Perception & Psychophysics
1989, 46 (6), 505-512

Effect of speaking rate on the perceptual structure of a phonetic category

JOANNE L. MILLER and LYDIA E. VOLAITIS
Northeastern University, Boston, Massachusetts

When listeners process temporal properties of speech that convey information about the phonetic segments of the language, they do so in a rate-dependent manner. This is seen as a shift in the location of the phonetic category boundary along a temporal continuum toward longer values of the acoustic property in question, as speech is slowed. In a series of experiments, we found that the adjustment for rate is not confined to the region of the category boundary, but extends throughout the phonetic category. Specifically, a change in rate modified the range of stimuli identified as members of a phonetic category, as well as which stimuli were overtly judged to be good exemplars of the category. These findings suggest that the listener's adjustment for speaking rate entails a comprehensive perceptual remapping between acoustic signal and phonetic structure.

Many of the acoustic properties of speech that specify the phonetic structure of an utterance—the sequences of consonants and vowels that define the lexical items of the language—are temporal in nature. When the property in question is short, it specifies one phonetic segment and when it is long, it specifies another.

A potential complication arises, however, in that when speakers talk, they do not maintain a constant rate of speech (e.g., see Goldman-Eisler, 1968). Importantly, the change in rate involves not only a change in the number and duration of pauses within and between utterances, but a change in the acoustic structure of the speech signal itself. For example, Miller, Grosjean, and Lomanto (1984) found that speakers being interviewed on a radio talk show changed speaking rate frequently and substantially while answering questions. For 29 of the 30 speakers tested, the average duration of a syllable (measured over a stretch of pause-free speech) changed during the interview by as much as 100 msec, and for 20 of the 30 speakers, by as much as 300 msec.

The critical issue for perception is how such alteration in speaking rate affects the temporal properties that specify phonetic structure. Take as an example the distinction in English between the voiced stop consonants (/b/, /d/, /g/) and their voiceless counterparts (/p/, /t/, /k/). One of the major properties distinguishing the two classes of consonants is voice-onset time (VOT). VOT is defined articulatorily as the time between the release of the consonant and the onset of vocal fold vibration, and is manifested in the acoustic signal as the time between the initial release burst and the onset of high-amplitude, quasiperiodic energy. It is well established that voiced consonants have shorter VOT values than do voiceless consonants, and that listeners can use this difference to identify a given consonant as voiced or voiceless (e.g., see Lisker & Abramson, 1964, 1970). However, a change in speaking rate can substantially alter the specific VOT values that are produced for a given consonant (e.g., see Summerfield, 1975). Miller, Green, and Reeves (1986) recently examined this effect in some detail for the contrast between /b/ and /p/. Three speakers were asked to produce the syllables /bi/ and /pi/ at a wide range of rates. The duration of the syllables varied between approximately 100 and 700 msec, values encompassing the range of syllable durations typically found in conversational speech. Miller et al. found that as syllable duration increased, there was only a slight increase in the VOT values for /b/, but there was a substantial increase in those for /p/. This pattern of change is readily seen in Figure 1.

To assess the implications of this type of change for perception, Miller et al. (1986) grouped the syllables by syllable duration (in 50-msec bins) and then determined, for each group of syllables, the VOT value that would yield maximal categorization performance if all syllables below that value were labeled /b/ and all those above were labeled /p/. The main finding was that as the syllables systematically became longer, so too did the optimal VOT value that distinguished the contrast. This means that if listeners used VOT to categorize /b/ and /p/ in the most effective manner, they would take account of syllable duration when doing so, systematically altering the location of the perceptual category boundary as syllable duration changed.

This research was supported by NIH Grant NS 14394 and HEW Biomedical Research Support Grant RR 07143. Some of the results were first reported at the 1986 Fall meeting of the Acoustical Society of America, and form the basis of a master's thesis by the second author under the direction of the first author. The software for the speech-processing facility used for stimulus preparation was developed at Northeastern University by Thomas Erb and Ashish Tungare and is based in part on the BLISS system developed by John Mertus at Brown University. We thank Peter D. Eimas for valuable suggestions throughout the course of the research and critical comments on an earlier version of this paper. Requests for reprints may be addressed to Joanne L. Miller, Department of Psychology, Northeastern University, Boston, MA 02115.

Indeed, there is considerable evidence that listeners do perceive speech in such a rate-dependent manner. For example, Summerfield (1981) synthesized /biz/-/piz/ VOT series that differed from each other in overall syllable duration and, in an identification task, asked listeners to label the stimuli from these series as beginning with /b/ or /p/. The main finding was that as the syllables became longer, the perceptual category boundary shifted toward a longer VOT value, just as predicted from the production data. Boundary shifts of this type are widespread; they have been found for a number of different acoustic properties specifying a variety of phonetic contrasts (see Miller, 1981, for a review).

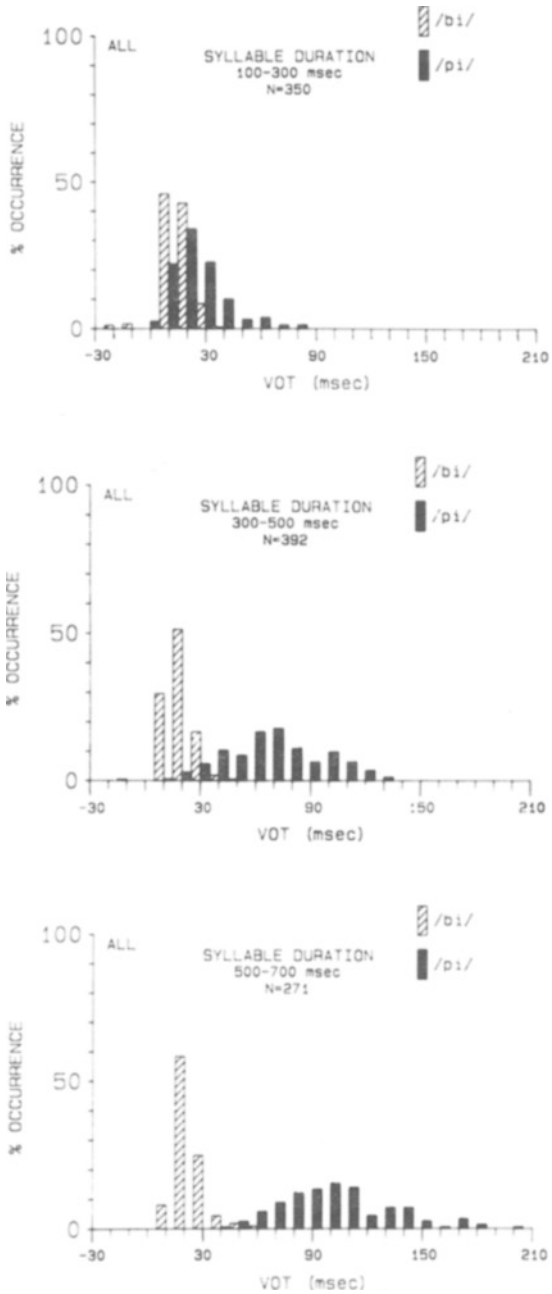

Figure 1. Percentage of /bi/ and /pi/ syllables with VOT values within successive 10-msec intervals, for three different syllable-duration intervals, for all three talkers (ALL) in the Miller, Green, and Reeves (1986) study. The syllables are classified as /bi/ and /pi/ according to the intentions of the talkers. (From "Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast," by J. L. Miller, K. P. Green, and A. Reeves, 1986, *Phonetica*, 43, 106-115. Copyright 1986 by S. Karger AG, Basel. Reprinted by permission.)

To date, the primary focus in the literature has been on the nature of such boundary shifts—the conditions under which they occur, the stage in processing at which they arise, and the underlying mechanism involved. However, a closer look at the way in which acoustic properties such as VOT vary with rate suggests that the listener's adjustment for rate may be more comprehensive than a shift in boundary location. Consider again the Miller et al. (1986) data for /bi/ and /pi/, shown in Figure 1. As we said earlier, VOT—especially for /p/—systematically increased as the syllables became longer. But there was another major change in the /p/ distribution as well. As the syllables became longer, an increasingly wider range of VOT values was produced. If listeners are sensitive to this effect of a change in rate, they should identify stimuli with a wider range of VOT values as members of the /p/ category when the syllables are long, relative to when they are short. In other words, the effect of rate should not be confined to the /b/-/p/ boundary region, but should extend throughout the /p/ category, substantially altering the range of stimuli perceived as members of the category. In Experiment 1, we tested this hypothesis.

EXPERIMENT 1

To assess how the range of stimuli identified as members of the /p/ category changes with speaking rate, we adopted a technique used by Liberman, Harris, Eimas, Lisker, and Bastian (1961), in which extended speech series are presented to listeners for identification. The basic stimuli for our experiment were two speech series that differed from each other in speaking rate, as specified by overall syllable duration. The syllables in one series were 125 msec in duration and in the other series, 325 msec in duration. The stimuli within each series ranged in small increments of VOT from a value appropriate for /bi/, to one appropriate for /pi/, to a very extreme value beyond that appropriate for /pi/. The syllables were randomized and presented to listeners for identification as /bi/, /pi/, or */pi/, where */pi/ was defined as an exaggerated, breathy version of /pi/. We expected to obtain the standard shift at the /b/-/p/ boundary, such that the boundary would be located at a longer VOT value for the 325- than for the 125-msec series. The critical question was whether we would also obtain a shift in the /p/-*/p/ boundary and, if so, whether the magnitude of this shift would be greater than that for the /b/-/p/ boundary, yielding a wider /p/ category for the longer syllables.

Method

Subjects

Ten native English speakers from the Northeastern University community served as subjects in the experiment. All passed a hearing screening test and were paid for their participation.

Stimuli

The preparation of the stimuli involved four steps, as follows.

Step 1. We synthesized a standard, 11-member /bi/-/pi/ series using the cascade version of the Klatt (1980) software synthesizer implemented on a DEC PDP 11/44 computer. Each stimulus was 700 msec in duration and consisted of a 5-msec release burst followed by 5 msec of silence, a segment of formant transition (20 msec in duration for F1, 45 msec in duration for F2, F3, and F4), and a 645 msec steady-state segment. The release burst was centered at 1800 Hz, and the steady-state values of the five formants (in hertz) were 330 (F1), 2200 (F2), 3000 (F3), 3600 (F4), and 3850 (F5). For the endpoint /bi/ stimulus, voicing began at the onset of the transitions—that is, after the 10-msec period of burst-plus-silence—such that the VOT value (time between onset of burst and onset of voicing) was 10 msec. The starting formant frequencies for this stimulus (in hertz) were 180 (F1), 1800 (F2), 2600 (F3), and 3200 (F4); F5 remained constant at 3850 Hz throughout the syllable. The F0 contour rose from 100 Hz at the onset of the transitions to 125 Hz at the end of the 45-msec transition segment, and then fell to 80 Hz over the remainder of the syllable. Beginning with this endpoint /bi/, we incremented VOT in 5-msec steps across the series until the endpoint /pi/ stimulus had a VOT value of 60 msec. The change in VOT was effected by eliminating virtually all energy in the region of F1 through an increase in F1 bandwidth, and by exciting the higher formants with a noise source instead of a periodic source for the appropriate duration. For example, for the second stimulus in the series, which had a 15-msec VOT value, these two modifications were in effect for the first 5 msec of the transitions; this 5 msec, plus the 10 msec for the period of burst-plus-silence, yielded the VOT value of 15 msec. Since the parameter values for the F1 transition were not altered across the stimuli within the series, as voicing onset was delayed, the F1 onset frequency became higher until, at 25-msec VOT, there was no F1 transition. Similarly, since the F0 parameters were also kept constant across the series, as VOT increased, F0 onset frequency also changed.

Step 2. We next extended the series beyond 60 msec VOT by incrementing VOT in 10-msec steps to the extreme value of 320 msec. This yielded 37 stimuli in all, with the first 12 stimuli incremented in 5-msec steps and the last 25 stimuli incremented in 10-msec steps.

Step 3. Using a waveform editing/display program on the PDP 11/44, we cut back the steady-state segments of the original syllables so as to yield two matched series, one with syllables 125 msec long (perceived as having a relatively fast speaking rate) and one with syllables 325 msec long (perceived as having a slower speaking rate). The 125-msec series contained 17 stimuli, with VOT values ranging from 10 to 120 msec, whereas the 325-msec series contained 37 stimuli, with VOT values ranging from 10 to 320 msec.

Step 4. The stimuli were digitally transferred to a DEC LSI 11/23 computer for on-line presentation to the subjects. Sequence protocols were created to control the order and timing of presentation during the various phases of the experiment (see below). During all practice and test trials, the stimuli were presented with an intertrial interval of 2 sec, measured from the subject's response to the onset of the next stimulus. All stimuli were output at a 10-kHz sampling rate and low-pass filtered at 4.8 kHz. The stimuli were presented to the subjects over matched Yamaha YH-1 earphones at a comfortable listening level, which remained constant throughout the experiment.

Procedure

All subjects were tested individually in a sound-attenuated booth and participated in three sessions, conducted on separate days. For half of the subjects, the 125-msec series was tested on Day 1 and the 325-msec series was tested on Days 2 and 3; for the other half, the 325-msec series was tested on Days 1 and 2, with the 125-msec series tested on Day 3.

As indicated above, one of the three sessions was devoted to the 125-msec series. This session consisted of three phases. First, a familiarization sequence was played to the subjects, in which the 17 stimuli within the series were played in order from the shortest to the longest VOT value. The subjects were instructed to listen to the series of stimuli, without making any overt response. They were informed that the stimuli were computer-generated syllables, and that, across the series, the syllable would change from /bi/ to /pi/ to an exaggerated, breathy version of /pi/, which was designated */pi/. All subjects indicated that they heard this progression. A practice phase followed familiarization, during which two randomized blocks of the 17 stimuli (34 stimuli in all) were presented. The subjects were instructed to identify each stimulus as /bi/, /pi/, or */pi/ by pressing an appropriately labeled key on a computer terminal (B, P, *P). They were told that it was essential to respond on every trial. Finally, the test phase of the experiment began, during which 20 randomized blocks of the 17 stimuli (340 in all) were presented for identification, with the same instructions as in the practice phase. A short break was given halfway through the test sequence.

The first of the two sessions devoted to the 325-msec series consisted of three phases. First, a familiarization sequence was presented; this consisted of the 37 stimuli in order from shortest to longest VOT value. This phase was followed by a practice phase, which consisted of one randomized block of the 37 stimuli. Finally, during the test phase, 10 randomized blocks of the 325-msec stimuli (for a total of 370) were presented for identification. During Day 2 of the 325-msec sessions, the practice sequence was again presented, followed by a different set of 10 randomized blocks of test stimuli (370 in all), so that across the 2 days of testing, 20 responses were obtained to each of the 325-msec stimuli. The instructions for the 325-msec series sessions were identical to those for the 125-msec series session and, again, a short break was given halfway through the test sequence.

Results and Discussion

Figure 2 displays the group identification functions for the 125-msec (top panel) and 325-msec (bottom panel) series. It is clear that identification performance was orderly for both series; stimuli with short VOT values were identified predominately as /bi/, those with longer VOT values as /pi/, and those with yet longer VOT values as */pi/. It is also clear that syllable duration had a large effect on the location of the category boundaries: as syllable duration increased, so too did the boundary values.

To assess the statistical significance of these effects, we calculated for each series, for each subject, the location of both the /b/-/p/ and the /p/-*/p/ boundary.[1] The boundary values (in milliseconds), averaged across subjects, were as follows: /b/-/p/ boundary = 35.61 for the 125-msec series and 43.89 for the 325-msec series; /p/-*/p/ boundary = 80.99 for the 125-msec series and 167.95 for the 325-msec series.

The individual boundary scores were submitted to a repeated measures ANOVA, contrast (/b/-/p/ vs.

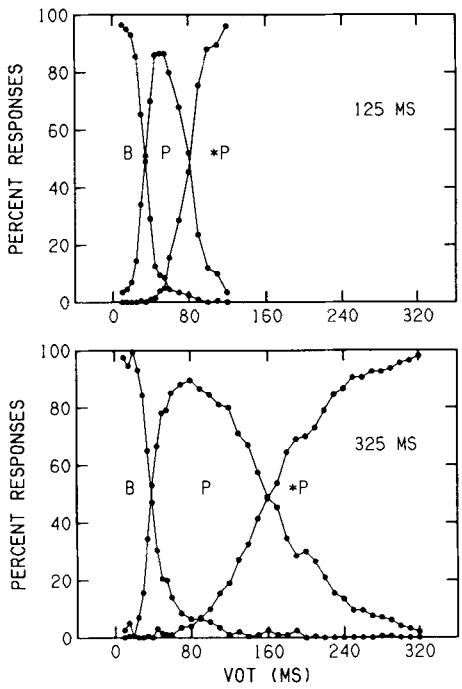

Figure 2. Group three-choice (/b/, /p/, */p/) identification functions in Experiment 1 for extended VOT series with syllable durations of 125 msec (top panel) and 325 msec (bottom panel).

/p/-*/p/) × series (125-msec vs. 325-msec). The main effects of contrast [$F(1,9) = 266.93$] and series [$F(1,9) = 50.36$] were both highly reliable, as was the contrast × series interaction [$F(1,9) = 74.24$], $p < .001$ in each case. The interaction was due to the effect of series being larger for the /p/-*/p/ than the /b/-/p/ contrast, although the effect was reliable for both contrasts [$t(9) = 8.06$, $p < .001$, and $t(9) = 2.11$, $p < .05$, respectively].

Three main findings emerge from this experiment. First, we replicated the shift in /b/-/p/ boundary location with a change in syllable duration: as syllable duration increased, so too did the VOT value that distinguished these phonetic segments. Second, the effect of syllable duration was not confined to the /b/-/p/ boundary region, but extended to the /p/-*/p/ boundary location as well, such that this boundary was also located at a longer VOT value for the longer syllables. Finally, the influence of syllable duration at the two boundary locations was asymmetric, the shift being considerably larger for /p/-*/p/ than for /b/-/p/. The consequence was that as syllable duration increased, there was a substantial increase in the range of VOT values considered to be members of the /p/ category. Taken together, these findings indicate that a change in syllable duration produces a comprehensive alteration in the perceptual mapping of VOT to phonetic category, in line with the changes in VOT distributions that occur as speech is produced at different rates.

EXPERIMENT 2

The rationale for this experiment is based on the notion that phonetic categories have internal perceptual structure. That is to say, some members of a given category are considered better exemplars—more prototypical—than others (e.g., Eimas & Corbit, 1973; Oden & Massaro, 1978). Evidence for such structure comes from a variety of sources, including dichotic listening experiments (e.g., Miller, 1977; Repp, 1977), selective adaptation studies (e.g., Miller, Connine, Schermer, & Kluender, 1983; Samuel, 1982), and duration-judgment tasks (e.g., Klatt & Cooper, 1975).

Informally, internal category structure can be demonstrated by listening to the members of a speech series in order. As the stimulus moves away from the boundary region, it becomes an increasingly better exemplar of the category in question. However, there is a limit to this effect, such that at some point along the continuum, the stimuli begin to sound like poorer members of the category. In other words, all stimuli within a category are not perceived as equivalent, but vary in category goodness.

The three-choice identification data from Experiment 1 provide indirect evidence for such internal category structure for the voicing contrast. This can be seen most readily by focusing on the /p/ category, as shown in Figure 2. It is apparent, for both series, that all stimuli within the /p/ category—defined as the range of stimuli between the /b/-/p/ and /p/-*/p/ boundaries—did not receive an equal percentage of /p/ responses. Rather, moving away from the /b/-/p/ boundary toward longer VOT values, there was first an increase in the percentage of /p/ responses and then a decrease. These data suggest that the stimuli within the category were not perceived as equally good /p/s. A comparison of the 125- and 325-msec identification functions indicates further that the stimuli that were most consistently identified as /p/ varied with syllable duration: for the longer syllables, stimuli with longer VOT values received the highest percentage of /p/ responses. Note too that for the longer syllables, there was a wider range of stimuli that received a high percentage of /p/ responses. These data suggest that phonetic categories have internal structure and, furthermore, that this structure can be influenced by syllable duration.

There is also one earlier finding in the literature suggesting that syllable duration can have an influence on internal category structure. Miller et al. (1983) used a selective adaptation procedure to assess perception of stimuli from a /bae/-/wae/ continuum, where the /b/-/w/ distinction, specified by the duration of the initial formant transitions, is processed in a rate-dependent manner: the /b/-/w/ boundary moves toward a longer transition duration as the syllables become longer (Miller & Liberman, 1979). Their experiment was based on the finding that repeated exposure to a member from a stimulus se-

ries (the adapting stimulus) alters the perception of members of that series, such that the location of the category boundary moves toward the value of the adapting stimulus. Internal category structure is revealed through this technique in that the magnitude of adaptation depends on the adapting stimulus (e.g., Miller et al., 1983; Samuel, 1982). As the adaptor moves away from the category boundary, there is an initial increase in magnitude of adaptation and then a decrease, such that only stimuli from a limited range within the category produce the maximal adaptation. With respect to the issue of speaking rate, Miller et al. (1983) found that as the syllables from the /bae/-/wae/ series were lengthened from 100 to 300 msec, the adapting stimulus that was maximally effective shifted toward a longer transition duration. This indicates that at the level of processing tapped by selective adaptation, internal category structure is determined, in part, by syllable duration.

In the present experiment, we used the 125- and 325-msec VOT series from Experiment 1 to examine whether internal category structure, as directly reflected in overt judgments of category goodness, is dependent on syllable duration, as is suggested by the identification data of Experiment 1. The current experiment consisted of two parts. The first part was a two-choice identification task for the stimuli from the 125- and 325-msec series ranging in VOT from 10 to 60 msec. Its purpose was to replicate the standard /b/-/p/ boundary shift with a change in syllable duration for this group of subjects.

In the second part of the experiment, listeners were asked to judge the goodness of each stimulus as a member of the /p/ category, using a 1-10 scale, with higher numbers designating a better /p/. For this task, the full range of stimuli from both series was included, 10- to 120-msec VOT for the 125-msec series and 10- to 320-msec VOT for the 325-msec series. We expected that for each series, the judgment score would initially increase as VOT increased but, at some point, would begin to decrease as VOT became more extreme and approached values that occur rarely or not at all during production. In other words, we expected an inverted U-shaped function for each series. Furthermore, we expected that the range of stimuli receiving the highest judgment scores would differ for the two functions, mirroring the difference found in the identification data for /p/ in Experiment 1. More specifically, the increase in syllable duration should produce both a shift in the location of the range of good exemplars toward longer VOT values and a widening of this range.

Method

Subjects

Ten different members of the Northeastern University community served as paid subjects. As in the first experiment, all were native speakers of English who passed a hearing screening test.

Stimuli

The stimuli consisted of the same 125- and 325-msec series that were used in the first experiment, again presented on-line from a DEC LSI 11/23 computer. New familiarization, practice, and test protocol sequences were created for use in the present experiment. As before, for all practice and test sequences, there was an intertrial interval of 2 sec, measured from the subject's response to the onset of the next stimulus.

Procedure

Each subject was tested individually in five sessions, conducted on separate days. For all subjects, Day 1 was a preliminary session that consisted of three phases. First, the subjects were familiarized with the stimuli from the two series by presenting to them the 17-member 125-msec series and the 37-member 325-msec series. For each series, the stimuli were presented in order from shortest to longest VOT value. Before each series was presented, the subjects were informed that the stimuli were computer-generated syllables, and that across the series, the syllable would change from /bi/ to /pi/ to an exaggerated, breathy version of /pi/, which we called */pi/. All subjects indicated that they heard this progression for each of the series. The subjects were next given practice on the /bi/-/pi/ identification task. This consisted of identifying five randomized, mixed blocks of the stimuli from the two series with VOT values ranging from 10 to 60 msec, in 5-msec steps (22 stimuli per block). The subjects were instructed to label each stimulus as /bi/ or /pi/ by pressing the appropriately labeled button on the computer terminal (B or P). Finally, the subjects were given practice on the judgment task, first for the 125-msec stimuli and then for the 325-msec stimuli. In each case, five randomizations of the stimuli from the series were presented for judgment. For the 125-msec series, this consisted of the 12 stimuli with VOT values from 10 to 120 msec, in 10-msec steps. For the 325-msec series, this consisted of the 32 stimuli with VOT values from 10 to 320 msec, in 10-msec steps. For each series, the subjects were instructed to judge the goodness of each stimulus as a member of the /p/ category using a scale of 1 to 10, where 10 designated a very good /p/ and 1 designated a very poor /p/. The subjects responded by pressing labeled keys on a computer terminal. They were encouraged to use the full range of the scale and, for each series, to judge the goodness of the stimulus in relation to the other stimuli within that series.

Each of the next four sessions consisted of two phases: identification and judgment. For the identification task, the subjects labeled the stimuli from the 125- and 325-msec series with VOT values of 10 to 60 msec (in 5-msec steps) mixed together. Five randomizations of these 22 stimuli were identified each day, such that over the 4 days of testing, 20 identification responses were obtained for each stimulus. For the judgment task, the 125- and 325-msec stimuli were tested separately. Half of the subjects were tested on the 125-msec stimuli on the first 2 days of testing and the 325-msec stimuli on the last 2 days; the other half were tested in the reverse order. On each of the 125-msec days, the subjects judged 10 randomizations of the stimuli with VOT values from 10 to 120 msec (in 10-msec steps), for a total of 120 trials. On the 325-msec days, they judged 10 randomizations of the stimuli with VOT values from 10 to 320 msec (in 10-msec steps), for a total of 320 trials (two short breaks were given during this test sequence). Thus, a total of 20 judgment responses was obtained for each stimulus.

Results and Discussion

From the identification data, individual /b/-/p/ boundary locations were calculated for each subject, for each series. The expected shift in boundary value with series was obtained, with the mean boundary value for the 125-msec series located at 31.14 msec, and that for the 325-msec series located at 33.58 msec. This difference was statistically reliable [$t(9) = 2.45$, $p < .025$].[2]

The group judgment data for both the 125- and the 325-msec series are displayed in Figure 3. Both functions are

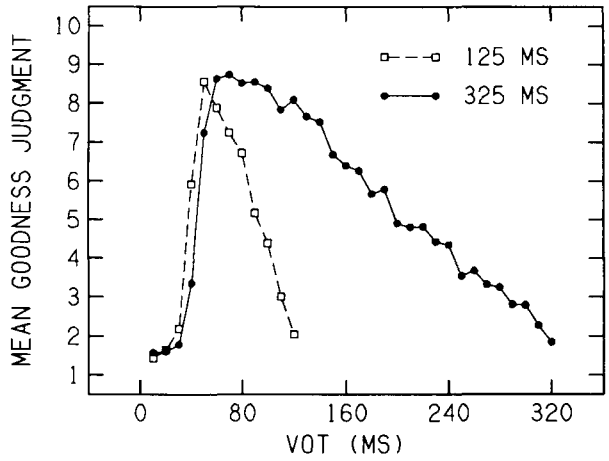

Figure 3. Mean category-goodness judgments in Experiment 2 for the /p/ category for extended VOT series with syllable durations of 125 and 325 msec.

orderly, with the judgment value first increasing with an increase in VOT, and then decreasing. Moreover, as predicted, the 325-msec function is considerably broader than the 125-msec function, and it is displaced toward longer VOT values.

To quantify this change, we determined for each subject, for each series, the location of the range of good exemplars of the /p/ category. This was accomplished as follows. We located the stimulus within the series that received the highest judgment score, and defined the range of good exemplars as the stimuli surrounding this VOT value that corresponded to a judgment score of at least 90% of the maximum. (When the calculated "90% of maximum score" fell between the obtained judgment scores for adjacent VOT values, we used linear interpolation to determine the VOT value that would correspond to the calculated score.) Averaged across subjects, the good exemplars for the 125-msec series ranged from a lower limit of 44.89 msec to an upper limit of 64.38 msec VOT (a span of 19.49 msec); those for the 325-msec series ranged from 54.93 to 114.19 msec (a span of 59.26 msec).

A repeated measures ANOVA on the individual data was performed, limit (lower vs. upper) × series (125-msec vs. 325-msec). The main effects of limit [$F(1,9) = 160.52$] and series [$F(1,9) = 40.90$] were both highly reliable, as was the limit × series interaction [$F(1,9) = 73.49$], $p < .001$ in each case. The interaction arose because the effect of series was larger for the upper than for the lower limit, although for both limits, the effect was reliable [$t(9) = 7.57$, $p < .001$, and $t(9) = 3.00$, $p < .01$, respectively]. Thus, as syllable duration increased, there was a substantial change in the VOT values that were judged to be good exemplars of the /p/ category: the range of good exemplars both moved toward longer VOT values and became wider, in accord with the changes that occur during speech production.[3]

Before discussing the implication of these findings, one possible methodological objection should be considered. We have interpreted the change in judgment function across series to be due to the change in syllable duration. However, the change in series not only involved syllable duration, but also the range of exemplars presented for judgment. For the 125-msec series, the stimuli ranged from 10 to 120 msec VOT, whereas for the 325-msec series, they ranged from 10 to 320 msec VOT. Inasmuch as stimulus range is known to influence perception, it is conceivable that it was the change in range, and not syllable duration, that produced the displacement of the two judgment functions (see Brady & Darwin, 1978). This possibility was tested in Experiment 3.

EXPERIMENT 3

Method

The stimuli for this experiment were identical to those used in Experiment 2, except that only the syllables from the 325-msec series that ranged in VOT value from 10 to 120 msec were included in the stimulus set used for the judgment task. In this way, the two series were equated for range of VOT values, while still differing in syllable duration.

A new group of 10 listeners was tested on the 125-msec and the modified 325-msec series, using the same instructions and procedures as in the previous experiment. That is, each subject participated in a preliminary session and four test sessions and, within each test session, performed both an identification task and a judgment task. As before, the stimuli for the identification task were the 125- and 325-msec stimuli that varied in VOT from 10 to 60 msec, in 5-msec steps. The stimuli for the judgment task were the syllables within each series that ranged from 10 to 120 msec VOT, in 10-msec steps. At no time during the experiment were the subjects exposed to the 325-msec syllables with VOT values beyond 120 msec.

Results

As expected, the data from the identification task were orderly, yielding a reliable shift in boundary value from 30.14 to 33.50 msec VOT [$t(9) = 5.31$, $p < .001$]. The critical judgment data are shown in Figure 4, and indi-

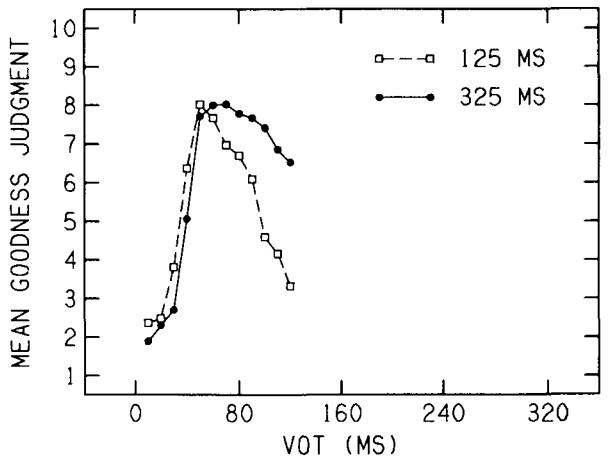

Figure 4. Mean category-goodness judgments in Experiment 3 for the /p/ category for modified, extended VOT series with syllable durations of 125 and 325 msec.

cate that equating stimulus range did not eliminate the effect of rate. As before, an increase in syllable duration from 125 to 325 msec produced both a widening of the judgment function and a displacement of the optimal stimuli toward longer VOT values.

Averaged across subjects, the good exemplars (defined as in Experiment 2) for the 125-msec series ranged between 43.36 and 69.92 msec VOT (a span of 26.56 msec VOT). It was not possible to obtain a valid measure of the range of good exemplars for the 325-msec series because for 2 of the 10 subjects, the upper limit of the good-exemplar range fell beyond the longest VOT value (120 msec) in the stimulus set. Nonetheless, we computed a conservative estimate of the good-exemplar range by arbitrarily assigning an upper limit for these 2 subjects of 120 msec VOT. Averaged across subjects, the good exemplars for the 325-msec series ranged between 50.37 and 93.13 msec (a span of 42.76 msec VOT).[4]

A repeated measures ANOVA on the individual data, limit (lower vs. upper) × series (125-msec vs. 325-msec), revealed significant main effects of limit [$F(1,9) = 62.85$, $p < .001$] and series [$F(1,9) = 24.08$, $p < .001$], as well as a limit × series interaction [$F(1,9) = 9.42$, $p < .025$]. The basis for the interaction was the larger effect of series for the upper, relative to the lower, limit, although the effect for both limits was reliable [$t(9) > 4.22$, $p < .005$ in each case].

GENERAL DISCUSSION

The present experiments provide clear evidence that a change in speaking rate, as reflected in overall syllable duration, produces a comprehensive alteration in the perceptual mapping between acoustic structure and phonetic category. This remapping entails a change in the range of stimuli identified as belonging to a given category and a change in the specific stimuli overtly judged to be good exemplars of the category. These perceptual effects mirror the systematic acoustic variation that occurs as rate is altered during speech production, and thus enable the listener to accommodate for contextual variation due to rate during speech perception.

The mechanism underlying this perceptual remapping is currently not known. One possibility is that phonetic perception is accomplished by an articulatory-based, specialized speech-processing system (Liberman & Mattingly, 1985), and that the adjustment for rate is part of this system (Miller & Liberman, 1979). By virtue of being based on articulatory principles, the processing system has information about the acoustic consequences of producing speech at different rates, and operates in terms of this information during perception. In this way, the mapping between acoustic structure and phonetic category is tightly coupled in production and perception. The finding of a close correspondence between the manner in which a change in speaking rate affects the VOT distribution for /p/, and the manner in which it alters the perceptual structure of the /p/ category, is certainly in accord with this view.

An alternative possibility is that the perceptual adjustment arises at an auditory level of processing. In other words, there is no adjustment for articulatory rate, per se. Instead, general perceptual effects account for the shifts in boundary location that occur as syllables become longer. Support for this view comes from experiments in which the standard category boundary shifts for speech series as a function of syllable duration are also found for nonspeech analogues of these series (e.g., Diehl & Walsh, 1989; Pastore, Harris, & Kaplan, 1982; Pisoni, Carrell, & Gans, 1983). It will be of considerable interest to determine whether the type of comprehensive acoustic-to-phonetic remapping that we have found in the current experiments is also found for nonspeech analogues of our stimuli.

Other evidence relevant to the issue of underlying mechanism comes from selective adaptation experiments. As we noted earlier, these experiments reveal internal category structure and, moreover, reveal that such structure can be affected by a change in syllable duration (Miller et al., 1983). The critical question with respect to the mechanism issue is the level of processing tapped by selective adaptation. Although the issue is far from resolved, current evidence suggests that adaptation may tap a prephonetic, auditory level of processing (see Miller et al., 1983, for a discussion of this claim). If this is so, then the adaptation data provide evidence that internal category structure, and perhaps the effect of syllable duration on this structure, might have an auditory, rather than a speech-specific basis. However, it is important to keep in mind that if an explanation of the rate effect in terms of general auditory processing is to be viable, it must account for the close correspondence between the effects of speaking rate in perception and the pattern of change effected by an alteration in rate during production. In other words, what must be explained is how the listener alters the acoustic-to-phonetic mapping during perception in just the way required to accommodate changes in production—and at the current time, it is not obvious how to do this without recourse to constraints imposed by production.

Finally, we consider the ontogeny of the type of comprehensive rate-dependent speech processing our data reveal for adult listeners. It has been known for some time that young infants process speech in a remarkably sophisticated manner, categorizing stimuli along speech continua in much the same way that adults do (see Aslin, Pisoni, & Jusczyk, 1983, for a review). Two findings are of particular interest. First, Grieser and Kuhl (1989) have recently reported that infants do not respond equivalently to all members of a vowel category, but rather respond as if some members are better category exemplars than others. This suggests that the perceptual categories of infancy that serve as precursors for adult phonetic categories may themselves have internal structure. Second, it has been found that infants, like adults, are sensitive to contextual variation produced by a change in speaking rate, in that the location of the infant's category boundary shifts with a change in syllable duration (Eimas & Miller, 1980).

A critical question now becomes whether the infant's adjustment for rate goes beyond this shift in boundary location; that is to say, whether the comprehensive remapping between acoustic signal and phonetic category in adults finds its roots in the early perceptual processing of infancy.

REFERENCES

ASLIN, R. N., PISONI, D. B., & JUSCZYK, P. W. (1983). Auditory development and speech perception in infancy. In M. M. Haith & J. J. Campos (Eds.), *Carmichael's manual of child psychology: Vol. 2. Infancy and the biology of development* (4th ed., pp. 573-687). New York: Wiley.

BRADY, S. A., & DARWIN, C. J. (1978). Range effect in the perception of voicing. *Journal of the Acoustical Society of America*, **63**, 1556-1558.

DIEHL, R. L., & WALSH, M. A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *Journal of the Acoustical Society of America*, **85**, 2154-2164.

EIMAS, P. D., & CORBIT, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, **4**, 99-109.

EIMAS, P. D., & MILLER, J. L. (1980). Contextual effects in infant speech perception. *Science*, **209**, 1140-1141.

GOLDMAN-EISLER, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. New York: Academic Press.

GRIESER, D. L., & KUHL, P. K. (1989). The categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology*, **25**, 577-588.

KLATT, D. H. (1980). Software for a cascade-parallel formant synthesizer. *Journal of the Acoustical Society of America*, **67**, 971-995.

KLATT, D. H., & COOPER, W. E. (1975). Perception of segment duration in sentence contexts. In A. Cohen & S. Nooteboom (Eds.), *Structure and process in speech perception* (pp. 69-89). New York: Springer-Verlag.

LIBERMAN, A. M., HARRIS, K. S., EIMAS, P., LISKER, L., & BASTIAN, J. (1961). An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. *Language & Speech*, **4**, 175-195.

LIBERMAN, A. M., & MATTINGLY, I. G. (1985). The motor theory of speech perception revised. *Cognition*, **21**, 1-36.

LISKER, L., & ABRAMSON, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, **20**, 384-422.

LISKER, L., & ABRAMSON, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics. *Proceedings of the Sixth International Congress of Phonetic Sciences* (pp. 563-567). Prague: Academia.

MILLER, J. L. (1977). Properties of feature detectors for VOT: The voiceless channel of analysis. *Journal of the Acoustical Society of America*, **62**, 641-648.

MILLER, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39-74). Hillsdale, NJ: Erlbaum.

MILLER, J. L., CONNINE, C. M., SCHERMER, T. M., & KLUENDER, K. R. (1983). A possible auditory basis for internal structure of phonetic categories. *Journal of the Acoustical Society of America*, **73**, 2124-2133.

MILLER, J. L., GREEN, K. P., & REEVES, A. (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica*, **43**, 106-115.

MILLER, J. L., GROSJEAN, F., & LOMANTO, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, **41**, 215-225.

MILLER, J. L., & LIBERMAN, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, **25**, 457-465.

ODEN, G. C., & MASSARO, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, **85**, 172-191.

PASTORE, R. E., HARRIS, L. B., & KAPLAN, J. K. (1982). Temporal order identification: Some parameter dependencies. *Journal of the Acoustical Society of America*, **71**, 430-436.

PISONI, D. B., CARRELL, T. D., & GANS, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception & Psychophysics*, **34**, 314-322.

REPP, B. H. (1977). Dichotic competition of speech sounds: The role of acoustic stimulus structure. *Journal of Experimental Psychology: Human Perception & Performance*, **3**, 37-50.

SAMUEL, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics*, **31**, 307-314.

SUMMERFIELD, A. Q. (1975). *Aerodynamics versus mechanics in the control of voicing onset in consonant-vowel syllables*. (Speech Perception Report No. 4). Belfast, Northern Ireland: Department of Psychology, Queen's University of Belfast.

SUMMERFIELD, [A.] Q. (1981). On articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 1074-1095.

NOTES

1. The individual boundary locations reported in this paper were determined by fitting a regression line to the data in the boundary region and taking the boundary to be the VOT value equivalent to 50% responses. For Experiment 1, the regression analysis was based on the percentage of B responses for the /b/–/p/ boundary and the percentage of *P responses for the /p/–*/p/ boundary. For Experiments 2 and 3, the analysis was based on the percentage of B responses.

2. It should be noted that for both the 125- and the 325-msec series, the /b/–/p/ boundary was located at a longer VOT value in Experiment 1 than in Experiments 2 and 3. This was most likely due to the substantially different ranges of stimuli presented for identification in the respective experiments (see Brady & Darwin, 1978).

3. An even stronger prediction about the relation between perception and production is that for any given syllable duration, the VOT values of the stimuli judged to be the best exemplars will correspond precisely to the modal production values (cf. Eimas & Corbit, 1973). Unfortunately, only limited production data are currently available, and the existing data indicate variability in the precise location of VOT distributions across subjects (see Miller et al., 1986). Thus, a meaningful test of this prediction is currently not possible.

4. Because a valid measure of the good-exemplar range for the 325-msec series could not be computed in this experiment for all subjects, a direct comparison of ranges across Experiments 2 and 3 is not meaningful. As a consequence, it is not possible to determine whether any part of the effect found in Experiment 2 was due to different ranges of VOT values. What is most important, however, is the finding that a change in syllable duration, by itself, can alter the stimuli within a series that are considered to be good exemplars of the category.

(Manuscript received February 6, 1989;
revision accepted for publication June 30, 1989.)