Perception & Psychophysics
2000, 62 (2), 285-300

Perceptual normalization for speaking rate II: Effects of signal discontinuities

JAMES R. SAWUSCH and ROCHELLE S. NEWMAN
State University of New York, Buffalo, New York

In a series of experiments, we examined how rate normalization in speech perception is influenced by segments that occur after the target. Perception of the syllable-initial target was influenced by the durations of both the adjacent vowel and the segment after the vowel, even when the identity of the talker was changed during the syllable. These results, together with earlier findings of a temporal window that follows a target phoneme within which segment duration influences perception of the target, help to resolve apparently conflicting results that have been reported previously. Overall, the results fit within a theoretical framework in which the rate at which events take place is extracted early in processing, prior to segregating voices, and the use of this information is obligatory in subsequent processing.

The process of perceiving speech involves achieving perceptual constancy in the face of substantial stimulus variation. In order to recognize a word, the listener must deal with variation between talkers and variation that occurs within the speech of a single talker. Between-talker differences include the length of the vocal tract, vocal fold mass and tension, articulatory habits, intrinsic speaking rate, and dialect. Within-talker effects include the emotional tone of the speech signal, the context in which segments occur, and the rate of speech.

Our focus here is on how listeners achieve constancy in the face of variation in the rate of speech. The rate of speech varies both across talkers (Byrd, 1992; Crystal & House, 1988c) and within the speech of a single talker (Miller, Grosjean, & Lomanto, 1984). Furthermore, the durations of speech segments change with speaking rate (Crystal & House, 1982, 1990; or see Miller, 1981, for a review of earlier work). Changes in speaking rate are important because the perception of some segmental distinctions can be altered when their durations are altered. For instance, differences in duration are sufficient cues for certain vowel contrasts such as /i/–/ɪ/ in English (see Ainsworth, 1972). Similarly, the /b/–/w/ manner contrast can be cued by differences in duration alone. Shorter formant transitions at syllable onset are generally identified as more "b-like" and longer transitions as more "w-like" (Miller & Liberman, 1979). Nevertheless, in rapid speech one still perceives the brief transitions of /w/ as /w/ and in slow speech one hears the long transitions of /b/ as /b/. The key issue appears to be how duration information is processed. The perception of duration, as a cue to phonetic identity, is relative to the rate of speech (see Miller, 1981, for a review). Put another way, listeners appear to normalize for the speaking rate during perception.

In the course of research on the influence of speaking rate on speech perception, a number of issues have been addressed. One question has been whether the effects that have been observed are somehow related to a speech mode of processing (see Green, Stevens, & Kuhl, 1994; Miller & Liberman, 1979). Subsequent studies have demonstrated parallel "rate normalization" effects in nonspeech processing (Diehl & Walsh, 1989; Pisoni, Carrell, & Gans, 1983). However, this issue has been extremely difficult to resolve because parallel results in two domains indicate only that a common mechanism of processing is possible (see also Fowler, 1990). Another issue has been the extent to which these effects occur in the perception of fluent natural speech and whether they are simply a result of using highly impoverished, reduced stimuli (see Miller & Wayland, 1993; Shinn, Blumstein, & Jongman, 1985). The relative automaticity of this effect and the locus in perceptual processing of normalization for the rate of events have also been addressed (see Lotto, Kluender, & Green, 1996; Miller & Dexter, 1988), as has the influence of speaking rate on the internal structure of phonetic categories (Flege & Schmidt, 1995; Miller & Volaitis, 1989), and whether the effects are based on phonetic segment durations or syllable durations (see Miller & Liberman, 1979).

More recently, studies have examined which types of changes in the acoustic signal mark the boundaries of speech segments (or break the acoustic signal into multiple perceptual streams or groups) and how they alter the influence of segment duration on perception (Green et al., 1994; Lotto et al., 1996). Finally, the relative roles of segments that occur immediately around the target and seg-

This research was supported by NIDCD Grant R01-DC00219 given to SUNY at Buffalo. The authors thank Aaron Armstrong, Douglas Brenner, Page Chapman, Randy Bennis, and Steve Piatek for their assistance with data collection. We also thank Andrew Lotto and two anonymous reviewers for their comments on an earlier draft of this research. R.S.N. is now at the Department of Psychology, University of Iowa. Comments may be sent to J. R. Sawusch, Department of Psychology, Park Hall, State University of New York, Buffalo, NY 14260 (e-mail: jsawusch@acsu.buffalo.edu).

—Accepted by previous editor, Myron L. Braunstein

ments at a greater distance from the target (both before and after) have been addressed (Kidd, 1989; Miller & Liberman, 1979; Newman & Sawusch, 1996; Summerfield, 1981). The last two topics—that is, the proximity of information that indicates speaking rate for the target segment and how changes in the acoustic composition of the waveform alter normalization for speaking rate—are the focus of the present research.[1]

When the information on speaking rate precedes the target, there appear to be two components to this influence. One component is set by the rate of speech for the phrase prior to the target. The other is based on the segment (or segments) that occur immediately prior to the target (see Kidd, 1989; Summerfield, 1981; or Miller, 1981, for a review). The duration of a segment that immediately follows the target can also influence listeners' perceptions of a duration-based distinction. Miller and Liberman (1979) had listeners identify tokens that ranged from /ba/ to /wa/, with the phonetic distinction carried by the duration of the initial formant transitions. Across different series, the syllable duration (a change in vowel duration) was varied. The category boundary between /b/ and /w/ changed when the vowel duration that followed varied. When a longer vowel duration occurred, the category boundary was displaced toward the /w/ end of the series relative to the series that was followed by a shorter vowel duration. That is, for a slow speaking rate (long vowel), listeners reported more /b/ responses and longer duration transitions needed to be present to cue the percept of /w/. Put another way, phonetic contrasts that incorporate duration cues seem to be perceptually normalized, or scaled, to the rate of speech in the local environment.

Two additional results reported by Miller and Liberman (1979) are also relevant. First, in addition to manipulating the vowel duration, they also manipulated the syllable duration while holding the vowel duration constant. They accomplished this by adding a /d/ to the end of their /ba/–/wa/ series. In this case, the short-vowel /bad/–/wad/ series had the same vowel duration as did the short-vowel /ba/–/wa/ series. However, the overall syllable duration of the short-vowel /bad/–/wad/ series was the same as that of a longer vowel /ba/–/wa/ series. The listeners responded to the short-vowel /bad/–/wad/ series as if it had been spoken at a rapid rate. As for the short-vowel /ba/–/wa/ series, fewer /b/ responses and more /w/ responses were reported. Thus, it appears that syllable duration did not influence the listeners' perceptions. Rather, the duration of the individual phonetic segments appears to have been the dominant factor influencing listeners' responses.

The second result of Miller and Liberman (1979) was that when a second syllable (/da/) was added to the /ba/–/wa/ series and the duration of this second syllable was varied, the duration of the second syllable also influenced listeners' identifications of the initial consonant as /b/ or /w/. Thus, it is not just the segment immediately following the target that can influence perception of the target.

Newman and Sawusch (1996) proposed a set of possible principles that might govern how information that follows the target influences the perception of the target. In a series of experiments, they examined whether adjacency, similarity, phonetic identity, or phonotactics could influence rate normalization. Their results revealed a very consistent pattern. The phonetic identity, phonotactics, and acoustic similarity of the subsequent information had no influence on the rate normalization effects of subsequent segments on the target. If the temporal distance between the target and the segment duration (speaking rate) information that follows was short, the perception of the target was affected. For example, in a /dlos/–/tlos/ series, both a long /l/ and a long /o/ resulted in more /d/ responses and a later category boundary than did the series with a short /l/ or a short /o/. Thus, in this example, the fact that the /dl/ and /tl/ phoneme sequences do not occur in English did not appear to have precluded an influence of either /l/ or /o/ duration on the perception of the /d/–/t/ distinction.

Conversely, when the temporal distance between the initial consonant target and a subsequent vowel or consonant was long, the variation in duration of the subsequent phoneme had no effect on the listeners' responses. Newman and Sawusch (1996) proposed that processing of the target took place over a limited *temporal window*. Subsequent segments that occurred while target processing was taking place could have influenced the processing of the target. This is similar to the proposal of Miller and Dexter (1988), who suggested that the processing of rate information is obligatory and that any such processing that takes place before processing of the target is finished will influence the perception of the target. A summary of the idea of a temporal window and how it can account for the earlier data can be found in Newman and Sawusch.

Before one accepts the idea of a temporal window, some further issues need to be resolved. First, Summerfield (1981) manipulated the duration of both the adjacent /i/ vowel and the nonadjacent, final /z/ in a /biz/–/piz/ ("bees"–"peas") series. The duration of the adjacent vowel had the usual effect: For the long-vowel series, listeners reported more /b/s and fewer /p/s relative to what they reported for the short-vowel series. However, the variation in the final /z/ had no effect. Although null effects must be interpreted with caution, this result appears to be different from the expectation that is based on the data of Newman and Sawusch (1996) and the data of Miller and Liberman (1979). Some reconciliation of Summerfield's data with the temporal window proposal and the data of Miller and Liberman is needed.

One possibility, suggested by Newman and Sawusch (1996), was that the variation in the duration of the final /z/ occurred outside the temporal window within which information about rate is extracted. In Summerfield's (1981) stimuli, the syllables with the short /z/ were 205–255 msec in duration, whereas the long /z/ syllables were 255–305 msec in duration. Newman and Sawusch pro-

posed that the temporal window for rate information extended to about 250 msec from the onset of the target. On the basis of this estimate, most of the variation in /z/ duration in Summerfield's stimuli would lie outside the temporal window and would not have influenced listeners' perceptions of the syllable-initial voicing contrast. In addition to this explanation, there are others. Of particular interest is Summerfield's report that his long /z/ actually sounded like /s/. He raised the possibility that listeners interpreted the variation in final consonant duration as cuing a phonetic contrast between /z/ and /s/ rather than as a variation in speaking rate. This is an interesting proposal that has not yet been tested. However, the focus of our experiments is on the first alternative, proposed by Newman and Sawusch, that segment information within a short temporal window of a phoneme target can influence the perception of the target.

A second issue that must be addressed is related to the first. None of the experiments that have shown an influence of variation in the duration of a segment that follows the target have shown an influence of a segment that was nonadjacent and not a vowel. That is, the only nonadjacent, subsequent segments that have been found to influence the target have been vowels. Newman and Sawusch (1996) have argued that earlier null results are due to the fact that none of these nonadjacent segments were within the temporal window that influences rate normalization. Testing this explanation requires that we examine whether segment durations that follow the vowel can alter the perception of a target that precedes the vowel. Two of the experiments reported here were designed to be a direct test of this.

Finally, the results of Green et al. (1994) and Lotto et al. (1996) raised an interesting question about the process of normalizing for speaking rate. In their experiments, a synthetic /b/–/p/ series was combined with short and long following vowels to investigate the influence of vowel duration (a cue to speaking rate) on classification of the initial phoneme. In both studies, the expected effect was found; more /p/ responses were reported for the shorter vowel and more /b/ responses were reported for the longer vowel. That is, for a longer vowel, a longer voice onset time (VOT, a sufficient cue to the /b/–/p/ contrast; see Lisker & Abramson, 1964) was necessary to cue /p/. The more interesting results occurred when the stimuli were altered part way through the vowel. In Green et al., either a sudden change in $F0$ or in the formant frequencies was introduced into the long-vowel stimuli after a brief interval. These changes were similar to what would occur if the identity of the person who was speaking was changed part way through the vowel (e.g., from female to male). In Lotto et al., similar changes were made. In both cases, the spectral/$F0$ discontinuity was introduced in such a way that the vowel duration before the discontinuity was the same as that of the vowel duration in the short-vowel series. The question was whether the long vowel with a discontinuity would behave like a single long vowel or like a short vowel. The general pattern of results was that the long, discontinuous vowel behaved like a short vowel. That is, in the listeners' data for the short-vowel series, the /b/–/p/ category boundary occurred at a relatively short VOT. In the discontinuous, long-vowel series, the category boundary was at a similar, short VOT. In contrast, the category boundary occurred at a longer VOT for the continuous long-vowel series.

These data indicate that a spectral discontinuity, such as occurs with a change in the talker, triggers a segmentation process. Consequently, the vowel portion that occurs before the discontinuity, which is short, has approximately the same effect as a short vowel that has no spectral discontinuity. Lotto et al. (1996) argued that only speaking-rate information from the same perceptual group is used to normalize for speaking rate. In effect, they proposed that the acoustic signal was segregated into streams, or perceptual groups, according to source characteristics or to other principles of grouping before the process of rate normalization occurs. An alternative interpretation would be that the change in source characteristics does, indeed, trigger segmentation. However, the information in the subsequent part of the signal can still contribute to speaking-rate normalization. In this case, the syllable with the spectral discontinuity has three segments in the same, short span of time as does the long, continuous vowel syllable. This indicates a faster speaking rate, so that the listeners' responses to the discontinuous vowel condition would be similar to the short, continuous vowel condition.

Lotto et al. (1996) addressed this alternative interpretation by including one additional condition in which the formant frequencies were changed abruptly as a means of introducing a spectral discontinuity. The change was "relatively" small, but large enough to induce some change in the perceived vowel quality (see Lotto et al. for details). The net effect of this manipulation was the same as that for a continuous vowel of the same duration. Thus, even though this series contained a spectral change sufficient to cue a change in phonetic identity and led to the perception of three segments on nearly 50% of the trials, listeners treated it as if it was equivalent to a long-vowel stimulus that had two segments. Consequently, these data suggest that it is not the number of phonetic segments, per se, that is critical to the perceived speaking rate. Rather, it may be that the relatively small change in the formant frequencies was not sufficient to trigger the stimulus-driven segmentation process. Only when the changes in the spectrum or source qualities are appropriate to multiple events will the segmentation process (perceptual grouping) be triggered. As an explanatory construct, Lotto et al. described their results as being consistent with the result of a rate-normalization process that occurs after a perceptual grouping process (cf. Bregman, 1990). One of the roles of such a perceptual grouping is to segregate speech that comes from different talkers. A spectral or source change in the speech signal that is sufficient to cue a different talker will also trigger the grouping and segmentation process.

Newman and Sawusch (1996) did not consider a case in which the source characteristics or the talker was changed in the middle of the syllable, but their basic proposal was that all segments in the processing window of the target would influence perception of the target. Talker discontinuity may be processed in essentially the same fashion as any other spectral discontinuity, including those that can occur between phonetic segments in the speech of a single talker. If so, the part of the vowel before the talker change would act just like a vowel of comparable duration in a speech signal spoken by a single talker. In addition, the duration of the vowel (or other signal segment) after the talker change (spectral discontinuity) would also influence perception. The important distinction here is that perceptual grouping or stream segregation (Bregman, 1990) may be only partially complete before (or concurrent with) the extraction of segment durations that contribute to event-rate normalization. In this case, the segment durations occurring subsequently to a spectral discontinuity should influence the perception of the initial target, even though listeners "hear" the speech signal as representing two talkers. Two of our experiments were a direct test of this.

The focus of the experiments that follow is on the influence of vowel and postvowel segment duration on the perception of a preceding target when the subsequent segment is within the temporal window proposed by Newman and Sawusch (1996). To the extent that the processing of segment duration and its use in normalizing perception for the rate of events is an early and relatively automatic process (cf. Miller & Dexter, 1988), we expected that all posttarget segments that were in close temporal proximity to the target would influence the perception of the target. Furthermore, this might even occur when an abrupt spectral discontinuity, such as when the talker changes, occurred between the target and the subsequent segment.

EXPERIMENT 1

The purpose of this experiment was to determine whether a distal (nonadjacent) consonant would have the same type of effect on the perception of a syllable-initial duration-based contrast as an adjacent vowel does. In earlier studies, Newman and Sawusch (1996) and Summerfield (1981) found no evidence that a postvocalic consonant could influence the perception of an initial phoneme distinction. However, as Newman and Sawusch pointed out, these earlier results must be interpreted with caution. There appears to be a temporal window around the target phoneme. The duration of segments that fall within this temporal window can influence the perception of the target. Once beyond this temporal window, succeeding segments do not influence the perception of the target. If the temporal window is assumed to be about 250 msec in duration (see Table 8 in Newman & Sawusch), the nonadjacent, postvocalic consonant in Newman and Sawusch (Experiment 1) would have been outside the temporal window. Similarly, in Summerfield's study, the variation in consonant duration (the difference between a short final /z/ and a long final /z/) may have occurred largely outside the temporal window. Any influence of the duration variation in the /z/ consonant may have been too small to observe in the experiment.

In Experiment 1, four series, which varied from /bʌlz/ to /pʌlz/, were created by digital waveform editing. In this set, two series differed in the duration of the adjacent /ʌ/ vowel and two series differed in the duration of the nonadjacent, or distal, /l/ consonant. For the two series that differed in vowel duration (short vs. long), the nonadjacent /l/ duration remained constant. Conversely, for the pair of series that differed in /l/ duration, the vowel duration remained constant. Syllables with the vowel /ʌ/ were chosen because it is intrinsically short in duration (see Crystal & House, 1988a). Consequently, it was possible to create series in which the variation in distal consonant (/l/) duration would fall within the temporal window that follows the target identified by Newman and Sawusch (1996).

If Newman and Sawusch (1996) are correct and any segment that occurs within the temporal processing window after the target can influence perception of the target, then both the variation in the adjacent vowel duration and the variation in the distal consonant duration should influence perception of the /b/–/p/ distinction. Specifically, a short following segment (/ʌ/ or /l/) should cue a rapid rate of speech. This should cause the voicing boundary for the /b/–/p/ distinction to occur at a shorter VOT, closer to the /b/ end of the series. In other words, more of the stimuli in this series should be labeled /p/. Conversely, long /ʌ/ and /l/ segment durations should lead to a /b/–/p/ category boundary at longer VOTs and lead to more /b/ responses. If this pattern of results is not found, other factors besides a temporal window around the target will be implicated in rate normalization.

Method

Participants. The listeners were 19 students from an introductory psychology course at the State University of New York at Buffalo who participated in the experiment for class credit. All listeners were native speakers of English and reported no history of a speech or hearing impairment. The data of 1 listener were omitted from the analysis. This individual fell asleep during the experiment and did not respond on a substantial portion of the trials. This left a total of 18 participants to identify the /bʌlz/–/pʌlz/ stimuli.

Stimuli. A female native speaker (R.S.N.) of midwestern American English recorded the syllables /bʌlz/ and /pʌlz/ in the context of fluent speech. The stimuli were amplified, low-pass filtered at 9.4 kHz, digitized via a 16-bit, analog-to-digital converter at a 20-kHz sampling rate, and stored on computer disk. The syllables were excised from the carrier sentence, "Norton said _____ to me." Spectral analysis of the /bʌlz/ syllable was used to determine the approximate boundaries between adjacent phonetic segments. The release burst and the first 14 vocal pulses (approximately 60 msec) were considered to be the initial /b/. The next 12 pulses (60 msec) were the vowel /ʌ/, and the 10 pulses (47 msec) following that were the liquid consonant /l/. The boundary between the /ʌ/ vowel and the /l/ consonant was placed at the start of the upward movement of the third formant and the downward movement of the first for-

mant after a short period of little formant movement. The remaining vocal pulses all showed traces of aperiodicity and were considered part of the final /z/. The duration of the final /z/ was 85 msec.

An eight-member series that ranged from /b/ to /p/ was created by replacing successively longer sections from the onset of the original /b/, up to the zero crossing that marked the onset of a vocal pulse that had similar sections from the aperiodic /p/ onset. Details of the waveform editing process for creating voicing series from natural speech can be found in Ganong (1980). This editing process results in a natural speech-based VOT continuum in which the voicing in the initial /b/ is gradually replaced by the aspiration of the initial /p/. The voice pitch was not constant over the initial part of the /b/. Consequently, because editing always involves whole pitch pulses, the increment in VOT between stimuli was approximately 10 msec. The actual VOT values were 10, 10, 19, 29, 39, 49, 58, and 67 msec. The first 10-msec stimulus represented the original, natural /b/. The second stimulus in the series (also 10-msec VOT) was created by removing the /b/ release burst and replacing it with the /p/ release burst. Stimulus 3 (19-msec VOT) was created by removing the release burst and the first two vocal pulses of /b/ and replacing them with the corresponding duration of release burst plus aspiration from the /p/. Each succeeding stimulus was created by removing an additional two vocal pulses from the onset of /b/ and replacing them with the corresponding duration of release burst plus aspiration from the onset of /p/. All waveform editing was done at zero crossings to avoid the introduction of clicks into the waveform.

The remainder of the syllable, /ʌlz/, was edited to create four new syllables: /ʌlz/ with a short /ʌ/ vowel, /ʌlz/ with a long /ʌ/, /ʌlz/ with a short /l/ consonant, and /ʌlz/ with a long /l/. For the two short /ʌ/ vowel series, every other vocal pulse from the vowel (1st, 3rd, ... 11th) was digitally removed (a total of 6 of the 12 pulses). The long /ʌ/ series were created by reduplicating each of the 12 vocal pulses of the vowel. For both the short- and the long-vowel series, no changes were made to the /l/ or the /z/. The short /l/ series were made by digitally removing 4 vocal pulses (the 2nd, 4th, 7th, and 9th), and the long /l/ series resulted from the reduplication of each of the 2nd through 10th vocal pulses of the /l/. No editing of the first vocal pulse of the /l/ was done, because it represented the boundary between the vowel and the consonant. No changes to the original /ʌ/ or the /z/ were made in the short and long /l/ series. The resulting /ʌ/ and /l/ durations were within the range found in fluent speech (Crystal & House, 1988a, 1988c, 1988d). The approximate durations for the vowel and liquid portions of these syllables are given in Table 1. The four tokens of /ʌlz/ were then spliced to the ends of each of the eight members of the /b/–/p/ VOT continuum, resulting in four /b/–/p/ series (32 different syllables).

Procedure. The listeners were tested individually. Each listener heard all four of the series. The stimulus presentation and response collection were controlled by an Apple Macintosh computer. The stimuli, which were stored on disk, were converted to analog form in real time by a 16-bit, digital-to-analog converter at a 20-kHz sampling rate, low-pass filtered at 9.0 kHz, amplified, and presented binaurally through TDH-39 headphones. The syllables were presented in random order. The listeners were asked to rate the quality of the initial phoneme on a 6-point scale, ranging from 1, *a good "b,"* to 6, *a good "p."* The use of ratings allowed us to pick up subtle differences within a phonetic category and may have yielded more accurate estimates of the phonetic category boundary (see Sawusch, 1976). After the stimulus had been presented, the listeners responded by pressing the appropriate button on a computer-controlled response box. The presentation pace depended on the listener's response speed. The next syllable was presented as soon as the listener had responded or after an interval of 4.0 sec had elapsed, whichever came first. Responses from the first block of 64 trials (two repetitions of each of the 32 items) were considered as practice and were not included in subsequent data analysis. After the practice set, stimuli were presented in blocks of 96 trials (three repetitions of each of the 32 items), and all listeners received five blocks. This resulted in a total of 15 responses to each of the 32 stimuli for each listener.

Table 1
Approximate Adjacent Vowel and Nonadjacent Consonant Durations (in Milliseconds) for the /bʌlz/–/pʌlz/ Series in Experiment 1

Series	/ʌ/	/l/
Short /ʌ/	31	47
Long /ʌ/	117	47
Short /l/	61	28
Long /l/	61	89

Results and Discussion

For each listener, a mean rating was computed for each stimulus in each series. The /b/–/p/ category boundaries were then determined for each listener's data for each series by linear interpolation between the rating responses for the two stimuli on either side of a neutral (3.5) response. As a check on the results, we also tabulated the total percentages of "b" responses given by each listener to all of the stimuli in each series.[2] The movement of the category boundary should indicate changes in the perception of ambiguous stimuli. However, the overall percentage of "b" responses to the series as a whole included any changes away from the boundary of the series as well as those at the boundary. Consequently, this overall percentage measure may be a more sensitive index of changes in perception (see Samuel, 1986). The effect of vowel duration and consonant duration on the placement of the /b/–/p/ category boundary was evaluated using *t* tests. The percent "b" data were evaluated in a similar fashion.[3]

The mean rating functions for the /bʌlz/–/pʌlz/ series are shown in Figure 1. The short- and long-vowel series data are shown on the left and the short- and long-consonant data are shown on the right. As expected, there was a significant effect of the variation in vowel duration on the voicing category boundary [$t(17) = 4.44$, $p < .001$, for the 7.0-msec VOT (0.70 stimulus unit) difference]. The effect of vowel duration was also reliable in the percent /b/ response data [$t(17) = 6.06$, $p < .001$, for the 9.8% difference]. The category boundary occurred at a shorter VOT, and listeners gave more /p/ responses (i.e., fewer /b/ responses) to the short-vowel series than to the long-vowel series. Our results replicated Summerfield's (1981) results for a similar series and represents the usual effect of variation in the adjacent vowel duration on a syllable-initial consonant contrast (see Miller & Liberman, 1979). The location of the category boundary in each series and the percent "b" response, as well as their respective standard deviations, are shown in Table 2.

For the varying /l/ duration, there were also significant effects in both the location of the category boundary [$t(17) = 4.83$, $p < .001$, for the 2.8-msec VOT difference] and the percent /b/ data [$t(17) = 5.31$, $p < .001$, for the 4.7% "b" difference]. These results can easily be

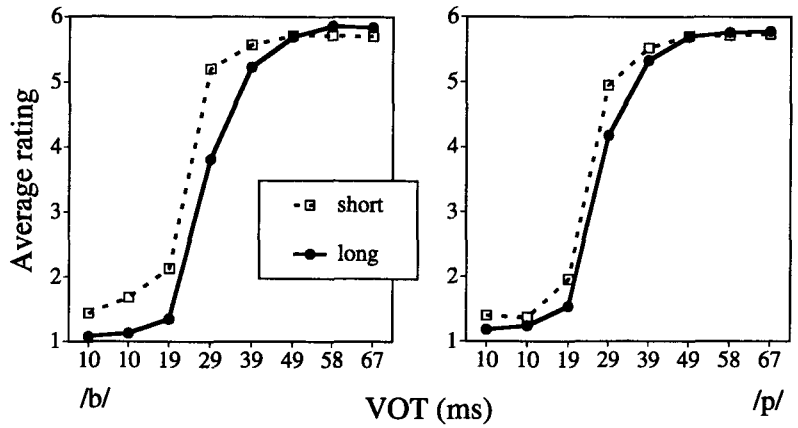

Figure 1. Group rating functions for the natural /bʌlz/–/pʌlz/ series. The series with /ʌ/ vowel duration varying are on the left and the series with /l/ consonant duration varying are on the right.

seen in the right panel of Figure 1, where the rating function for the long /l/ series is displaced to the right, relative to the function for the short /l/ series. Again, this difference was as expected (if the variation in /l/ duration represented a change in speaking rate and were to influence perception of the /b/–/p/ contrast).

The variation in the duration of the adjacent vowel influenced the perception of the initial /b/–/p/ contrast. Variation in the duration of the nonadjacent consonant had a similar effect.[4] Consequently, these data appear to show that variation in the duration of any segment that occurs within a short temporal window that follows the target can influence perception of the target. This is consistent with the earlier work of Newman and Sawusch (1996) and supports their conclusion that temporal proximity (occurrence within a brief temporal window of the target) is the major determinant of rate-normalization effects in speech.

Why then did Summerfield (1981) not find an effect when he varied the duration of the final /z/ in /biz/–/piz/ series? The simplest explanation is that the variation in the duration of the final /z/ was too remote from the initial phoneme. That is, part of the final /z/ and all of the variations in the duration of the final /z/ occurred outside a brief temporal window that followed the initial phoneme target. In the present experiment, a shorter vowel between the initial target consonant and the nonadjacent consonant allowed us to move the duration variation of the nonadjacent consonant closer to the target. Consequently, it seems likely that a single principle of rate normalization, in which information that occurs within a brief temporal window after the target can influence the perceptual processing of the target, can uniformly explain all of the data on the effects of succeeding phonemes (both adjacent and nonadjacent).

Table 2
Category Boundary Locations (in Milliseconds Voice Onset Time) and Percent "b" Responses for the /bʌlz/–/pʌlz/ Series in Experiment 1

Series	Boundary Location		Percentage "b" Responses	
	M	*SD*	*M*	*SD*
/ʌ/ Vowel Varying				
Short	22.3	3.97	35.1	4.69
Long	29.3	4.90	44.9	6.19
/l/ Consonant Varying				
Short	24.7	2.66	37.6	4.27
Long	27.5	4.49	42.3	6.07

EXPERIMENT 2

The results of Experiment 1 are consistent with a single explanation for both the presence and absence of rate-normalization effects for nonadjacent consonants. However, there are also differences between the stimuli used by Summerfield (1981) and those of our Experiment 1 that could be responsible for the differences in the results. One difference is that the final consonant in Summerfield's study contained aperiodic information (frication), whereas the nonadjacent /l/ consonant in Experiment 1 was voiced and contained no frication. Thus, the change from adjacent vowel to final consonant in Summerfield was accompanied by a change in the source quality of the sound. In contrast, the change from /ʌ/ vowel to /l/ consonant in Experiment 1 contained no such change. Instead, only a smooth change in the formant frequencies occurred.

To test this possibility, a new set of stimuli, which contained four series varying from /bʊʃ/ to /pʊʃ/ ("bush" to "push"), was created. If the change in source is a critical factor, variation in the duration of the adjacent vowel should alter perception of the initial /b/–/p/ contrast, whereas variation in the duration of the final consonant should have no effect because the fricative /ʃ/ is voiceless and aperiodic. However, if the proposal of Newman and Sawusch (1996) is correct, then as long as the vari-

ation in the final /ʃ/ occurs within the temporal processing window of the initial /b/–/p/ contrast, the processing of the initial consonant should be influenced. This experiment differs in two important ways from Summerfield's (1981) study. First, because the lax vowel /ʊ/ is intrinsically short in duration, the variation in the duration of the final fricative was closer, in time, to the initial consonant. Second, the participants were asked, after the listening session, about what the stimuli had sounded like. None of the listeners reported hearing anything other than "bush" and "push." This eliminated the possibility that the variation in final consonant duration could have cued different phonemes rather than different speaking rates (see Summerfield).

Method

Participants. The listeners were 19 volunteers at the State University of New York at Buffalo who were paid $5 for their participation. All listeners were native speakers of English and had no reported history of a speech or hearing impairment.

Stimuli. A male native speaker (J.R.S.) of midwestern American English recorded the syllables /bʊʃ/ and /pʊʃ/ in the context of fluent speech. The stimuli were amplified, low-pass filtered at 9.4 kHz, digitized via a 16-bit, analog-to-digital converter at a 20-kHz sampling rate, and stored on computer disk. The syllables were excised from the carrier sentence, "Norton said _____ to me." The base /bʊʃ/ and /pʊʃ/ syllables differed in their durations, with /pʊʃ/ being longer. The syllable /bʊʃ/ had a VOT of 0 msec and was followed by 10 vocal pulses (95 msec), and then by 120 msec of frication. The original /pʊʃ/ had a VOT of 66 msec, followed by eight vocal pulses (75 msec) and then by 130 msec of frication. Spectral analysis showed that the initial formant transitions in the /b/ were approximately 25 msec in duration. In creating the /b/–/p/ series, we decided to partially preserve the difference in stimulus duration that existed in the base /bʊʃ/ and /pʊʃ/ syllables. A seven-stimulus continuum was created in a fashion similar to that used in Experiment 1. Stimulus 1 was the base /bʊʃ/ syllable. Stimulus 2 was created by splicing the 10-msec release burst from /pʊʃ/ to the beginning of /bʊʃ/. For Stimuli 3, 4, 5, and 6, vocal pulses were removed from the onset of the /bʊʃ/ syllable (1, 2, 3, and 4, respectively) and replaced by the burst plus equivalent duration of aspiration from /pʊʃ/ (19, 30, 42, and 51 msec, respectively). Finally, for Stimulus 7, the first four vocal pulses of /bʊʃ/ were removed and replaced by burst and aspiration totaling 62 msec from /pʊʃ/. This resulted in a /b/–/p/ series with VOTs of 0, 10, 19, 30, 42, 51, and 62 msec for the seven stimuli.

In the long- and short-vowel series, the vocalic portion was edited. For the short-vowel series, the number of vocal pulses in the seven stimuli were reduced to 7, 7, 6, 5, 4, 3, and 3 for Stimuli 1–7, respectively. This was accomplished by removing three nonadjacent vocal pulses (numbers 5, 7, and 9 from Stimulus 1) from each of the seven stimuli. For the long-vowel series, each of the vocal pulses that were numbered 4, 5, 6, 7, 8, 9, and 10 in the /bʊʃ/ endpoint were reduplicated. This resulted in 17, 17, 16, 15, 14, 13, and 13 vocal pulses in the seven stimuli.

For the long and short final /ʃ/ series, the final 120-msec frication was edited. The short /ʃ/ was created by digitally removing the final 45 msec of frication (leaving a 75-msec frication) and tapering the amplitude of the remaining frication to zero over the last 25 msec. The long /ʃ/ was created by reduplicating nonadjacent 5-msec portions of the /ʃ/ frication. Both of these series left the number of vocal pulses as described above. The /ʊ/ vowel and /ʃ/ consonant segment durations for the four series are all shown in Table 3. These segment durations are within the range reported by Crystal and House (1988a, 1988b) for fluent speech. The editing process resulted in 28 different syllables: 7 in each of four series. Each listener heard all of the stimuli.

Table 3
Approximate Adjacent Vowel and Nonadjacent Consonant Durations (in Milliseconds) for the /bʊʃ/–/pʊʃ/ Series in Experiment 2

Series	/ʊ/	/ʃ/
Short /ʊ/	38	120
Long /ʊ/	130	120
Short /ʃ/	72	75
Long /ʃ/	72	195

Procedure. The procedure was identical to that used in Experiment 1. All participants listened to a practice block containing two occurrences of each of the 28 syllables in random order. This was followed by five blocks with three repetitions of each stimulus per block. Thus, by the end of the experiment, each listener had provided 15 rating responses to each of the 28 stimuli.

Results and Discussion

The basic data analysis procedure was the same as in Experiment 1. For each listener, a mean rating was computed for each stimulus in each series. The /b/–/p/ category boundary for each listener was then determined for each series, and paired *t* tests were used to evaluate the effect of segment duration for each pair of series (vowel duration varying or consonant duration varying). As in Experiment 1, equivalent analyses were also run on the percentage of "b" responses given to each series. The group rating functions for the four series are shown in Figure 2. The data for the two series in which the vowel duration varied are shown on the left, and the /ʃ/ consonant-duration–varying data are shown on the right. The means for the locations of the category boundaries, along with the comparable information for the percent "b" responses and their standard deviations, are shown in Table 4.

The two series that differed in their vowel durations yielded a reliable difference in both the location of the category boundary and the percentage of "b" responses. The mean difference of 8.0-msec VOT in the locus of the category boundary was significant [$t(18) = 6.50$, $p < .001$], and the difference of 9.1% in the overall "b" responses was also significant [$t(18) = 7.06$, $p < .001$]. These effects were in the expected direction. As is shown in the left panel of Figure 2 (and the top of Table 4), the short-vowel series yielded an earlier category boundary and fewer "b" responses than did the long-vowel series. This replicates the findings of Experiment 1 and of previous studies.

The effect of variation in the duration of the final /ʃ/ was also reliable. In the category boundary data, the boundary for the short /ʃ/ series was 1.7-msec VOT earlier (toward /b/) than that for the long /ʃ/ series [$t(18) = 3.09$, $p < .01$]. In the percent "b" data, fewer "b" responses were given to the short /ʃ/ series relative to the long /ʃ/ series. The difference of 3.3% was significant [$t(18) = 3.97$, $p < .001$]. These effects can be seen in the right panel of Figure 2 as the displacement of the short /ʃ/ se-

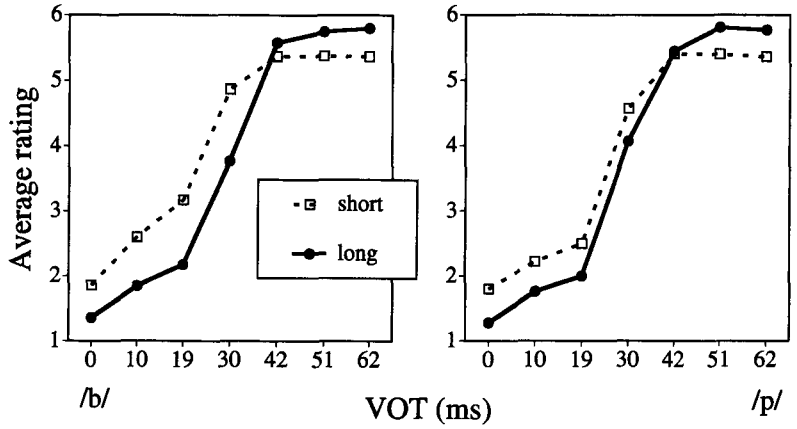

Figure 2. Group rating functions for the natural /bʊʃ/–/pʊʃ/ series. The series with /ʊ/ vowel duration varying are on the left and the series with /ʃ/ consonant duration varying are on the right.

ries rating function to the left of the series, relative to the rating function for the long /ʃ/ series.

Overall, the results of Experiment 2 show a small, but consistent, effect of variation in the final /ʃ/ duration on the perception of the initial /b/–/p/ contrast. In the present study, the identity of the final phoneme did not change when its duration was manipulated, and the variation in final consonant duration was largely within 250 msec of the initial consonant. It is important to note that Summerfield's (1981) proposal, that a change in segment duration that is interpreted as a change in phoneme identity might eliminate rate-normalization effects, remained untested. The present data show that variation in final fricative duration can affect the perceived identity of an initial target under certain circumstances. Since an effect of the final fricative duration was found, the present data are consistent with the proposal of Newman and Sawusch (1996) that segment durations within a brief temporal window after the target can alter the perceived duration (and phonetic identity) of the target.

On the basis of the results of Experiments 1 and 2, we can safely rule out any role for a vocalicness principle in rate-normalization effects (cf. Newman & Sawusch, 1996). That is, as long as the segment occurs within a brief temporal window of the target, the segment duration influences the perception of the target. Across the studies of Newman and Sawusch (1996), and in Experiments 1 and 2 here, it does not appear to matter whether the segment is a consonant or a vowel, adjacent or nonadjacent, phonotactically legal or illegal, voiced or voiceless. Only the temporal distance between the target and the subsequent segment seems to matter.

Table 4
Category Boundary Locations (in Milliseconds Voice Onset Time) and Percent "b" Responses for the /bʊʃ/–/pʊʃ/ Series in Experiment 2

	Boundary Location		Percentage "b" Responses	
Series	*M*	*SD*	*M*	*SD*
/ʊ/ Vowel Varying				
Short	19.6	5.70	35.3	6.6
Long	27.6	5.75	44.4	6.5
/ʃ/ Fricative Varying				
Short	24.8	3.55	41.2	4.7
Long	26.5	3.46	44.5	4.4

EXPERIMENT 3

If temporal distance is the critical variable, what would happen if we introduced an abrupt spectral discontinuity, such as that created by an inharmonic change in the fundamental frequency or by a change in the identity of the speaker, within the brief temporal window? To briefly reiterate, one possibility would be that normalization for speaking rate would occur after the speech signal had undergone perceptual grouping, and only the components of the signal belonging to a single stream (perceptual group) would influence the normalization process. This is similar to the proposal of Lotto et al. (1996; see also Green et al., 1994). The alternative would be that while some perceptual grouping or stream segregation (see Bregman, 1990) might occur before speaking-rate normalization, the signal duration after the source discontinuity would still influence the perception of the target. As long as the segment duration was within the temporal window following the target, it would influence the perception of the target.

The results from Experiment 2 show that the change in voicing between /ʊ/ and /ʃ/ does not preclude the duration of the /ʃ/ from having an effect on the initial target phoneme. However, this /ʃ/ has coarticulatory information indicating continuity in talker and continuity or coherence in the phonetic stream (cf. Remez, Rubin, Berns, Pardo, & Lang, 1994). It is entirely plausible that rate normalization occurs within a coherent perceptual group and that perceptual grouping based on continuity

of the source or talker occurs before this rate-normalization process. In particular, if the effect that we have been exploring is one of "speaking-rate normalization," it may occur only after the speech stream has been grouped by talker. Alternatively, speaking-rate normalization may be an instance of a more general case of event-rate normalization (cf. Fowler, 1990), which may or may not be based on prior grouping of the speech stream according to event source.

Experiment 3 was designed to test these two alternatives using synthetic stimuli that were similar to those used by Green et al. (1994) and Lotto et al. (1996). Four synthetic /bi/–/pi/ continua were created. In two of them, a continuous fundamental frequency (*F*0) was present. One of these continua had a long vowel and one had a short vowel. In the second pair of continua, the *F*0 in the vowel was changed (from 200 to 117 Hz) instantaneously, as if speech from two different sources had been spliced together. Overall, the design and the stimuli in this experiment were very similar to those used in Green et al. and Lotto et al. However, those previous studies included short- and long-vowel continuous *F*0 series but only one series with a discontinuity in *F*0. Experiment 3 differed from the earlier work in the presentation of two different vowel durations after the *F*0 discontinuity. This allowed us to examine the effect of the segment duration that occurred after the discontinuity, in addition to the effect of duration that occurred before the discontinuity.

If the series with long and short vowels that occurred after the discontinuity produced effects that paralleled those of the long and short vowels with no discontinuity, we would have evidence that the normalization process occurs prior to (or concurrently with) perceptual grouping according to *F*0. Conversely, if we found no effect of variation in vowel duration that occurred after the discontinuity, it would be consistent with a speaking-rate normalization process that occurred after perceptual grouping by source characteristics. Synthetic stimuli were used in this experiment for two reasons. First, it facilitated a comparison of our results with those of Green et al. (1994) and Lotto et al. (1996), who also used synthetic stimuli in their experiments. Second, given the work of Miller and Wayland (1993; also Shinn et al., 1985), the use of lower quality synthetic speech might amplify the magnitude of any effects observed. We felt that it was prudent to try the *F*0 discontinuity manipulation with synthetic speech before attempting a talker-discontinuity manipulation with natural speech.

Method

Participants. The listeners were 50 students from an introductory psychology course at the State University of New York at Buffalo who participated in the experiment for class credit. All listeners were native speakers of English and reported no history of a speech or hearing impairment. Twenty-six participants listened to the two sets of /bi/–/pi/ stimuli with a continuous *F*0, and 24 listened to the two series with a discontinuity in *F*0.

Stimuli. Two sets of series were generated with the use of a cascade synthesizer described by Klatt (1980) and waveform editing. One pair of series consisted of a synthetic /b/–/p/ continuum that varied VOT in 5-msec steps from a 10-msec VOT /bi/ to a 50-msec VOT /pi/. The second pair of series was identical to the first except that part way through the vowel, the *F*0 changed from 200 to 117 Hz. For both pairs of series, one series in each pair had a long vowel and the other series had a short vowel.

The base syllable from which all other syllables were generated was 340 msec in duration and consisted of a natural, 10-msec release burst (spliced from a natural speech /p/) followed by a synthetic 330-msec /bi/. The amplitude of the release burst was 12 dB below the peak amplitude of the vowel. In the synthetic syllable, the first formant (*F*1) started at 313 Hz and made a 20-msec linear transition to 388 Hz, where it remained for the rest of the syllable. Similarly, the second and third formants had onset frequencies of 1773 and 2503 Hz, which were followed by 25-msec linear transitions to the steady-state values of 1954 and 2775 Hz, which were maintained through the balance of the syllable. The fourth and fifth formants were set to 3328 and 4113 Hz for the entire syllable and the bandwidths of the first five formants were 50, 95, 134, 205, and 229 Hz, respectively, for the entire syllable. The fundamental frequency was set to 200 Hz for the entire syllable. The amplitude of voicing was 56 dB at onset, jumped to 64 dB at 5 msec, and then changed linearly to 60 dB at 30 msec into the syllable. The value of 60 dB was then maintained until the syllable amplitude was ramped off over the last 50 msec. Finally, the natural 10-msec release burst from the syllable /pi/, which was spoken by a female talker (R.S.N.), was digitally attenuated by 9 dB and spliced to the onset of the synthetic syllable.

The /b/–/p/ series was generated by replacing the voicing source with an aperiodic source (aspiration) for a short period of time at synthesis onset. For each syllable, the amplitude of voicing was set to zero for the duration of the aspiration and the aspiration amplitude was set to 66 dB. This change was done for the first 5, 10, 15, 20, 25, 30, 35, and 40 msec of the base syllable. Simultaneously, the bandwidths of *F*1, *F*2, and *F*3 were set to 400, 150, and 190 Hz, respectively. Finally, as with the base syllable, the 10-msec release burst was digitally spliced to the onset of each series syllable. The amplitude of the burst was increased by 1 dB for each 5-msec step in VOT (relative to the amplitude in the 10-msec VOT /bi/). This yielded a nine-stimulus continuum from a 10-msec VOT /bi/ to a 50-msec VOT /pi/ in which VOT changed in 5-msec steps. This series will be referred to as the long-vowel, continuous *F*0 series. The short-vowel, continuous *F*0 series was generated from the long-vowel series by digitally removing all of the pitch pulses from the steady-state vowel between 65 msec from onset and 280 msec from onset. The difference between the two series was the duration of the steady-state vowel (290 and 75 msec for the 50-msec VOT /p/ end of each series).

In the discontinuous *F*0 series, most of the steady-state vowel from each of the long-vowel, continuous *F*0 stimuli was digitally removed and replaced with a synthetic vowel with a different *F*0. The new vowel was generated using the base /bi/ syllable and changing *F*0 to 117 Hz for the entire syllable. All other synthesis parameters remained the same. This resulted in the two discontinuous *F*0 series. In each stimulus, the initial 115 msec was retained, leaving 65 msec of vowel with a 200-Hz *F*0 (for the 50-msec VOT /p/ end of each series). This was followed by either 60 msec (short-vowel series) or 270 msec (long-vowel series) of the same synthesis parameters with the lower (117 Hz) *F*0. Thus, the vowel duration prior to the change in *F*0 in the discontinuous *F*0 series was approximately the same duration as the short vowel in the short-vowel, continuous *F*0 series. The /i/ vowel durations, both before any discontinuity in *F*0 and after, are shown in Table 5 for each of the four series.

Procedure. The procedure was identical to that used in Experiments 1 and 2. One group of 26 listeners heard the two continuous-talker series, whereas the other group of 24 listeners heard the two discontinuous-talker series. All participants listened to a practice

Table 5
Approximate /i/ Vowel Duration (in Milliseconds) Both Before and After Any Discontinuity in *F*0 for the /bi/–/pi/ Series in Experiment 3

Series	Before Discontinuity	After Discontinuity
Continuous short /i/	74	–
Continuous long /i/	290	–
Discontinuous short /i/	65	60
Discontinuous long /i/	65	270

block that contained two occurrences of each of the 18 syllables in their set in random order. This was followed by blocks of four repetitions of each stimulus in the set. All listeners participated in four blocks of test trials. Thus, by the end of the experiment, each listener had provided 16 rating responses to each of the 18 stimuli in their set.

Results and Discussion

The basic data analysis procedure was the same as that in Experiments 1 and 2. For each listener, a mean rating was computed for each stimulus in each series. The group rating functions for the four series are shown in Figure 3. The data for the two continuous *F*0 series, in which the vowel duration varied, are shown on the left, whereas the data for the variation in vowel duration after the discontinuity in *F*0 (the discontinuous *F*0 series) are shown on the right. The means for the locations of the category boundaries, as well as the comparable information for the percent "b" responses and their standard deviations, are shown in Table 6.

The two continuous *F*0 series, which differed in their vowel durations, yielded a reliable difference in both the location of the category boundary and the percentage of /b/ responses. The mean difference of 3.1-msec VOT in the locus of the category boundary was significant [$t(25) = 7.20, p < .001$] and the difference of 9.1% in the overall percentage of "b" responses was also significant [$t(25) = 9.11, p < .001$]. This is the expected effect of vowel duration on the perception of the initial /b/–/p/ contrast. As can be seen on the left side of Figure 3, for the long vowel, listeners rated the intermediate stimuli as more /b/-like, which yielded a category boundary at a longer VOT and more "b" responses overall than for the short-vowel series.

For the discontinuous *F*0 series, in which the vowel duration varied after the discontinuity in *F*0, reliable differences were found in both the location of the category boundary and the percentage of "b" responses. Here, the mean difference of 1.1-msec VOT in the location of the category boundary was significant [$t(23) = 4.17, p < .001$] and the 5.2% difference in overall "b" responses was also significant [$t(23) = 5.56, p < .001$]. As is shown on the right side of Figure 3 (see also Table 6), the effect was small. However, the discontinuous *F*0 series are, in many ways, comparable to the series with remote duration differences in Experiments 1 and 2. This is because the two discontinuous *F*0 series contained the same short-vowel segment between the initial consonant and the final vowel segment after the discontinuity. This is similar to the short /ʊ/ vowel between the initial consonant and the final /ʃ/ in the /bʊʃ/–/pʊʃ/ series of Experiment 2. Since the effects of variation in the duration of remote, nonadjacent phonemes were generally small (see Newman & Sawusch, 1996; Experiments 1 and 2, above), we were not surprised by the small effect found here.

Finally, for comparison with the earlier data of Green et al. (1994) and Lotto et al. (1996), we compared the influence of the long, discontinuous *F*0 vowel to the short and long continuous *F*0 vowels. In Lotto et al.'s first experiment, the inharmonic *F*0 change was most comparable to our discontinuous talker, long-vowel condition. Their results showed that the inharmonic vowel, in spite of having an overall vowel duration comparable to that of the long continuous vowel, produced a category boundary at a significantly shorter VOT than did the long vowel. Their inharmonic vowel series category boundary occurred at a slightly longer VOT than did their short-vowel series category boundary, but the difference was not significant.

In our data, similar effects were found. The locus of the /b/–/p/ category boundary for the long, discontinu-

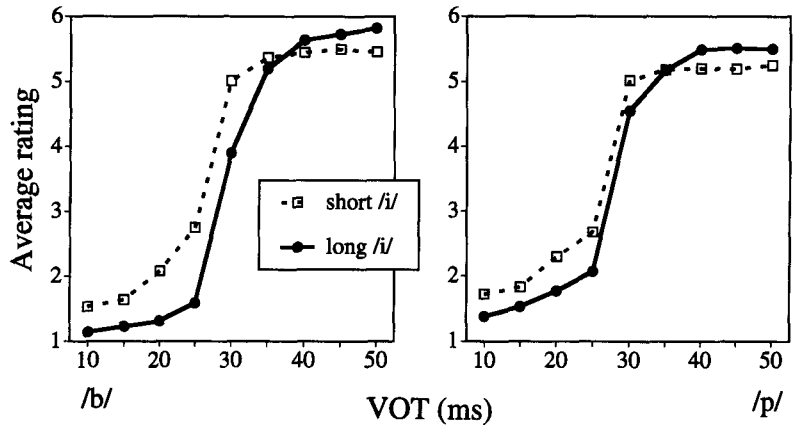

Figure 3. Group functions for the synthetic /bi/–/pi/ series. The series with a continuous *F*0 are on the left and the discontinuous *F*0 series data are on the right.

Table 6
Category Boundary Locations (in Milliseconds Voice Onset Time) and Percent "b" Responses for the /bi/–/pi/ Series in Experiment 3

	Boundary Location		Percentage "b" Responses	
Series	*M*	*SD*	*M*	*SD*
Continuous Talker				
Short	26.3	1.89	40.6	5.59
Long	29.4	2.46	49.7	6.29
Discontinuous Talker				
Short	26.2	2.07	40.5	6.21
Long	27.3	2.53	45.7	6.52

ous $F0$ occurred at a significantly shorter VOT than that for the long, continuous $F0$ series [$t(48) = 3.31$, $p < .005$].[5] The difference in location of the category boundary between the long, discontinuous $F0$ and the short, continuous $F0$ series was marginally significant [$t(48) = 1.70$, $p < .10$]. That is, the long, discontinuous $F0$ vowel had an effect that was substantially different from that of the long, continuous $F0$ vowel. The category boundary for the long, discontinuous $F0$ vowel occurred at a slightly longer VOT value than that of the short, continuous $F0$ vowel. This parallels the comparable data of Lotto et al. (1996).

It appears that the $F0$ change triggers a segmentation process. The category boundaries for the discontinuous $F0$ series (both long and short following vowel) were both more similar to those of the short continuous $F0$ vowel series. Consequently, the introduction of a discontinuity in $F0$ seems to have triggered a segmentation process that resulted in a short segment following the target. This is similar to the continuous $F0$, short-vowel series. The effect of variation in the duration of the segment following the $F0$ discontinuity was then layered over the influence of the short segment preceding the discontinuity.

Overall, the results of Experiment 3 appear to indicate that even when the $F0$ of a speech signal changes abruptly, the duration of the signal segments after the $F0$ change influence the perception of phonetic segments before the $F0$ change. However, this does not, necessarily, imply that similar results would have been found if the identity of the talker had been changed. Whereas the $F0$ was changed in a single pitch pulse, from 200 to 117 Hz, the formant frequencies were not changed. Thus, it is possible that a listener might have interpreted these stimuli as having a pitch change but not a talker change.

There is one additional qualification to note. It is possible that the vowel after the $F0$ discontinuity could have been heard as a different phoneme, or qualitatively different from, the vowel before the $F0$ change. If this were the case, the discontinuous $F0$ series would differ from the continuous $F0$ series in the number of perceived phonetic segments. The greater number of segments would cue a faster speaking rate. Although this is possible, it does not alter the basic conclusion that information about segment duration after an abrupt, inharmonic change in $F0$ can influence the perception of the initial target. Furthermore, as described previously, Lotto et al. (1996) presented data showing that the number of perceived phonetic segments per se does not appear to be the controlling factor in speaking-rate normalization.

EXPERIMENT 4

In the last experiment in this series, we sought to replicate and extend the results of Experiment 3 by using natural speech. If a change in talker from a female voice to a male voice is made with digital waveform editing on natural, high-quality speech, it should be readily apparent to listeners that the identity of the talker has changed midway through the stimulus. If the segment duration following the discontinuity in talker still has an effect on the stimulus initial target, we can be fairly certain that the normalization process that is taking place either precedes, or is concurrent with, the process of grouping by source quality or talker identity. This is not to imply that rate normalization precedes all perceptual grouping. Rather, this experiment was a test of one type of perceptual grouping, one based on talker identity, or voice continuity.

Two sets of stimuli were chosen for this experiment. One set consisted of natural tokens of the syllables /bi/ and /pi/ as spoken by both female and male talkers. The discontinuity was introduced during the vowel. Thus, these stimuli represented a natural speech analogue to the synthetic stimuli of Experiment 3 and introduced the signal discontinuity in the middle of a phonetic segment, just as in Experiment 3. The second set of stimuli was based on the /bʌlz/–/pʌlz/ series from Experiment 1. In this case, the discontinuity in talker was introduced between phonetic segments so that the initial CV was in the female voice and the final CC was in the male voice. If both sets of stimuli produced analogous results, we could be fairly certain that the nature of the location where the discontinuity was introduced was not a critical factor. The /bʌlz/–/pʌlz/ series offerred one additional benefit. If an influence of /l/ duration was found in the new, discontinuous talker series, then we could further examine the influence of talker discontinuity by comparing the magnitude of the effect of the continuous-talker /l/ duration variation from Experiment 1 with the results of the discontinuous-talker /l/ duration. To the extent that they were the same, it would provide further support for the proposal that all segmental information within a brief temporal window around the target serves as a reference for the rate at which events occur.

Method

Participants. The listeners were 40 students from an introductory psychology course at the State University of New York at Buffalo who participated in the experiment for class credit. All listeners were native speakers of English and reported no history of a speech or hearing impairment. Twenty participants listened to the two sets of /bi/–/pi/ stimuli with a discontinuous talker, and 20 listened to the two /bʌlz/–/pʌlz/ series with a discontinuous talker. One listener in the /bʌlz/–/pʌlz/ group was unable to classify the ends

of their series consistently, and that person's data were omitted from further analysis.

Stimuli. Two sets of discontinuous-talker stimuli were created through digital editing of natural speech. In both cases, the initial portion was spoken by a female and the final portion by a male.

The discontinuous /bʌlz/–/pʌlz/ series was based on the stimuli that were used in Experiment 1. In this new series, the initial /bʌ/ and /pʌ/, which were from the original series, were spoken by a female and the /lz/ was spoken by a male.

A male talker (J.R.S.) recorded the syllable /bʌlz/ in sentential context, as described previously. On the basis of spectral analyses, the boundary between the vowel /ʌ/ and the liquid /l/ was placed at a point where the first formant moved downward and the third formant moved upward (see also Experiment 1). The initial /bʌ/ was digitally removed (up to the last negative value that preceded the onset of the first vocal pulse of the /l/). This left an /lz/ that comprised an /l/ which consisted of the 12 vocal pulses before the onset of aperiodicity (124 msec) and an 81-msec final /z/. The *F*0 at the onset of the /lz/ for the male was 108 Hz, and the *F*0 at the offset of /bʌ/ for the female was 200 Hz. The duration of this /lz/ was 205 msec, which was approximately the same as the corresponding /lz/ segment in the original series that had the base vowel and long /l/. Consequently, the natural /lz/ of the male talker was used to create the long /l/ series.

In each of the eight items from the long /l/ series in Experiment 1, spoken by a female, the final /lz/ was digitally removed. The precise cut point varied with the VOT of the stimulus, but the average duration of the /bʌ/ was 112 msec (see Experiment 1 for a description of the determination of the cut point). The /lz/ of the male talker was then appended to each stimulus to generate the discontinuous-talker, long /l/ series. The corresponding short /l/ series was generated by deleting 6 of the 12 pitch pulses of the /l/ (the 2nd, 4th, 6th, 8th, 10th, and 12th). This resulted in an /lz/ with a duration of 142 msec (including an /l/ duration of 61 msec). As with the long /l/ series, the short /lz/ of the male talker was digitally appended to the /bʌ/–/pʌ/ series of the female talker to create the discontinuous-talker, short /l/ series. The /ʌ/ vowel and /l/ consonant durations for these two series are shown in Table 7.

The discontinuous /bi/–/pi/ series was based on the natural syllables /bi/ and /pi/, spoken by a female native speaker (M.E.S.) of midwestern American English and the syllable /bi/ spoken by a male talker (J.R.S.). The recordings were done as described previously. From the natural tokens of /bi/ and /pi/ spoken by the female talker, a /b/–/p/ series was created by digitally removing successively longer portions of /bi/ from stimulus onset and replacing them with the corresponding segments of /pi/. The editing process was the same as that used in Experiments 1 and 2. The natural /bi/ had a VOT of 12 msec and an *F*0 of approximately 217 Hz. Consequently, the /bi/ end of the series had a VOT of 12 msec. The remaining seven stimuli had VOTs of 12, 18, 23, 27, 32, 37, and 42 msec.

Each of these eight items was then edited to remove all vocal pulses after approximately 87 msec. This meant that for the /pi/ end of the series that had a VOT of 42 msec, there were still 10 vocal pulses that represented the vowel /i/ (approximately 46 msec). The /bi/ from the male talker was edited so that only the final 274 msec of the vowel remained. Spectral analysis showed that the formant transitions of the initial stop were complete before this part of the syllable. The *F*0 at the onset of this vowel was 104 Hz. This segment was digitally appended to the eight items in the female voice to form the discontinuous-talker, long-vowel series. The short vowel was created by the deletion of all pitch pulses from the long vowel after the first six and by digitally tapering off the amplitude of this short vowel (58 msec) over the last two pitch pulses (20 msec). Again, the short vowel was appended to each of the eight items in the female voice to create the discontinuous-talker, short-vowel series. These vowel durations (before and after the talker change) were similar to the vowel durations that occurred before and after the *F*0 change in Experiment 3. The /i/ vowel durations for these series, both before and after the change in talker, are shown in Table 7.

Table 7
Approximate Segment Durations Before and After the Discontinuity in Talker (in Milliseconds) for the /bʌlz/–/pʌlz/ and /bi/–/pi/ Series in Experiment 4

Series	Before Discontinuity	After Discontinuity
/bʌlz/–/pʌlz/ short /l/	61	61
/bʌlz/–/pʌlz/ long /l/	61	124
/bi/–/pi/ short /i/	46	58
/bi/–/pi/ long /i/	46	274

For both sets of series, this method of editing resulted in stimuli in which there was a clear change in talker. During preliminary tests, some listeners described the stimuli as splitting or streaming into two distinct components and all listeners reported hearing two different talkers. No listener reported a vowel quality (identity) change for the /bi/–/pi/ series or the presence of a phoneme other than those originally intended by the talkers.

Procedure. The procedure was identical to that used in Experiments 1, 2, and 3. One group of 20 listeners heard the two /bi/–/pi/ series while the other group of 20 listeners heard the two /bʌlz/–/pʌlz/ series. All participants listened to a practice block containing two occurrences of each of the 16 syllables in their set in random order. This was followed by four blocks of trials which contained four repetitions of each stimulus in the set. Thus, by the end of the experiment, each listener had provided 16 rating responses to each of the 16 stimuli in their set.

Results and Discussion

The basic data analysis procedure was the same as in Experiments 1, 2, and 3. For each listener, a mean rating was computed for each stimulus in each series. The group rating functions for the four series are shown in Figure 4. The data for the two /bʌlz/–/pʌlz/ series, in which the /l/ duration varied, are shown on the left; the data for the two /bi/–/pi/ series, in which vowel duration varied after the discontinuity in talker, are shown on the right. The means for the locations of the category boundaries, as well as the comparable information for the percent "b" responses and their standard deviations, are shown in Table 8.

For the two /bi/–/pi/ discontinuous-talker series in which the vowel duration varied after the discontinuity, reliable differences were found for both the location of the category boundary and the percentage of "b" responses. Here, the mean difference of 1.3-msec VOT in the location of the category boundary was significant [$t(19) = 2.43, p < .05$], and the 4.3% difference in overall "b" responses was also significant [$t(19) = 2.44, p < .05$]. As is shown on the right side of Figure 4 (see also Table 8), the change in listeners' ratings was such that intermediate stimuli near the category boundary were more likely to be classified as /b/ in the environment of a long vowel and as /p/ in the environment of a short vowel. Thus, the natural /bi/–/pi/ discontinuous-talker series seem to have produced an effect similar to that found with the synthetic discontinuous *F*0 series in Experiment 3. Consequently, it does not appear to matter whether the discontinuity that was introduced in the stimuli reflects a change just in *F*0 or a change in *F*0 and the formant frequencies. The duration of the vowel segment

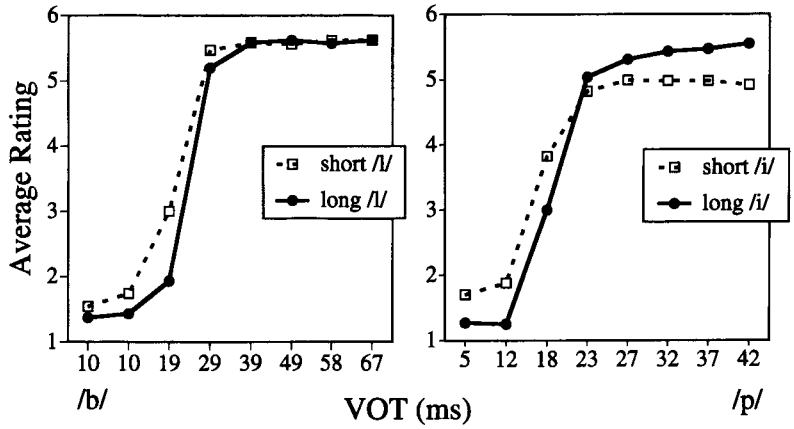

Figure 4. Group rating functions for the discontinuous natural /bʌlz/–/pʌlz/ series (left) in which /l/ duration varied and the discontinuous natural /bi/–/pi/ series (right) in which vowel duration varied.

that followed the discontinuity influenced the listeners' labeling of the initial /b/–/p/ contrast.

The two /bʌlz/–/pʌlz/ series, which differed in their /l/ duration, yielded a reliable difference in both the location of the category boundary and the percentage of "b" responses. The mean difference of 3.5-msec VOT in the locus of the category boundary was significant [$t(18) = 6.18, p < .001$], and the difference of 5.4% in the overall /b/ responses was also significant [$t(18) = 8.23, p < .001$]. As can be seen on the left side of Figure 4, for the long /l/, listeners rated the intermediate stimuli as more /b/-like, which yielded a category boundary at a longer VOT and more "b" responses overall than for the short /l/ series.

Finally, it may be instructive to compare the magnitude of the effect of the short and long /l/ in the discontinuous /bʌlz/–/pʌlz/ series with the magnitude of the effect of /l/ duration variation in the analogous continuous-talker series in Experiment 1. The short and long /l/ durations in these two sets of stimuli were roughly similar. In spite of the change in talker in the discontinuous-talker series, the magnitude of the effect of variation in /l/ duration was approximately the same across the two experiments (2.8- vs. 3.5-msec VOT for the shift in the category boundary for the continuous- and discontinuous-talker series). The difference in the magnitude of the effect for these two series was not significant [$t(35) = -0.89, p > .25$ for the category boundary data, and $t(35) = -0.72$, $p > .5$ for the percent "b" response data] (see note 5). If anything, the effect of variation in /l/ duration for the discontinuous-talker series was greater than for the continuous-talker series. Consequently, it appears that the process of rate normalization treats all of the stimulus information that is within a brief temporal window of the target equivalently, even when the target and a following segment are from two different talkers.

Table 8
Category Boundary Data (in Milliseconds Voice Onset Time) and Percent "b" Responses for the Discontinuous-Talker /bi/–/pi/ Series (Top) and /bʌlz/–/pʌlz/ Series (Bottom) in Experiment 4

	Boundary Location		Percentage "b" Responses	
Series	*M*	*SD*	*M*	*SD*
/i/ Vowel Varying				
Short	21.7	2.14	31.2	6.16
Long	23.0	1.69	35.5	7.08
/l/ Consonant Varying				
Short	19.8	2.14	31.5	5.06
Long	23.3	3.23	36.9	4.93

GENERAL DISCUSSION

The overall pattern of results seems to be most readily explained by the Newman and Sawusch (1996) proposal that rate normalization occurs within a limited time window after presentation of the target. Anything within that window can have an effect, but anything after that window cannot. In the experiments reported here, both the vowel that immediately followed the target and the consonant that followed the vowel influenced the perception of the target. In all cases, the increase in duration of the vowel or the consonant led to more "b" responses to the syllable-initial phoneme. This result supports the idea that the perception of phonetic contrasts that are mediated by acoustic correlates varying in duration is relative to the rate of speech. Furthermore, even the change in talker during the syllable did not appear to alter the influence of the subsequent segment on the perception of the initial contrast.

The results of Summerfield (1981) and Newman and Sawusch (1996), in which an effect of the syllable final consonant was not found, can be readily explained. In both studies, the duration variation in the final consonant occurred too long after the initial contrast. By the time the duration of the final consonant had been processed, perception of the initial consonant was already complete. Thus, there was no opportunity for the duration of the

final consonant to influence the perception of the initial phoneme contrast. The present results, and the results of Newman and Sawusch (1996), show that variation in the duration of virtually any type of phoneme can produce this rate normalization. Stops, liquids, fricatives, and vowels have all been shown to vary in their duration with changes in speaking rate (Crystal & House, 1990). Furthermore, when the duration of any of these types of phonemes is varied and the phoneme occurs in close proximity to the target, perception of the target is altered.

On the basis of the earlier results of Newman and Sawusch (1996), the temporal processing window within which these effects occur seems to extend to about 250 msec following the onset of the target phoneme. However, as Newman and Sawusch note, this estimate is based on high-quality synthetic and natural speech stimuli. Miller and Wayland (1993; see also Shinn et al., 1985) have shown that the size of rate normalization effects varies with the stimulus quality. This led Newman and Sawusch to speculate that the temporal window within which rate normalization effects occur may also vary with the quality of the stimulus. To the extent that processing of a phonetic contrast is slower with lower quality stimuli, there might be a longer period of time in which other segments could occur and still influence processing of the target. Consequently, seemingly conflicting results from various experiments can be explained by the existence of a common mechanism. Newman and Sawusch reasoned that the temporal window may have been larger for some of the highly stylized, synthetic stimuli used by Miller and Liberman (1979) but shorter for the stimuli used by Summerfield (1981). In Miller and Liberman, variation in the duration of a second syllable (/da/ in a /bada/–/wada/ series) altered perception of the initial phonetic distinction. In the data of Summerfield and Newman and Sawusch, variation in the duration of a syllable-final fricative had no effect on an initial target. Furthermore, in Miller and Liberman the final fricatives were temporally as close as, or closer than, the second syllable. However, if the temporal processing window is longer for lower quality, synthetic stimuli, these results all fit within the processing-window framework proposed by Newman and Sawusch.

On the basis of previous results and the present results, it seems reasonable to outline a process model of rate normalization. At a relatively early point in auditory perceptual processing, the signal is segmented. The segmentation process envisioned here would involve demarcating the points in the signal at which substantial changes in the spectrum occur. The segmentation process does not involve dividing the signal into phonemes (although a phoneme would generally correspond to one or more segments, a segment could also contain information about adjacent phonemes). This segmentation is followed by a process in which segment durations are computed and by a process in which a running average of segment durations is computed. The running average influences how segment durations are mapped onto events, regardless of whether that event is a speech gesture, feature, phoneme, or other unit.

The data from Green et al. (1994) and Lotto et al. (1996) and from Experiments 3 and 4 allow us to specify part of the process of segmentation. This process is apparently triggered by discontinuities in the acoustic signal as it is transduced by the auditory system. The data from Lotto et al., for instance, indicate that in order for an abrupt change in a formant frequency to trigger this segmentation, the change in formant frequency must be such that the first formant falls on different harmonics of the fundamental. Similarly, in order for a change in fundamental frequency to trigger the segmentation, the change should not be a simple octave change (doubling or halving) of the fundamental. This is consistent with using a model of the auditory system, such as that of Patterson, Holdsworth, and Allerhand (1992), which is based on periodicity processing. For a change, or discontinuity, to trigger the marking of a segment boundary, the pattern of output in the filter bank would have to change. Simply changing a formant frequency would not be sufficient unless the change in formant frequency caused the resonance and the resulting peak in the acoustic spectrum to fall onto a different harmonic and thus alter the output of the auditory filterbank (see Lotto et al., 1996).

The segmentation could be done explicitly by a segmentation process or implicitly as a consequence of some other processing operation. In an implicit segmentation model, whenever the pattern matching element that most closely corresponds to the input changes, a segment boundary is implicitly established (see Pisoni & Sawusch, 1975, for such a proposal). An explicit model would have a mechanism that monitored the input representation for discontinuities and explicitly marks their occurrence. For our purposes, either model is sufficient, because both are capable of providing the information that is necessary for the extraction of segment duration.

The next step is to determine the duration of an event. The acoustic correlate to duration is the time interval between "segment boundaries." The computed duration is used in the normalization process and is, itself, normalized for use in determining segment or event identity. The use of segment durations in the normalization process involves maintaining a "running average" of segment durations. The weighting scheme for computing this average is unknown, but might be akin to a normal curve in which the segments immediately adjacent to the target receive the greatest weight. The segments outside of a relatively brief temporal window would be weighted zero, which would indicate that these remote segments had no effect. The window might be fixed in size, but might continue accumulating the average over time. In this case, segments that were processed and were within the window prior to the conclusion of processing of the target could all influence the target. If, for some reason, processing of the target were to complete rapidly, the "effective temporal window" would appear to be short, because only the segment duration information up to that point in time

could influence processing of the target. This is the basic point made by Miller and Dexter (1988).

Finally, we can begin to specify where, in the time course of processing, the segmentation and duration-extraction processes occur, at least relative to other processes. Given the results of Experiments 3 and 4, these two processes clearly do not require prior perceptual grouping or streaming of the signal based on talker identity or *F*0 continuity. That is, the proposal of Lotto et al. (1996)—that the effects of speaking-rate normalization occur after perceptual grouping by the acoustic correlates of sound source—is incorrect, at least with respect to the present results. However, the reverse conclusion—that segmentation and duration extraction occur prior to perceptual grouping—would be very tenuous.

There are a number of principles of auditory stream formation (Bregman, 1990), and our data relate to only some of them. Clearly, simple continuity in the fundamental is not required for the rate normalization process. Abrupt, inharmonic changes in *F*0 and in *F*0 and the formant frequencies in Experiments 3 and 4 did not alter the influence of a subsequent segment on the event-rate normalization process. However, there are other bases for perceptual grouping, including those that result from sounds originating from different locations in space, that may operate prior to segmentation and duration computation (see Bregman, 1990). In addition, Remez et al. (1994) have proposed principles of coherence that may be used to group disparate signal elements together, prior to phonetic analysis, even when the basic auditory principles enumerated by Bregman appear to fail. Since our discontinuous talker stimuli were not designed to test the proposal of Remez et al., we cannot specify at this time whether the segmentation and duration extraction process precedes the grouping preparatory to phonetic analysis. A third alternative, described by Liberman and Mattingly (1985), is that the grouping of basic elements would be based on phonetic analysis. Since all our stimuli contained coarticulatory information, which would support the idea of such a grouping process, our results are consistent with Liberman and Mattingly's proposal. Finally, it is also possible that if perceptual grouping by talker had already been established *before* the onset of the target, information from the second talker might not influence perception of the target. Our stimuli were not designed to test this possibility. Instead, our data simply show that the process of speaking-rate normalization can take place concurrently with the process of segregating the speech signal according to talker.

A related issue concerns the role of duration information in perception. Segment durations vary as a function of the context within which they occur and of segment identity, as well as of speaking rate (see Crystal & House, 1988b; Miller, 1981). If variation in the duration of a segment is interpreted as being a cue to the phonetic identity of the segment, does this preclude the same duration variation from producing a "speaking-rate" effect? For example, it is possible to create synthetic vowels whose formant frequencies are ambiguous between two neighboring vowels, such as /i/ and /ɪ/. Long versions of such a vowel are identified as /i/ and short versions as /ɪ/ (see Ainsworth, 1972). Would a /b/–/p/ pair, with long and short versions of such a vowel, produce a change in stop classification? To be sure, if the duration variation is due to the intent of the talker to produce different phonemes, it is not a manifestation of a change in speaking rate. However, the issue regarding perception is when, in perceptual processing, the "event-rate normalization" process takes place and the scope of information that drives this process. If the rate-normalization process occurs early in perception, it may use all variations in segment duration that occur in the processing window, regardless of the articulatory reason for the variation. Alternatively, even though rate normalization is a relatively early, obligatory process (see Miller & Dexter, 1988), it may interact with other phonetic processes in a manner that recovers the intent of the talker (cf. Liberman & Mattingly, 1985). Summerfield (1981) made an essentially similar proposal when he suggested that the reason he had found no effect of duration variation for a final fricative on an initial contrast was that the final duration variation was interpreted by listeners as being a cue to a change in the phonetic identity of the final fricative (from /z/ to /s/). However, as noted earlier, no extant data either confirm or disconfirm this conjecture.

In summary, the process of normalizing for the rate at which events occur in perception seems to involve a temporal window around the target segment. All segments that occur within this brief temporal window around the target influence the perception of the target. This occurs in spite of changes in periodicity in the signal or spectral discontinuities, even the relatively large acoustic change that occurs with a change in talker. The results reported here support the notion of there being an autonomous event-rate normalization process that occurs early in perception.

REFERENCES

AINSWORTH, W. A. (1972). Duration as a cue in the recognition of synthetic vowels. *Journal of the Acoustical Society of America*, **51**, 648-651.

BREGMAN, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.

BYRD, D. (1992). Preliminary results on speaker-dependent variation in the TIMIT database. *Journal of the Acoustical Society of America*, **92**, 593-596.

CRYSTAL, T. H., & HOUSE, A. S. (1982). Segmental duration in connected-speech signals: Preliminary results. *Journal of the Acoustical Society of America*, **72**, 705-716.

CRYSTAL, T. H., & HOUSE, A. S. (1988a). The duration of American-English vowels: An overview. *Journal of Phonetics*, **16**, 263-284.

CRYSTAL, T. H., & HOUSE, A. S. (1988b). A note on the durations of fricatives in American English. *Journal of the Acoustical Society of America*, **84**, 1932-1935.

CRYSTAL, T. H., & HOUSE, A. S. (1988c). Segmental durations in connected-speech signals: Current results. *Journal of the Acoustical Society of America*, **83**, 1553-1573.

CRYSTAL, T. H., & HOUSE, A. S. (1988d). Segmental durations in con-

nected-speech signals: Syllabic stress. *Journal of the Acoustical Society of America*, **83**, 1574-1585.

CRYSTAL, T. H., & HOUSE, A. S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America*, **88**, 101-112.

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.

FLEGE, J. E., & SCHMIDT, A. M. (1995). Native speakers of Spanish show rate-dependent processing of English stop consonants. *Phonetica*, **52**, 90-111.

FOWLER, C. A. (1990). Sound-producing sources as objects of perception: Rate normalization and nonspeech perception. *Journal of the Acoustical Society of America*, **88**, 1236-1249.

GANONG, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception & Performance*, **6**, 110-125.

GREEN, K. P., STEVENS, E. B., & KUHL, P. K. (1994). Talker continuity and the use of rate information during phonetic perception. *Perception & Psychophysics*, **55**, 249-260.

KIDD, G. R. (1989). Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception & Performance*, **15**, 736-748.

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

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

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

LOTTO, A. J., KLUENDER, K. R., & GREEN, K. P. (1996). Spectral discontinuities and the vowel length effect. *Perception & Psychophysics*, **58**, 1005-1014.

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., & DEXTER, E. R. (1988). Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception & Performance*, **14**, 369-378.

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.

MILLER, J. L., & VOLAITIS, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, **46**, 505-512.

MILLER, J. L., & WAYLAND, S. C. (1993). Limits on the limitations of context-conditioned effects in the perception of [b] and [w]. *Perception & Psychophysics*, **54**, 205-210.

NEWMAN, R. S., & SAWUSCH, J. R. (1996). Perceptual normalization for speaking rate: Effects of temporal distance. *Perception & Psychophysics*, **58**, 540-560.

PATTERSON, R. D., HOLDSWORTH, J., & ALLERHAND, M. (1992). Auditory models as preprocessors for speech recognition. In M. E. H. Schouten (Ed.), *The auditory processing of speech: From the auditory periphery to words* (pp. 67-83). Berlin: Mouton de Gruyter.

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.

PISONI, D. B., & SAWUSCH, J. R. (1975). Some stages of processing in speech perception. In A. Cohen & S. G. Nooteboom (Eds.), *Structure and process in speech perception* (pp. 16-34). New York: Springer-Verlag.

REMEZ, R. E., RUBIN, P. E., BERNS, S. M., PARDO, J. S., & LANG, J. M. (1994). On the perceptual organization of speech. *Psychological Review*, **101**, 129-156.

SAMUEL, A. (1986). Red herring detectors and speech perception: In defense of selective adaptation. *Cognitive Psychology*, **18**, 452-499.

SAWUSCH, J. R. (1976). Selective adaptation effects on end-point stimuli in a speech series. *Perception & Psychophysics*, **20**, 61-65.

SHINN, P. C., BLUMSTEIN, S. E., & JONGMAN, A. (1985). Limitations of context conditioned effects in the perception of [b] and [w]. *Perception & Psychophysics*, **38**, 397-407.

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

NOTES

1. In previous research, a number of different terms have been used to refer to the influence of speaking rate on perception. Lotto et al. (1996), among others, have referred to the influence of duration variation in the succeeding vowel on perception of a syllable initial consonant as a "vowel length effect." Miller (1981) refers to the effects as "speaking rate." Fowler (1990) used the more general terminology "event rate," to emphasize that all perception of events was normalized for the rate of events. Since all of the experiments that are reported here used speech stimuli and the segments whose durations were varied included consonants as well as vowels, we have chosen the terms *speaking rate* and *speaking-rate normalization*. These terms also highlight the ecological rationale for how some of the variation in segment duration comes about in spoken language and its relevance to the listener.

2. The percentage data are the result of collapsing across rating responses within each phonetic category. Ratings of 1, 2, and 3 were treated as "b" responses, and ratings of 4, 5, and 6 were treated as "p" responses.

3. Two-tailed, dependent *t* tests were used in all cases except between-group comparisons in Experiments 3 and 4, in which two-tailed, independent *t* tests were used.

4. The results of Experiment 1 were replicated with /bʌl/–/pʌl/ series that were created by digitally removing the final /z/ from the stimuli in Experiment 1. Since the data for these two sets of stimuli precisely parallel one another, only the one set of results is reported here.

5. Two-tailed, independent *t* tests were used here for the between-groups comparison.

(Manuscript received July 7, 1997;
revision accepted for publication December 6, 1998.)