Journal of Experimental Psychology:
Human Perception and Performance
1985, Vol. 11, No. 6, 799-813

0096-1523/85/$00.75

Perceptual Coherence of Speech: Stability of Silence-Cued Stop Consonants

Bruno H. Repp
Haskins Laboratories, New Haven, Connecticut

A series of experiments was conducted to examine the perceptual stability of stop consonants cued by silence alone, as when [s] + silence + [læt] is perceived as *splat*. Following a replication of this perceptual integration phenomenon (Experiment 1), attempts were made to block it by instructing subjects to disregard the initial [s] and to focus instead on the onset of the following signal, which was varied from [plæt] to [læt]. However, these instructions had little effect at short silence durations (Experiment 2), and they reduced stop percepts for only 2 subjects at longer silence durations (Experiment 3). That is, subjects were generally unable to voluntarily dissociate the [s] noise from the following signal and thus to perceive the silent interval as silence rather than as a carrier of phonetic information. A low-uncertainty paradigm facilitated the task somewhat (Experiment 4). However, when the [s] frication was replaced with broadband noise (Experiment 5), listeners had no trouble at all in the selective-attention task, except at very short silence durations (<40 ms). This last finding suggests that, except for the shortest durations, the effect of silence on phonetic perception does not arise at the level of psychoacoustic stimulus interactions. Rather, the results support the hypothesis that perceptual integration of speech components, including silence, is a largely obligatory perceptual function driven by the listener's tacit knowledge of phonetic regularities.

When listening to speech we perceive a coherent stream of sound, not a sequence of clicks, whistles, buzzes, and hisses. In view of the many abrupt changes of excitation and spectral structure that take place in normal speech, this apparent auditory coherence might seem like a remarkable perceptual accomplishment. However, it may well reflect the fact that the ordinary listener's attention is not focused on the detailed physical properties of the speech signal but on the underlying, linguistically relevant information. That is, auditory coherence of speech may be *inferred* from the perceived actual continuity of certain underlying articulatory events. If so, then there may be a more analytic level of perception that is sensitive to physical discontinuities in the speech signal.

This research was supported by National Institute of Child Health and Human Development Grant HD-01994 and Biomedical Research Support Grant RR-05596 to Haskins Laboratories. I am grateful to Richard E. Pastore for extensive critical comments that helped improve the article.

Requests for reprints should be sent to Bruno H. Repp, Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511-6695.

Speech does possess certain acoustic features that promote auditory coherence of otherwise disparate signal portions. For example, formant transitions have been considered to provide a kind of "perceptual glue" that holds successive sounds together and helps preserve their temporal order (Cole & Scott, 1973; Dorman, Cutting, & Raphael, 1975). This can hardly be the whole story, however. If perceptual coherence and integration were determined entirely by properties of the acoustic signal and their auditory transforms, it would be impossible for a listener to deliberately decompose the speech signal into its components. Nevertheless, this is possible, at least to a certain extent, by focusing one's attention on the level of auditory qualities (see e.g., Pilch, 1979). For example, it is not difficult even for a naive listener to selectively attend to the series of high-pitched hisses that represent repeated occurrences of [s] in the speech stream. Under special conditions, the perceptual isolation of such auditory components may be facilitated: Cole and Scott (1973) rapidly repeated the syllable [sa] over and over, and listeners soon reported hearing two separate streams of sounds, one consisting of hisses (the fricative noises)

and the other of syllables sounding like [ta] (the vowel with its initial formant transitions). In this unnatural situation, the segregation may take place at a relatively early perceptual stage; similar "streaming" can be induced in repetitive multicomponent nonspeech signals (Bregman, 1978).

Under more natural circumstances, the perceptual integration of certain disparate acoustic components of speech may still not be completely obligatory, though it reflects the normal mode of speech perception. If perceptual integration of these speech components could be disengaged by manipulating listeners' interpretation of the stimulus, this would suggest that the normally perceived coherence of the speech signal is contingent on a nonobligatory, central function characteristic of phonetic perception. If the integrative function proved difficult to disengage, and if low-level psychoacoustic interactions can be ruled out as the cause of the integration, then the conclusion would be that perceptual integration of speech components is not only a characteristic but also an obligatory function of phonetic perception.[1]

Evidence in favor of the hypothesis that certain types of perceptual integration are speech specific has been obtained in several recent studies concerned with "trading relations" among acoustic cues. Thus, Best, Morrongiello, and Robson (1981) have shown that, in noise-plus-sinewave analogs of utterances of the type *say* versus *stay,* the silent closure interval following the noise and the onset frequency of the tone mimicking the first formant (F1) both contribute to a stop consonant percept as long as the stimuli are perceived as speech; however, when the stimuli are perceived as nonspeech, the two acoustic cues are no longer integrated and are perceived as unrelated auditory properties. In another study, Repp (1981) trained subjects to discriminate the pitch of fricative noises preceding different vowels containing one of two sets of formant transitions. There was no effect of the vocalic context on the subjects' pitch judgments, even though the phonetic identification of the fricative consonant was influenced by both vowel quality and formant transitions. Furthermore, Dorman, Raphael, and Liberman (1979) and Rakerd, Dechovitz, and Verbrugge (1982) experimented with utterances whose precise phonetic interpretation depended on the duration of a silent closure interval occurring at a syllable boundary. When either fundamental frequency (Dorman et al., 1979) or the intonation contour (Rakerd et al., 1982) was changed abruptly across syllables, the silence lost its perceptual effect. Although spectral discontinuity could have played a role here, circumstantial evidence suggests that subjects' perception of one versus two speakers or utterances was responsible for the effect. Thus, the studies cited all provide evidence for a central level of perceptual integration that can be disengaged in at least three ways: by leaving the speech mode altogether, by selectively attending to specific auditory properties of the speech signal, or by perceiving a change of source or of linguistic structure.

In the present research, the focus is on the perceptual integration occurring in [spl] clusters. Acoustic cues to the perception of a labial stop consonant in this context include, first and foremost, an interval of silence following the [s] noise (Bastian, Eimas, & Liberman, 1961; Fitch, Halwes, Erickson, & Liberman, 1980), but also spectral changes in the fricative noise and the amplitude contour at noise offset (Summerfield, Bailey, Seton, & Dorman, 1981), the duration of the [s] noise (Repp, 1984c), the presence and amplitude of a release burst following the silent closure (Repp, 1984b, 1984d), formant onset frequencies and transitions in the following voiced portion (Fitch et al., 1980; see also Bailey & Summerfield, 1980), and the duration and possibly the amplitude envelope of the voiced portion (Repp, 1984c). Of special interest here is the the finding (Dorman et al., 1979) that a percept of

[1] The question posed here is similar in many ways to that underlying categorical perception research (see Repp, 1984a), but the methodology is different. Categorical perception experiments examine subjects' ability to discriminate stimulus differences within phonetic categories; here, the focus is on listeners' ability to ignore one part of a stimulus (a skill that may play a role in some discrimination tasks). Both tasks are difficult because listeners tend to adhere to their habitual mode of phonetic perception, which is categorical and integrative. No claim is made here that this *type* of perceptual mode is specific to speech; it is called "phonetic" only because the stimuli happen to be speech. That being so, however, many specific instances of perceptual integration may indeed be speech specific, simply because they have no parallels in other domains of experience.

split can be elicited by simply concatenating an [s] noise and a [lIt] syllable, with an appropriate interval of silence (about 100–300 ms) in between; in other words, in this context silence alone can be a sufficient cue for the perception of a *p,* as long as there are no contradictory cues from the surrounding signal portions. Because neither of the energy-carrying signal portions in isolation contains sufficient cues to a *p,* and the silence by itself naturally does not either, the stop consonant percept in this case is a pure product of perceptual integration over time and thus constitutes an ideal test case for our purposes. (Throughout this article, phonetic symbols in brackets denote stimuli or the speaker's intentions, whereas orthographic symbols in italics refer to responses or the listeners' percepts.)

The question addressed in the present study is how robust is this perceptual integration effect—that is, can a listener deliberately avoid the stop consonant percept and hear the stimulus components the way they sound in isolation, for example, as *s* followed by *lit?* This question is not unreasonable because a stop cued by silence alone does not sound perfectly natural and might be expected to be perceptually unstable, almost an illusion. The answer to the question also bears on two contrasting hypotheses that have been put forward to account for perceptual integration and cue trading relations in phonetic perception (see Pastore, 1981; Repp, 1982): If these phenomena are a function of purely psychoacoustic stimulus properties that emerge in peripheral auditory processing, then it should be extremely difficult to disengage them through acts of selective attention or linguistic restructuring. If they are a function of speech-specific mechanisms, however, it might be possible to change them by manipulating listeners' interpretation of the stimulus, without necessarily leaving the speech mode. A positive result would simultaneously refute the psychoacoustic hypothesis and support the existence of a special integrative level of perception, whereas a negative result, to be interpretable, would require an additional demonstration that psychoacoustic interactions are not the cause of the subjects' difficulty.

Accordingly, this article reports several attempts to "get rid of the stop" in subjects' perception of [s] + silence + [lIt] = *split* type utterances by directing their attention to the stimulus portion following the silence. A replication of the basic phenomenon of silence-cued stop consonant perception (Experiment 1) is followed by experiments that investigate the effect of selective attention instructions for stimuli with different absolute silence durations (Experiments 2 and 3), and with some subsequent changes in test format to reduce stimulus uncertainty (Experiment 4). Because, as will be seen, the stop consonant percepts proved unexpectedly resistant to these manipulations, the last experiment (Experiment 5) aimed at ruling out psychoacoustic interactions as the cause of the silence-cued stop percept. On the assumption that this last study succeeded in its aim, the conclusion is that perceptual integration of speech components, in this instance at least, is a relatively compulsory function of phonetic perception.

Experiment 1

Experiment 1 was an attempt to replicate an earlier striking demonstration of the perceptual integration phenomenon of interest, owing to Dorman et al. (1979, Experiment 3). These authors concatenated natural [s] and [lIt] utterances that had been recorded in isolation and that were considered to contain no traces of any [p]. When the silent interval between the stimulus components was shorter than 60 ms, listeners uniformly reported *slit.* At silent intervals between 80 ms and 450 ms, however, listeners reported predominantly *split,* with a maximum of over 90% around 300 ms of silence. This optimal closure interval was much longer than a typical [p] closure in this context (about 90 ms; see Morse, Eilers, & Gavin, 1982); moreover, it took as much as 650 ms of silence before subjects uniformly reported hearing *s-lit* (i.e., *s* followed by *lit*), rather than *split.* Because the *p* percepts in such stimuli are sometimes not very convincing, a replication of the Dorman et al. study seemed advisable, to verify that their subjects' *p* percepts were not just phantoms.

The long optimal closure duration (300 ms) in the Dorman et al. experiment may have been due to perceptual compensation for the absence of cues to stop manner in the surrounding signal portions. However, there is also the possibility that the use of a wide range of closure durations (0–650 ms), combined with

a higher relative frequency of short intervals, promoted a bias toward reporting *split* at atypically long closure durations. Therefore, two different stimulus ranges were employed here to assess the effect of this variable on the *sl–spl* and *spl–s-l* boundaries. The stimuli in this part of the experiment (1A) began with a fricative noise that contained some positive stop-manner cues and that was also used in Experiments 2–4. To approximate the conditions of the Dorman et al. (1979) study even more closely, the test employing a wide range of closure durations was later repeated (1B) using a fricative noise without positive stop manner cues.

Method

Subjects. Nineteen paid volunteers served as subjects, 10 in Experiment 1A and 9 in 1B. They were Yale undergraduates and native speakers of American English.

Stimuli. A female speaker recorded several repetitions of the utterance [splæt] (*splat*). One good token was low-pass filtered (−3 dB at 9.6 kHz, −55 dB at 10 kHz) and digitized at a 20 kHz sampling rate. Because this speaker's fricative noises contained significant energy at frequencies above 10 kHz, which caused some digitization artifacts, digitization and subsequent recording of audio tapes were done at half speed. The [s] noise was 125 ms long. The silent closure interval and the initial 11.5 ms of the following stimulus portion, corresponding to the labial release burst (and perhaps including a weak first glottal pulse), were removed. The remaining portion in isolation elicited over 90 percent *lat* responses (see Experiments 2 and 3, pretest). Thus it did not seem to contain any sufficient cues to a preceding labial stop. The fricative noise from [splæt], however, may have contained such cues. Therefore, Experiment 1B used a fricative noise derived from an utterance of [slæt] produced by the same speaker, 190 ms in duration.[2]

Two identification tests were assembled for Experiment 1A. In one, the [s] noise was followed by the [læt] portion at each of 14 different closure durations: 0, 20, 40, 60, 80, 100, 150, 200, 250, 300, 400, 500, 600, and 700 ms. This test was also duplicated in Experiment 1B with the different [s] noise. In the other test used in Experiment 1A, only the 9 closure durations up to 250 ms were included. Each test contained 10 successive randomizations of the stimuli, with interstimulus intervals (ISIs) of 2.5 s and interblock intervals of 6 s. The stimulus sequences were recorded at half speed on audio tape using high-quality equipment, with closure durations and ISIs at twice their nominal values; thus they had the intended values at playback speed.

Procedure. The subjects listened individually or in small groups over TDH-39 earphones in a quiet room. They identified each stimulus in writing as beginning with *sl*, *spl*, or *s-l* (i.e., *s* followed by silence and *lat*).

Results and Discussion

The average percentage of stop (i.e., *spl*) responses is plotted in Figure 1 as a function of closure duration (on a logarithmic scale). Filled and open circles represent the data from the two conditions of Experiment 1A. It is evident that stimuli with short closure intervals were perceived as beginning with *sl*. The *sl–spl* boundary fell at about 70 ms of closure duration. *Spl* responses were obtained for closure intervals ranging from 60 ms to 300 ms, with the peak occurring at 100–150 ms of silence. At longer closure durations, an increasing number of *s-l* responses was obtained.[3] Truncation of the stimulus range did not affect the *sl–spl* boundary but shortened the *spl–s-l* boundary by about 80 ms. At closure intervals of 200 and 250 ms combined, there were significantly fewer *spl* responses in the narrow-range than in the wide-range condition: one-way repeated-measures ANOVA, $F(1, 9) = 26.25$, $p = .0006$. The *spl–s-l* distinction is not very categorical and was expected to be affected by stimulus range. The fixed *sl–spl* boundary, on the other hand, suggests that the silence-cued *p* percepts at closure durations below 150 ms were relatively stable and insensitive to range effects.

The results from Experiment 1B are represented by the triangles in Figure 1. They confirm that the fricative noise in Experiment 1A contained some positive stop-manner cues. The *sl–spl* boundary was at a longer silent interval here (close to 100 ms), the maximum of *spl* responses was less pronounced and occurred at longer silences (150–250) ms), and the subjects experienced more uncertainty at the longest intervals, giving more *spl* responses here than in Experiment 1A. All these differences are at least in part due to the longer duration of the fricative noise used in Experiment

[2] To be sure, the [s] noise must not be too long, and its offset and the [l] onset not too gradual; otherwise, no stop percepts will be obtained. The presence of stop-manner cues in the [s] noise was irrelevant in Experiments 2–4, because subjects' attention was directed toward the stimulus portion following the silence. As far as that portion is concerned, it was sufficient that it not elicit any stop percepts in isolation. No claim is being made that either signal portion contained no cues whatsoever to stop consonant perception (see also Footnote 5).

[3] Some subjects, especially in Experiment 1A, spontaneously gave *s*^{p}*-l* responses, indicating that they detected stop manner cues in the frication, while at the same time perceiving a gap between the [s] and the rest of the stimulus. These responses were treated as equivalent to *s-l;* thus they are not included in the *spl* percentages plotted in Figure 1.

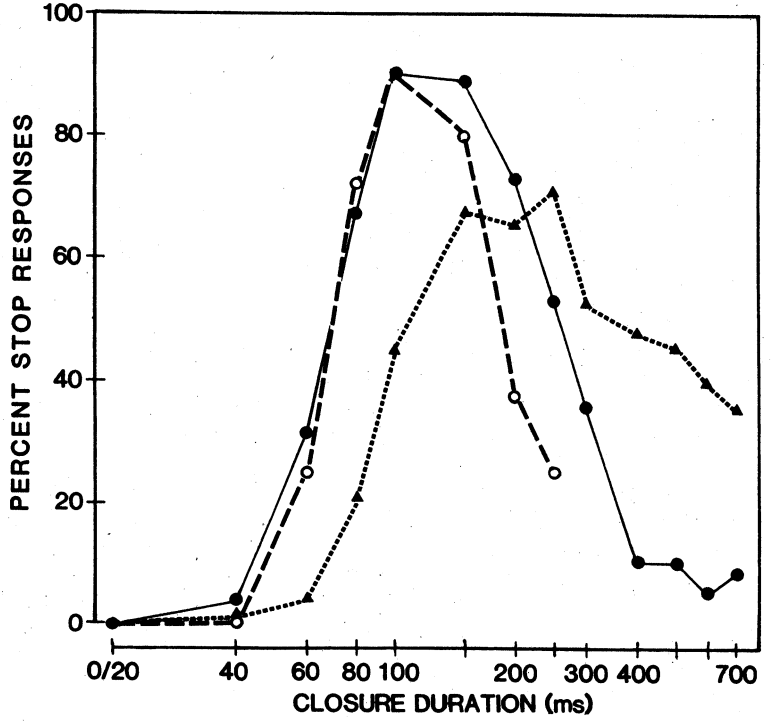

Figure 1. Percent stop (i.e., *spl*) responses as a function of closure duration (in milliseconds) in Experiments 1A (filled and open circles) and 1B (triangles). (The open circles represent the results from the condition with a reduced range of closure durations.)

1B (cf. Repp, 1984c), but spectral differences at noise offset may also have played a role.

The general pattern of these results is consistent with the findings of Dorman et al. (1979). That is, even without any strong stop-manner cues in the surrounding signal portions, *p* percepts are obtained in a certain range of closure durations. The 70-ms boundary separating *sl* from *spl* responses in Experiment 1A is very close to that obtained by Dorman et al. The results of Experiment 1B resemble the Dorman et al. findings in terms of the optimal closure duration for hearing *p;* they suggest that listeners need exceptionally long closure intervals for stop perception when closure duration is the sole stop-manner cue, perhaps to compensate for the absence of other cues. The optimal closure duration in Experiment 1A, however, is shorter than in the Dorman et al. study, and so is the longest closure at which *p* percepts were still obtained. These results are somewhat closer to reflecting the typical closure durations observed in natural speech.

Experiment 2

Even though Experiment 1 demonstrated the perceptual reality of silence-cued stop consonants, it did not tell us how obligatory these percepts are. The fact that the percentage of *spl* responses did not reach 100% at any closure duration suggests a certain amount of ambiguity. Subjects may also have felt compelled to apply the *spl* response category supplied by the experimenter. How easy would it be to convince listeners that what they are hearing is really *s* followed by *lat* (either *slat* or *s-lat*), and not *splat?* The technique adopted to investigate this issue in the following experiments was to construct a continuum from [plæt] to [læt], to prefix it with an [s] noise plus a varying silent interval, and to instruct listeners either to identify the whole stimulus (integrative condition) or to ignore the [s] and identify only the part following the silence (analytic or selective-attention condition). Because the test included clear [splæt] (i.e., [s] + silence +

[plæt]) stimuli, there was no pressure to give any stop responses to [s] + silence + [læt] stimuli. On the contrary, contrast among stimuli in the test should reduce any such tendencies. The analytic instructions were reinforced by the use of the response *b* (actually, *bl*) for the syllable-initial labial stop, if one was perceived, as contrasted with *p* (actually, *spl*) in the integrative condition.[4] Note that the analytic instructions required a perceptual reinterpretation within the linguistic domain, without leaving the speech mode (although thinking of the [s] as some extraneous noise might help). If the instructions were effective, fewer stop responses should be obtained in the analytic than in the integrative condition at closure durations beyond 100 ms, particularly for those stimuli whose final portion was perceived as beginning with *l* in isolation.

The *stop generation effect* discussed so far—the introduction of a stop percept by appropriate amounts of silence in the absence of any other sufficient cues—may be contrasted with a *stop suppression effect* due to an absence of a sufficient interval of silence in the presence of other sufficient cues. Thus, earlier observations (e.g., Fitch et al., 1980; Mann & Repp, 1980) lead to the expectation that stimuli perceived as beginning with *bl* in isolation will lead to *sl* responses when preceded by an [s] noise with little or no silence in between. If this stop suppression effect reflected the same higher level, integrative mechanisms as the stop generation effect, and if analytic listening instructions were effective, then *more* stop responses should be obtained in the analytic than in the integrative condition at short closure durations, particularly for those stimuli whose final portion was perceived as beginning with *bl* in isolation.

Thus, the strongest prediction for Experiment 2 is that silent closure duration will have a marked effect on stop perception in the integrative listening condition but no effect at all in the analytic condition: Stimuli should be labeled as if there were no preceding [s]. However, apart from the fact that it is more realistic to expect only a more or less pronounced tendency in the predicted direction, the stop generation and suppression effects may well be differentially sensitive to attentional strategies. The stop suppression effect, which results from signal components occurring in close succession, is much more likely to involve auditory interactions (such as forward masking) than is the stop generation effect, which results from components that are more widely separated in time. If this notion is correct, then the prediction should be that selective attention instructions, if effective, will lead to a reduction of stop percepts at longer silences but not to an increase of stop percepts at short silences.

Method

Subjects. The same 10 subjects as in Experiment 1A participated.

Stimuli. A continuum from [plæt] to [læt] was constructed from the source utterance used in Experiment 1A, [splæt]. The original 11.5-ms labial release burst was truncated by 0, 2, 4, 7.5, or 11.5 ms, yielding five stimuli intended to range perceptually from *blat* to *lat* in the absence of a preceding [s].[5] The cutpoints were placed at zero-crossings in the digitized waveform. A brief *pretest* was assembled in which these five stimuli (without any preceding [s]) occurred 10 times in random sequence, with ISIs of 2.5 s.

Two additional identification tests were assembled. In one, designed for integrative listening, each stimulus from the [plæt]–[læt] continuum was preceded by [s] at silent intervals of 0, 40, 80, 120, and 160 ms, for a total of 25 stimuli that were recorded 10 times in random sequence with ISIs of 2.5 s. The other test, designed for analytic listening, contained 10 random sequences of the same 25 stimuli plus 10×2 replications of the 5 stimuli without a preceding [s] interspersed among them, resulting in ten 35-item blocks. The no-[s] stimuli were intended to remind the subjects of the stimulus portion to attend to and perhaps to facilitate selective attention.

Procedure. All subjects listened first to the tapes of Experiment 1A. Subsequently, in the same session, the integrative listening test was presented. As in Experiment 1A, the task was to label the stimuli as beginning with *sl* or *spl*. The pretest followed, with instructions to label the stimuli as beginning with *bl* or *l*. Finally, the analytic listening test was presented, in which the labels *bl* and *l* were again to be used. Subjects were told to ignore the [s], if

[4] The phonetic symbol [p] represents a voiceless unaspirated labial stop consonant, which in English orthography is rendered as *p* in some contexts (e.g., following a voiceless fricative in the same syllable) but as *b* in others.

[5] For the author and most subjects, excision of the natural labial release burst in [plæt] resulted in elimination of the stop percept. Some listeners, however, still claimed to hear a *b*, which may reflect a special sensitivity to weak coarticulatory cues in the [l] portion. These coarticulatory cues may reside in spectral or amplitude properties of the signal immediately following the release burst or, perhaps more likely, in the shorter duration of the [l] as compared to one articulated in absolute utterance-initial position. One additional subject in Experiment 2 and 2 additional subjects in Experiment 3 were excluded because they perceived all stimuli from the [plæt]–[læt] continuum as *blat*.

present, to the best of their ability. They were informed about the structure of the stimuli and about the perceptual effect to be avoided.

Results and Discussion

The [plæt]–[læt] continuum was perceived as intended. In the pretest, the average percentages of *bl* responses to the 5 stimuli were 100, 100, 90, 9, and 3, respectively. (Note the listeners' remarkable sensitivity to the 3.5-ms release burst cutback occurring between Stimuli 3 and 4; for comparable results, see Repp, 1984d, Experiment 1.) The same no-[s] stimuli interspersed in the analytic listening test received 99, 99, 92, 24, and 20 percent *bl* responses, respectively. Thus, Stimuli 4 and 5 were sometimes perceived as beginning with *bl* in this environment, but they still were clearly distinguished from Stimuli 1, 2, and 3, which sufficed for the purposes of this experiment.

In both the integrative and analytic listening conditions, stimuli with no closure silence at all never elicited labial stop responses. Clearly, analytic listening instructions were totally ineffective here—not an unexpected result. Therefore, those data were excluded from further analysis, reducing the number of closure durations to four. Figure 2 shows the percentages of labial stop responses in the two listening conditions as a function of closure duration and of stimulus number on the continuum. The responses to no-[s] stimuli in the analytic test are plotted on the far right.

It is evident that the response patterns in the integrative and analytic conditions were highly similar. A repeated-measures ANOVA showed the expected significant main effects of closure duration and stimulus continuum, and also an interaction between these factors (all $ps < .0001$), but no significant main effect of conditions. The conditions by closure duration interaction was significant, $F(3, 27) = 5.45$, $p < .005$, due to a slight reduction in labial stop percepts at the shorter closure durations in the analytic condition relative to the integrative condition, and a relative increase at the longest closure duration, where perceptual segregation of the [s] noise from the rest of the stimulus might have been expected to be relatively easier. This pattern of results is the opposite of the predicted one. Thus there is no evidence that the analytic listening instructions had the desired effect. Instead of selectively attending to the stimulus portion following the silence, the subjects apparently responded by parsing off the *s* and changing the *p* to *b* in their phonological (or orthographic) representation of the whole stimulus.

The peak rate of labial stop responses to Stimuli 4 and 5 preceded by [s] (about 70% at 120 ms of silence in both conditions) clearly exceeded that for Stimuli 4 and 5 in isolation, but was lower than that in Experiment 1A (about 90%). This may suggest unstable *p* percepts, but the results of the analytic condition do not bear this out. That is, the instability was only in the choice of response from one

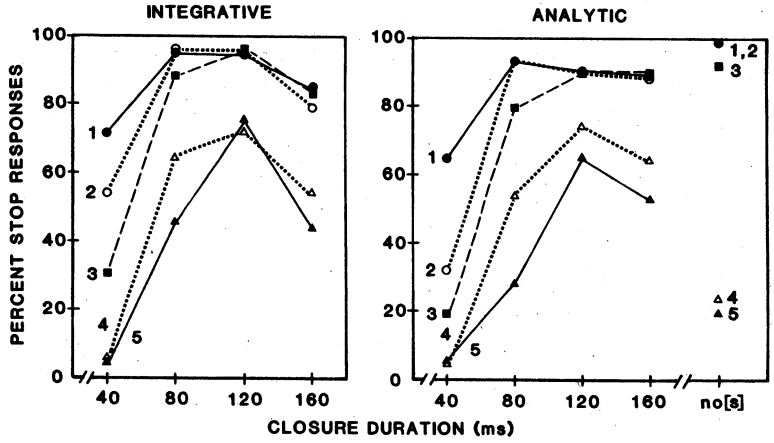

Figure 2. Percent stop responses in the integrative and analytic conditions of Experiment 2, separately for the five stimuli from the [plæt]–[læt] continuum. (Data for the 0-ms closure duration are omitted.)

trial to the next, not in the percept on which it was based.

It is interesting to note that Stimuli 1, 2, and 3, which tended to give very similar results at longer closures and in isolation (probably due to a ceiling effect), elicited different response rates at the 40-ms closure duration. In fact, an orderly trading relation can be seen between stimulus number (i.e., degree of release burst truncation) and silent closure duration, as previously demonstrated by Repp (1984d, Experiment 1) for alveolar stops in the *say–stay* contrast. The *sl–spl* boundary (50% intercept) ranged from approximately 30 ms (Stimulus 1, extrapolated) to over 90 ms of silence (Stimulus 5)—a remarkable range, considering that the release burst being truncated was only 11.5 ms long. A lot of silence was needed to compensate for the loss of a small piece of plosive noise.

Experiment 3

The results of Experiment 2 suggest that, at least without special training, subjects are unable to perceptually dissociate an [s] noise from the following speech signal. In part, this may have been due to the relatively short silent intervals used. In Experiment 3 the same issue was examined at longer closure durations, where selective attention to the stimulus portion following the [s] might be facilitated by the increased temporal separation and the consequent reduction of any potential auditory stimulus interactions across the silence. Only an analytic listening condition was used in Experiment 3, with the integrative identification data of Experiment 1A taken for comparison. Because the closure intervals used were all in the range beyond the stop suppression effect, the expectation was that stop responses would be reduced relative to Experiment 1A and would approximate the percentages for no-[s] stimuli.

Method

Subjects. Ten paid volunteers participated, 4 of whom had taken part in Experiments 1A and 2.

Stimuli. The test sequence contained the five stimuli from the [plæt]–[læt] continuum preceded by the [s] noise at silent intervals of 100, 150, 200, 250, 300, 400, and 500 ms. The resulting 35 stimuli were augmented by four repetitions of the 5 stimuli without preceding [s], and all 55 stimuli were recorded in five randomized orders with ISIs of 2.5 s. The pretest of Experiment 2 (no-[s] stimuli only) was also used.

Procedure. Six of the subjects first listened to the pretest, labeling each stimulus as beginning with *bl* or *l*. (The 4 remaining subjects had received the pretest in an earlier session in connection with Experiment 2.) Following the pretest, all subjects went through Experiment 4 (described below) before embarking on Experiment 3. The instructions were to ignore the initial [s], if present, and to label each stimulus as beginning with either *bl* or *l*. The subjects were informed about the purpose of the experiment and about the nature of the stimuli.

Results and Discussion

The average percentages of labial stop responses to the five stimuli in the pretest were 100, 100, 89, 16, and 10, respectively. For the same stimuli in the analytic identification test, subjects' average percentages were 99, 99, 78, 13, and 8. Unlike Experiment 2, there was no increase in *bl* responses to Stimuli 4 and 5 in the environment of stimuli with initial [s], perhaps because there were no contextual stimuli that sounded like *slat*.

Figure 3 plots *bl* responses to stimuli preceded by [s] as a function of silent closure duration. The response percentages for the interspersed no-[s] stimuli are plotted on the far right. Several patterns are evident in the results: (a) Stimuli 1, 2, and 3 elicited fewer stop responses when preceded by [s] than when pre-

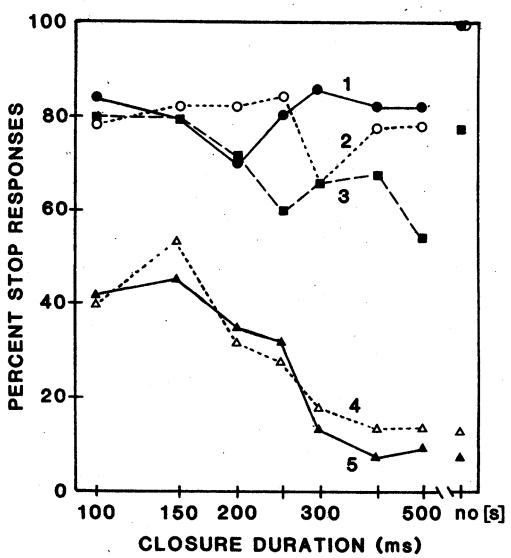

Figure 3. Percent stop responses as a function of closure duration (in milliseconds) in the analytic task that constituted Experiment 3.

sented in isolation. (b) At closure durations shorter than 300 ms, Stimuli 4 and 5 elicited more stop responses when preceded by [s] than when presented in isolation. (c) The percentage of stop responses increased as closure duration decreased, reaching a peak at 150 ms for Stimuli 3, 4, and 5. Responses to Stimuli 1 and 2, on the other hand, were not sensitive to changes in closure duration. In the analysis of variance, this was reflected in a significant closure duration by stimulus number interaction, $F(24, 216) = 2.09$, $p < .005$.

The main result of this study is the increase in stop responses when [læt]-like stimuli were preceded by [s] at closure durations of less than 300 ms. This increase resembles the results of Experiment 1A, obtained with Stimulus 5 in a standard (integrative) labeling task. Thus, as in Experiment 2, subjects were not able to get rid of stop percepts by ignoring the [s] precursor and focusing their attention on the onset of the stimulus portion following the closure silence. Some measure of success in the selective-attention task is indicated, perhaps, by the fact that stop responses to Stimulus 5 preceded by [s] reached a maximum of only 50%, whereas the same stimulus elicited as much as 90% stop responses in Experiment 1A. However, in the integrative condition of Experiment 2, there was also a relatively low percentage of stop responses to Stimulus 5 at comparable closure durations (about 60%). Moreover, because subjects had been told that a preceding [s] tended to generate labial stop percepts that were to be avoided, a bias against responding *bl* may have operated. This is strongly suggested by the lowered rate of *bl* responses (around 80%) to Stimuli 1 and 2 preceded by [s], which certainly would have been labeled *spl* 100% of the time in an integrative task. Thus, the effect of the selective-attention instructions on perceptual organization may actually have been rather small (see discussion of Figure 5 below).

This conclusion must be qualified immediately, however, because closer inspection of the data revealed considerable individual differences (in contrast to Experiment 2). In particular, there were 2 (out of 10) subjects who appeared to be totally successful in ignoring the [s] precursor, whose labeling responses were not influenced by closure duration, and who exhibited no response bias.[6] Four or 5 other subjects showed patterns of which Figure 3 is representative, and the remaining subjects exhibited idiosyncratic patterns and showed large response biases against *bl*. These individual differences are reminiscent of those observed by Repp (1981) in a study that required listeners to perceptually dissociate a fricative noise from a following vocalic portion. The success of 2 subjects in the present study suggests that analytic listening to speech components is not an impossible task, at least not when the closure durations are fairly long. These observations are consistent with the hypothesis that silence-induced stop percepts are products of a higher level integrative process, and not of psychoacoustic interactions among stimulus components. Nevertheless, the fact remains that the perceptual strategy for performing the selective attention task was not available to most listeners, even though they had received a moderate amount of training by performing the low-uncertainty task of Experiment 4 before Experiment 3.

Experiment 4

Experiments 2 and 3 have provided only very limited evidence that subjects can perceptually dissociate the two stimulus components, even at relatively long temporal separations. In part, subjects' difficulties in carrying out the selective-attention instructions may reflect ingrained habits of integrative phonetic processing when listening to speech. At very short temporal separations, however, psychoacoustic interactions among the stimulus components may come into play, and these interactions may be truly impossible to disengage by acts of selective attention or other perceptual strategies. To investigate this issue further,

[6] One of these 2 subjects had participated in Experiments 1A and 2. In the labeling task of Experiment 1A, which used Stimulus 5 of the [plæt]–[læt] continuum, she gave 90% stop responses at closure durations of 100 and 150 ms. In Experiment 2, for Stimuli 4 and 5 with 120 and 160 ms of silence, she gave 63% stop responses in the integrative condition, 70% in the analytic condition, and 0% when there was no preceding [s] noise. In Experiment 3, however, she gave not a single stop response to the same stimuli with silent intervals of 100 ms and 150 ms. Clearly, she had discovered an effective selective attention strategy in Experiment 3, perhaps as a result of going through the task of Experiment 4 (where she likewise did not give any stop responses in the comparable stimulus conditions).

when a burst of white noise is substituted for the [s] frication, provided that the energy of the white noise is not substantially below that of the frication. From the viewpoint of phonetic perception, however, the white noise is less speechlike and therefore should be more easily filtered out in a selective-attention task. If the *sl–spl* boundary does not rest on a psychoacoustic interaction, subjects should be more successful in identifying *blat* and *lat* when white noise replaces the [s] precursor.

Method

Subjects. The same 9 subjects as in Experiment 1B participated.

Stimuli. The five stimuli from the [plæt]–[læt] continuum were again used. Instead of a natural [s] noise, however, a burst of white noise was used as a precursor. The white noise was recorded from a General Radio 1390-A random noise generator, low-pass filtered and digitized at half speed at a 20 kHz sampling rate. It differed from the [s] noise used previously (Experiment 1A and Experiment 2–4) in three respects: (a) Its duration was 200 ms, versus 125 ms for the [s] noise. (b) It was gated on and off abruptly, whereas the [s] noise had gradual on- and offsets. (c) It had a flat spectrum, whereas the spectrum of the [s] noise had a pronounced peak at about 8.6 kHz, which projected by about 20 dB above a relative energy plateau ranging from 4 to 10 kHz. The spectral energy of the white noise matched that of the plateau; its energy was higher than that of the [s] noise below 4 kHz and above 10 kHz, and lower between about 8–9 kHz. Its energy at offset was considerably higher than that of the fricative noise across the whole spectrum. All these differences led to the expectation that the white noise would have a more pronounced forward masking effect than the [s] noise, if such a psychoacoustic effect is involved at all. On the other hand, relatively long duration, abrupt offset, and flat spectrum are all uncharacteristic of natural fricative noises preceding a stop closure.[9]

The stimulus tape matched that of the analytic condition in Experiment 2. That is, silent intervals ranged from 0 to 160 ms, and no-noise stimuli were interspersed.

Procedure. All subjects listened first to the tape of Experiment 1B (an integrative labeling task) and then to the pretest, as used in Experiments 2 and 3 (stimuli without preceding noise). Instructions for the main test were the same as in the analytic condition of Experiment 2: Ignore the noise and label the stimuli as beginning with *bl* or *l*.

Results and Discussion

Figure 6 shows the results, which are strikingly different from those of Experiment 2 (cf. Figure 2, right-hand panel). Over the range from 40 to 160 ms of silence, the white noise precursor had no effect at all on subjects' ability to identify the stimuli from the [plæt]–[læt] continuum, except for introducing a slight bias against stop responses.[10] In particular, the

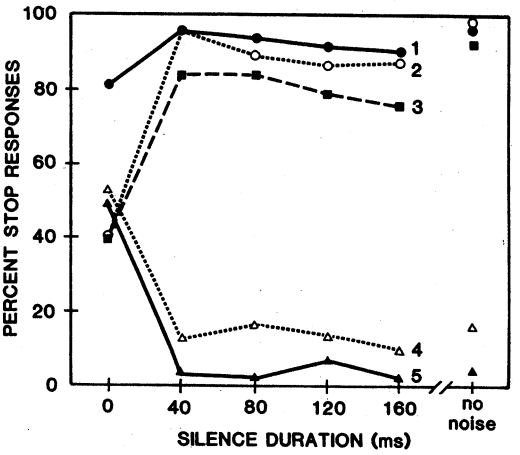

Figure 6. Percent stop responses as a function of silence duration (in milliseconds) in Experiment 5.

white noise did not induce any stop percepts when it preceded Stimuli 4 and 5. Only when there was no silent interval between the noise and the speech did the noise exert a perceptual effect, rendering Stimuli 2–5 indiscriminable, while Stimulus 1 continued to receive a higher rate of stop responses. Note also that, in this condition, subjects were equally willing to respond *bl* or *l*, whereas in the corresponding condition of Experiment 2 (not shown in Figure 2) responses were exclusively *l*. This suggests that the subjects in Experiment 5 considered the white noise as an extraneous signal that might obscure stop consonant cues present in the speech signal, whereas the subjects in Experiment 2 perceived the [s] noise as part of the utterance, even when asked not to do so, and thus were unwilling to consider the possibility of an inaudible stop consonant.

It seems extremely unlikely that spectral or other properties of the white noise were responsible for its reduced masking power, because it was a more powerful signal than the [s] noise by most acoustic criteria. Although

[9] Of course, the white noise did not sound like a fricative noise (at best, it sounded remotely [f]-like). For this reason, an integrative listening condition, in which subjects try to interpret the noise as a fricative, was not considered. The point here is that, if psychoacoustic interactions are involved, they should not depend on the speechlikeness of the noise.

[10] This tendency, as well as its apparent increase with closure duration, was due to 2 subjects' data only.

the [s] noise was more intense between 8 and 9 kHz, the spectral peaks of the labial release burst were in a region (below 4.5 kHz) where the white noise exceeded the [s] noise in energy. Therefore, the results suggest that psychoacoustic interference (i.e., forward masking) was involved only at the very shortest closure intervals (less than 40 ms). Consequently, the reduction in stop responses when an [s] noise precedes [plæt] stimuli by 40–80 ms (see Figure 2) probably does not represent psychoacoustic interference, but rather a specifically phonetic effect reflecting the listener's tacit knowledge about the minimal permissible duration of stop consonant closures in this context. Apparently, listeners are compelled to apply this knowledge as long as they perceive a coherent stream of speech. This conclusion is consistent with that reached by Pastore et al. (1984), and it suggests that the two effects of closure silence (stop suppression at short durations, stop generation at longer durations) can be accounted for within a single theoretial framework, that of perception in the "speech mode" (Liberman, 1982; Repp, 1982).

Summary and Conclusions

The present series of studies addressed the question of the origin of the auditory coherence of speech by focusing on one particularly striking phenomenon—that of silence-cued labial stop consonants in fricative-liquid context. This phenomenon illustrates both the coherence of acoustically heterogeneous speech components in general and the perceptual integration of disparate cues to the perception of a particular phonetic contrast. Between the fricative noise and the resonances resulting from production of the liquid consonant, there is an abrupt change in the nature and location of the sound source (from voiceless and dental to voiced and laryngeal) and in spectral composition (from higher to lower frequencies). Nevertheless, with or without an intervening brief silent interval, listeners usually perceive both sounds as part of a coherent speech stream. This coherence in turn gives rise to a stop consonant percept when a silent interval of appropriate duration (roughly, 80–200 ms) is present. Thus the silence itself becomes part of the speech stream; rather than interrupting the continuity and contributing to the perceptual segregation of acoustically disparate signal components, the silence functions as a carrier of phonetic information. Only when the silence duration clearly exceeds the acceptable limits of a stop consonant closure does it lead to perceptual segregation of the signal components.

It was hypothesized that the integrative function that gives rise to these phenomena is a characteristic of perception in the speech mode—that is, of perceiving the information that is most useful for linguistic communication. One way of testing this hypothesis would be to lead listeners to perceive the same stimuli as either speech or nonspeech. Some evidence favoring the hypothesis has already been obtained using variants of that method (Best et al., 1981; Repp, 1981). A somewhat different approach was taken here. It was argued that, if perceptual integration of the form studied here is a speech-specific function, it might be possible to influence its operation by directly manipulating the listeners' interpretation of the speech stimulus, staying entirely within the speech mode. The success of this approach was not guaranteed, of course, because manipulation of listeners' strategies through instructions may simply be ineffective. In the absence of a convincing psychoacoustic explanation for the perceptual integration of speech components, however, negative findings may tell us that certain perceptual strategies are not easily modified or abandoned, not that they are not speech specific.

In a series of experiments (Experiments 2–5) following a basic demonstration of silence-cued stop consonants (Experiment 1), it was attempted to alter subjects' interpretation of the stimulus by instructing them to mentally separate the fricative noise from the following signal portion. The relative ineffectiveness of the selective-attention instructions with stimuli of seemingly minimal acoustic coherence is interpreted as evidence for the relative stability of the perceptual integration function. Experiment 3 indicated, however, that some subjects can be successful in this task, and Experiment 4 showed that a low-uncertainty paradigm also facilitates selective attention. These results parallel those obtained in studies of categorical perception (see Repp, 1984a, for a review), where subjects frequently need to disengage or ignore another basic function of the speech mode, that of phonetic classification, in order

to discriminate speech stimuli. In these studies, it seems that success in within-category discrimination often requires perceptual strategies that operate outside the speech mode. The present task, too, could in principle have been accomplished by listening specifically for the release burst, though there was no evidence that the subjects used this "auditory" strategy. Rather, the few successful subjects appeared to be able to do what the instructions asked for: to ignore the fricative noise and listen to the remainder of the stimulus as speech—a skill that trained phoneticians presumably would have in their repertoire.

One way of ignoring a fricative noise is to think of it as a nonspeech hiss arising from a source outside the speaker's vocal tract. That this strategy could be effective is clear from Experiment 5, which, by substituting a nonspeech noise for the frication, actually created the situation that subjects otherwise might try to imagine. The ease with which the subjects carried out the selective-attention instructions in this situation argues against a psychoacoustic account of perceptual integration and of the effect of the silent interval on stop consonant perception. This latter effect has two aspects, which were termed *stop suppression* (short intervals) and *stop generation* (longer intervals). On the basis of the results of Experiment 5, it was concluded that both of these effects are likely reflections of speech-specific perceptual criteria, with only the suppression effect at extremely short closure silences having a psychoacoustic origin.[11]

In conclusion, then, the results of the present experiments are consistent with a theoretical view of speech perception that postulates a number of specific—though not necessarily unique—functions. These perceptual functions, which include the perceptual integration of speech components, are assumed to be driven by an internal representation of the regularities of spoken language. How this representation should be characterized and how it is acquired are fundamental questions for future research.

[11] Another possible auditory interaction that was not considered seriously here, but that may warrant some further investigation, is auditory short-term adaptation (see Delgutte & Kiang, 1984). The [s] precursor should adapt high-frequency neurons more than low-frequency neurons, so that the auditory response to the following signal portion would be more vigorous in the low-frequency regions, which might favor labial stop percepts. There are several problems with that hypothesis, however: (a) The long temporal range of the stop generation effect (Experiment 1) exceeds the range of auditory adaptation. (b) The stop suppression effect remains unexplained. (c) The ability of some subjects to disengage the stop generation effect argues against peripheral auditory factors. (d) The [s] noise spectrum is not differentiated enough in the low-frequency region to substantially alter the shape of the "auditory spectrum" at the onset of the following signal. (e) The stop generation effect is reduced by an increase in fricative noise duration (Experiment 1B; Repp, 1984c).

References

Bailey, P. J., & Summerfield, Q. (1980). Information in speech: Observations on the perception of [s]-stop clusters. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 536–563.

Bastian, J., Eimas, P., & Liberman, A. M. (1961). Identification and discrimination of a phonemic contrast induced by silent interval. *Journal of the Acoustical Society of America, 33,* 842(Abstract).

Best, C. T., Morrongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception & Psychophysics, 29,* 191–211.

Bregman, A. S. (1978). The formation of auditory streams. In J. Requin (Ed.), *Attention and performance VII* (pp. 63–76). New York: Wiley.

Cole, R. A., & Scott, B. (1973). Perception of temporal order in speech: The role of vowel transitions. *Canadian Journal of Psychology, 27,* 441–449.

Delgutte, B., & Kiang, N. Y. S. (1984). Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics. *Journal of the Acoustical Society of America, 75,* 897–907.

Dorman, M. F., Cutting, J. E., & Raphael, L. (1975). Perception of temporal order in vowel sequences with and without formant transitions. *Journal of Experimental Psychology: Human Perception and Performance, 1,* 121–129.

Dorman, M. F., Raphael, L. J., & Liberman, A. M. (1979). Some experiments on the sound of silence in phonetic perception. *Journal of the Acoustical Society of America, 65,* 1518–1532.

Fitch, H. L., Halwes, T., Erickson, D. M., & Liberman, A. M. (1980). Perceptual equivalence of two acoustic cues for stop consonant manner. *Perception & Psychophysics, 27,* 343–350.

Liberman, A. M. (1982). On finding that speech is special. *American Psychologist, 37,* 148–167.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]–[s] distinction. *Perception & Psychophysics, 28,* 213–228.

Morse, P. A., Eilers, R. E., & Gavin, W. J. (1982). The perception of the sound of silence in early infancy. *Child Development, 53,* 189–195.

Pastore, R. E. (1981). Possible psychoacoustic factors in speech perception. In P. D. Eimas & J. L. Miller (Eds.),

Perspectives on the study of speech (pp. 165–205). Hillsdale, NJ: Erlbaum.

Pastore, R. E., Szczesiul, R., & Rosenblum, L. (1984). Does silence simply separate speech components? *Journal of the Acoustical Society of America, 75,* 1904–1907.

Pilch, H. (1979). Auditory phonetics. *Word, 29,* 148–160.

Rakerd, B., Dechovitz, D. R., & Verbrugge, R. R. (1982). An effect of sentence finality on the phonetic significance of silence. *Language and Speech, 25,* 267–282.

Repp, B. H. (1981). Two strategies in fricative discrimination. *Perception & Psychophysics, 30,* 217–227.

Repp, B. H. (1982). Phonetic trading relations and context effects: New evidence for a phonetic mode of perception. *Psychological Bulletin, 92,* 81–110.

Repp, B. H. (1984a). Categorical perception: Issues, methods, findings. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 10, pp. 243–335). New York: Academic Press.

Repp, B. H. (1984b). Closure duration and release burst amplitude cues to stop consonant manner and place of articulation. *Language and Speech, 27,* 245–254.

Repp, B. H. (1984c). Effects of temporal stimulus properties on perception of the [sl]–[spl] distinction. *Phonetica, 41,* 117–124.

Repp, B. H. (1984d). The role of release bursts in the perception of [s]-stop clusters. *Journal of the Acoustical Society of America, 75,* 1219–1230.

Summerfield, Q., Bailey, P. J., Seton, J., & Dorman, M. F. (1981). Fricative envelope parameters and silent intervals in distinguishing 'slit' and 'split.' *Phonetica, 38,* 181–192.

Received November 20, 1984
Revision received July 1, 1985 ■