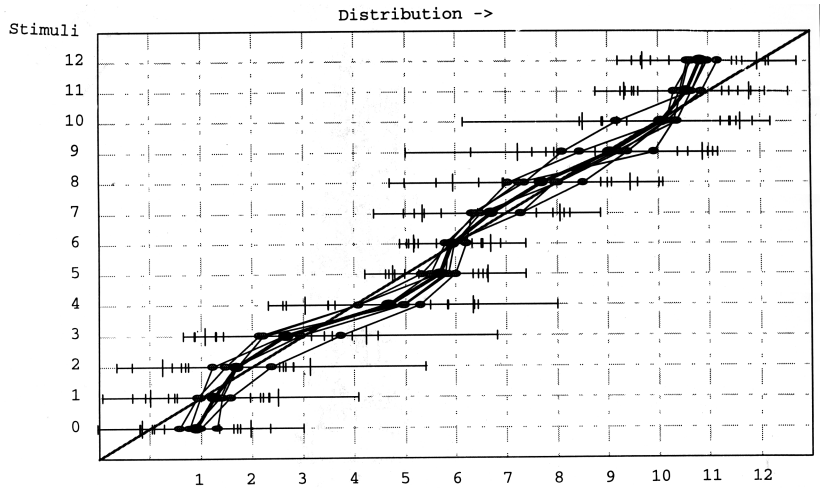

FIG. 9. Experiment IIa, stop consonants. Distribution means for AI-2 (the last absolute identification session). The composition of this figure is the same as that of Fig. 1, except that the ordinate here represents a spectral continuum from /p/ (stimulus 0) via /t/ (stimulus 6) to /k/ (stimulus 12). Stimuli are spectral interpolations between /p/, /t/, and /k/.

/t/, was only used as a reference. In experiment IIb, stimulus 6 was not used as a reference and was included as part of the continuum, which now contained 11 stimuli.

Sampling frequency was 20 kHz, with 16-bit resolution. All stimuli sounded entirely natural—as if they had been pronounced by the original speaker.

2. Subjects

Five of the six subjects in experiment IIa were the same as in experiment I; one female student was replaced by a male student. In experiment IIb, the four subjects were the same as those in experiment Ib.

3. General procedure

It soon turned out that prolonged exposure to the same speech sounds caused nearly all our subjects to hallucinate. They increasingly heard all sorts of sounds that were just not there, but, fortunately, all of them had different experiences (there was nothing wrong with our equipment). We therefore decided to restrict training rather severely: instead of five, there were only two training sessions in experiment IIa: AI-1 and AI-2. These two absolute-identification sessions were preceded and followed by non-numerical magnitude-estimation sessions: ME-1 and ME-2. Experiment IIb consisted of just a single magnitude-estimation session.

4. Procedure for magnitude estimation (ME-1 and ME-2)

Procedure was exactly the same as for the intensity stimuli in experiment I, except that the appropriate points of the response bar were now marked ''p'' and ''k'' instead of ''soft'' and ''loud.'' A second difference was that in both experiments IIa and IIb accuracy was now rewarded financially in view of the limited amount of training that could be given, although subjects did not receive feedback about this after each trial, since it was based on their average rating of the stimuli and on the standard deviations around these ratings.

5. Procedure for absolute identification (AI-1 and AI-2)

Procedure was exactly the same as for the intensity stimuli in experiment Ia, except that the response bar was now divided into 13 segments, marked with the numbers 1–13 (the middle reference stimulus now occupied a segment of its own).

B. Results and discussion

1. Experiment IIa

Figure 9 presents the results of the second absolute identification session (AI-2); stimuli 0, 6, and 12 are the reference stimuli. Accuracy is obviously rather low here: subjects seem to distribute their responses almost randomly over a group of likely candidates, getting increasingly frustrated at the amount of ''negative'' feedback (and the resulting loss of earnings). It is hard to tell whether more training would have been beneficial. Every single subject declared that he or she much preferred the straightforward clarity of experiment I, where hallucinations did not occur and practice really helped performance.

The magnitude-estimation results are presented in Fig. 10 for ME-2 (after training). The most striking aspect of the results is that subjects exhibit a large degree of categorical perception even after training: they appear to give only three different responses (apart from random variation). Before training (not shown), stimuli 1, 2, and 3 belonged to the first category, stimuli 4, 5 (6), 7, and 8 belonged to the second category, and stimuli 9, 10, and 11 to the third. The category in the middle was quite narrow and coincided with the reference /t/; the /p/ and /k/ categories at either end showed much more variance and are severely truncated. After training (Fig. 10), some subjects seem to have acquired a bimodal distribution in the middle of the range, as can be seen in the second peak for stimuli 7 and 8 in Fig. 10. The cause of this is probably that training has taught them to identify stimuli 7 and 8 more accurately, while repeated presentation of the anchor in the middle of the range continues to exert its pull.

Figure 9 shows that perception is not absolutely categorical: each data point is at least slightly to the right of its

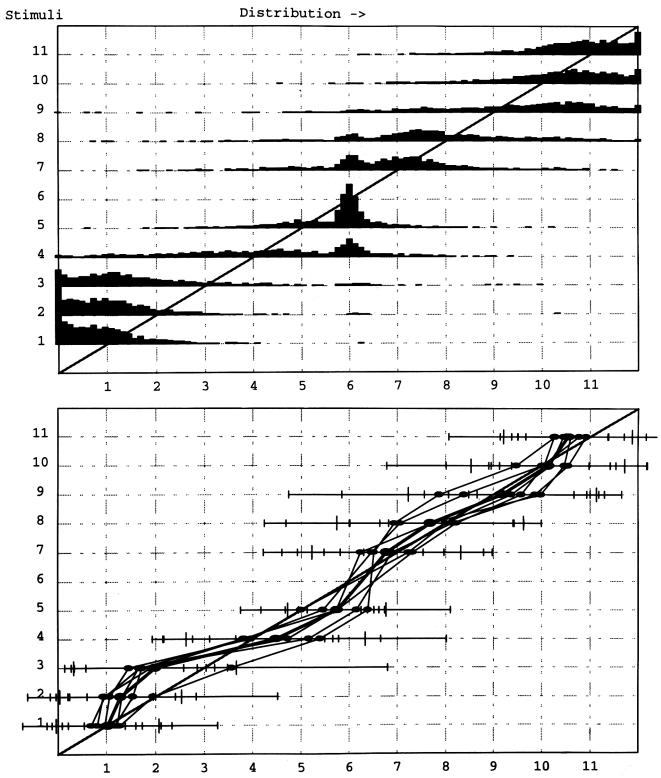

FIG. 10. Experiment IIa, stop consonants. Response distributions for ME-2 (magnitude estimation after training). The composition of this figure is the same as that of Fig. 3, except that the ordinates here represent a spectral continuum from /p/ (stimulus 0) via /t/ (stimulus 6) to /k/ (stimulus 12), and that these three stimuli (0, 6, 12) served only as references and were not used to elicit responses.

lower-number neighbor, so there are audible differences within each of the categories.

Figure 11 displays the overall standard deviations of the distributions from ME-1 (before training) and ME-2 (after training). There is no data point for stimulus 6, since this stimulus served only as a reference. Again, as in Fig. 4, the differences between the stimuli are due in large measure to the truncated nature of many of the distributions, but, also, in the case of stimuli 5 and 7, to proximity to a phoneme category and/or a reference stimulus. The most important aspect of Fig. 11 is, therefore, the significant reduction in variance as a result of identification training ($p<0.05$).

The individual subjects' standard deviations have been

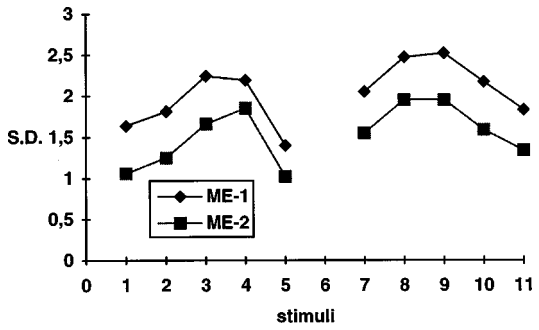

FIG. 11. Experiment IIa, stop consonants. Standard deviations in magnitude estimation before (ME-1, diamonds) and after (ME-2, squares) identification training. Stimuli are spectral interpolations between /p/, /t/, and /k/.

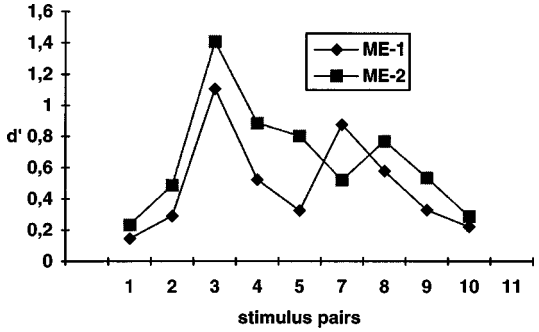

FIG. 12. Experiment IIa, stop consonants. Magnitude-estimation d' values before (ME-1, diamonds) and after (ME-2, squares) identification training, calculated by dividing the difference between two distribution means by their averaged standard deviations. Stimuli are spectral interpolations between /p/, /t/, and /k/.

used to determine the d' values displayed in Fig. 12 (d' equals the difference between the means of the distributions over their averaged standard deviations). Note that the data points denote the perceptual distance between stimuli n and $n+1$ (n being a number along the abscissa), except in the case of 5, where it is the distance between stimuli 5 and 7 that is displayed. In spite of this, before training (ME-1, diamonds), d' is very low for this within-category comparison, almost as low as it is for the other within-category comparisons 1–2 and 10–11. Training has a significant effect here: although it does not affect discriminability of the stimuli at the ends of the range, it does teach subjects to tell stimuli 5 and 7 apart. On the other hand, stimuli 7 and 8 become more similar through training.

Experiment IIb was set up for exactly the same reasons as experiment Ib: to avoid truncations by widening the response range, and to avoid the effect of an anchor in the middle of the range. The subjects were the same as in experiment Ib, as was the procedure.

2. Experiment IIb

The results of the magnitude-estimation session involving stop consonants are shown in Fig. 13, separately for each of the four subjects. Subjects 1, 2, and 3 had three response categories; the one on the left and the one in the middle were well separated, but there was some uncertainty about the demarcation between the middle category and the one on the right, leading to some bimodality in the response distributions. Subject 4 did not have a middle category: he divided the /t/-like stimuli into two classes. The patterns for all four subjects had remained the same with respect to the middle stimuli since experiment IIa had been run nearly a year before, but this need not mean that the effect of training had persisted over the intervening period: it is inevitable that, if only three categories are heard, they come to be positioned the way they are here by subjects 1, 2, and 3.

Apart from some bimodality in the distributions of subjects 2 and 3, mainly evoked by stimuli 7 and 8, distributions in Fig. 13 are fairly normal. As long as we calculate d' for individual subjects, therefore, there is not much that can go wrong. This was done in Fig. 15 on the basis of the average calculated standard deviations shown in Fig. 14.

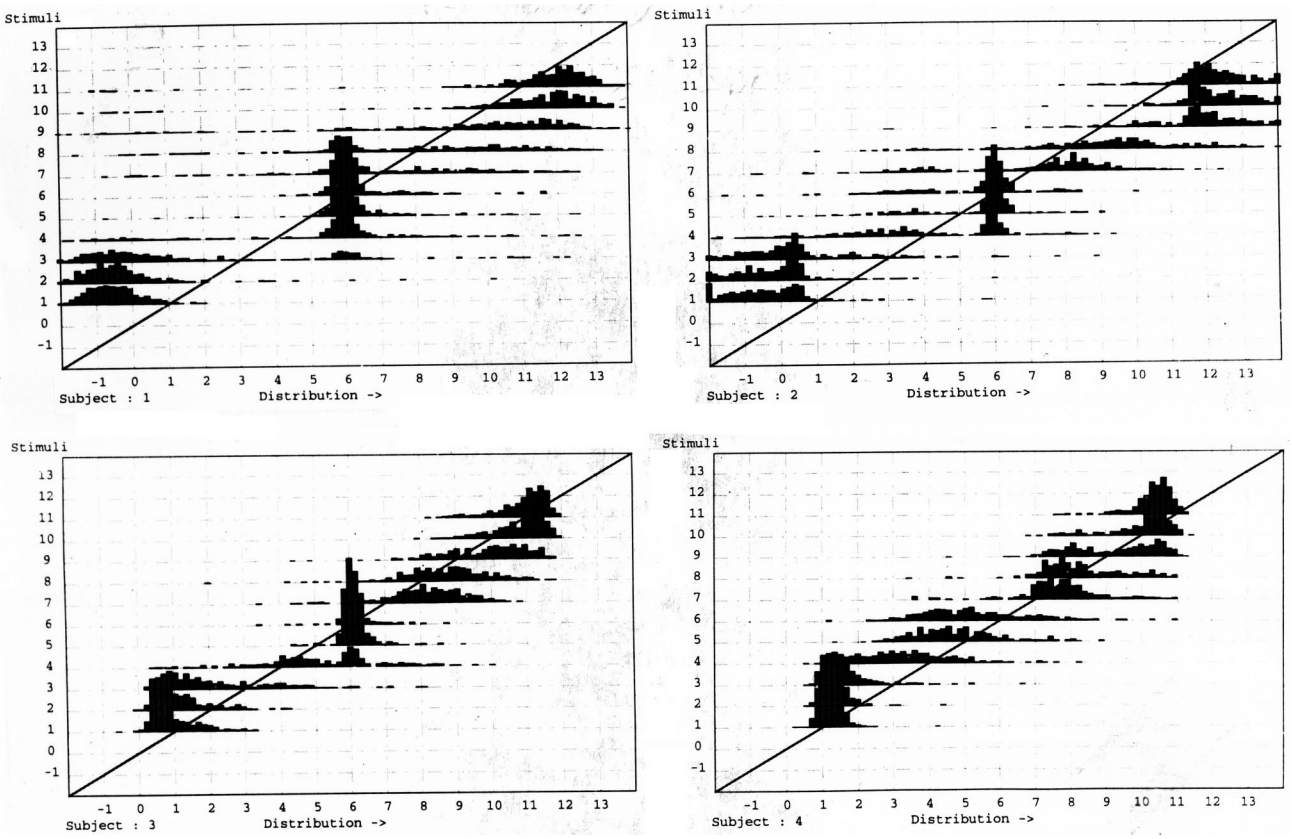

FIG. 13. Experiment IIb, stop consonants. Individual magnitude-estimation response distributions. Stimuli are spectral interpolations between /p/, /t/, and /k/.

III. GENERAL DISCUSSION

The assumption behind all the experiments reported in this paper is that it is possible, in principle, to obtain a good impression of the way the stimuli of a series are represented along a subject's hypothetical internal decision axis, by performing magnitude estimation or absolute identification. Both procedures should, we assume, yield the same response means for the stimuli but different variances, since part of the variance of any response, but not its mean, is determined by the task. Given the same conditions, i.e., the same knowledge about the stimuli, the means should always be at the same points on the decision axis, regardless of the task, but the average distribution variance should vary systematically from one task to another. If this is true, we can determine response mean and variance for each stimulus in one magnitude-estimation or absolute-identification experiment, and, using an appropriate equation for task-dependent variance, we can calculate the variance, and thus predict d', for each stimulus in each task. However, in practice it is usually impossible to separate task factors and knowledge of the stimuli. An example of what is meant here can be obtained by comparing Figs. 1 and 3. In principle, the prior conditions are the same in these two figures: subjects have been trained extensively in the identification of these 1000-Hz tones, and now they are asked to identify them (Fig. 1) and to estimate their magnitude (Fig. 3) again. We would therefore expect the same response means for each subject over the two tasks. Figures 1 and 3 show us, however, that, if any subjects behave like this, certainly not all of them do. The reason is, presumably, that no feedback was given during magnitude estimation, so that after a little while, stimulus knowledge fell behind that in absolute identification, where feedback was given after each response. This is a difference between the tasks, but it is mainly a difference in stimulus knowledge (the difference can be turned into a purely task-related one by omitting feedback in identification).

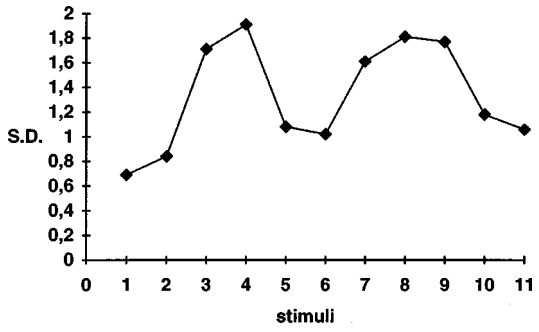

FIG. 14. Experiment IIb, stop consonants. Averaged standard deviations in magnitude estimation. Stimuli are spectral interpolations between /p/, /t/, and /k/.

Despite the differences in response means between identification and magnitude estimation, we feel that it is useful to stick to a model in which each stimulus has its own task-independent mean position on the decision axis, plus an amount of variance that is partly stimulus related, and may vary from stimulus to stimulus, and partly task related. The main advantage of this model is that it makes it possible to talk about the position of a stimulus along a decision axis, even where this position is not measured directly, such as in

any discrimination experiment. In addition, if we treat stimulus- and task-related variance as two independent noise sources, it becomes possible to determine the relative variances in the distributions caused by two stimuli. We need to know this if we want to decide whether these variances are near enough to yield a ROC with a slope of unity; if they are not, we will have to determine d' directly, i.e., by expressing stimulus distance along the decision axis in terms of some form of averaged standard deviation.

On the basis of these assumptions, which were not strongly contradicted by any of the findings, we tested the hypothesis that the distributions caused by pure-tone stimuli differing only in level are equal in variance and Gaussian or at least unimodal, but that a series of speech stimuli, differing in much more complicated ways from each other, and influenced by long-term memory categorization, might not cause equal, Gaussian distributions. We did not know what to predict with respect to the shape of the response distributions for speech sounds, but we did expect relatively large differences in variance, with stimuli from the center of a phoneme category leading to much narrower distributions than stimuli at or near a phoneme boundary, which could be heard as belonging to different categories from one presentation to another, and thus lead to much wider, perhaps even bimodal, response distributions. In short, we expected to confirm that d' for simple psychoacoustic stimuli can safely be calculated in the traditional way, using just one pair of z-transformed hit and false-alarm probabilities, but that the d' values that have up to now been calculated for speech sounds are much less valid.

The experiments have not confirmed the expectations. In both experiments, we have found evidence that, as long as distributions are not affected by lack of response space, they tended to be Gaussian, both for the tones and for the speech stimuli. However, in neither set of stimuli were they equal. This can be seen in Figs. 4, 7, 11, and 14. Tone stimuli (Figs. 4 and 7) from the middle of the range are more difficult to identify than stimuli at either end, which agrees with the anchor effect described by Braida *et al.* (1984): subjects construct their own references, which usually coincide with the end points of the stimulus range, and use a ''noisy ruler'' to measure the distance between each stimulus and these anchors. For speech stimuli, such anchors do not have to be constructed: provided the stimuli are close enough to natural speech sounds, the anchors exist already—they form part of the ''permanent context'' (Schouten and van Hessen, 1992). Figures 11 and 14 show that, as expected, stimuli that are close to such a permanent anchor are easier to position on the decision axis than stimuli that are further away from one; the transition between these two states is, moreover, relatively sudden or ''categorical.'' The d' values calculated on the basis of the standard deviations in Figs. 7 and 14 are shown in Figs. 8 and 15, respectively. One might expect a negative correlation between standard deviation and d', but there seems to be no correlation between Figs. 7 and 8 at all, whereas there appears to be a positive correlation between the speech data in Figs. 14 and 15: a high standard deviation seems to be associated with a high d'. This is due to the rather special, categorical nature of the perception of well-known speech sounds: stimuli belonging to a single category are positioned very closely together, and this is done very consistently.

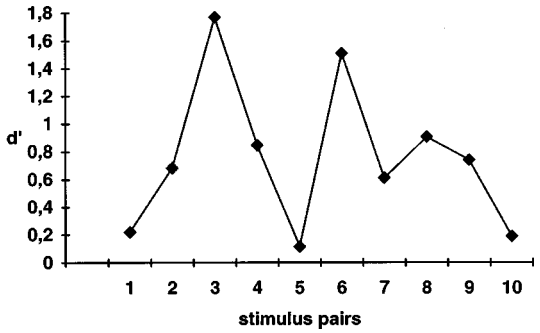

FIG. 15. Experiment IIb, stop consonants. Magnitude-estimation d' values, calculated by dividing the difference between two distribution means by their averaged standard deviations. Stimuli are spectral interpolations between /p/, /t/, and /k/.

What lesson should be learned from all this in relation to the calculation of d'? How serious is the deviation from equal variance, i.e., how much does the slope of the various ROC curves deviate from unity? The answers for intensity perception and speech perception are different. As Figs. 4 and 7 show, standard deviations do not change much from one intensity stimulus to the next, so if stimulus comparisons are restricted to nearest neighbors, one pair of $z(\mathrm{H})$ and $z(\mathrm{FA})$ estimates will produce a d' that is very close to the ''real,'' underlying d'. The deviation will become more serious as more distant stimuli are compared in an experiment.

For speech stimuli (see Figs. 11 and 14) the situation is much more serious: only for stimuli that unambiguously belong to the same category can d' be based on just one $z(\mathrm{H})-z(\mathrm{FA})$ pair; in all other cases we must expect a severe departure from equality of variance. Does this conclusion invalidate all speech d' values that have been collected so far? Fortunately, at least in the present speech data, the negative effects are compensated for: in regions of high variance, perceptual distances are great, due to relatively great distances between mean positions. Although such a compensation should not be taken for granted under all circumstances, it does seem to indicate that the standard procedure for calculating d' is, in most cases, robust enough.

ACKNOWLEDGMENTS

This research was supported by the Foundation for Language, Speech, and Logic (TSL), which is subsidized by the Netherlands Organisation for Scientific Research (NWO). The authors would like to thank Dick Pastore and Neil Macmillan for helping us to get a better grip on our own concerns.

Braida, L. D., and Durlach, N. I. (**1972**). ''Intensity perception. II. Resolution in one-interval paradigms,'' J. Acoust. Soc. Am. **51**, 483–502.

Braida, L. D., Durlach, N. I., Lim, J. S., Berliner, J. E., Rabinowitz, W. M., and Purks, S. R. (**1984**). ''Intensity perception. XIII. Perceptual anchor model of context coding,'' J. Acoust. Soc. Am. **76**, 722–731.

Cowan, N., and Morse, P. A. (**1986**). ''The use of auditory and phonetic memory in vowel discrimination,'' J. Acoust. Soc. Am. **79**, 500–507.

Macmillan, N. A., Braida, L. D., and Goldberg, R. F. (**1987**). ''Central and peripheral processes in the perception of speech and nonspeech sounds,'' in *The Psychophysics of Speech Perception*, edited by M. E. H. Schouten (Martinus Nijhoff, The Hague), pp. 28–45.

Macmillan, N. A., Goldberg, R. F., and Braida, L. D. (**1988**). ''Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua,'' J. Acoust. Soc. Am. **84**, 1262–1280.

Macmillan, N. A., Kaplan, H. L., and Creelman, C. D. (**1977**). ''The psychophysics of categorical perception,'' Psychol. Rev. **84**, 452–471.

Pisoni, D. B. (**1973**). ''Auditory and phonetic memory codes in the discrimination of consonants and vowels,'' Percept. Psychophys. **13**, 253–260.

Rosner, B. S. (**1984**). ''Perception of voice-onset-time continua: A signal detection analysis,'' J. Acoust. Soc. Am. **75**, 1231–1242.

Samuel, A. G. (**1987**). ''Lexical uniqueness effects on phonemic restoration,'' Journal of Memory and Language **26**, 36–56.

Schouten, M. E. H., and van Hessen, A. J. (**1992**). ''Modeling phoneme perception. I: Categorical perception,'' J. Acoust. Soc. Am. **92**, 1841–1855.

Swets, J. A., Tanner, W. P., and Birdsall, T. G. (**1961**). ''Decision processes in perception,'' Psychol. Bull. **68**, 301–340.

Uchanski, R. M., Millier, K. M., Reed, C. M., and Braida, L. D. (**1992**). ''Effects of token variability on resolution for vowel sounds,'' in *The Auditory Processing of Speech: From Sounds to Words*, edited by M. E. H. Schouten (Mouton De Gruyter, Berlin), pp. 291–302.

van Hessen, A. J., and Schouten, M. E. H. (**1992**). ''Modeling phoneme perception. II. A model of stop consonant discrimination,'' J. Acoust. Soc. Am. **92**, 1856–1868.