The effect of discrimination training on speech perception: Noncategorical perception

ARTHUR G. SAMUEL

University of California, San Diego, La Jolla, California 92093

Three subjects were given extensive practice in discriminating syllables which differed in voice onset time. For these subjects, there were two major findings. First, discrimination of speech follows normal psychophysical laws: long-onset-time stimuli require larger differences than shorter ones for comparable discrimination. Second, the shape of the discrimination function for experienced subjects is more like a leaning W than an inverted V, the usual shape for naive subjects. The data support a model of speech perception with both an acoustic and a phonetic component. The phonetic component is best characterized as a prototype matching process, with the prototype including information on the simultaneity of formant onset.

For the last 20 years, the laws governing speech perception have been thought to differ from the laws governing psychophysical perception. In most psychophysical experiments, subjects can discriminate many more stimuli than they can identify. In most speech perception studies, discrimination seems to be bounded by identification; subjects can only discriminate two speech items if they can give them different phonetic labels. Speech perception appears to be categorical. Largely on the basis of this finding, researchers at the Haskins Laboratories (Liberman, 1970; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman, Harris, Hoffman, & Griffith, 1957; Liberman, Cooper, Harris, & MacNeilage, Note 1) have argued that a "specialized speech mode" of perception is invoked to process incoming speech. Since identification and discrimination functions in speech studies differ from those in nonspeech studies, the speech processor is not believed to be subject to normal psychophysical laws.

Support for this position has come from several studies of voicing, the phonological feature which distinguishes /p,t,k/ from /b,d,g/ in English. The three classic studies of this feature were done by Liberman, Delattre, and Cooper (1958), Liberman, Harris, Kinney, and Lane (1961), and Lisker and Abramson (1964). Through speech synthesis and

spectrographic analysis, these studies established that two acoustic features are most important in the transition from voiced to voiceless stops. In voiced stops, the first formant (F1) begins at the same time as the higher formants. Removing the initial portion of the first formant, and thereby delaying its onset, leads to the perception of voicelessness. A more realistic continuum is obtained if the higher formants are aspirated (energized by a noise source) during the period of F1 cutback.

The Haskins researchers have generally tested discrimination with the ABX paradigm. In this paradigm, subjects hear three syllables per trial. The first two (A and B) always differ from each other, while the third (X) is identical to one of the first two. The subject's task is to determine if X is the same as A or B.

Liberman et al. (1961) used this paradigm in their study of voicing. The authors synthesized a continuum of speech syllables which varied in voice onset time (VOT) by varying the F1 cutback and aspiration cues in 10-msec steps. The other parameters were appropriate for an alveolar consonant followed by the vowel /o/, yielding a continuum perceived as /do/ at one end (0-msec VOT) and as /to/ at the other (60-msec VOT). The data generally indicate better discrimination between phonetic categories than within them, an example of categorical perception.

Early Haskins papers (e.g., Liberman et al., 1967; Liberman et al., Note 1) cited this finding as evidence for a motor theory of speech perception. In more recent work (e.g., Liberman, 1970), no specific mechanism has been offered which would produce the categorical results, but the general position of a special speech mode has been maintained. As Liberman (1970) puts it, "The [speech] decoder is not merely an extension of our auditory system, but is, more properly, an integral part of the

This research was supported by National Science Foundation Grant GB 32235X to David Rumelhart and National Institute of Mental Health Grant MH 15828 to the Center for Human Information Processing. I wish to thank Dr. Dennis Klatt and Dr. Kenneth Stevens for constructing the stimuli. I would also like to thank Dave Rumelhart, Don Norman, Elissa Newport, and the LNR Research Group for their help in all phases of this project. Reprint requests may be sent to Arthur G. Samuel, Department of Psychology C-009, University of California at San Diego, La Jolla, California 92093.

mechanisms that underlie our use of language'' (p. 252).

In the last several years, a number of studies have yielded data which suggest that a reevaluation of categorical perception is called for. These results may be divided into two types: (1) the observation of categorical perception of nonspeech stimuli, and (2) the observation of noncategorical perception of speech. Cutting and Rosner (1974) constructed a continuum in which they varied the rise time of a sawtooth wave. When they presented their stimuli to subjects for identification (as "plucked" or "bowed" tones) and ABX discrimination, they obtained data totally analogous to the speech findings. The labeling function showed a sharp crossover, discrimination was near chance within categories, and between-category discrimination was excellent.

More recently, Miller, Wier, Pastore, Kelly, and Dooling (1976) varied the relative onset time of a noise burst and buzz, and tested identification ("noise" or "no noise") and discrimination. Their data also showed sharp labeling crossovers, with appropriate peaks and troughs in the discrimination function. A similar study by Pisoni (in press), using the relative onset time of two tones, yielded similar results—categorical perception.

Several demonstrations of noncategorical speech perception exist, but none is definitive. Pisoni and Lazarus (1974), using a VOT continuum, provide the most convincing data available. Using a discrimination test which minimized memory demands, the authors found limited noncategorical perception when subjects were instructed to attend to stimulus differences rather than phonetic identity. However, with one possible exception, all of the within-category syllable pairs which were discriminated at better than chance levels contained an item near the phoneme boundary. For some subjects, these may not have been true within-category pairs. Pisoni and Tash (1974) conducted a reaction time study in which subjects made same-different judgments of items from a /ba/-/pa/ continuum. They reported reaction times which varied as a function of physical similarity within phonetic categories. However, due to the nature of reaction time studies and the data collected, it is difficult to measure the degree of noncategorical perception present.

In Barclay's (1972) study of the place feature, subjects classified syllables from a /bae/-/dae/-/gae/ continuum into either two (b,g) or three (b,d,g) categories. When two categories were used, /d/ items near the /g/ boundary were grouped with /g/, while those /d/ items which were physically similar to /b/ were classified as /b/, indicating within-category discrimination. However, as with Pisoni and Tash (1974), the data provide little information on the

underlying psychophysical process. Carney and Widen's (Note 2) study, in which within-category same-different judgments were above chance, is subject to the same criticism. Take together, these studies strongly suggest that noncategorical perception of speech is possible. However, they are all essentially demonstrations rather than systematic investigations of the perceptual process.

Several recent models of speech perception have emphasized acoustic properties more than phonetic ones, an approach supported by the studies just cited. For example, Stevens and Klatt (1974) have argued that the voiced/voiceless distinction is cued by the presence or absence of a significant transition in the appropriate frequency range. Since present/absent is a binary feature, categorical perception is to be expected.

Miller et al., (1976) and Divenyi, Sachs, and Grant (Note 3) have also taken strong acoustic positions. In both studies, the stimuli were tones or noise bursts in which a timing feature was varied. While neither set of stimuli was perceived as speech, both yielded discrimination functions which resemble those for speech. Based on this, Divenyi et al. and Miller et al. argue that categorical perception (at least of voicing) is simply a function of normal psychophysical processing; phonetic experience may induce minor changes in boundary location and sharpness, but it is not a prerequisite for categorical perception.

Fujisaki and Kawashima (1969, 1970) and Pisoni (1973) have argued for a compromise position in which both acoustic and phonetic processing take place. In this type of model, the incoming waveform is stored in both an acoustic memory and a phonetic one, with each store having different properties. Most importantly, the acoustic code decays rapidly, leaving only the phonetic information in tasks which impose a large memory load, leading to categorical perception.

In the present study, I use a somewhat different procedure than previous investigators have used. Subjects in earlier studies were given no more than a few hours' practice at the novel discrimination task. My subjects were given extensive discrimination training, with feedback. The data I collected were chosen to bear on two questions: (1) What does the discrimination function look like as a function of VOT? and (2) How does this function change (if at all) as the listener becomes more practiced? More specifically, Can within-category discrimination become as good as between-category discrimination?

The answer to these questions may help select among current theories of speech perception. For example, since the basic function of a special speech decoder is to categorize phonemes, the Haskins theory predicts little or no effect of practice on dis-

crimination performance. The learning data from the present study can show whether this theory is tenable.

A second purpose of this study is more empirical than theoretical. This goal is simply to provide useful data that are not currently available. In virtually all of the experiments on voicing, the experimenters have chosen one or more VOT step sizes (usually 10, 20, or 30 msec) and measured ABX discrimination along the continuum for these step sizes. In the present study, I measure discrimination using both this method and one borrowed from psychophysics. In the psychophysical method, discriminability is held constant at 75% correct ($d' = 1$), and the VOT step size needed to maintain this performance level is measured at various points along the continuum. The two measures are complementary and, together, give a better picture of the underlying structure than either one does alone.

METHOD

Subjects

There were three subjects in the experimental group. All were right-handed native speakers of English with no known hearing problems. I served as one of the three (subject A.G.S.), and the other two were recruited through a sign-up sheet. The first two applicants who met the above criteria and were willing to participate in a long-term experiment were chosen; there was no preselection of subjects. The subjects were told that the purpose of the experiment was to see how well they could discriminate certain speech sounds through practice. Neither of the recruited subjects had had any previous experience with synthetic speech, and each was paid \$2/h of participation.

Stimuli

The stimuli used in all training and tests were synthetic CV syllables 300 msec long. They were constructed on the vocal tract analogue synthesizer at the Research Laboratory of Electronics, Massachusetts Institute of Technology.¹ Due to the superior performance of this synthesizer, the stimuli used in this study were extremely realistic. For all syllables, the parameters were appropriate for an alveolar consonant followed by /a/. To construct a voicing continuum, F1 cutback was varied in 3-msec steps. The higher formants were aspirated during the period of F1 cutback. The continuum ranged from 0-msec VOT (/da/) to 81-msec VOT (/ta/), for a total of 28 different items. The stimuli were digitized at a 10-kHz sampling rate and stored on computer magnetic tape.

Apparatus

All training and tests were run under computer control (the PDP-9 in the LNR Research Laboratory, University of California at San Diego), with the digitized syllables stored on disk files. Each subject sat in an acoustically isolated booth which contained a set of stereo headphones, a television monitor, and a keyboard. Syllables were fed into a 12-bit digital-to-analog converter at a 10-kHz sampling rate, low-pass filtered at 5 kHz, amplified, and presented through headphones at a comfortable listening level (approximately 78 dBA). The subject responded by pressing one of two labeled keyboard buttons, and the computer presented feedback on the television monitor.

Procedure

The core of the present study involved extensive discrimination training along the VOT continuum. I chose five points (*anchors*)

at which to train subjects. These anchors were at 0-, 15-, 42-, 57-, and 81-msec VOT. These values were intended to span the whole continuum with relatively even spacing, without putting an anchor at the expected phoneme boundary (20-40 msec). Each anchor was assigned a variable *mate*. Mates were initially 27 msec VOT away from their anchors, displaced towards the phoneme boundary, yielding values of 27-, 42-, 15-, 30-, and 54-msec VOT.

An ABX paradigm with feedback was used for training. Table 1 lists the events on a single trial during training. For each trial, both the order of anchor and mate presentation (first or second) and identity of X (anchor or mate) were determined by a random number generator.

One training session was run per day, with four sessions generally run per week. A session consisted of 400 trials. Including a 5-10-min break after 200 trials, a session lasted approximately 1½ h. Each block of 200 trials consisted of 40 consecutive trials at each of the five anchors, with the order of anchors determined randomly for each block of 200 trials.

The distance (in milliseconds VOT) between each anchor and its mate was adjusted every fourth trial in order to keep performance at 75% correct. The adjustment routine followed these rules: After each block of four trials, (1) if the subject got 3 of the last 4 trials correct, make no change; (2) if the subjects got 0, 1, or 2 correct, make the task easier by moving the mate 3-msec VOT away from the anchor; (3) if the subject got all 4 trials correct, move the mate 3-msec VOT closer to the anchor. The adjustment routine also included provisions to prevent the mate from "passing through" the anchor or from going off either end of the continuum. Since adjustments were made after every four trials, each anchor-mate distance was updated 10 times per block of 40 trials and 20 times per training session.

For each session, the initial anchor-mate distances were set to the last values from the previous session. Two subjects completed 30 training sessions, and the third completed 20.²

Data Collection

Three types of data were collected:

(1) Five *distance scores* were collected daily, one for each anchor. Each day's distance scores were simply the mean distances between the anchors and their mates, rounded off to the nearest stimulus value (3-msec step). These scores were based on 20 non-independent scores, the 20 values which the mate took on after each adjustment.

(2) At roughly weekly intervals, I collected *standard 1-step ABX and identification* data. These two tests were designed to be comparable to the tests in the literature. For identification, 10 randomizations of the 28 stimuli were presented, and subjects labeled them as /da/ or /ta/. Subjects were given 2.5 sec to respond to each item.

The stimuli in the standard 1-step ABX test were the 0-, 9-, 18-, 27-, 36-, 45-, 54-, 63-, 72-, and 81-msec VOT syllables. Adjacent items (those that differed by 9 msec) were presented for ABX discrimination, without feedback. Table 2 lists the events during a single ABX test trial.³ Twenty randomizations of the nine one-step pairs were presented, with AB ordering and X identity determined randomly for each trial.

Table 1
Timing of Events During a Training Trial

Event	Duration (msec)
1. Present Stimulus A	300
2. Silent interstimulus interval	350
3. Present Stimulus B	300
4. Silent interstimulus interval	550
5. Present Stimulus X	300
6. Subject responds	2,700
7. Present correct answer on screen	1,700
8. Wait for next trial	4,500

Table 2
Timing of Events During a Test Trial

Event	Duration (msec)
1. Display "READY" on screen	1,000
2. Clear screen	300
3. Present Stimulus A	300
4. Silent interstimulus interval	450
5. Present Stimulus B	300
6. Silent interstimulus interval	650
7. Present Stimulus X	300
8. Subject responds	3,000
9. Wait for next trial	2,000

(3) At the end of training, a set of final tests was given, consisting of the final identification and standard 1-step ABX tests plus a 2-step ABX test. The 2-step test was identical to the 1-step test except that the AB pairs were separated by 2 steps (18-msec VOT) along the continuum.

RESULTS

The Distance Scores

Figure 1 presents the mean distance scores for each subject as a function of training. Each point represents the mean distance (on the VOT continuum) between an anchor and its mate during five training sessions.⁴

I conducted a separate analysis of variance on each subject's scores. A separate analysis for each subject was indicated by the extremely large individual differences in performance level and phoneme boundary location, particularly the latter. The two

factors analyzed were placed along the VOT continuum (five anchors) and period of training (six five-session periods for A.G.S. and M.D.S., four for M.C.B.). The five sessions within each training period were treated as replications.

The results for the three subjects were quite similar. The main effect of anchors was highly significant [A.G.S: $F(4,120) = 22.4$, $p < .001$; M.D.S: $F(4,120) = 241.9$, $p < .001$; M.C.B: $F(4,80) = 82.2$, $p < .001$], reflecting the vertical spread of the anchors apparent in Figure 1. Except in those cases where a subject's phoneme boundary fell near an anchor (see Figure 2 for the boundaries), the ordering of the anchors is quite straightforward: the shorter the voice onset time of the anchor, the better the performance. These results are quite similar to those obtained in a study of the discriminability of a nonspeech timing feature (Divenyi & Danner, 1977). Divenyi and Danner attribute the "degrading performance as the base [time] intervals become longer and longer" (p. 134) to the operation of Weber's law. It appears that the perception of a timing feature in speech follows the same psychophysical law.

The main effect of training was also significant [A.G.S: $F(5,120) = 19.3$, $p < .001$; M.D.S: $F(5,120) = 8.8$, $p < .001$; M.C.B: $F(3,80) = 23.0$, $p < .001$]. This result reflects the generally negative slope of the learning curves during training.

For some of the anchors, there was clearly no improvement during training. The selective effect of

Figure 1. Difference between anchor and mate (in msec VOT) needed to maintain 75% correct discrimination. Each point represents the mean of the distance scores of five training sessions (see Footnote 4).

practice on some anchors and not on others is reflected in a significant Anchors by Training interaction [A.G.S: $F(20,120) = 3.6$, $p < .001$; M.D.S: $F(20,120) = 2.5$, $p < .001$; M.C.B: $F(12,80) = 9.2$, $p < .001$]. Inspection of Figure 1 indicates that performance on the anchors at the ends of the continuum (0- and 81-msec VOT) improved most; the other curves are generally flat.

Figures 2 and 3 illustrate this effect graphically. Figure 2 presents the anchor-mate distances at the beginning (Sessions 2-7) and end (the last six sessions) of training. The arrows indicate anchor locations, with each mate joined to its anchor. With this representation, the width of a bar joining an anchor and its mate is proportional to the mean distance between them. Therefore, a shorter bar indicates better discrimination. The phoneme boundary for each subject (estimated from labeling tests) is represented by a dashed vertical line.

Two aspects of the "before-training" data are of most interest. First, different phonetic labels are sufficient, but not necessary, for discrimination. All three subjects clearly demonstrate within-category discrimination before extensive training; mates and anchors need not be on opposite sides of the phoneme boundary. However, when a mate approaches the boundary, discrimination quickly reaches the $d' = 1$ criterion, illustrating the sufficiency of phonetic differences. The second point of interest is that the VOT difference needed for discrimination increases with the VOT of the anchors. This is just another representation of the psychophysical effect of Weber's law.

For each subject's last six training sessions (25-30 for A.G.S. and M.D.S., 15-20 for M.C.B.), the distances are generally smaller, but the pattern is similar. Subjects still require greater differences in the /ta/ range than for /da/s. Within the /ta/ range, however, there is no longer any evidence of Weber's law. This reflects the interaction between anchors and training: the 81-msec VOT anchor showed more improvement than the others.

Figure 3 depicts each subject's change in discrimination performance as a result of training. Each bar represents the mean anchor-mate distance at the end of training (the last six sessions) minus the mean anchor-mate distance during Sessions 2-7. The data are mostly as expected by this point: most improvement occurred at the continuum ends. It might be objected that a more appropriate measure of improvement is the proportion of possible improvement. The greater improvement at the continuum ends might simply reflect more room for improvement, since they are furthest from the phoneme boundary. When these proportions (change in distance/initial distance) are computed, however, they show the same pattern as the absolute differ-

Figure 2. Mean distances (in msec VOT) between anchors and their mates at the beginning (Sessions 2-7) and end (the last six sessions) of training. Anchors (indicated by arrows) are joined to their mates. The width of the bar joining an anchor to its mate represents the mean difference necessary to maintain performance at 75% correct. Each subject's phoneme boundary is represented by a dashed line.

ences: discrimination improves most near the continuum ends.

The 1-Step ABX and Identification Tests

Figure 4 illustrates the labeling and standard ABX discrimination functions for the three subjects. The number at the bottom right corner of each graph gives the number of training sessions that preceded

Figure 3. Difference (in msec VOT) between anchor-mate distance during the last six training sessions and Sessions 2-7. Each subject's phoneme boundary is represented by a dashed line. A descending bar indicates worse performance after training.

the test. The dashed curves represent the discrimination functions expected under the extreme assumption of categorical perception.⁵

I believe that the simplest way to interpret the graphs is to divide them into two stages. In the first stage, the task is in general too difficult, leading to noisy, near-chance behavior. The first test for M.D.S. and the first three for M.C.B. fit this pattern. As discrimination improves, a more systematic pattern appears. Subject A.G.S. shows a consistent pattern starting with his first ABX test (after six sessions). M.D.S. is somewhat more variable, but clearly nonrandom, beginning with his second test (after Session 4). M.C.B. is the last to show stable results (after Session 18). Figure 5 presents the average 1-step ABX and labeling curves for each subject's Stage 2 data.

The data in Table 3 support the preceding interpretation. The table presents mean distance scores during training, divided into five-session blocks. These values represent the differences necessary to maintain 75% correct performance in the blocked procedure used for training. When the mean distance is greater than about 15 msec, the obtained ABX function for stimulus pairs differing by 9 msec will be noisy and near chance. Below this value, useful data may be obtained.

For each subject, I conducted a two-way analysis of variance on the 1-step ABX data. Only the second

stage data were used in these analyses (tests run after Session 18 for M.C.B., after Session 4 for M.D.S., and all for A.G.S.). One factor had two levels (obtained vs. predicted discrimination), while the second had nine (the nine 1-step comparisons along the VOT continuum). For all three subjects, the obtained discrimination was superior to the predicted [A.G.S: $F(1,108) = 227.7$, $p < .001$; M.D.S: $F(1,108) = 13.3$, $p < .001$; M.C.B: $F(1,18) = 7.1$, $p < .025$]. The main effect of voice onset time comparison, reflecting the curve's peaks and troughs, was similarly significant [A.G.S: $F(8,108) = 34.0$, $p < .001$; M.D.S: $F(8,108) = 10.5$, $p < .001$; M.C.B: $F(8,18) = 4.8$, $p < .01$]. The significant interaction [A.G.S: $F(8,108) = 16.2$, $p < .001$; M.D.S: $F(8,108) = 4.0$, $p < .001$; M.C.B: $F(8,18) = 3.8$, $p < .01$] reflects the fact that some points along the continuum conform more closely to the predictions of the categorical perception model than others, notably those points in the /ta/ range.

I conducted a separate one-way analysis of variance on the obtained scores for each subject to test the reliability of the observed peaks and troughs. The main effect of place along the VOT continuum was significant [A.G.S: $F(8,54) = 29.6$, $p < .001$; M.D.S: $F(8,54) = 6.4$, $p < .001$; M.C.B: $F(8,9) = 4.5$, $p < .025$]. The question of interest is whether the troughs are significantly lower than the peaks near the continuum ends. I tested these differences

Figure 4. Labeling and 1-step ABX data for the three subjects. The number in the lower right corner of each panel is the number of training sessions which preceded the test.

Figure 5. Composite labeling and 1-step ABX data for the three subjects. The composite curves are based on stage 2 data (140 observations per point for A.G.S. and M.D.S., 40 per point for M.C.B.).

with the Newman-Keuls test of specific comparisons (Keppel, 1973). For A.G.S., the /da/ range trough (the 27-36-msec comparison) is significantly lower than the short VOT peak (0-10-msec comparison), $p < .01$. A similar result is obtained in the /ta/ range, with the 54-63-msec pair significantly below the 63-72-msec comparison, $p < .01$. However, while the /da/ trough for M.D.S. approached significance, neither peak-trough comparison reached the .05 level of significance for M.D.S. or M.C.B.

The Final Tests

Figure 6 presents the results of the final tests for each subject. These tests include the last 1-step ABX test, the last labeling function, and a 2-step ABX test.⁶ The 2-step test provides a means of equating the performance of M.C.B. and M.D.S. with that of A.G.S. That is, if the 2-step test matches the underlying psychophysical structure of M.C.B. and M.D.S. in the same way that the 1-step test maps A.G.S.'s, then it is possible to see if all three underlying structures are of essentially the same form. If

so, we may expect to find more pronounced peaks and troughs in the 2-step test than in the 1-step for M.C.B. and M.D.S. The obtained functions support this hypothesis, particularly in the /ta/ range. The phoneme boundaries are too low (about 25 msec) to allow a major trough between peaks in the /da/ range.

To assure that the results obtained in this study were not due to some artifact of procedure, stimuli, or subjects, I ran five control subjects through the final set of tests. Figure 6 presents the results for the three subjects with the highest percentage correct on the two ABX tests. Two facts should be noted. First, the experimental group's performance far exceeds that of the control group, as it should if the training was effective. Second, the performance of the control group is very much in line with the data in the literature (e.g., Liberman, Harris, Kinney, & Lane, 1961), ruling out most explanations by artifact.

DISCUSSION

There were two main findings in the present study. The first of these was that a normal psychophysical law operates as one component of subjects' discrimination functions for speech stimuli. Figures 2 and 5 illustrate this observation: in order to maintain the same level of performance ($d' = 1$), subjects require larger differences between stimuli with long voice onset times than for short VOT stimuli. Pisoni and Lazarus (1974) reported a similar asym-

Table 3
Mean Distance (in msec VOT) Between Anchors and Mates
as a Function of Training

Subject	Sessions					
	1-5	6-10	11-15	16-20	21-25	26-30
A.G.S.	13.32	8.04	8.40	9.00	6.84	6.96
M.D.S.	18.00	15.72	14.40	14.76	13.20	15.24
M.C.B.	21.96	16.08	15.36	15.24

Figure 6. The final tests: labeling, 1-step ABX, and 2-step ABX tests given after training (trained subjects) or without training (control group).

metry in discrimination of synthetic /ba/-/pa/ syllables. In fact, the individual subjects' data in the Liberman, Harris, Kinney, and Lane (1961) study follow much the same pattern. For the 2-step comparisons, 9 of the 13 subjects show better performance on short VOT stimuli than on long ones, and 11 of the 13 do so for the 3-step pairs.

In a concurrent (but independent) study similar to the present one, Sachs and Grant (Note 4) also found that short VOT stimuli were more discriminable than long ones. These investigators reported that their experienced subjects required larger VOT differences in the /ka/ range (30-60 msec VOT) than in the /ga/ range (10-30 msec VOT).

Pisoni and Lazarus (1974) suggested that the observed asymmetry may reflect subjects' use of two different cues for discrimination. In the short VOT range, subjects may decide on the basis of first formant duration and/or onset frequency. For longer VOT stimuli, this cue is unavailable, causing subjects to respond on the basis of the absolute duration of voice onset time. This hypothesis corresponds very closely to my introspective belief of how I was responding to the test and training items. In the /da/ range, the stimuli sounded more "ta-like" as they got closer to the phoneme boundary. To discriminate two /da/s, I tried to scale them as being more or less "ta-like." In the /ta/ range, the situation was very different. Near the phoneme boundary, /ta/s did not sound more "da-like." The cue I used to dis-

criminate /ta/s was duration of aspiration, which covaried with F1 silence. These cues are very similar to those suggested by Pisoni and Lazarus.

The second major finding of the present study concerns the shape of the discrimination function rather than its slope. The usual description of this function is an inverted V, or a single peak between two troughs. While this description may be accurate for naive subjects, my data indicate that a leaning W better describes the discrimination function of experienced subjects. Figure 3 illustrates how the inverted V may become a W—training has its greatest effect at the continuum ends, leaving troughs between the areas of improved discrimination.

A striking aspect of the data was the large between-subject variability. There were very large individual differences in overall level of performance (see Table 3), as well as in subjects' phoneme boundaries. These results support Lane's (1965) call for the illustration of individual subject's data in speech perception studies. The interaction of these individual differences with the type of test given is of even greater importance. In the present study, similar discrimination functions were obtained for all three subjects when each subject was given a test which matched his level of competence (i.e., 9-msec pairs for A.G.S., 18-msec pairs for M.C.B. and M.D.S.). The moral seems to be that speech perception data must be interpreted very cautiously, with attention paid to the possibility of such interactions between individual differences and testing paradigms.

What implications do the findings of the present study have for current theories of speech perception? The data are incompatible with the notion of a special speech mode (e.g., Liberman, 1970). Even without extensive training, all three subjects could clearly discriminate stimuli drawn from the same phonetic category. These results are not overly surprising, given the suggestive results of Barclay (1972), Pisoni and Lazarus (1974), Pisoni and Tash (1974), and Carney and Widen (Note 2), but they do seem to be the clearest demonstration of noncategorical perception of stop consonants (and the only such demonstration using the ABX paradigm).

In recent years, some of the Haskins researchers (e.g., Studdert-Kennedy, Liberman, Harris, & Cooper, 1970) have argued for a weaker version of their earlier position, due to the common finding that obtained discrimination functions do not perfectly match their predicted values. As they put it, "In practice, the hypothesis is seldom fully supported: the obtained function almost always lies somewhat above the predicted, indicating that there remains some basis for discrimination, however marginal, between stimuli that are placed in the same category" (Studdert-Kennedy et al., 1970, p. 236). The improvement of subjects through practice could be

accounted for by substantial growth of the system responsible for within-category discrimination.

A modified version of the Haskins theory in which the input is matched to a "prototype phone" can account for an interesting aspect of the data. The persistence of troughs at certain places along the VOT continuum would be expected if such prototypes were used. Stimuli which are most similar to the prototype might not be discriminable, while those further away from the category center might become discriminable. The sloping nature of the discrimination function is difficult to fit into any version of the Haskins theory.

The models of Miller et al. (1976) and Pisoni (in press) are quite similar. Both attribute categorical perception to fundamental psychophysics. Pisoni posits the existence of a simultaneity detector which would produce three natural voicing categories, leading, simultaneous, and lagging. Miller et al. propose a similar breakdown of voicing into categories based on the existence of one vs. two psychophysical events. The sloping discrimination function obtained in the present study seems very compatible with these models. It indicates that speech perception is based firmly in normal psychophysics, a fact not evident from much previous data.

The discrimination function's W shape is somewhat more difficult to explain, however. Several reasonable interpretations of the acoustic theory yield incorrect predictions. For example, it might be supposed that training would increase the temporal resolution of a simultaneity detector. However, this would lead to the phoneme boundary shifting towards 0 msec, which does not occur. Another possible interpretation is that training should not affect the detector: 20 msec is the minimum time needed for discriminating two events. This is also incorrect; subjects could discriminate items differing by as little as 3 msec.

Despite these problems, the discrimination data may be compatible with the acoustic position. Miller et al.'s comparison of voicing perception to the psychophysical task of judging the timing of two tones provides the basis for this interpretation. They state "As the high tone is started more and more in advance of the low, one may notice a change from apparent simultaneity to nonsimultaneity with uncertain order, to a definite Gestalt-like sequence of high leading low, to a clear temporal separation wherein the high-tone and low-tone onsets are heard as distinct events with the high starting ahead of the low" (Miller et al., 1976, p. 415). A possible explanation of the triple peaks involves the use of three different "detectors." In most cases, subjects may rely on the "Gestalt-like sequence" level as input for phonetic processing. With training, however, they may use the cruder "nonsimultaneity with un-

certain order" and finer "clear temporal resolution" levels to classify previously nondiscriminable items. The problem with this approach is similar to that of the Haskins position: it involves the rapid development of systems which were previously little used and crude.

I believe the model which best fits the data is one which uses the best features of both the acoustic and phonetic positions within the framework suggested by Fujisaki and Kawashima (1969, 1970) and Pisoni (1973). This is a two-component model, with both acoustic and phonetic systems. The acoustic component is indicated by the slope of the discrimination function and the level of within-category perception. This component should include a relatively fast decaying memory (Pisoni, 1973). While the acoustic image is available, subjects may extract information on F1 duration and/or onset frequency and length of aspiration. The rapid improvement of within-category discrimination with practice suggests that these processes are available for use in normal speech perception. The phonetic processor is indicated by the peak at the phoneme boundary and the persistence of troughs within phonetic categories. A possible mechanism for phonetic processing is matching the input to a prototype phone. This prototype might include the expected value of a simultaneity detector's output and other invariant properties of phonemes (cf. Cole & Scott, 1974a, 1974b). While these properties are actually acoustic, they differ from other features in the short time needed for their extraction and the larger weight accorded them in the phonetic decision process.

Much remains to be done in testing and specifying this model. Evidence is needed on the notion of prototypes. Evidence is needed on the issue of acoustic processing. Nevertheless, it appears that we have narrowed the theoretical possibilities and may realistically look forward to understanding categorical perception reasonably soon.

REFERENCE NOTES

1. Liberman, A., Cooper, F., Harris, K., & MacNeilage, P. *A motor theory of speech perception*. In C. Fant (Ed.), *Proceedings on the Speech Communication Seminar*. Unpublished Report, Speech Transmission Laboratories, Royal Institute of Technology, Stockholm, 1963.
2. Carney, A., & Widen, G. *Acoustic discrimination within phonetic categories*. Paper presented at the Acoustical Society of America conference, Washington, D.C., 1976.
3. Divenyi, P., Sachs, R., & Grant, K. *Stimulus correlates in the perception of voice onset time [VOT]: I. Discrimination of the time interval between tone bursts of different intensities and frequencies*. Paper presented at the Acoustical Society of America conference, San Diego, 1976.
4. Sachs, R., & Grant, K. *Stimulus correlates in the perception of voice onset time [VOT]: II. Discrimination of speech with high and low stimulus uncertainty*. Paper presented at the Acoustical Society of American conference, San Diego, 1976.

REFERENCES

- BARCLAY, J. R. Noncategorical perception of a voiced stop: A replication. *Perception & Psychophysics*, 1972, 11, 269-273.
- COLE, R., & SCOTT, B. The phantom in the phoneme: Invariant cues for stop consonants. *Perception & Psychophysics*, 1974, 15, 101-107. (a)
- COLE, R., & SCOTT, B. Toward a theory of speech perception. *Psychological Review*, 1974, 81, 348-374. (b)
- CUTTING, J., & ROSNER, B. Categories and boundaries in speech and music. *Perception & Psychophysics*, 1974, 16, 564-570.
- DIVENYI, P., & DANNER, W. Discrimination of time intervals marked by brief acoustic pulses of various intensities and spectra. *Perception & Psychophysics*, 1977, 21, 125-142.
- FUJISAKI, H., & KAWASHIMA, T. On the modes and mechanisms of speech perception. *Annual Report of the Engineering Research Institute*, Vol. 28, Faculty of Engineering, University of Tokyo, Tokyo, 1969, 67-73.
- FUJISAKI, H., & KAWASHIMA, T. Some experiments on speech perception and a model for the perceptual mechanism. *Annual Report of the Engineering Research Institute*, Vol. 29, Faculty of Engineering, University of Tokyo, Tokyo, 1970, 207-214.
- KEPPEL, G. *Design and analysis: A researcher's handbook*. Englewood Cliffs, N.J.: Prentice-Hall, 1973.
- LANE, H. The motor theory of speech perception: A critical review. *Psychological Review*, 1965, 72, 275-309.
- LIBERMAN, A. Some characteristics of perception in the speech mode. In *Perception and its disorders* (Vol. XLVIII). The Association for Research in Nervous and Mental Disease, 1970.
- LIBERMAN, A., COOPER, F., SHANKWEILER, D., & STUDDERT-KENNEDY, M. Perception of the speech code. *Psychological Review*, 1967, 74, 431-461.
- LIBERMAN, A., DELATTRE, P., & COOPER, F. Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech*, 1958, 1, 153-167.
- LIBERMAN, A., HARRIS, K., HOFFMAN, H., & GRIFFITH, B. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 1957, 54, 358-368.
- LIBERMAN, A., HARRIS, K., KINNEY, J., & LANE, H. The discrimination of relative onset time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology*, 1961, 61, 379-388.
- LISKER, L., & ABRAMSON, A. Cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 1964, 20, 384-422.
- MILLER, J., WIER, C., PASTORE, R., KELLY, W., & DOOLING, R. Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America*, 1976, 60, 410-417.
- PISONI, D. Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, 1973, 13, 253-260.
- PISONI, D. Identification and discrimination of the relative onset of two component tones: Implications for voicing perception in stop consonants. *Journal of the Acoustical Society of America*, in press.
- PISONI, D., & LAZARUS, J. Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, 1974, 55, 328-333.
- PISONI, D., & TASH, J. Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, 1974, 15, 285-290.
- STEVENS, K., & KLATT, D. Role of formant transitions in the voiced-voiceless distinction for stops. *Journal of the Acoustical Society of America*, 1974, 55, 653-659.
- STUDDERT-KENNEDY, M., LIBERMAN, A., HARRIS, K., & COOPER, F. Motor theory of speech perception: A reply to Lane's critical review. *Psychological Review*, 1970, 77, 234-249.

NOTES

1. Drs. Dennis Klatt and Kenneth Stevens constructed the stimuli. I thank them for their generous help.
2. The third subject's training was terminated early at her request.
3. If the subject did not respond in the allotted 3 sec, the response was considered wrong. This presumably accounts for several points being slightly below chance.
4. After subject A.G.S. performed at 100% correct on the 0-msec VOT anchor for seven consecutive sessions, the anchor was moved to 9-msec VOT and the mate was initialized to 18-msec VOT. The change was made after 13 sessions, leaving 17 training sessions for the new anchor. Therefore, for the 0-msec anchor, the point labeled "Sessions 11-15" is actually the average of Sessions 11-13. Similarly, for the 9-msec anchor, the point labeled "Sessions 16-20" is the average of Sessions 16-17.
5. To compute the predicted discrimination, I used the formula:

$$\begin{aligned} \text{Pred. Discrimin.} = & .5[d_A t_B d_A + t_A d_B t_A + d_A t_B t_B + t_A d_B d_B \\ & + .5(t_A t_B t_A + d_A d_B d_A + t_A t_B d_A + d_A d_B t_A \\ & + t_A t_B t_B + d_A d_B d_B + t_A t_B d_B + d_A d_B t_B)], \end{aligned}$$

where $d_A(d_B)$ is the percentage of time stimulus A (B) was labeled /da/, and similarly for t_A and t_B . The actual order of A and B and the identity of X were determined randomly on line, whereas this formula is exact only when each pair is presented equally often in each order and X is A half of the time. The slight deviation from these assumptions due to the random determination of AB order and X identity should have a minor effect on the predictions.

6. The 2-step test for M.C.B. was conducted several weeks after training terminated. To assure an accurate estimate of her phoneme boundary at that time, an identification test was run, and is presented in Figure 6, rather than the labeling test given immediately after training.

(Received for publication February 18, 1977;
revision accepted June 16, 1977.)