Perception & Psychophysics
1998, 60 (4), 533-543

Effects of token variability on our ability to distinguish between vowels

ROSALIE M. UCHANSKI and LOUIS D. BRAIDA
Massachusetts Institute of Technology, Cambridge, Massachusetts

Even when the speaker, context, and speaking style are held fixed, the physical properties of naturally spoken utterances of the same speech sound vary considerably. This variability imposes limits on our ability to distinguish between different speech sounds. We present a conceptual framework for relating the ability to distinguish between speech sounds in single-token experiments (in which each speech sound is represented by a single wave form) to resolution in multiple-token experiments. Experimental results indicate that this ability is substantially reduced by an increase in the number of tokens from 1 to 4, but that there is little further reduction when the number of tokens increases to 16. Furthermore, although there is little relation between the ability to distinguish between a given pair of tokens in the multiple- and the 1-token experiments, there is a modest correlation between the ability to distinguish specific vowel tokens in the 4- and 16-token experiments. These results suggest that while listeners use a multiplicity of cues to distinguish between single tokens of a pair of vowel sounds, so that performance is highly variable both across tokens and listeners, they use a smaller set when distinguishing between populations of naturally produced vowel tokens, so that variability is reduced. The effectiveness of the cues used in the latter case is limited more by internal noise than by the variability of the cues themselves.

Our ability to distinguish between speech sounds is determined by the physical properties of the sounds and by limitations on auditory resolution and memory for salient differences between the sounds. The long-term goal of this research is to understand the limitations on our ability to identify and discriminate between speech sounds associated with natural variations in the production of different utterances of the same speech sound. Although the difficulties posed by such variation for automatic speech recognition systems have increasingly been appreciated (e.g., Perkell & Klatt, 1986; Pisoni, 1990), the effects of such variation on human speech reception have not been studied systematically. While we are seldom aware of these effects in quiet communication situations, they may be more problematic when speech reception is restricted by noise or hearing loss. Under conditions that restrict the salience of certain physical differences between speech sounds, the variations in the utterances may consume a proportionally larger fraction of the available perceptual space.

Ultimately this research will consist of three components: (1) determining the perceptual effects of utterance to utterance variations (token variability); (2) characterizing the variability in the acoustic properties of the speech sounds; and (3) developing a model that relates perceptual effects to the acoustical properties and variability of the stimuli. This report addresses the initial component of our research program.

Our approach to determining the effects of token variability on auditory discrimination focuses on measuring how well listeners can distinguish between speech sounds as a function of the number of tokens that represent the sound in a given task. Since we are ultimately interested in relating the effects of variability to those associated with basic auditory resolution and memory, we use a generalized measure of sensitivity, *d′* (see, e.g., Braida, 1991; Green & Swets, 1966) to characterize the ability to distinguish between sounds. This methodology and the generalized measure are thus potentially applicable to both speech and nonspeech sounds and to both discrimination tasks involving a pair of speech sounds and identification tasks involving an arbitrary number of speech sounds. In previous work on the effects of such variability on the identification of a set of 10 vowel sounds, Uchanski, Millier, Reed, and Braida (1992) found only modest reductions in sensitivity as the number of tokens was increased from 1 to 4 (*d′* reduced by roughly 25%) and then further increased to 16 (*d′* reduced by roughly 33% in comparison with the 1-token case). This report analyzes the effect of token variability on pairwise discrimination tasks for the same stimuli used in the previous 10-vowel identification tasks.

Pairwise discrimination experiments that use one token of each speech sound can be regarded as complex-sound analogues of simple psychoacoustic discrimination tests (see, e.g., Durlach & Braida, 1969) that use a fixed pair of sounds. Such tests are typically used to measure the ability

The first author is now at the Central Institute for the Deaf in St. Louis. This work was supported by NIH Grants R01-DC00117 and P01-DC00361. The authors have benefited from thoughtful discussions with J. Besing, H. S. Colburn, R. Gilkey, C. B. Huang, J. W. Maney, K. Millier, D. B. Pisoni, C. M. Reed, and T. Hanna. We are grateful for the comments of T. M. Neary, B. S. Rosner, and an anonymous reviewer. Correspondence concerning this article should be sent to L. D. Braida, 36-747 MIT, Cambridge, MA 02139 (e-mail: braida@mit.edu).

to discriminate changes in elementary physical characteristics (e.g., amplitude, frequency, etc.). Researchers have introduced stimulus variability into psychoacoustic experiments for three main purposes. Variations in *irrelevant* characteristics are introduced in discrimination tests employing complex sounds to control the relative effectiveness of the cues available to the listener. For example, in tests of the ability to resolve differences in pitch of the two-tone complex having frequencies nf and $(n + 1)f$, Houtsma and Goldstein (1972) randomly varied the parameter n to reduce the effectiveness of cues based on spectral locus. In tests of the ability to distinguish between different spectral shapes, random variations in overall level are often introduced to reduce the effectiveness of cues related to loudness differences (e.g., Farrar et al., 1987; Green, Kidd, & Picardi, 1983). Variations in *relevant* characteristics are introduced when one attempts to equate the "memory load" in different tasks. For example, Pollack (1956), Berliner and Durlach (1973), and Berliner, Braida, and Durlach (1977) randomly varied the overall level of tone pairs in a two-interval intensity discrimination task to match the range of intensities used in an intensity identification task. In a third type of variation, small controlled *perturbations* of selected elements of complex sounds are introduced to assess the relative weights assigned by the listener to each of these elements. Such perturbations are assumed to introduce differential changes in discriminability that can be estimated with the use of a COSS (conditioned on single stimuli) analysis (e.g., Berg & Green, 1990).

Studies of the ability to distinguish between speech sounds rarely have employed only one token of each item. Simple pairwise discrimination abilities are seldom studied. Irrelevant variations have been introduced to reduce the effects of possible artifacts in synthetic stimuli (see, e.g., Macmillan, Goldberg, & Braida, 1988). More typically, speech discrimination experiments have included trial-to-trial variation in the relevant parameter. This variation is intended to assure that the range of stimuli spans a continuum from one phonetic category to another. The perturbation technique and associated COSS analysis are not routinely employed with speech stimuli, although the procedure developed by Kuhl (1991) to study the perceptual magnet effect would seem to be congenial to such an analysis.

THEORY

The analysis of experiments that use complex stimuli such as speech is problematic because several physical characteristics of complex stimuli may be relevant to a given task. Moreover, the set of relevant characteristics is likely to depend on the strategy used by the observer. In the analysis that follows, we assume that the strategies used differ only with respect to the weights given various stimulus characteristics. We develop relations between measures of expected performance in different tasks when the same weights are used in all tasks. This is essentially a null hypothesis which assumes that the decision process is independent of the stimulus context.

We assume that the perceptually relevant physical characteristics of vowel sounds are described by a K-dimensional vector of cues $\vec{Q} = (q_1, q_2, \ldots, q_K)$; see Figure 1. Decisions are made on the basis of the vector of observations, $\vec{O} = \vec{Q} + \vec{N}$, where the components of the additive internal noise vector $\vec{N} = (n_1, n_2, \ldots, n_K)$ are identically distributed, independent Gaussian random variables with zero mean and variance σ^2 that are independent of the stimulus wave form. The observer is assumed to distinguish between vowels by forming a decision variable X that is a weighted sum of the observations:

$$X = \sum_{k=1}^{K} \alpha_k o_k .$$

For convenience, we assume

$$\sum_{k=1}^{K} \alpha_k^2 = 1 .$$

When an observer is distinguishing between one fixed token V_i and W_j of each of the vowels V and W, the observer's sensitivity $d'_F(V_i, W_j)$ is

$$d'_F(V_i,W_j) = \sum_{k=1}^{K} \alpha_k \frac{v_{ik} - w_{jk}}{\sigma}, \tag{1}$$

where v_{ik} and w_{jk} are the values of the kth component of $\vec{Q}$ for tokens V_i and W_j, respectively. In Figure 1, $d'_F(V_i, W_j)$ is proportional to the length of the projection of the line segment connecting the points representing V_i and W_j onto a line whose direction cosines are given by the weights α_k. The highest value of d'_F, $d'_{F,\max}$ is achieved by matching the set of α_k to the direction cosines of the line segment that connects a given pair of tokens, by using α_k, that are proportional to $v_{ik} - w_{jk}$, yielding

$$d'_{F,\max}(V_i,W_j) = \frac{1}{\sigma} \sqrt{\sum_{k=1}^{K} (v_{ik} - w_{jk})^2} . \tag{2}$$

In this case, $d'_F(V_i, W_j)$ is proportional to the length of the line segment connecting the points representing V_i and W_j.

When more than one utterance is used to represent each vowel, the listener is assumed to attempt to distinguish between populations of utterances, applying a fixed decision rule to the vector of observations, rather than attempting to remember the detailed characteristics of each utterance of each vowel. The population of tokens V_i and W_j of vowels V and W is described probablistically, by assuming that the set of v_{ik} and w_{jk} are independent Gaussian random variables with means m_{Vk} and m_{Wk} and variance τ_k^2 that is the same for both vowels. Figure 1 illustrates these assumptions for the case of $K = 2$ cues and $\tau_1 \approx 5\tau_2$. As in the 1-token case, we assume that decisions are based on a weighted sum of observations. We express the limits imposed by internal noise and cue variation from token to token in the quantity d'_P, which, together with the observer's criterion, determines the listener's overall score in the multiple-token experiment, ignoring the identity of the

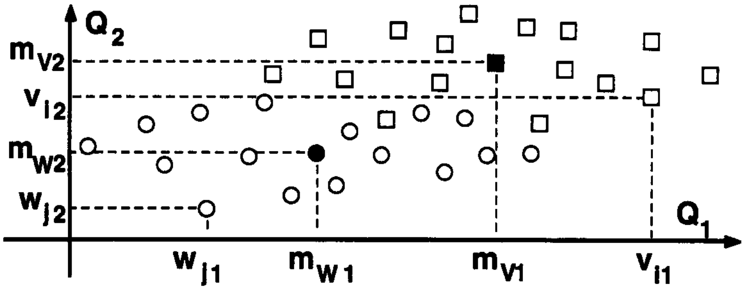

Figure 1. Two-dimensional (Q_1, Q_2) cue space for a hypothetical discrimination experiment. The open squares represent sample tokens of vowel *V*; the open circles represent sample tokens of *W*. Coordinates (v_{i1}, v_{i2}) denote the cue values for a specific token (wave form) V_i of the vowel *V*. The filled symbols represent the coordinates of the population means of the cue values for each vowel type. The coordinates (m_{V1}, m_{V2}) denote the mean cue values for vowel *V* and (m_{W1}, m_{W2}) denote the mean cue values for vowel *W*. The weights used by the observer in formulating the decision variable correspond to the selection of a specific direction in this space. Since internal noise is assumed to be equal for all coordinates, an observer's sensitivity in discriminating a single token of *V* from a single token of *W* is proportional to the projected distance between the cue values along this direction.

tokens. The quantity d'_P, the listener's sensitivity in distinguishing between the populations of *V* and *W*, is given by

$$d'_P(V,W)=\frac{\sum_{k=1}^{K}\alpha_k(m_{Vk}-m_{Wk})}{\sqrt{\sigma^2+\sum_{k=1}^{K}\alpha_k^2\tau_k^2}}. \qquad (3)$$

In the multiple token case, the highest value of d'_P, $d'_{P,\max}$, is achieved if each weight α_k is proportional to $(m_{Vk}-m_{Wk})/(\sigma^2+\tau_k^2)$, yielding

$$d'_{P,\max}(V,W)=\sqrt{\sum_{k=1}^{K}\frac{(m_{Vk}-m_{Wk})^2}{\sigma^2+\tau_k^2}}. \qquad (4)$$

In the context of this theory, d'_P, together with a specification of the listener's criterion, determines the probability, averaged over the population, that tokens of *V* and *W* will be identified correctly as representatives of their respective vowels.

The relation between average measures of performance on tests in which each vowel is represented by a single token and tests in which multiple tokens are used is complex, depending on both the characteristics of the stimuli and the strategies used by the observer. In general, the variation in cue values across the population of tokens causes $d'_F(V_i, W_j)$ to vary from token pair to token pair, particularly if the optimum choice of weights is made for each pair. However, if one *fixed* set of weights is used for all tests (consistent with the null hypothesis), it is possible to relate the mean (E_{IJ}) and variance (V_{IJ}) of d'_F across the population of token pairs to d'_P.

$$E_P\left[d'_F\right]=E_{IJ}\left[d'_F(V_i,W_j)\right]=\frac{1}{\sigma}\sum_{k=1}^{K}\alpha_k(m_{Vk}-m_{Wk}), \qquad (5)$$

and, because the variation in properties over the population is uncorrelated,

$$V_{IJ}\left[d'_F(V_i,W_j)\right]=\frac{2}{\sigma^2}\sum_{k=1}^{K}\alpha_k^2\tau_k^2=2\gamma^2. \qquad (6)$$

The relation of these quantities to d'_P is then

$$d'_P=\frac{E_P\left[d'_F\right]}{\sqrt{1+\gamma^2}}, \qquad (7)$$

where $E_P[d'_F]$ is the expected value of d'_F over the population of token pairs. If the weights are not constant across experiments, additional assumptions are required to relate d'_P to measures of d'_F.

The quantity γ^2 is a relative measure (with respect to the sensation variance σ^2) of the weighted average variance of the cue components over the vowel tokens, using weightings that are relevant to the task. Because we have assumed that the squared weight values sum to one, it is clear that

$$\frac{\tau_{\min}}{\sigma}\le\gamma\le\frac{\tau_{\max}}{\sigma}, \qquad (8)$$

where $\tau^2_{\min}$ and $\tau^2_{\max}$ are the minimum and maximum values of cue variance across components.

An important special case of this analysis applies when there is only one perceptual dimension: $K=1$. In this case, there is effectively only one choice of weights, $\alpha_1=1$, and the form of many of the results can be substantially simplified. It should be noted, however, that in this case, it is not possible for the listener to change the weights as the number of tokens is varied, so the null hypothesis and Equation 7 must necessarily hold:

$$d'_P=\frac{E_P\left[d'_F\right]}{\sqrt{1+\tau_1^2/\sigma^2}}. \qquad (9)$$

According to the framework that we have described, cue variability and internal noise have similar effects on resolution. A technique used to analyze binaural detection experiments with "frozen noise" maskers (Siegel & Colburn, 1989) can be used to estimate γ from data collected in a multiple-token experiment. This technique uses a modest sample of tokens to represent the population of tokens. Each time token V_i is presented in such an experiment, the decision variable X has a Gaussian distribution with mean

$$X_i = \sum_{k=1}^{K} \alpha_k v_{ik}$$

and variance σ^2. If the observer's responses are determined by comparing the decision variable X to a fixed criterion C, the probability of response R to stimulus V_i is given by

$$\Pr(R \mid V_i) = \Pr(X < C \mid V_i) = \int_{-\infty}^{C-X_i} \frac{1}{\sqrt{2\pi}} e^{-x^2/2\sigma^2}, \quad (10)$$

so that the z-score of the probability of response R to token V_i is given by

$$z(V_i) = z\big[\Pr(R \mid V_i)\big] = z\big[\Pr(X < C \mid V_i)\big] = \frac{C - X_i}{\sigma}. \quad (11)$$

According to the theory presented above, the mean and variance of this quantity, and its counterpart $z(W_j)$, over the population of tokens, determine the average single-token sensitivity corresponding to the weights used in the multiple-token experiment, and the value of γ corresponding to these weights:

$$E_J[z(W_j)] - E_I[z(V_i)] = E_{IJ}[d'_F(V_i, W_j)] = E_P[d'_F], \quad (12)$$

and

$$V_I[z(V_i)] = V_J[z(W_j)] = \gamma^2. \quad (13)$$

If the weights used to form the decision variable are independent of the composition of the token set, there is a further constraint on the z scores. Specifically, when the same stimuli are presented in different experiments, the z scores corresponding to these stimuli in the two experiments should differ at most by an additive constant, corresponding to the difference in criteria for the two experiments. As a result, the z scores corresponding to these stimuli in one experiment should show strong correlations with those in the other experiment.

METHOD

Speech Materials

The speech sounds consisted of vowels produced in CVC context. Sixteen tokens of each of 10 American English monophthongs /æ/, /a/, /ɔ/, /i/, /ɛ/, /ɝ/, /ɪ/, /u/, /ʊ/, /ʌ/ in six consonant contexts (initial consonants /b/, /p/, /h/; final consonants /t/, /d/) were spoken in random order using citation style by one female of Middle American dialect over a 2-day period. The speaker was a teacher of the deaf skilled in producing cued speech (Cornett, 1967). Recordings of her productions were digitized (with 16-bit resolution) at a 20-kHz sampling rate. The utterances were then excised and stored in computer files. A single distorted version of each utterance was created by the process described below. The wave forms used in the experiments were then reconstructed using a 9-kHz antialiasing filter.

The average vowel duration across this corpus was 264 msec, and the average F_0 for this speaker was 214 Hz. Preliminary analysis of our recordings suggests that the coefficient of variation for these parameters is roughly 9% (Maney, 1989). In quiet, these vowels have been identified with better than 99% accuracy by listeners with normal hearing. The experiments reported in this paper test the ability of listeners to distinguish between vowels in /h/V/d/ syllables—specifically, the four vowel pairs found to be most highly confused in 10-vowel identification experiments (Uchanski et al., 1992): /i/, /u/; /æ/, /ɑ/; /ʌ/, /ɛ/; and /ɪ/, /ʊ/.

Distortion

Since our purpose was to study the effects of token variability under difficult listening conditions, each utterance was distorted by a multiplicative process (Schroeder, 1968). Unlike additive noise, which can mask different utterances by different amounts, the multiplicative distortion is self-adjusting relative to the levels of the utterances. In the Schroeder distortion, each sample of the distorted speech signal, $r(t)$, was related to the corresponding sample of the undistorted speech signal, $s(t)$, by

$$r(t) = \frac{s(t) + \kappa n(t)}{\sqrt{1+\kappa^2}}, \quad (14)$$

where

$$n(t) = \epsilon(t) \times s(t), \quad (15)$$

and $\epsilon(t)$, was a random pulse train with values 1 or -1 occurring with equal probability. The parameter κ specified an equivalent instantaneous signal-to-noise ratio (SNR);

$$\mathrm{SNR} = 10\log\frac{S}{N} = 20\log\frac{1}{\kappa}. \quad (16)$$

The distorted signals were normalized to the same overall RMS level. The signals used in the present experiments, including the distortions, were identical to those used in our previous 10-vowel identification study (Uchanski et al., 1992). Given the pattern of errors observed in that study, at an SNR of -12 dB ($\kappa \approx 4.0$), the primary physical characteristics used to identify the vowel sounds appeared to be F_0, F_1, and duration.

Procedures

Three normal-hearing listeners, including the first author (S3) and two young adult M.I.T. students (S1, S2) were tested. The stimuli were presented monaurally over TDH-39 headphones at a level of roughly 74 dB SPL to the listeners, who were seated in a sound-treated room. The tests used a one-interval binary-response paradigm without feedback. On each trial a single wave form was presented and listeners assigned one of two labels to the single sound presented.

Initially, subjects were trained (with feedback) to distinguish between the vowel pair /ɔ/, /ɝ/, which had been found to be easily distinguished by trained listeners in earlier identification experiments (Uchanski, et al., 1992). Two 1-token tasks using different wave forms were performed, followed by a single 64-trial 4-token experiment and a 64-trial 16-token experiment. These initial training runs were designed to familiarize the listeners with the overall procedures and specifically with the one-interval task.

To control the effects of learning, the number of tokens used to represent the vowels was progressively increased from 1 to 4 to 16 as the testing proceeded, as in the 10-vowel identification experiment (Uchanski et al., 1992). The listeners were tested together as a group and thus experienced the same order of presentation of the stimuli.

For each of the four pairs of vowels (/i/, /u/; /æ/, /ɑ/; /ʌ/, /ɛ/; and /ɪ/, /ʊ/), several training trials were presented with feedback prior to each test run in order to ensure that listeners used the intended "label" consistently for each vowel.[1] The familiarization with labels provided during this period facilitated combining results across con-

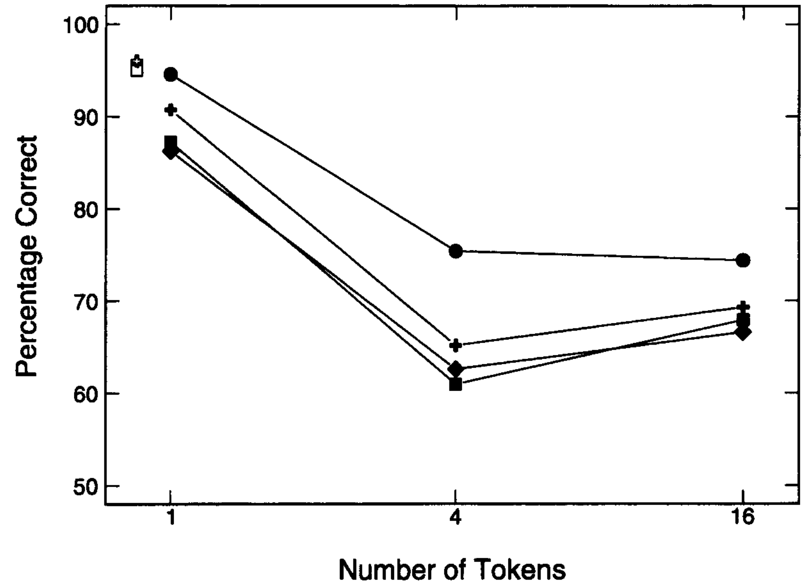

Figure 2. Filled symbols are percentage correct scores for across-vowel discrimination averaged across 3 listeners as a function of number of tokens for the four vowel pairs /i/, /u/ (crosses); /æ/, /ɑ/ (circles); /ʌ/, /ɛ/ (squares); and /ɪ/, /ʊ/ (diamonds). Averaged over vowel pairs, the standard deviations of the scores across listeners were 3.5, 6.6, and 7.2 points for the 1-, 4-, and 16-token conditions, respectively. Open symbols are estimates of percentage correct scores for within-vowel discrimination for the four vowels /i/, /u/, /ɪ/, and /ʊ/. These scores have been averaged across the pairs /i/, /u/ (cross) and /ɪ/, /ʊ/ (square) for Listeners S1 and S3.

ditions. For the 1-token conditions, an 8-trial training run (with feedback) was performed before each of the three 64-trial test runs (without feedback). For the 4-token conditions, a 16-trial training run was performed before the first 64-trial test run, and 8-trial training runs were performed before the two subsequent 64-trial test runs. Finally, for the 16-token condition, a 32-trial training run (with feedback) was performed before the first two 64-trial test runs, but no training runs were performed prior to the remaining ten 64-trial test runs. Responses obtained from trials during which feedback was provided are not included in the data analyses.

In the main experiments, the stimuli consisted of wave forms drawn from the set of 16 utterances of each of two vowels in each run. In the 1-token tests, 8 tokens of each vowel were selected at random from the set of 16 utterances. Eight pairs were then created by random selection without replacement in order to create the stimuli tested in the 1-token tests. In the 4-token tests, the 16 utterances of each vowel were partitioned into four groups of four vowels each. Each group contained two different pairs of tokens that had been used in the 1-token tests and two additional pairs of tokens that had not been used in these tests. Thus, each of the 16 tokens of each vowel sound were included in one of the groups. The number of presentations of each token was 96 in the eight 1-token experiments, and 24 in each of the four 4-token experiments and the 16-token experiment.

Auxiliary experiments were conducted to determine how well listeners could distinguish between pairs of tokens of the same vowel. In these experiments, the stimuli consisted of randomly selected pairs of tokens of a single vowel. Six contrasts using four tokens of each of the vowels /i/, /u/, /ʌ/, and /ɛ/ were tested. Only Listeners S1 and S3 participated in these experiments. They were instructed to report which of the two tokens (specified by numerals) in the presentation set was presented on each trial. Each token was presented 96 times in these tests.

RESULTS

Test Scores

An initial analysis was performed to determine the principal effects of token variability on the ability to distinguish between pairs of vowel sounds. For each vowel pair, the responses provided by each listener were combined into a 2 × 2 confusion matrix, ignoring the identity of the token representing each vowel. This matrix was then analyzed to determine the percentage of correct responses. For all vowel pairs, the resulting scores (Figure 2) were much higher when comparisons were restricted to a single token of each speech sound in a given run than when multiple tokens were used to represent each speech sound. By comparison, there were only small differences between scores in the 4- and 16-token conditions. Similar results were obtained from all 3 listeners, although there were minor interactions between listeners and vowel pairs in the 4- and 16-token conditions.

Examination of more detailed results of the one-token experiments provides an observation critical to further analysis. When only one token was used to represent each vowel sound, most pairs of tokens were distinguished with high accuracy, but scores were much lower for a few token pairs. For example, Listener S1 was able to respond correctly on roughly 95% of the trials for the vowel pair /æ/, /ɑ/, but for a particular pair of tokens of these vowels S1 responded correctly on only 73% of the trials.[2] Similar effects were seen for all vowel pairs and all listeners: Specific pairs of tokens proved much more difficult for a given listener to distinguish than the other pairs, and a pair that proved difficult for a given listener was generally easily distinguished by at least one other listener, although not necessarily by all listeners.

In the auxiliary experiments involving pairs of tokens of a single vowel, scores pooled across tokens were also very high, ranging from 93% to 98% correct. However, as in the case of tests using single tokens of two different vowels, some token pairs proved much more difficult to distinguish than others. For example, in the case of the vowel /ʊ/, for which both listeners achieved average scores of roughly 93%, one pair of tokens yielded scores of only 59% and 77%.

Taken together, the relative ease with which most pairs of tokens of a single vowel sound can be distinguished, the inhomogeneity of scores across token pairs, and the existence of pairs that were difficult for some listeners but not for others suggest that a plurality of token-dependent cues can be used to distinguish between these complex natural sounds. As can be seen in Figure 1, cue variability across the population of utterances of a single vowel sound implies that, although all listeners will find some pairs to be difficult to distinguish, most pairs of tokens can be easily distinguished if appropriate cues are used. If a given listener uses cues inappropriate for a given pair of tokens, the pair will prove difficult to tell apart, whereas another listener who uses different cues may be able to distinguish between the two tokens with high accuracy. In the remainder of the paper, we consider whether the ability to distinguish between pairs of vowels can be accounted for by assuming that a fixed set of weights is used independent of the number of tokens.

Data Analysis

Results were analyzed for each subject separately. For each N-token experiment, we formed a $2N \times 2$ confusion matrix, segregating the responses to each vowel by token. We then computed the relative frequencies of producing one of the two available responses and used these frequencies to derive estimates of $z(V_i)$ and $z(W_j)$, restricting the computations to frequencies that were both greater than 0 and less than 1 so that z scores were finite. As suggested by the analysis developed by Siegel and Colburn (1989), we then computed the mean and variance of these estimates to estimate $E_P[d'_F]$ and $V_{IJ}[d'_F]$, using Equations 12 and 13. These estimates were restricted to the same eight token pairs used in the one-token experiments. As might be expected on the basis of the high average scores seen in Figure 2, this analysis proved problematic for the one-token experiments, however. Estimates of model parameters were derived from the data obtained in those experiments using an iterative method described in the Appendix.

Effect of Varying the Number of Tokens

Consider, first, the dependence of $E_P[d'_F]$ on number of tokens shown in Figure 3. This dependence is consistent with that seen previously for percentage correct scores. Sensitivity in the single-token experiments is roughly three times as large as in the multiple-token experiments. However, there is very little difference in sensitivity for the 4- and 16-token conditions. Similar patterns were seen for all vowel pairs and all 3 listeners.

Second, consider the dependence of γ on the number of tokens. In the single-token experiments, γ^2 (Equation 6) was estimated according to the procedure described in the Appendix. In the multiple-token experiments, the mean and variance of z score estimates associated with the token pair matrices were computed for each vowel of the pair. The average of the variances[3] was used as an estimate of γ^2. Whereas the 16-token experiments provided a single estimate of γ, the 4-token experiments provided in principle four estimates of this quantity. These four estimates were averaged to derive a single estimate of γ for the 4-token condition.

The parameter γ exhibits a dependence on the number of tokens which is similar to that for $E_P[d'_F]$ (Figure 4). In particular, γ is largest for the case of a single token, and substantially smaller, but roughly the same, for stimulus sets consisting of 4 and 16 tokens. Since γ^2 measures weighted average variance of the cue components relative to the sensation variance, this variation in γ would be unexpected if listeners used the same set of cue weights in all experiments.

Estimates of d'_P

According to the theory presented previously, if listeners use the same set of weights independently of the number of tokens, then different experiments should yield consistent estimates of d'_P, computed using Equation 7. To test the constant-weight hypothesis more precisely, estimates of $E_P[d'_F]$ and of γ were computed as described previously, but only tokens that were used in all experiments were considered. (Inclusion of tokens was also subject to the requirement of finite z scores in the multiple-token experiments.) For each listener and each vowel pair, a single estimate of d'_P was thus obtained for each experiment. The estimates of d'_P for the 1- and 4-token conditions are compared in Figure 5; these estimates are compared in Figure 6 for the 1- and 16-token conditions. In both cases, it is clear that the estimate of d'_P derived from the single-token experiments is much higher than that derived from the multiple-token experiments. (Simulations showed that our method of estimating d'_P in the 1-token case had negligible bias over the range of d'_F and γ encountered.) Moreover, there is little correlation of pairs of d'_P estimates across listeners and vowel contrasts between the single- and multiple-token cases. By contrast, estimates of d'_P de-

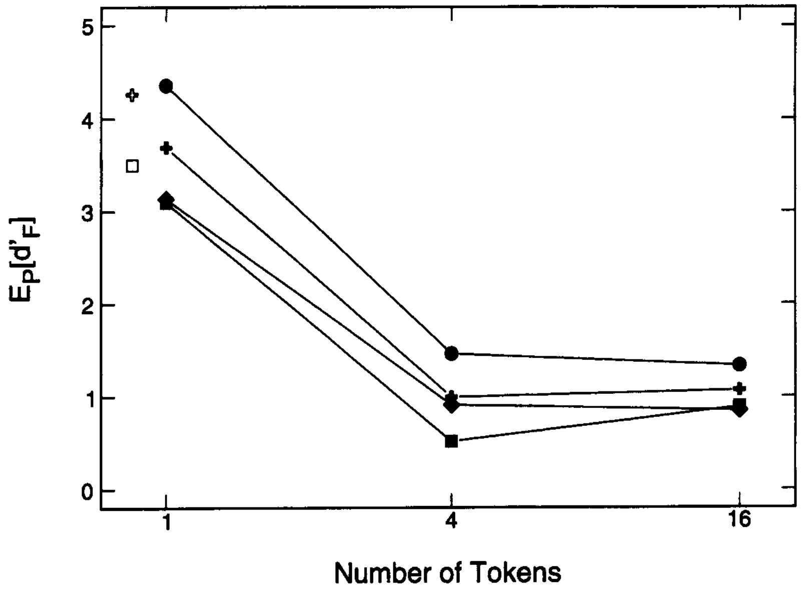

Figure 3. Filled symbols are estimates of $E_P[d'_F]$ for across-vowel discrimination averaged across 3 listeners as a function of number of tokens for the four vowel pairs used in the main experiments. Symbols are as defined in Figure 2. Averaged across vowels, the standard deviations of these estimates across listeners were 0.32, 0.22, and 0.32 in the 1-, 4-, and 16-token conditions, respectively. Open symbols are estimates of $E_P[d'_F]$ for within-vowel discrimination for the four vowels /i/, /u/, /ɪ/, and /ʊ/, averaged across the pairs /i/, /u/ (cross) and /ɪ/, /ʊ/ (square) for Listeners S1 and S3. Averaged across vowels, the standard deviation of these estimates across listeners was 0.18. Estimates of $E_P[d'_F]$ were obtained using the techniques described in the Appendix in the 1-token case, and by conventional techniques in the 4- and 16-token cases, averaging over token groups in the 4-token case.

rived from the 4- and 16-token experiments (Figure 7) are both fairly similar in size and show that average values of sensitivity are moderately correlated ($\rho \approx .7$) across listeners and vowel contrasts. These results indicate that the cue weights used in the 4- and 16-token experiments were fairly similar, but that they differed from those used in the 1-token experiments.

Correlation of Sensitivity Across Experiments

We also compared the ability to distinguish between the specific pairs of tokens that were used in both the single- and multiple-token experiments. In the multiple-token case, sensitivity is estimated using the Siegel–Colburn analysis with $d'_F(V_i, W_j) = z(V_i)-z(W_j)$. In the single-token case, d'_F was estimated from the proportion of correct responses q, assuming that there was no response bias, $d'_F = 2z(q)$. Comparisons were restricted to cases in which z scores were finite in the multiple-token experiment and $0 < q < 1$ in the single-token case.

No comparisons of sensitivity between the 1- and 4-token or between 1- and 16-token experiments yielded correlation coefficients that were different from 0 at the .01 level of significance. Our failure to find a correlation between measures of sensitivity in the single- and multiple-token experiments is partially due to the small number of possible comparisons (eight pairs per vowel per listener), the difficulty of estimating high values of d', and the relatively high variability of estimates of z scores in the multiple-token experiments (each z score is computed from responses to 24 stimulus presentations; differencing z scores to estimate d' doubles the variance of the estimate). Nevertheless, the finding of low correlation is consistent with our failure to find significant correlation for values of d'_P for these conditions and suggests that the same set of cue weights was not used in the single- and multiple-token experiments.

There was some correlation between measures of sensitivity (as measured by differences in z scores) for the token pairs in the 4- and 16-token experiments. When correlations were restricted to the eight token pairs used in the 1-token experiments, significant correlations were observed for /i/, /u/ ($\rho = .84$) and /ʌ/, /ɛ/ ($\rho = .97$) contrasts for listener S2. When all 16 pairs were included, three additional comparisons were found to be significant: /æ/, /ɑ/ for Listener S1 ($\rho = .99$) and /ɪ/, /ʊ/ for Listeners S1 ($\rho = .67$) and S2 ($\rho = .86$).

Finally, we tested the hypothesis that listeners used the same cue weights in the 4- and 16-token experiments, by determining the correlation between the z scores corresponding to response frequencies for individual utterances. With fixed weights, z scores for the same utterance in two different experiments should differ at most by an additive constant, corresponding to differences in response criteria, and this constant should be the same for all the utterances

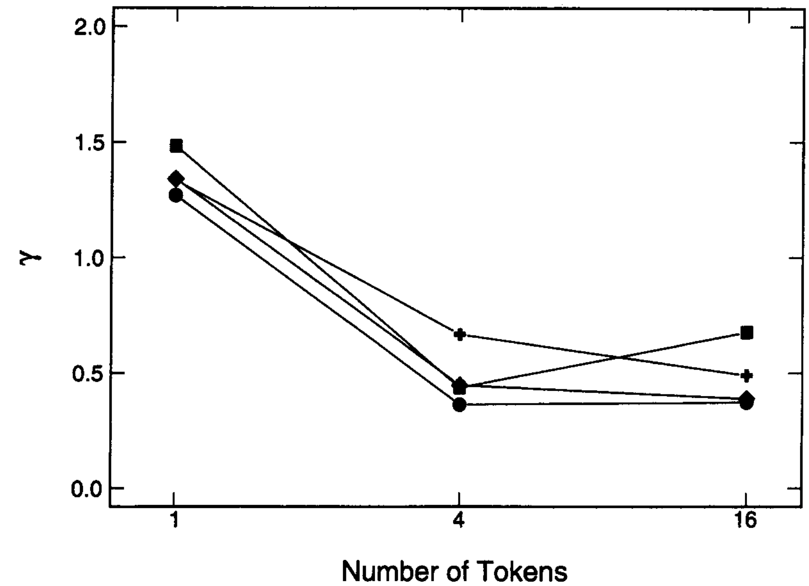

Figure 4. Estimates of γ averaged across the 3 listeners for across-vowel discrimination as a function of number of tokens for the four vowel pairs used in the main experiments. Symbols are as defined in Figure 2. Averaged across vowel pairs, the standard deviations of these estimates across listeners were 0.30, 0.12, and 0.14 in the 1-, 2-, and 16-token conditions, respectively. Estimates of γ in the multiple-token experiments were derived from the same set of eight token pairs used in the 1-token experiments.

used in the two experiments. Because estimates of z scores are meaningful only when response frequencies are greater than zero and less than unity, we combined estimates obtained across the four 4-token subexperiments for each vowel pair in order to have sufficient data. Since each of the 4-token subexperiments was assumed to have a criterion value C that might differ from that in the 16-token experiment, we estimated the criterion for each 4-token subexperiment by determining the value of C that gave response frequencies closest (in a chi-squared test) to those observed for the same tokens in the 16-token experiment for each listener and each vowel pair. The resulting estimated criterion value was added to all z scores in each subexperiment. With this correction,[4] there was evidence for correlations that were different from 0 at the .01 level of significance ($.52 < \rho < .86$) for each vowel pair for each listener. This analysis provides additional evidence that listeners used similar cue weights in the 4- and 16-token experiments.

DISCUSSION

The variability of our stimuli results from both the utterance to utterance variation in the production of the syllables and the wave form to wave form variability of the distortions applied to the utterances (Equation 14). Since each utterance was distorted only once, the experiments reported in this paper do not permit us to separate the effects of these two types of variability. This separation is essential to our goal of relating perceptual measures of variability to variations in the physical characteristics of the utterances of vowels. To this end, we plan in future work to measure resolution under conditions that include controlled variation in the distortion as well as across tokens. Nevertheless, we believe that the major effects of token variability on the ability to distinguish between vowels are not artifacts associated with the specific distortion and would have been seen if, for example, a nonstochastic distortion (e.g., low-pass filtering or spectrum reversal) had been used instead. In related work, Ronan (1992) found that the relative ease with which specific pairs of distorted tokens can be distinguished is fairly well accounted for by measurements of the physical properties of the *undistorted* tokens. Also, we repeated our auxiliary experiments using a few pairs of undistorted tokens of the vowels /i/, /ʌ/, /ɛ/, and /u/, and we found roughly the same pattern of within-vowel scores as that for the distorted tokens. Additional evidence that limits the possible role of the noise distortion as a source of discrimination cues comes from Hanna's (1984) study of the ability to determine whether two 400-msec bursts of wideband noise separated by 500 msec were the same or different. The ability of highly practiced listeners to compare the wave forms correctly ($d' \approx 1.1$) was much poorer than the ability of our listeners to discriminate two vowel wave forms when each was subject to a single noise distortion ($d' > 3$).

For a given vowel, the utterance-to-utterance variability studied in these experiments is only one component of the variation that occurs in natural speech. The physical characteristics of a given vowel are known to vary substantially across speakers (see, e.g., Peterson & Barney, 1952) and consonant contexts (e.g., House & Fairbanks, 1953; Klatt,

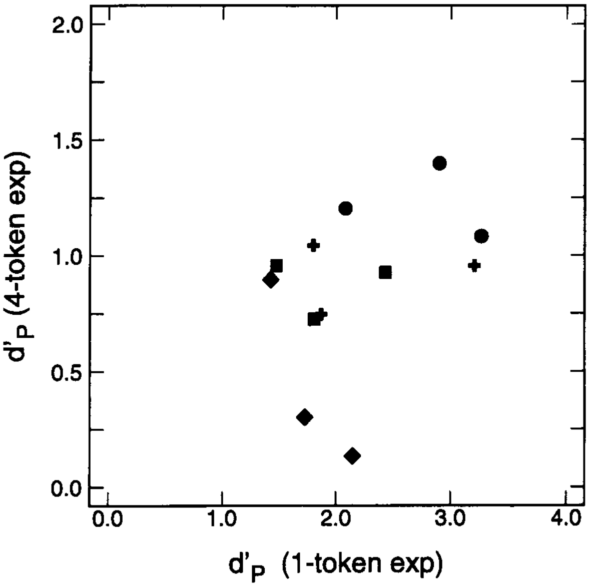

Figure 5. Estimates of d'_P in the 1- and 4-token experiments. Symbols are as defined in Figure 2. All estimates are derived from the token pairs used in the 1-token experiments. Multiple data points for each of the four vowel pairs are separate results for each of three listeners.

1976), as well as speaking style and lexical stress (e.g., Huang, 1991; Picheny, Durlach, & Braida, 1986). Nevertheless, this type of variability is fundamental to all styles of speech production, and its study contributes to development of methods that may prove useful when other types of variability are present.

Our results show, however, that even this restricted utterance-to-utterance variability can have substantial perceptual effects. When listeners attempted to distinguish between pairs of single tokens of these vowels selected randomly, they nearly always were able to do so, whether the utterances represented the same vowel or different vowels that were difficult to distinguish in a 10-vowel identification experiment. Each listener found some pairs of tokens difficult to distinguish, although other listeners could usually distinguish the same pairs fairly easily. The effect of variability in this case was to provide a plurality of cues that listeners could use to distinguish between the utterances, although not all listeners appeared to base decisions on the most salient differences. When the number of tokens was increased to 4 and 16, and listeners were required to distinguish between the populations of utterances, performance decreased markedly relative to the single-token case. The effect of variability in this case was to reduce the utility of cues that were effective in distinguishing between individual tokens but were not relevant to general differences between the populations.

It is unlikely that the decreased ability to distinguish between vowel populations simply reflects the uniform increase in internal noise for the same decision variable X used in the single-token experiments. Such an increase would reduce the ability to distinguish token pairs uniformly. In particular, the ordering of difficulty with which token pairs can be distinguished would be the same as the number of tokens changed. Had the values of d' for specific token pairs in the 1-token experiment been correlated with the values in the 4- and 16-token experiments, the effect of token variability could be attributed to a uniform change in internal noise, but such correlations were not observed. By contrast, there was evidence for modest correlation for the 4- versus 16-token conditions. This is consistent with the notion that listeners modified the weights applied to the available cues as the number of tokens used to represent each vowel changed. When the number of tokens is small, weights appropriate for specific tokens can be used. When the number of tokens is large, the weights approximate those appropriate for the whole population of tokens.

Our estimates of the value of the parameter γ shed some light on the relationship between auditory perceptual resolution and speech production variability. In the 4- and 16-token experiments, $d' \approx 1$ and $\gamma \approx 0.4$. This indicates that the populations were relatively difficult to distinguish and that the difficulty resulted more from auditory perceptual limitations (or internal noise) than from relevant cue variability (or external noise). It is interesting to compare our estimates of γ in the multiple-token experiments with estimates derived by Siegel and Colburn (1989) for the case of binaural detection of a tone in additive low-pass Gaussian noise. Their estimates (for three listeners and two listening conditions) were in the range $0.4 < \gamma < 1.1$. By com-

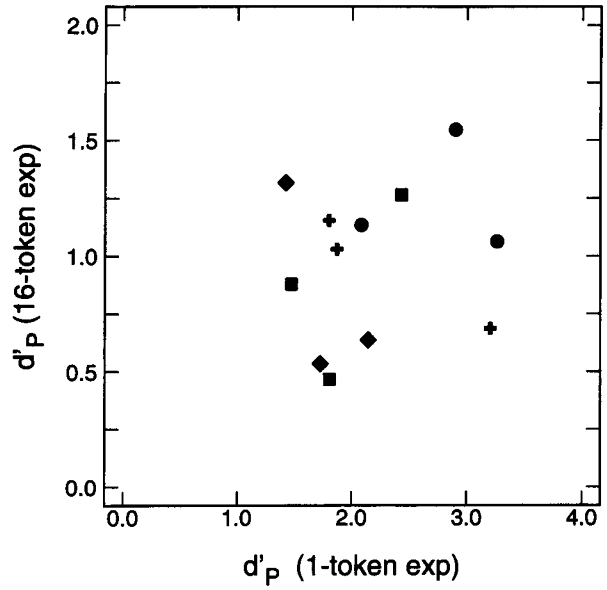

Figure 6. Estimates of d'_P in the 1- and 16-token experiments. Symbols are as defined in Figure 2. All estimates are derived from the token pairs used in the 1-token experiments. Multiple data points for each of the four vowel pairs are separate results for each of three listeners.

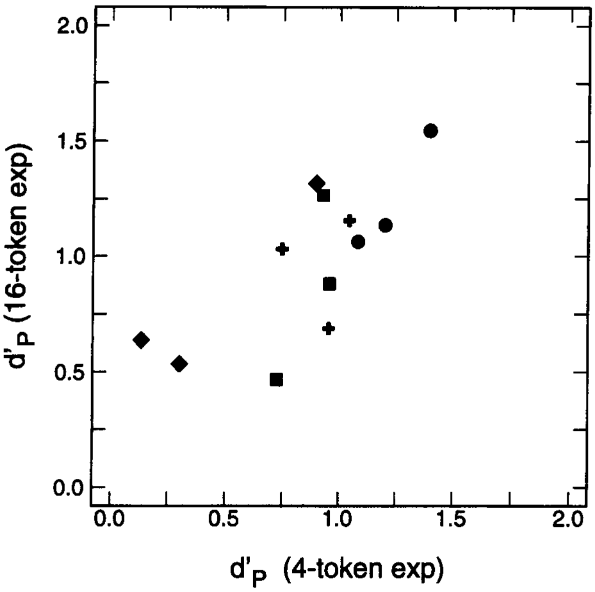

Figure 7. Estimates of d'_P in the 4- and 16-token experiments. Symbols are as defined in Figure 2. All estimates are based on the same token pairs used in the 1-token experiments. Multiple data points for each of the four vowel pairs are separate results for each of 3 listeners.

parison, our estimates were in the range $0.2 < \gamma < 0.8$. Clearly there is substantial overlap between these ranges.

Although different listeners were tested and different tokens of the vowel sounds were employed, our results can be compared with the data of Uchanski et al. (1992) on vowel identification. In the 1-token case, resolution for all four vowel pairs was greater (by a factor of 1.5–2.0) in the present 2-vowel task than in the 10-vowel identification task. By contrast, in the 4- and 16-token cases, sensitivity was roughly the same in the 2- and 10-vowel tasks. Thus while the present results are similar to those found for identification (Uchanski et al., 1992) in that the effect of increasing the number of tokens from 1 to 4 was greater than that from 4 to 16, the effect of increasing the number of tokens from 1 to 4 was much larger for 2- than for 10-vowel tests.

The 2- and 10-vowel tasks allow the listener different degrees of flexibility in selecting the various cues used to perform the task. According to the model of discrimination experiments, when only two vowels are to be distinguished, the listener forms a single decision variable as the linearly weighted sum of the available cues. This weighted sum, even if optimal for a particular vowel pair, is unlikely to be adequate for the 10-vowel identification task. To achieve high identification scores, listeners must attend to characteristics relevant to all of the vowels used in the more complex task. These characteristics are more likely to be similar to those used to make phonetic distinctions in ordinary situations.

Depending on instructions, listeners can make distinct estimates of the "psychophysical" and "phonetic" distance between synthetic vowel sounds (Klatt, 1979). Psychophysical distances are manifest when subjects judge the difference between pairs of vowel sounds (see, e.g., Carlstrom, Grantstrom, & Klatt, 1979). Phonetic distances are manifest when subjects consider vowel identity explicitly and disregard such factors as harshness and speaker identity. Klatt (1979) found that different parameters affected the two types of judgments. For example, altering the phase relations among voicing source harmonics strongly affected the judgment of psychophysical distance but not the judgment of phonetic distance. It is also noteworthy that, although different groups of listeners were used in the two rating tasks, the average estimates of phonetic distances were generally much smaller than average estimates of psychophysical distances. Although Klatt (1979) did not examine the effect of utterance to utterance variations in vowel production, these results suggest that components that contribute to the ability to distinguish between vowels as sounds may have little bearing on the ability to distinguish between vowel sounds as speech elements.

The existence of utterance to utterance variability in speech production is likely to have substantial effects on studies of the relationship between physical and perceptual properties of speech sounds. The traditional psychophysical strategy of varying specific stimulus characteristics and examining the effects on the listener's ability to distinguish between stimuli, like our 1-token experiments, may have limited interpretative value because the perceptual results may not depend on stimulus properties that have general phonetic value. A more promising approach would introduce random variation in some characteristics while focusing on controlled differences in others, in an effort to encourage listeners to generalize over irrelevant variations. Although this would seem likely to create situations in which the ratio of external to internal variance was small, as in our multiple-token experiments, it is noteworthy that listeners appear to adopt relatively consistent cue weightings when a large set of tokens of each utterance is used. However, in this case γ is relatively small, so it may be more difficult to relate perceptual results to physical characteristics, because they may be dominated by internal variability rather than the characteristics of the stimuli.

REFERENCES

Berg, B. G., & Green, D. M. (1990). Spectral weights in profile listening. *Journal of the Acoustical Society of America*, **88**, 758-766.

Berliner, J. E., Braida, L. D., & Durlach, N. I. (1977). Intensity perception: VII. Further data on roving level discrimination and the resolution and bias edge effects. *Journal of the Acoustical Society of America*, **61**, 1256-1267.

Berliner, J. E., & Durlach, N. I. (1973). Intensity perception: IV. Resolution in roving-level discrimination. *Journal of the Acoustical Society of America*, **53**, 1270-1287.

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology*, **43A**, 647-677

Carlstrom, R., Grantstrom, B., & Klatt, D. H. (1979). The relative perceptual salience of selected acoustic manipulations. *Journal of the Acoustical Society of America*, **66**, S86.

Cornett, R. O. (1967). Cued speech. *American Annals of the Deaf*, **112**, 3-13

Durlach, N. I., & Braida, L. D. (1969). Intensity perception: I. Preliminary theory of intensity resolution. *Journal of the Acoustical Society of America*, **46**, 372-383.

Farrar, C. L., Reed, C. M., Durlach, N. I., Zurek, P. M., Ito, Y., & Braida, L. D. (1987). Spectral-shape discrimination: I. Results from normal-hearing listeners for stationary broadband noise. *Journal of the Acoustical Society of America*, **81**, 1085-1092.

Gourevitch, V., & Galanter, E. (1967). A significance test for one parameter isosensitivity functions. *Psychometrika*, **32**, 25-33.

Green, D. M., Kidd, G., & Picardi, M. C. (1983). Successive versus simultaneous comparison in auditory intensity discrimination. *Journal of the Acoustical Society of America*, **73**, 639-643.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hanna, T. E. (1984). Discrimination of reproducible noise as a function of bandwidth and duration. *Perception & Psychophysics*, **36**, 409-416.

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America*, **25**, 105-113.

Houtsma, A. J. M., & Goldstein, J. L. (1972). Central origin of complex-tone pitch. *Journal of the Acoustical Society of America*, **51**, 520-529.

Huang, C. B. (1991). *An Acoustic and Perceptual Study of Vowel Formant Trajectories in American English*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, **59**, 1208-1221.

Klatt, D. H. (1979). Perceptual comparisons among a set of vowels similar to /æ/, some differences between psychoacoustic distance and phonetic distance. *Journal of the Acoustical Society of America*, **66**, S86.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, **50**, 93-107.

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

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

Maney, J. W. (1989). *Token variability of intra-speaker speech*. Unpublished bachelor's thesis, Massachusetts Institute of Technology.

Perkell, J. S., & Klatt, D. H. (Eds.) (1986). *Invariance and variability in speech processes*. Hillsdale, NJ: Erlbaum.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, **24**, 175-184.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech & Hearing Research*, **29**, 434-446.

Pisoni, D. B. (1990). Effects of talker variability on speech perception: Implications for current research and theory. In *Research on speech perception* (Progress Rep. No. 16, pp. 169-191). Bloomington: University of Indiana, Department of Psychology, Speech Research Laboratory.

Pollack, I. (1956). Identification and discrimination of components of elementary auditory displays. *Journal of the Acoustical Society of America*, **28**, 906-909.

Ronan, D. E. (1992). *Effects of token variability on vowel intelligibility*. Unpublished bachelor's thesis, Massachusetts Institute of Technology.

Schroeder, M. R. (1968). Reference signal for signal quality studies. *Journal of the Acoustical Society of America*, **44**, 1735-1736.

Siegel, R. A., & Colburn, H. S. (1989). Binaural processing of noisy stimuli: Internal/external noise ratios for diotic and dichotic stimuli. *Journal of the Acoustical Society of America*, **86**, 2122-2128.

Uchanski, R. M., Millier, K. M., Reed, C. M., & Braida, L. D. (1992). Effects of token variability on vowel identification. In M. E. Schouten (Ed.), *The auditory processing of speech. From sounds to words* (pp. 291-302). Berlin: Mouton de Gruyter.

NOTES

[1] These labels corresponded to the names of the intended vowels in the main experiments and to numerals in the case of the auxiliary experiments.

[2] Listener S2 also found this pair difficult to discriminate, but Listener S3 was able to distinguish the pair without error.

[3] The raw variance estimates were first corrected to account for the Bernoulli variability of the estimate of the response frequency. An estimate of the variance associated with this variability (Gourevitch & Galanter, 1967) was subtracted from the computed variance of the z scores.

[4] The number of pairs available to estimate the correlation then ranged from 13 to 28.

[5] Specifically, in 12 of the 96 across-vowel tests and 5 of the 48 within-vowel tests.

APPENDIX

In the 1-token experiments, there were an inadequate number of confusions for many token pairs to allow d' to be estimated conventionally (see, e.g., Macmillan & Creelman, 1991). Rather than exclude such cases from our analysis, or assign them an arbitrarily selected value of d', we make use of the fact that for a fixed set of weights, the distribution of d'_F over populations of token pairs is Gaussian with mean μ and variance η^2. We then used μ as the estimate of $E_P[d'_F]$ and η^2 as the estimate of $V[d'_F] = 2\gamma^2$. We estimated μ and η from the measured distribution of proportion correct (P_C) scores. In most cases $P_C < 1.0$, and we made use of the relation $P_C = \Phi(d'_F/2)$. In the small number[5] of cases in which $P_C = 1.0$, we modified the observed value of P_C by assuming that a response error would have occurred on the next trial. Finally, with the following procedure, we improved the estimates of μ and η obtained from these values of P_C: We simulated fixed token experiments for values of μ and η in the range $0.5 < \mu < 5.0$ and $0.0 < \eta < 3.0$ with the same number of trials per token pair as in the real experiments. We then used our procedures to estimate μ and η from the results of the simulated experiments. Improved estimates were derived by linearly regressing the simulated values of μ and η on the estimated values. Additional simulations that were conducted to test the adequacy of these regressed estimates indicate that over the relevant range of parameter values the procedure yields estimates of μ and η that show little bias (less than 10% of nominal values), with standard deviations of roughly 20% of nominal values.

(Manuscript received January 6, 1994;
revision accepted for publication April 23, 1997.)