Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua

Neil A. Macmillan,* Rina F. Goldberg,* and Louis D. Braida
Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 6 January 1988; accepted for publication 23 June 1988)

Discrimination and identification experiments were performed for a vowel continuum (/i/-/I/-/e/) and two consonant continua (/ba/-/pa/ and /ba/-/da/-/ga/). The results were interpreted in terms of a generalization of a theory of intensity resolution [N. I. Durlach and L. D. Braida, J. Acoust. Soc. Am. 46, 372–383 (1969)] that makes precise the distinction between basic sensitivity (sensory-based resolution) and context coding (labeling processes). On the vowel continuum, basic sensitivity increased gradually across the range, whereas, for both consonant continua, sensitivity peaked between phonetic categories. All speech continua were found to have small ranges (measured in jnd’s); context memory was good, and better for consonants than for vowels. The stimuli that could be labeled most reliably were near the category boundaries on the vowel continuum, but near good phonetic exemplars for consonants. Introduction of a standard in identification primarily altered response bias, not sensitivity.

PACS numbers: 43.71.Es, 43.71.An, 43.66.Ba, 43.66.Lj

INTRODUCTION

A fundamental psychophysical enterprise explores resolving power, the ability to distinguish similar stimuli. Much psychoacoustic work, in both speech and nonspeech perception, maps this terrain by varying physical characteristics of stimuli and measuring correlated differences in resolution.

This approach can be successful when the processes limiting performance are primarily sensory. Resolution can change for reasons unrelated to stimulus characteristics, however, when performance is heavily influenced by memory and decision factors. Such cognitive processes are pervasive in the perception even of simple stimuli, and are of interest in their own right. Investigators have attempted to distinguish sensory from cognitive effects by assuming qualitative differences between subjects (e.g., humans versus animals, adults versus infants) or stimuli (speech versus nonspeech, words versus nonwords), but a thorough understanding of either sensation or cognition requires a systematic theory that incorporates both. In this article, we present a quantitative model that explicitly relates resolution to sensory and memory effects, and use it to study the perception of synthetic vowels and consonants.

Our experiments measure resolution along speech sound continua using a variety of discrimination and identification paradigms. An early experiment of this type was conducted by Liberman et al. (1957), who interpreted their data as demonstrating "categorical perception." According to this hypothesis, discrimination between stimuli depends on their being assigned to different categories. Thus stimuli from a synthetic /be/-/de/-/ge/ continuum falling on either side of a phonetic boundary are better discriminated than those falling in the same category, and discrimination "peaks" are observed between /be/ and /de/, and again between /de/ and /ge/. Much subsequent research has studied the categorical perception phenomenon for different sets of stimuli.

The present work focuses not on categorical perception per se (in Sec. IV, we suggest that the hypothesis blurs some important distinctions), but on a number of issues that have arisen in research following the categorical perception tradition. Specifically, these questions are: (1) What is the pattern of sensory resolution on different continua? (2) How good are listeners at labeling sounds on speech dimensions? (3) Which specific stimuli are easy to label? and (4) Does the presentation of standards influence labeling behavior by affecting sensitivity or response proclivity? Presently available answers to these questions are ambiguous, we believe, for lack of sufficiently detailed analysis. Our model provides a language for stating the issues, motivates techniques for addressing them experimentally, and leads to some new (and some old) conclusions.

A. Basic auditory sensitivity

Several authors have interpreted observed patterns of discrimination in terms of sensory resolution. Peaked discrimination functions resembling those found by Liberman et al. (1957) for speech sounds have been reported by Miller et al. (1976) and Pisoni (1977) for nonspeech continua. Miller et al. interpreted their data as revealing a "psychoacoustic threshold," rather than mediation by labeling. The importance of sensory resolution is stressed in Pastore et al.’s (1977) common-factor theory, and by Stevens (1981).
Rosen and Howell (1987) provide a summary of attempts to interpret categorical perception data in terms of the basic resolution of the sensory system. In support of this view, proponents point to data from animals and infants that resemble adult human data, and argue that psychoacoustic explanations provide a more unified and parsimonious account than those based on labeling.

We consider the pattern of sensory resolution to be of interest in its own right, not just as an explanation of other phenomena. Our model specifies a procedure for measuring this pattern in adult human subjects. "Basic sensitivity" so obtained then provides a baseline for evaluating data from other experimental paradigms.

B. General labeling ability

With few exceptions, speech identification experiments have restricted responses to phoneme names. The subjects of Liberman et al. (1957), for example, could label a sound as "be," "de," or "ge." Identification performance has typically been considered ideal if, for each distinct sound on a continuum, the same distinct response is assigned on every presentation (Studdert-Kennedy et al., 1970).

If, however, we wish to measure the listener's optimal labeling ability, then the use of a small number of responses can lead to a severe underestimate of performance. An observer who is able to distinguish perfectly four points on a continuum cannot demonstrate this prowess with two or three responses. Macmillan et al. (1977) suggested the use of graded responses to guard against such low-fidelity outcomes, and we have used such response sets here. In particular, we have employed the "absolute identification" design, in which each stimulus has a corresponding "correct" response assigned by the experimenter. For purposes of comparing these experiments with conventional ones, we assume that observers would be able to partition their responses into phonetic categories if asked.

Speech researchers who conceive their goal to be an understanding of phonetic identification may see little purpose in providing more response categories than phonetic ones. An analogous approach to sensory psychology is that of Stevens (1975), who also argued that (a) resolution data are unhelpful in understanding perception and (b) perception can be equated with an observer's response. We find neither attitude congenial. Point (a) is ultimately empirical, and we believe that the resolution approach has proved valuable in understanding many sensory dimensions. Point (b) denies the power of decision-theory strategies for separating sensitivity from bias effects, and might fairly be called noncognitive.

The resolution approach to one-dimensional perceptual continua, which dates to Fechner (1860/1966), has been productive in many content areas; we see no reason to deprive speech perception of these tools.

C. Specific labeling ability

Some points on a continuum may be easier to label than others. The two most natural candidates for such reference points are "prototypes" (e.g., Oden and Massaro, 1978), typical members of a category; and the "boundaries" that divide categories (e.g., Diehl et al., 1980). Our model provides a method for locating stimuli that are memorable in this sense.

D. Context effects

In phonetic identification experiments (i.e., those in which the only possible responses correspond to phonetic categories), labeling can be affected by preceding the stimulus to be judged with a "standard" sound from the same continuum. A commonly observed context effect (found, for example, by Repp et al., 1979, for vowels, and Diehl et al., 1978, for consonants) is one of contrast: A sound near one end of a continuum decreases the likelihood of a response corresponding to that endpoint. Our model enables us to determine whether this effect is primarily a sensitivity or a response-bias change.

I. OUTLINE OF THE MODEL

A. General description

Our model is derived from Durlach and Braida's (1969) theory of intensity resolution. The original statement of the theory is in a series of articles in this Journal (most recently, Braida et al., 1984, which contains references to others). Concise recent summaries have been provided by Braida and Durlach (1988) and Macmillan (1987).

Although the theory was originally formulated for intensity perception, those postulates specific to intensity can be easily separated from general ones about memory and decision processes. The model requires only that stimulus sets to which it is applied vary perceptually on a single dimension. [For example, Searle et al. (1976) applied the model to auditory localization.] Stimuli generate distributions of values along this dimension that are normal, differing only in mean. Responses are determined by the location of adjustable criteria, as in signal detection theory (SDT; Green and Swets, 1974).

The primary performance statistic is the $d'$ of SDT, the normalized distance between distribution means, which measures sensitivity. A second useful statistic is criterion location, which is often termed response bias. The necessary data to apply the theory form confusion matrices. Subjects must perform less than perfectly; this requirement does not constrain the possible tasks greatly, but does lead us to avoid identification designs with small response sets. According to the theory (and the data of Braida and Durlach, 1972), resolution is independent of whether the number of responses is greater than, equal to, or less than the number of stimuli.

For a given experimental paradigm, the theory specifies the sources of variance that limit performance, and the method of their combination. All tasks are encumbered with sensory variance, the common irreducible variability that arises from such processes as neural transduction. In addition, one or both of two types of memory constraints may apply. Context variance limits the observer's ability to label stimuli, and increases with stimulus range. Trace variance limits the ability to compare two stimuli presented at different times, and increases with interstimulus interval (ISI). When context variance is small, so that context coding is the
best strategy, the listener is said to be in the context mode; when trace coding is more efficient, the listener is said to be in the trace mode.

A possible mechanism by which context variance depends on range has been recently described by Braida et al. (1984). Listeners compared sensations with perceptual anchors; on the intensity continuum, these are located at the extremes of the stimulus range. Comparisons are made using a "noisy ruler" (with a fixed number of measurement intervals) to estimate the distance between a stimulus event and an anchor; the greater this distance, the less accurately the stimulus is perceived. This postulate accounts for the observed decline in performance with increasing range, and also for local maxima in sensitivity near the extremes.

The variance parameters of the model can be estimated by comparing data from different discrimination and identification conditions. In the next section, we describe these tasks, and the predictions of the model, in detail; the tasks resemble paradigms used previously to study speech sound resolution, but differ in important ways from the methods most often employed in such work. After delineating our strategy for applying the model, we compare it with alternatives. In experiments I and II, we report data on vowels and consonant continua, respectively, to which the model can be applied.

B. Predictions for specific tasks

In these experiments, we used four tasks: fixed and roving discrimination, and identification with and without a standard. We describe these tasks in order of their complexity, as viewed through the model.

1. Fixed discrimination

In fixed discrimination, only two different stimuli, A and B, are presented during a block of trials. A single trial may contain one or more stimuli, depending on the discrimination paradigm used. In the present study, we used two interval forced choice (2IFC), in which the possible stimuli are (A,B) and (B,A); and same-different (SD), in which the possible stimuli are (A,A), (B,B), (A,B), and (B,A), and the listener responds "same" or "different."

Both paradigms, and indeed all tasks used in these experiments, can be analyzed in SDT terms to abstract $d'$. According to the Durlach–Braida model, fixed discrimination performance is determined, to a first approximation, only by stimulus characteristics and sensory variance. Context variance is small because the range is very small; we shall see below that, under these conditions, the magnitude of trace variance does not matter. Thus, if $\beta^2$ is the sensory variance and $\alpha$ is the mean sensation difference,

$$d'_{\text{fixed}} = \alpha / \beta.$$  \hfill (1)

2. Identification

In identification, a single stimulus is presented for judgment on each trial. The number of possible responses may be as small as two, or (as in the present experiments) as large as the number of stimuli. A large response set was dictated by a desire to measure our observer's optimal labeling ability (see the Introduction).

Identification is assumed to employ only the context mode, and the Durlach–Braida theory asserts that the variance limiting identification performance is the sum of sensory and context variance; that is,

$$d'_{\text{ident}} = \alpha / [\beta^2 + (GR)^2]^{1/2},$$  \hfill (2)

where $R$ is the stimulus range and $G$ is a constant.

The absolute magnitude of the sensory variance, $\beta^2$, plays the role of a scale factor in the theory, so a useful measure of context memory is the relative context variance $(GR/\beta)^2$. According to Eq. (2), data from identification and fixed discrimination can be used to estimate the relative context variance:

$$(GR/\beta)^2 = [d'_{\text{fixed}} / d'_{\text{ident}}]^2 - 1. \hfill (3)$$

We use measures of relative context variance in two ways, globally and locally. To answer the global question, how large is overall context variance, relative to sensory variance, Eq. (3) can be applied to total $d'$, the sum of $d'$ values across the range. Locally, sensitivity is greatest near perceptual anchors, which can therefore be located using Eq. (3). An anchor is thus operationally defined as a position on an internal continuum corresponding to a stimulus that is relatively easy to label consistently, compared to basic sensitivity in that region. In intensity perception, anchors lie at the edges of the stimulus range, but this need not be true on other continua.

3. Identification with a standard

In this task, the identification paradigm is modified by preceding each sound to be judged by a constant standard sound. In detection-theory terms, standards provide a context that could influence either sensitivity or response bias. A model for sensitivity changes developed by Khazatsky (1985) assumed that the observer estimates the perceptual distance of a sound from both the standard and the nearest edge anchor. The model shows that standards can improve performance, especially in the region of the standard, and accounts for the data of Berliner et al. (1978) on large-range intensity identification. Berliner et al. found that a standard in the middle of the range improved $d'$ overall by 28%, and locally by as much as a factor of 2 to 1.

Analysis of the data in detection-theory terms is the same as in no-standard identification. The theory makes no specific predictions about the effects of context on response bias, but does permit straightforward separation of sensitivity and bias.

4. Roving discrimination

In roving discrimination, the two stimuli to be discriminated vary from trial to trial within a block. If there are ten stimuli, the first trial may use stimuli 3 and 2, the second stimuli 7 and 8, etc. We used the same paradigms in roving discrimination as in fixed discrimination: 2IFC and same-different.

Either the context or the trace mode can be used in rov-
ing discrimination. In the trace mode, the listener in 2IFC compares the second stimulus with the “trace” of the first, finds (for example) that the second is greater, and responds “2.” In the context mode, the listener covertly labels the first sound as stimulus 4, the second as stimulus 5, and responds “2.” The model assumes that the two processes are combined in an optimal fashion, in which case (Durlach and Braida, 1969),

$$d'_{\text{roving}} = a/\{eta^2 + [(GR)^{-2} + (AT)^{-1}]^{-1}\}^{1/2} \, (4)$$

where $T$ is the ISI and $A$ is a constant (see section D below for more discussion of this relation).

Several implications of Eq. (4) are noteworthy. First, if the context variance $(GR)^2$ is much smaller than the trace variance $AT$, then the latter contributes little to the denominator, and Eq. (4) reduces to Eq. (2): Performance is in the context mode. Analogously, if the trace mode is more efficient, only trace and sensory variance matter. Finally, if the context variance is small relative to both sensory and trace variance, the denominator of Eq. (4) reduces to $\beta$; i.e., Eq. (1) applies. This condition is satisfied in fixed discrimination of small differences, as asserted above.

Assessing the trace mode in roving discrimination requires varying the ISI, a manipulation not used in the present experiments. (Pisoni, 1973, varied ISI in a vowel discrimination experiment in order to study memory for vowels, albeit not in the context of the Durlach–Braida theory.) Inference about context memory are therefore on much firmer ground, and will be accorded much more attention than those about trace coding. Watson et al. (1976) also argue for fixed paradigms in measuring basic sensitivity.

**C. Applying the model**

The theory motivates the following approach to studying vowel and consonant continua. First, we estimate basic sensitivity in a fixed discrimination task, across the continuum. The pattern of sensitivity reveals whether the continuum contains a “natural boundary,” that is, a region of high basic sensitivity. Second, we estimate the relative context variance globally by comparing total sensitivity in identification and in fixed discrimination. This statistic measures the ability of listeners to label stimuli on the continuum. Third, we examine relative context variance locally to determine the continuum locations (anchors) that are particularly easy to context code. Finally, we ask whether the effects of context in identification with a standard are due to response bias, sensitivity, or both.

The theory makes two specific, qualitative predictions. First, there should be an ordering of performance across conditions, with fixed discrimination best, followed by roving discrimination, then identification. This prediction follows directly from the equations in this section. The ordering need not be strict—fixed and roving discrimination can be equal if either memory variance is very small, and roving discrimination can be as good as identification if trace variance is much larger than context variance—but it cannot be worse. Second, the theory predicts a range effect in identification: Sensitivity should be better for smaller ranges. These predictions [and more detailed ones based on Eqs. (1)–(4)] have been supported in experiments on the auditory intensity continuum (see, for example, Braida and Durlach, 1972, and Berliner and Durlach, 1973).

**D. Comparison with other models**

Our model provides a framework for relating performance in different psychophysical tasks, and for asking questions about sensory and memory effects on perceptual continua. Its role is therefore similar to that often played, in interpreting data of this sort, by the quantitative statement of categorical perception offered by Liberman et al. (1957), the “Haskins model.” (For a summary of the Haskins model, see Macmillan, 1987.) Like our detection-theoretic account, the Haskins model prescribes a relation between identification and discrimination along a perceptual continuum. One difference between the models is the explicit role of memory in our approach, but a modified version of the Haskins model, the dual-process model (Fujisaki and Kawashima, 1970; Pisoni, 1975) does incorporate a memory process. (To complete the symmetry, a memoryless SDT model was proposed earlier by Macmillan et al., 1977. The importance of describing memory explicitly in these models seems now to be generally accepted; see Schouten, 1987.) In view of these similarities between our approach and the Haskins model, a shift away from the familiar requires justification.

In spite of the similarities, the models do make predictions that differ in detail. In general, these comparisons favor the detection-theory approach (Macmillan, 1987). We do not focus on such fine-grain analysis here, but suggest two broader advantages of the SDT account: its psychophysical plausibility and its utility in describing data.

**1. Psychophysical considerations**

The model of Liberman et al. (1957) is a “low-threshold” model of the class later proposed explicitly by Luce (1963a); it differs only in assuming the absence of response bias (Macmillan et al., 1977). In threshold theories, discrimination decisions are based on discrete intervening events—in this case, the phonetic categories to which each stimulus is covertly assigned.

In fact, the very use of proportion correct is now seen to have threshold implications (Snodgrass and Corwin, 1988; Macmillan and Kaplan, 1985; Swets, 1986a). Thus the common assumption that data expressed as proportion correct are less theory bound than those expressed by $d'$ is incorrect. Any measure of sensitivity is consistent with a decision process in which underlying distributions overlap on a decision axis. The use of $d'$ implies that these distributions are normal; the use of proportion correct implies that they are discrete (or uniform). Data in the form of ROC curves support normal distributions for a wide variety of perceptual and cognitive continua (Swets, 1986b).

The treatment of identification data raises a similar issue. When the response to a stimulus in identification is summarized by its average rating—the mean category scale (Luce and Galanter, 1963)—the ratings are assumed interval scaled; in addition, sensitivity and bias cannot be separated. The latter problem is also present if ratings are not used.
2. Considerations of utility

The psychophysical disadvantages of threshold theory can be significant, but are minimal in the important special case of unbiased responding. In many discrimination designs popular in speech research (e.g., same-different and ABX), response bias tends to be large, but in the 2IFC procedure used (for the most part) here it does not. With care, the distorting consequences of using threshold measures can be held in check. More important is the limited utility of such measures. Our detection-theory model offers at least three practical advantages: a dependent variable (d') with distance-measure characteristics; a more explicit connection with important experimental variables; and an account of memory processes that is both more general and more detailed.

That d' is a distance measure (Luce, 1963b) solves problems that arise within an experiment and in comparing different studies. (Both advantages are illustrated, in the present article, in experiment IIA.) Within an experiment on a single perceptual dimension, it is often difficult to find a constant stimulus difference such that performance is always above chance but less than perfect. When differences of varying size are used, comparisons are difficult: There is no way to determine whether, say, a proportion correct of 0.69 for a stimulus difference of three units is more or less impressive than a proportion of 0.84 for a difference of six units. But if d' values of 1.0 and 2.0 are measured (the equivalent of the yes–no proportions correct 0.69 and 0.84, for unbiased responding), then it is possible to compute d' per unit, which equals one-third in both cases. In comparing different experiments, which rarely choose the same exact stimulus spacing, this property of d' is especially useful.

The Durlach–Braida model refers to only two experimental variables, the stimulus range R and the interstimulus interval T. These variables are basic, however, and are not mentioned in threshold theories (the lack of stimulus models has been a weakness of threshold theory generally; see Luce, 1963b). The particular relations between the decision variance and these variables are theoretically motivated. Context coding variance is proportional to the square of the range (this follows from the assumptions of the model; see Durlach and Braida, 1969, and Braida et al., 1984). The linear relation between trace variance and T arises from a diffusion model of the trace mode (Kinchla and Smyzer, 1967). The systematic incorporation of these variables makes possible quantitative predictions about their effects. A special case of the dependence on range is the predicted difference between fixed and roving discrimination, a distinction not made by threshold theory.

Memory processing in the Durlach–Braida model is more general than the threshold version in being less tied to a specific information-processing sequence. In the dual-process model, memory influences the response only if sensory factors are ambiguous. Thus, in the ABX design (the only paradigm for which the model has been specified), memory processes are invoked only if the A and B stimuli lead to the same covert label. The model cannot be applied to designs like same–different, in which the two covert labels are always either the same or different, so that no ambiguity is possible. Attempts to apply dual-process ideas to the same–different paradigm (e.g., Repp et al., 1979) have been only qualitative.

Our memory model is more detailed than the dual-process model in two ways. First, it specifies two types of memory, each influenced by a different experimental variable. Second, it proposes a process—comparison with perceptual anchors—by which context memory operates. The anchor subtheory is the only quantitative model that addresses in any way the emergence of peaks in discrimination; other models concern themselves entirely with the discrimination–identification relation.

These advantages have paid dividends in the study of auditory intensity, an observation that suggests yet another advantage. Not only does the model permit comparisons across experiments, it permits comparisons across sensory dimensions. Later in this article, we compare the role of memory on vowel and consonant continua with its role on the intensity dimension.

To summarize, the data reported here are not designed to distinguish between quantitative models. Our model might, in fact, be viewed as a realization, in psychophysically plausible terms, of the ideas of the dual-process model. We offer a more systematic method of data analysis, not a new view of speech perception. The present model enables us to pose more precise questions (or to pose old questions more precisely). These experiments do not provide a test of whether the model is correct, but of whether it is useful.

II. EXPERIMENT I—VOWELS

A. Stimuli

Thirteen 256-ms steady-state vowel stimuli ranging from /i/ (stimulus 1) to /u/ (stimulus 7) to /e/ (stimulus 13) were synthesized using Klatt's (1980) software cascade synthesizer with glottal excitation pulses as described by Rabiner and Schafer (1978, p. 103). The starting center frequencies of the first three formants were set to the values given by Stevens et al. (1969), listed in Table I, and were approximately logarithmically spaced. The center frequen-

| TABLE I. Vowel stimuli in experiment I. | [Last column is the Euclidean distance between stimuli i and i + 1 in the space described by Miller and Chang (1986); coordinates are log (F1/F0), log(F2/F1), and log (F3/F2). Absolute distance is arbitrary.] |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| Stimulus No. | F1 | F2 | F3 | Distance |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| 1 | 269 | 2296 | 3019 | 1.008 |
| 2 | 285 | 2263 | 2955 | 0.755 |
| 3 | 327 | 2230 | 2912 | 1.082 |
| 4 | 315 | 2183 | 2829 | 1.112 |
| 5 | 336 | 2151 | 2769 | 0.982 |
| 6 | 354 | 2105 | 2709 | 1.000 |
| 7 | 375 | 2075 | 2670 | 1.060 |
| 8 | 397 | 2030 | 2632 | 0.988 |
| 9 | 420 | 2001 | 2567 | 0.905 |
| 10 | 442 | 1973 | 2557 | 1.200 |
| 11 | 472 | 1930 | 2539 | 1.005 |
| 12 | 500 | 1902 | 2520 | 1.072 |
| 13 | 530 | 1862 | 2484 | ⋮ |
cies of the fourth and fifth formants were 3500 and 4500 Hz. The bandwidths for formants one to five were 69, 200, 400, 250, and 200 Hz. Over the duration of the stimuli, F0 declined by 36% and intensity by 48%. Four tokens were generated for each stimulus by varying fundamental frequency (F0) from 100 to 130 Hz in 10-Hz steps. In addition, the intensity of the stimuli varied uniformly over a 10-dB range. Variation in F0 and intensity was introduced to reduce the likelihood of responses based on idiosyncratic speech-irrelevant characteristics of the stimuli. The range of F0 variation used is about four times as large as the intrinsic F0 difference between the vowels /i/ and /e/ spoken by males (Peterson and Barney, 1952). However, for the range of F0 used, very little effect of fundamental frequency on vowel quality is to be expected (e.g., Traunmüller, 1981; Syrdal and Gopal, 1986).

A PDP12 computer played the digitally stored waveforms through a D/A converter at a sampling rate of 10 kHz. The waveforms were then low-pass filtered at 4.5 kHz, attenuated, and presented at roughly 70 dB SPL through the right earphone of headphones to subjects sitting in a sound-treated room.

B. Design

Two subexperiments were conducted with these stimuli. In experiment IA, identification without a standard and fixed 2IFC discrimination were measured with the full set of 13 stimulus values. In experiment IB, stimuli 1–7 were used in a wider range of tasks: identification without a standard, identification with a standard (at stimuli 1, 4, and 7), and both fixed and roving discrimination. Discrimination performance was measured using both 2IFC and same–different paradigms. Table II lists the conditions comprising the experiment, and the approximate number of trials for each condition and stimulus comparison.

C. Subjects

Seven MIT undergraduates (six of them male) with no known hearing defects served as paid subjects, four in experiment IA, and three in experiment IB. None had previously participated in hearing experiments.

D. Procedure

1. Trial structure

In the discrimination paradigms, two sounds separated by 400 ms were presented; the listener had 2200 ms to respond by pressing a key on a response box, after which correct–response feedback was presented visually. Discrimination was measured for adjacent stimuli on the continuum.

In identification without a standard, one stimulus was played on each trial, and the subject had 2750 ms to identify it by pressing the button labeled with its stimulus number on the response box. In identification with a standard, the standard was presented 750 ms before the random stimulus. The number corresponding to the correct response was displayed as feedback.

2. Experiment structure

Experiment IA was conducted in 22 two-h sessions, the first 6 of which were practice. A session contained only one of the five possible tasks (identification with stimuli 1–13, 1–7, 4–10, or 7–13, or discrimination), and consisted of 10–12 blocks of 40–80 trials. The first block of each session and the first five trials of every block were considered practice and were not included in the data analysis.

Experiment IB was conducted in 34 two-h sessions, the first 7 of which were practice. The sessions alternated between identification and discrimination tasks. An identification session consisted of, alternately, 1-h periods of identification without a standard and identification with a standard at stimulus 1, or identification with standards at stimuli 4 and 7. Discrimination sessions alternated between the 2IFC and same–different paradigms. Each session consisted of roving discrimination, and fixed discrimination of two different stimulus pairs. Blocks of trials, including practice blocks, were structured as in experiment IA.

The design of the experiment assumes that transfer effects between conditions are slight. Other work with the same stimuli (Macmillan and Braida, 1985) has shown that learning effects are essentially complete in a few hours; the practice blocks were intended to eliminate any remaining sequential dependencies between conditions.

E. Data analysis

The data from each experimental condition for each subject were arranged in a separate confusion matrix. The primary analysis was in SDT terms: One-interval d’ values were calculated from each confusion matrix using the methods of Braida and Durlach (1972) for identification tasks,

<table>
<thead>
<tr>
<th>Exp.</th>
<th>Paradigm</th>
<th>Stimuli</th>
<th>Approximate No. of trials per subject</th>
<th>Approximate No. of trials per stimulus per subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>IA</td>
<td>Ident-no std</td>
<td>1–13</td>
<td>2156</td>
<td>166</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1–7</td>
<td>1755</td>
<td>251</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4–10</td>
<td>900</td>
<td>129</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7–13</td>
<td>1350</td>
<td>193</td>
</tr>
<tr>
<td>IA</td>
<td>Fixed discrim</td>
<td>1–13</td>
<td>2700</td>
<td>225</td>
</tr>
<tr>
<td>IB</td>
<td>Ident-no std</td>
<td>1–7</td>
<td>1462</td>
<td>209</td>
</tr>
<tr>
<td></td>
<td>Ident with std</td>
<td>1–7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Std = 1</td>
<td></td>
<td>1360</td>
<td>194</td>
</tr>
<tr>
<td></td>
<td>Std = 4</td>
<td></td>
<td>1212</td>
<td>173</td>
</tr>
<tr>
<td></td>
<td>Std = 7</td>
<td></td>
<td>1063</td>
<td>152</td>
</tr>
<tr>
<td>IB</td>
<td>Fixed discrim</td>
<td>1–7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2IFC</td>
<td></td>
<td>1827</td>
<td>304</td>
</tr>
<tr>
<td></td>
<td>Same–diff</td>
<td></td>
<td>2099</td>
<td>350</td>
</tr>
<tr>
<td>IB</td>
<td>Roving discrim</td>
<td>1–7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2IFC</td>
<td></td>
<td>1513</td>
<td>252</td>
</tr>
<tr>
<td></td>
<td>Same–diff</td>
<td></td>
<td>1444</td>
<td>327</td>
</tr>
</tbody>
</table>
Green and Swets (1974) for 2IFC discrimination, and Kaplan et al. (1978) for same-different discrimination. Note that all reported d' values are those theoretically expected in a one-interval task, even though both discrimination tasks have two intervals. Average d' values were calculated by taking the mean of the individual subjects' values. Standard errors were calculated using the methods of Gourevitch and Galanter (1967) for discrimination and Lippmann (1973) for identification (see the Appendix for details). In presenting the data, we display an average two-standard-error interval when all standard errors in a condition are comparable in magnitude; when differences are substantial, individual error bars are shown.

In threshold analyses, discrimination data are commonly presented in terms of proportion correct; identification data are presented as the proportion of trials on which one of the (phonetic) responses was used. To present the data in this way, we would need to reduce the 13 possible identification responses to three: for example, responses 1–4 might be considered /i/; responses 5–9, /l/; and responses 10–13, /e/. Then the proportion of /i/ /l/ and /e/ identifications could be plotted across the range. When such an analysis is performed, the data look very much like those of previous investigators. However, because we never asked our subjects to label stimuli phonetically, any response partition is arbitrary, and the data cannot be used to estimate a "category boundary." Accordingly, we do not usually present the data in this form, but instead plot sensitivity (or bias), which can be estimated from all paradigms. Identification with a standard is a special case, because the differences in responding across conditions do not strongly depend on the response partition chosen; in these conditions, data are analyzed in both proportion and SDT terms.

F. Results of experiment IA

Figure 1 shows the sensitivity pattern for identification and discrimination. Responses to stimuli of different fundamental frequencies were pooled because no significant differences were found among them. In identification, the same pattern of sensitivity was found for all values of F0. In roving discrimination, there was no difference between sensitivity when the two sounds had the same or different values of F0. In fixed discrimination, sensitivity was slightly (10%) but insignificantly greater when the same F0 was used (see Goldberg, 1986, for details).

There is a local minimum between stimuli 2 and 3; we argue in Sec. II H that this may reflect uneven physical spacing. There are no strong peaks in fixed discrimination, just a gradually increasing basic sensitivity across the range. Identification data do show peaks, for discriminating stimuli 4 vs 5 and 10 vs 11. These stimuli are intermediate between the best exemplars (as judged by the experimenters) of the underlying phonetic categories.

Identification in the limited-range conditions is shown in Fig. 2. Reducing the range serves to improve sensitivity, consistent with the assumption that context variance increases with range. Values of overall sensitivity (total d'), given in Table III, show that fixed discrimination (which may be viewed as identification with the smallest possible range) yields the best performance, followed by half-range and full-range identification of the same stimuli. For example, total sensitivity between stimuli 1 and 13 can be estimated in three ways: The sum of all fixed-discrimination d' values is 14.65; the sum of the 1–7 and 7–13 half-range identification values is 8.89; and total full-range identification d' is 7.34. The larger the stimulus range, the poorer the performance.

The apparent shifts in peak performance between half-range and full-range identification are a surprise. Because the variable plotted is sensitivity, not proportion identified, the effect is not the same as the "range effect" described by Rosen (1979).
TABLE III. Total $d'$ in experiment I.

<table>
<thead>
<tr>
<th>Exp.</th>
<th>Paradigm</th>
<th>Stimuli</th>
<th>Subrange</th>
<th>Total $d'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>IA</td>
<td>Ident-no std</td>
<td>1-13</td>
<td>1-13</td>
<td>7.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-7</td>
<td>1-7</td>
<td>3.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4-10</td>
<td>4-10</td>
<td>3.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7-13</td>
<td>7-13</td>
<td>3.85</td>
</tr>
<tr>
<td></td>
<td>Ident-no std</td>
<td>1-7</td>
<td>1-7</td>
<td>4.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4-10</td>
<td>4-10</td>
<td>6.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7-13</td>
<td>7-13</td>
<td>4.68</td>
</tr>
<tr>
<td>IA</td>
<td>Fixed discrim</td>
<td>1-13</td>
<td>1-13</td>
<td>14.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-7</td>
<td>1-7</td>
<td>5.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4-10</td>
<td>4-10</td>
<td>7.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7-13</td>
<td>7-13</td>
<td>8.79</td>
</tr>
<tr>
<td>IB</td>
<td>Ident-no std</td>
<td>1-7</td>
<td>1-7</td>
<td>3.58</td>
</tr>
<tr>
<td>IB</td>
<td>Ident with std</td>
<td>1-7</td>
<td>1-7</td>
<td>3.18</td>
</tr>
<tr>
<td></td>
<td>Std = 1</td>
<td>1-7</td>
<td>1-7</td>
<td>3.34</td>
</tr>
<tr>
<td></td>
<td>Std = 4</td>
<td>1-7</td>
<td>1-7</td>
<td>4.23</td>
</tr>
<tr>
<td></td>
<td>Std = 7</td>
<td>1-7</td>
<td>1-7</td>
<td>6.42</td>
</tr>
<tr>
<td>IB</td>
<td>Fixed discrim</td>
<td>1-7</td>
<td>1-7</td>
<td>7.11</td>
</tr>
<tr>
<td></td>
<td>2IFC</td>
<td>1-7</td>
<td>1-7</td>
<td>4.60</td>
</tr>
<tr>
<td></td>
<td>Same-diff</td>
<td>1-7</td>
<td>1-7</td>
<td>4.00</td>
</tr>
</tbody>
</table>

G. Results of experiment IB

1. Same–different versus 2IFC

We used two discrimination paradigms: 2IFC is often used in psychoacoustics; same–different is often used in speech perception. The same statistic $d'$ can be estimated from both paradigms; do the two conditions lead to the same value of sensitivity?

On the average, they do: The mean $d'$, across conditions, stimuli, and subjects, is 0.92 for 2IFC and 0.93 for same–different. There is a strong positive correlation between the two $d'$ values, as shown in Fig. 3, but there is also substantial scatter ($r = 0.69$). In addition, the figure reveals one systematic deviation: At low levels, same–different $d'$ is occasionally negative. These negative values of $d'$ correspond to below-chance performance, whether the data are summarized by $d'$ or proportion correct, and thus reflect nonoptimal listening strategies in this design. According to the Macmillan et al. (1977) model, $d'$ increases rapidly as performance edges above chance, so that even small random variation of observed proportions around values near chance can easily yield a substantial negative $d'$ in same–different. Apparently, then, the negative values of $d'$ cast a shadow on the same–different paradigm, not on the model. Because of these problems, we use only the 2IFC data in later analysis.

2. Sensitivity patterns

Average data for fixed and roving discrimination and identification without a standard are shown in Fig. 4. Peaks in sensitivity appear for roving discrimination and identif-
Identification with a standard at:  

**FIG. 5.** Threshold analysis of identification with a standard in experiment IB. Average of data from three subjects.

this partition is relatively unimportant. The figure clearly shows that the apparent boundary is affected by the presentation of a standard: Standards at stimuli 1, 4, and 7 yield boundaries of 3.7, 4.2, and 4.8, respectively. This result is a contrast effect of the type found previously (for example, by Repp et al., 1979).

The threshold approach leaves undecided whether the standards used in experiment IB led to sensitivity changes. Detection-theory analysis (Fig. 6; see also Table III) reveals that (except possibly for the /i/ standard) they did not. The effect of the /i/ standard is specific to its neighboring region. Presentation of a standard does have a noticeable effect on performance (Fig. 5), but sensitivity changes are slight (Fig. 6), so changes in response bias seem probable. The locations of the criteria, measured as a proportion of \( d' \), are shown in Fig. 7. A criterion value of zero means that the criterion is symmetrically located between two means; positive values indicate criteria above this point (toward the /i/ end of the continuum), negative values below. Values of ± 0.5 describe criteria located at the means of the adjacent distributions. Compared to the no-standard baseline, a standard at stimulus 4 has little effect. A standard at stimulus 1 moves most criteria toward /i/, that is, makes /i/ responses more probable; a standard at stimulus 7 has the opposite effect. Figure 5 shows that this is indeed the result to be explained; Figs. 6 and 7 together show that the effect is due primarily to response bias, not sensitivity.

**H. Discussion**

1. **Basic sensitivity**

Almost all previous studies of vowel discrimination have used roving designs (see, for example, Repp et al., 1979). The typical peaked functions reported in the literature are replicated in our roving condition. The comparison between the fixed and roving conditions in experiment IB shows that roving paradigms do not measure best-possible performance, and that discrimination functions may not even have the same shape in the two conditions. Basic sensitivity for the /i/-/I/ dimension has no peaks, but gradually improves along the continuum.

The exact form of the function describing sensitivity across the range depends, of course, on whether the physical differences between adjacent stimuli are equal. We calculat-
ed the spacing of our stimuli according to the model of Miller and Chang (1986), which states that perceptual distances between vowels can be measured as the vector sum of the logarithms of formant frequency ratios. The distances between our stimuli, shown in Table I, are clearly not uniform; in particular, stimuli 2 and 3 are very close together, consistent with the relatively low sensitivity for discriminating them. However, the distances predicted by the Miller and Chang model do not increase uniformly across the range, as do our basic sensitivity data, so their model does not account for the overall pattern of the results.

The stimuli were chosen to be as comparable as possible to those used by previous investigators (Stevens et al., 1969; and later Pisoni, 1973; Repp et al., 1979; and others). There was therefore no reason to expect a constant level of discriminability. Nor is there a reason to be disappointed in the result: The primary importance of fixed discrimination data is as a baseline for performance in other paradigms. Whether the model of Miller and Chang is intended to predict fixed discrimination data is not clear to us.

2. Memory coding

a. Global comparisons. As pointed out earlier, the Durham-Braida theory predicts an ordering of sensitivity across conditions: fixed discrimination, then roving discrimination, then identification. (The model does not explicitly consider identification with a standard.) The ordering arises because identification uses only the context coding memory process, roving discrimination supplements context coding with trace coding, and the fixed task is limited by neither. This pattern was observed for the total $d'$ values of all subjects. These three values of total $d'$ can be used to estimate the relative trace and context variance by using Eqs. (1)—(4).

In experiment IA, only relative context variance can be estimated, because roving discrimination was not performed. This statistic increases with stimulus range, equaling 0.94 for the /i/-/v/ continuum and 2.98 for /i/-/v/-/e/. Relative context and trace variance were comparable in size (2.22 vs 1.65) in experiment IB. In roving discrimination, our listeners apparently used both context and trace coding to make their decisions. The two estimates of relative context variance for the /i/-/v/ range, from experiments IA and IB, differ by a factor of more than 2. Although the experiments employed different listeners, most aspects of procedure were identical, and it is surprising that the two values are so different. On the other hand, both numbers are small: Clearly sensory variance is a major contributor to performance in vowel identification. We shall see in Sec. IV that context variance can be much greater for some non-speech continua.

b. Local comparisons (perceptual anchors). The differences in shape between fixed and roving discrimination must be attributed to context coding. (While it is true that trace coding also plays a larger role in roving than in fixed discrimination, its contribution is the same at all points along the continuum.) In particular, the peaks in our roving discrimination data may therefore be attributed to the context coding process.

Our model states that listeners are most successful in remembering stimuli near perceptual anchors. The basic strategy for determining the locations of such anchors is to seek local maxima in identification, relative to discrimination. In Fig. 8, we plot the ratio of identification to fixed discrimination $d'$ for both parts of experiment I. Standard errors of these ratios were calculated by a method described in the Appendix.

The curves in Fig. 8 suggest that anchors are located between phonetic categories: Listeners compare the sounds they hear to boundary stimuli on this dimension. Our estimates of anchor locations are not, however, very reliable because context variance is relatively small. Anchors produce only minor peaks in sensitivity under such conditions (Braida et al., 1984).

c. Response processes. Our conclusion that the main influence of standard stimuli is on response bias leaves other issues about such effects unresolved. Repp et al., 1979 conducted an "AX [same—different] identification" task in which pairs of stimuli were presented for identification. As in the present study, identification of the second stimulus was altered, contrastively, by the first. Repp et al. consider a number of mechanisms by which such contrast might occur, including one of context coding. In our model, the distinction between sensitivity and bias is orthogonal to coding mechanism. A more detailed understanding of the context-coding mechanisms might or might not lead to predictions about bias shifts.

III. EXPERIMENT II—CONSONANTS

A. Stimuli

1. VOT continuum

Nine consonant stimuli ranging from /ba/ (stimulus 1) to /pa/ (stimulus 9) were synthesized using Klatt's (1980) synthesis-by-rule-program. All stimuli had a total duration...
of 273 ms, divided between an aspirated segment (of duration equal to the VOT) and a voiced segment. A constant noise burst began each stimulus. Duration of the consonantal, aspirated segment ranged from 0–36 ms; the longest VOT was that of the syllable /pa/ when synthesized by the program. The asymptotic formant frequencies were 750, 1300, 2600, 3250, and 3700 Hz. Values of VOT were chosen to insure that performance would be neither chance nor perfect, and were not evenly spaced: 0, 6, 12, 15, 18, 21, 24, 30, and 36 ms.

Four tokens of each stimulus were produced by varying average F0 from 100 to 130 Hz in 10-Hz steps. As in experiment I, stimulus level was varied over a 10-dB range. The stimuli were stored digitally and played by a PDP12 computer at roughly 65 dB SPL.

2. Place of articulation continuum

Thirteen stimuli ranging from /ba/ (stimulus 1) to /da/ (stimulus 7) to /ga/ (stimulus 13) were synthesized using Klatt’s (1980) program. There was no noise burst at the plosive release because bursts at different places of articulation have different spectra (Blumstein and Stevens, 1979), and we wished not to provide another cue for discrimination. Asymptotic formant frequencies were the same as for the voicing continuum. The starting frequencies of F1 and F3 were fixed at 450 and 2350 Hz for all stimuli, while the starting frequency of F2 varied linearly across the continuum. The F0 variation used for the other continuum was not used for this one. The stimuli were stored digitally and played by a PDP12 computer at roughly 80 dB SPL. Absolute intensity was again jittered across trials over a 10-dB range.

B. Design

Two subexperiments were conducted, one with the VOT stimulus series, the other with the place series. In experiment IIA, identification without a standard, identification with a standard (at stimuli 1, 5, and 9), fixed 2IFC discrimination, and roving 2IFC discrimination were measured on the VOT continuum. In experiment IIB, the place continuum was investigated using all of these tasks except identification with a standard. Table IV lists the conditions comprising each experiment, and the approximate number of trials for each condition and stimulus comparison.

C. Subjects

One male and three female native English speaking MIT students, with no known hearing defects, served as subjects in experiment IIA. Three were undergraduates, who had never before participated in auditory experiments and who were paid for their participation. The fourth subject was one of the authors (RFG), an experienced listener. Two days a week the subjects participated in this experiment; another 2 days a week they participated in an experiment involving nonspeech stimuli.

One male and two female native English speaking MIT undergraduates, with no known hearing defects, served as subjects in experiment IIB. Two were paid, naive listeners; the third was an experimenter.
0 ms, 16.9 for 18 ms, and 19.2 for 36 ms. The response shift induced by standards is qualitatively the same for this consonant continuum as for the /i/-/I/ vowel continuum (Fig. 5), although the shift is smaller, as a proportion of stimulus range, for consonants. Healy and Repp (1982) conducted "AX identification" experiments on several auditory dimensions, including a consonantal one (/ba/-/da/) and an /i/-/I/ vowel continuum. This allowed them to compare the amount of contrast for consonants and vowels. Unfortunately for our purposes, the data (their Fig. 3, p. 77) are presented as proportion of responses of one type to a given stimulus for a range of standards, rather than the reverse. The size of the shift in proportion is greater for vowels than for consonants, but, as Healy and Repp (p. 76) point out, the data display a ceiling effect. Our data clarify this issue, providing a clear-cut comparison between stimulus domains.

As with vowels, the effects of standards on identification can be divided into sensitivity and bias changes. The effects on sensitivity, displayed in Fig. 11, are small and appear to be restricted to the region around the category boundary (15–21 ms). Figure 12 shows that criteria shift towards a standard, thus increasing the proportion of responses away from it. Criterion shifts are smaller than observed for vowels (cf. Fig. 7).

3. Token effects

There is a possible interaction between fundamental frequency and voicing perception that could have influenced

FIG. 11. Sensitivity in identification with and without a standard in experiment IIA. Average of data from four subjects.
our results. In natural speech, the fundamental frequency averaged across the duration of a vowel placed between voiceless plosives is about 5% higher than in a vowel placed between voiced plosives, with the greatest difference of about 15% occurring right after voice onset (Klatt, 1975). Also, syllables with ambiguous VOTs have been shown to be identified as /b/ when synthesized with a low-rising F0 contour, and /p/ with a high-rising contour (Haggard et al., 1970). Thus one might expect the stimuli having tokens corresponding to the lower F0’s to be perceived more as /ba/, and those having tokens corresponding to the higher F0’s to be perceived more as /pa/. In our identification experiments, we observed no such effect, although our ability to discern this interaction was limited by the small number of trials per token per subject (25–45). In discrimination, performance with different tokens was slightly better than with the same tokens.

F. Results of experiment IIB

For the /ba/-/da/-/ga/ continuum, the pattern of sensitivity measured by the three paradigms is similar, with peak performance near the category boundaries (Fig. 13). The ordering of overall performance (see Table V for values of total d') is the same as for the voicing continuum studied in experiment IIA: Fixed discrimination is best; roving and identification are indistinguishable. There are unaccounted small differences in the location of maximum sensitivity across conditions.

G. Discussion

1. Basic sensitivity

There have been a few other fixed discrimination studies of VOT, and the overall pattern of results is interesting. Both Sachs and Grant (1976) and, in a recent replication, Kewley-Port et al. (1988) reported monotonic, peakless discrimination functions. A comparison of the three data sets (Fig. 14) suggests that the disagreement is not over the shape of the function describing sensitivity as much as over the range of values investigated. The key requirement in comparing these data is the use of a consistent dependent variable. Although the various studies used different psychophysical methods, d’ can be abstracted from all. Because d’ is a distance measure, it is possible to recale observed values to correspond to constant (10-ms) differences in VOT.
Previous data also differ in the way the independent variable is reported; we plot each point at the value of VOT halfway between the two stimuli being discriminated. It is worth describing the exact calculations.

In reploting our own data from Fig. 9, we have simply multiplied $d'$ by 10/6 (for 6-ms differences) or 10/3 (for 3-ms differences). Although this seems obvious, there is an alternate strategy: $d'$ for adjacent stimulus pairs could be added to generate a $d'$ for an approximately 10-ms difference. Thus adding $d'$ for the 0- and 6-ms stimuli to that for the 6- and 12-ms pair yields a 12-ms $d'$, which must then be multiplied by 10/12 to be comparable to a 10-ms $d'$. We have used this strategy in the past (Macmillan et al., 1987, Fig. 3), but now believe it to be less accurate than simple normalization. (The qualitative patterns of results are extremely similar.)

Sachs and Grant (1976) used a same–different task, but not that used in experiment IB. Rather, the first of two intervals always contained the same stimulus. Macmillan et al. (1977) called this a fixed-standard procedure, in contrast to the variable-standard procedure described earlier. In fixed-standard designs, the observer’s decision rule need not take account of the first interval, and $d'$ can be computed as though the task were yes–no. Data comparing fixed same–different and 2IFC paradigms (Jesteadt and Sims, 1975; Creelman and Macmillan, 1979) suggest that this analysis is correct. The $d'$ values reported by Sachs and Grant for a velar continuum (published in Spiegel and Watson, 1981, Fig. 7) were calculated in this way, and have been simply rescaled in Fig. 14 to correspond to 10-rather than 5-ms differences. The VOT values over which they are plotted have also been changed. Sachs and Grant plotted sensitivity for discriminating, say, 15- and 20-ms stimuli at 15 ms; we plot this point at 17.5 ms.

Kewley-Port et al. (1988) modeled their experiment on that of Sachs and Grant, and used the same fixed-standard procedure. They measured sensitivity for three bilabial VOT pairs: (5,15), (35,45), and (65,75), and reported $p(c)$ max, a statistic that can be converted to $d'$ by applying a z transformation and doubling. We computed $d'$ in this way for the data presented in Kewley-Port et al., Fig. 4. In their figure, points are plotted over the lower endpoint; as with the Sachs and Grant data, we moved them to the midpoint instead.

Considering these three studies—two showing monotonic patterns and one showing a peak—there is clearly no inconsistency. Our study, by measuring sensitivity densely across the appropriate range, reveals a peak that the others have missed. The absolute levels of performance in the three studies are quite similar. All these data are consistent with the conclusion that there is a VOT peak in basic sensitivity, at about 15 ms. The remaining discrepancies among the discrimination patterns may be due to differences in the stimuli (labial versus velar, and details of stimulus generation).

The picture is clouded somewhat by a fourth study, that of Rosner (1984). He conducted three fixed (as opposed to roving) experiments on an alveolar continuum with the same subjects, an ABX study and two same–different experiments (ISI = 1 and 4 s). All paradigms were of the variable-standard type, and the Kaplan et al. (1978) tables were therefore used to calculate $d'$. We plot the same–different, 1-s ISI data because: (1) performance was best in that condition, and we are attempting to determine optimal sensitivity, and (2) same–different data were collected after ABX, and thus come from more experienced subjects. [Rosner found poorer performance for a 4-s than a 1-s ISI, in accordance with Eq. (4)] Rosner used 20-ms differences, so we divide his reported $d'$ values by 2. Rosner’s data do not appear to fit the overall pattern; they show generally poorer performance, and peak at about 30 ms. The discrepancy might be explained by the relative lack of training of Rosner’s observers, by differences in stimulus generation, or perhaps by the long ISI—in all other studies, ISI was 400 ms. The data resemble those for the “high-uncertainty” condition of Kewley-Port et al.; perhaps Rosner’s listeners in fixed paradigms were as uncertain as if the paradigm were roving. (This was very nearly the case in Rosner’s only direct comparison of fixed and roving designs, using the ABX paradigm.) Notice, however, that the critical region around 15 ms is not represented in Rosner’s data. A $d'$ of 3.5 at that value, and no change in the other values, would produce a pattern quite consistent with the other data in Fig. 14.

2. Memory coding

a. Global comparisons. Both consonant continua satisfy a special case of the predicted ordering of performance across conditions: Fixed discrimination is best, but performance in identification and roving discrimination is similar. This result implies that trace coding was ineffective in the roving discrimination task. Trace variance cannot be estimated from these data, being effectively infinite.

Context variance, a measure of the discrepancy between fixed discrimination and identification, is even smaller than for vowels: 0.65 times sensory variance for the voicing continuum, 1.05 for the (three-category) place dimension. Listeners have little difficulty in accurately labeling these consonantal sounds; the primary limitations on discriminating them are sensory.

b. Local comparisons (perceptual anchors). The strategy described for experiment I was used to estimate the locations of anchors, or best-remembered stimuli, with results shown in Fig. 15. For both continua, anchors appear to lie at or near prototypes, that is, good exemplars of the phonetic categories. Compared to the fixed-discrimination baseline, listeners do better in identifying /ba/ and /pa/ than neighboring stimuli. On the place continuum, the best-labeled stimuli are numbers 3 (/ba/) and 8 (/da/); a high plateau rather than a sharp peak characterizes stimuli in the /ga/ region (numbers 10–13).

IV. GENERAL DISCUSSION

Our primary conclusions about the perceptual processing of the vowel and consonant continua used here have already been stated. In this discussion, we relate our findings on basic sensitivity, context coding, and the effects of standards to other work that is similar either in methodology or in interpretation. At several points, we consider our data in relation to results from analogous experiments using a nonspeech continuum, auditory intensity. Finally, we address a
broader issue, the implications of the data for the categorical perception hypothesis.

A. Basic sensitivity

1. Other criteria for identifying sensory effects

The problem of distinguishing sensory from cognitive effects is an old one in speech research. Two influential approaches have compared adult human performance with that of other subjects, and data for speech sounds with data for nonspeech.

The two main subject comparison groups used as controls for adult human observers are infants and animals (see Kuhl, 1987, for a recent summary). Much animal speech perception research relies on the premise that, if animals show the same behavioral pattern of discrimination as humans, then the processes involved cannot be unique to speech, and the data should be taken to describe basic sensitivity. Similar infant and adult data are used sometimes to support a similar argument, sometimes to attribute speech processing to infants.

These interpretations generally neglect memory processes that can, as we have argued, significantly affect the pattern of sensitivity. It is difficult to take such processes into account, however, for lack of appropriate experimental comparisons. Performance is usually measured in only one paradigm; in infants, for example, identification has proved near-impossible to assess (Macmillan, 1985). Practical considerations, especially with infants, preclude the training necessary to guarantee that best performance has been reached, so it is difficult to know whether discrimination data should be viewed as fixed or roving, and whether discrimination peaks should be attributed to high basic sensitivity or perceptual anchors.

In the stimulus-oriented approach, the pattern of resolution on a speech continuum is compared with that on a nonspeech dimension that resembles it. Thus Miller et al.'s (1976) noise-buzz stimuli and Pisoni's (1977) "temporal-onset-time" continuum are both intended to mimic VOT. If similar discrimination patterns are observed, then both sets of data are assumed to result from a general auditory process not specific to speech.

Like animal and infant studies, nonspeech experiments have not compared experimental conditions to the degree that would be helpful, a situation that would not be difficult to rectify. Absent such comparisons, research in this tradition cannot be used to make inferences about basic resolution. These stimulus sets do, however, provide an entry point to psychoacoustic studies of speech, in which aspects of complex stimuli are systematically varied. This approach will be most successful, according to our model, if the designs employed involve only sensory variability.

2. Explaining patterns of basic resolution

Fixed discrimination data in the present experiments display peaks for two consonantal continua, but not for a vowel dimension. According to our model, basic sensitivity also follows these patterns. In assessing the role of memory, we use these patterns as a baseline. A question not addressed by the model is why these functions are what they are. It is worth considering what form an answer to this question might take.

In the first place, some researchers might view only the consonant data as requiring explanation, since the vowel data, to a first approximation, satisfy Weber's law. But Weber's law itself does not always hold. In auditory intensity discrimination, for example, it characterizes noise bursts (e.g., Penner, 1972), but not tone bursts (Rabinowitz et al., 1976).

One form of explanation for peaks in resolution invokes their adaptive value. Regions of high discriminability along dimensions that distinguish phonetic quality may have been useful in the development of phonetic categories (Stevens, 1981). Whether or not this statement is true, its converse appears to be false: The usefulness of a boundary, as evidenced by the presence of phonetically different stimuli on either side of it, does not mean that a "natural boundary" exists. The vowel continuum of experiment I provides a counterexample. That some discrimination peaks are sensory, even if for a good reason, does not mean that all are.

3. Comparisons of total sensitivity across continua

Significant as the presence or absence of peaks in the fixed discrimination function may be, our model assigns a more important role to another indicator: total d' across the range. It is interesting to compare our consonant and vowel continua on this variable, and to compare the speech-sound continua with auditory intensity.

If a just-noticeable difference (jnd) is defined as the stimulus difference needed for d' to equal 1, then total d' is the same as the range measured in jnd's. Ades (1977) speculated that some processing differences between vowels and consonants might arise from differences in this range. But the total number of jnd's between /ba/ and /pa/ (6.2 in experiment IIA) is about the same as between /i/ and /I/ (5.9 and 6.4 in experiments IA and IB). Although our esti-

FIG. 15. Ratio of d' in identification to d' in fixed discrimination in experiments IIA (top) and IIB (bottom). Average of data from four subjects in IIA and three subjects in IIB.

1276 J. Acoust. Soc. Am., Vol. 84, No. 4, October 1988

Macmillan et al.: Speech sound resolution 1276

Downloaded 24 Sep 2013 to 128.119.168.112. Redistribution subject to ASA license or copyright; see http://asadl.org/terms
mates of total $d'$ are somewhat arbitrary—the range of VOTs, in particular, is smaller than that used by most other investigators—the small consonant range hypothesized by Ades does not materialize, and thus cannot explain vowel-consonant differences in the pattern of discrimination.

In the context of intensity perception research, the values of total $d'$ obtained in the present experiments are quite modest. For example, Berliner and Durlach (1973) measured total $d'$ values approaching 50 for a 54-dB intensity range. It is not obvious how to compare 54 dB with any measure of the physical range spanned by our speech sounds, but the total $d'$ values are quite discrepant. As the stimulus range (and therefore total $d'$) is critical, according to the theory, in predicting the magnitude of context memory effects, we should expect that such effects will be quite different for these speech and nonspeech dimensions. We consider this issue next.

B. Context memory and anchors

1. Comparisons of context memory across continua

Our measure of memory load, the relative context variance, differs importantly between the consonant and vowel continua. For vowels, context variance is between one and three times sensory variance; for consonants, it is less than or (approximately) equal to this baseline. To paraphrase, one-half to three-quarters of the variability limiting identification performance is due to memory in the vowel experiments, one-third to one-half for consonants.

Whatever the reason for this difference, we have seen that it is not a difference in range (measured in jnd's). For the specific continua compared in the previous section (/ba/-/pa/ vs /i/-/I/), each of which spanned about six jnd's, relative context variance was 0.65 and (averaging two estimates) 1.6. Apparently, the jnd range is not the only factor influencing context memory for speech sounds; on the intensity continuum, on the other hand, the two variables are functionally related. Of our speech continua, the one that most closely resembles intensity in this respect is the vowel set. For 6 and 15 jnd's (the total $d'$ values in experiments IB and IA), relative context variance is virtually identical for vowels and intensity (see Macmillan et al., 1987, Fig. 6, for more detail).

An important remaining difference between intensity and vowels (or consonants) is the maximum jnd range available. The comparisons in the previous paragraph are with relatively small intensity ranges (less than 20 dB). The magnitude of anchor effects in intensity experiments depends heavily on the range (Braida et al., 1984). In particular, large sensitivity improvements near anchors in context-coding tasks are greatest for the largest stimulus range (54 dB). Direct application of the theory to our speech sounds, therefore, leads us to expect only small anchor effects. By the standard of intensity research, the two-to-one differences shown in Fig. 15 are surprisingly large.

2. Anchor locations

According to our theory, some stimuli along a dimension play a special role in providing a reference point that is used in context coding other stimuli. The location of these anchors is not specified by the theory, but can be determined by comparing identification and fixed discrimination data. The analysis shows that anchors on the vowel continuum lie approximately at category boundaries, while consonant anchors coincide approximately with prototypes (or, for two-category dimensions, endpoints).

Both category boundaries and prototypes have been assigned central roles by previous writers, although the exact hypothesized characteristics of these stimuli have varied in different treatments. Two points of view that stress the importance of the boundary are the common-factor theory of Pastore et al. (1977) and models based on Helson's (1964) adaptation-level theory (e.g., Diehl et al., 1980; Wilson, 1987). Evidence for the importance of prototypes has been adduced from selective adaptation experiments by Samuel (1982; see also Miller et al., 1983), and from cue-integration studies by Massaro (1987; Oden and Massaro, 1978).

If, as our data suggest, the best-remembered stimuli are near the category boundaries for some continua, but near prototypes or endpoints for others, then any debate about which type of stimulus plays the more central role must be couched in terms of specific stimulus continua. The intelligent listener presumably uses the most stable points as references. On the intensity continuum, there is almost no possibility except the endpoints. On speech dimensions, the choice appears to depend on very specific aspects of the continuum being sampled.

3. Is context memory phonetic? Long-term?

To perceive speech, the listener must ultimately translate the incoming auditory information into linguistic entities (phonemes, syllables, and/or words). The highest-level constructs of the present model are perceptual anchors. What is the relation between the context coding we have inferred and phonetic coding?

Phonetic identification differs from the task performed by our listeners in two ways: Fewer categories are used, and those that are used are stored for a long time. Many writers (e.g., Schouten, 1987) have stressed the importance of "long-term" memory in speech perception. Whether our postulated context-coding process could lead to such relatively permanent perceptual changes is uncertain.

The difference in the number of categories may result from an important limitation of the present experiments, the use of a single acoustic dimension. Consider the identification task we used: Each stimulus corresponded to its own response, and listeners were generally able to use more than two responses reliably (that is, the total $d'$ measured in identification was more than 2). These identifications are obviously not the same as phonetic labeling. But, if the stimulus space were multidimensional (and therefore more realistic), the number of categories on each dimension that could be reliably labeled might decrease substantially. At least that was the result of Pollack and Ficks (1954), who manipulated the number of dimensions of their nonspeech auditory stimuli from one to six. Total information transmitted rose with dimensionality; but information transmitted on a single
C. Effects of standards

In our experiments, the presentation of a standard stimulus in identification produced a change in response bias. The response criterion moved in the direction of the standard, so that more responses from other regions of the continuum were given. In addition, there were sometimes small improvements in sensitivity in the region of the standard. Evaluating discrimination/identification equivalence similarly requires deciding whether it is a fixed or a roving discrimination to which identification should be compared. The former is essentially the definition proposed by Macmillan et al. (1977); in terms of the present model, this amounts to the assertion that context variance is zero. By this definition, no continuum considered here (or, we suspect, anywhere else) is categorical. But *roving* discrimination equals identification whenever the trace mode is ineffective, a condition satisfied for both our consonant continua, but not for vowels.

ACKNOWLEDGMENTS

We thank Dennis Klatt for help and advice regarding stimulus generation. For valuable comments on the manuscript, we are grateful to W. M. Rabinowitz, B. Scharf, R. M. Uchanski, and especially to B. H. Repp, one of the reviewers, who encouraged us to draw more connections between our work and that of previous researchers. Special thanks are due to L. Stephens, who collected the data of experiment IIB (Stephens, 1986). Preliminary reports have been presented at meetings of the Acoustical Society of America (Goldberg et al., 1985, 1986). This work was supported by grants from NSF to Macmillan and Braid, and by a PSC-CUNY Research Award to Macmillan.

APPENDIX: CALCULATION OF STANDARD ERRORS

The standard errors used in plotting the figure reflect both the number of trials per subject and the number of subjects. For a single subject in a discrimination task, the variance of $d'$ has been shown by Gourevitch and Galanter (1967) to be approximately

$$\text{var}(d') = \frac{H(1-H)}{N(H)d'(H)} + \frac{F(1-F)}{N(F)d'(F)},$$

where $H$ and $F$ are the hit and false-alarm rates, $N(H)$ and $N(F)$ are the number of trials on which those rates are based, and $\phi$ is the normal density function. When there are $N(S)$ subjects, the variance of average $d'$ is the sum of the variances computed from Eq. (A1), divided by $N(S)^2$.

Lippmann (1973) extended this approach to identification data. Once $d'$ and the criterion $c$ have been estimated, apply the normal distribution function $\Phi$ to set $H = \Phi(d'/2 + c)$ and $F = \Phi(c - d'/2)$, then use Eq. (A1).

Group data are treated as for discrimination.

Because $d' = z(H) - z(F)$ and $c = -[z(H) + z(F)]/2$, $\text{var}(c) = \text{var}(d')/4$. In comparing two values of $c$, the variance of the difference is simply the sum of the variances.

The variance of a ratio $d_{num}/d_{den}$ can be approximated by the following:

$$\text{var} \left( \frac{d_{num}}{d_{den}} \right) = \left( \frac{d_{num}}{d_{den}} \right)^2 \left( \frac{\text{var}_{num}}{d_{num}^2} + \frac{\text{var}_{den}}{d_{den}^2} \right),$$

where num and den denote the numerator and denominator of the ratio of $d'$ values.
A summary of the document: Speech sound resolution


Pisoni, D. B. (1973). "Auditory and phonemic memory codes in the dis-