PERCEPTION AND PRODUCTION OF TWO-STOP-CONSONANT SEQUENCES

Bruno H. Repp

Abstract. The duration of the silent closure interval required to perceive two stop consonants in a VC₁C₂V sequence depends, to some extent, on their places of articulation. In production, too, the duration of the closure interval varies systematically with place. However, there appears to be little relation between the patterns of variability in production and in perception. Moreover, two analogous perceptual experiments—one using synthetic stimuli, the other, natural speech—yield quite different results. Thus, variations in the amount of closure required to perceive two successive stops seem to be governed by stimulus-specific acoustic factors, not by an internal representation of articulatory patterns or constraints. This conclusion is further supported by the unexpected finding that some listeners do not require any closure interval for accurate perception of both stops.

INTRODUCTION

Lisker (1957) first reported that, when the waveforms of naturally produced /ræɡ/ (with /g/ unreleased) and /bzd/ are abutted without any intervening silence (which serves to indicate oral closure), listeners hear /ræbzd/—that is, they fail to perceive the first (syllable-final) stop consonant. This effect was later rediscovered by Abbs (1971) and has, more recently, been investigated in considerable detail (Dorman, Raphael, & Liberman, 1979; Raphael & Dorman, in press; Repp, 1978, 1979a, 1979b, 1980; Rudnicky & Cole, 1978). These studies used both synthetic and natural speech, and a variety of stop-consonant combinations and vocalic contexts. Several studies assessed precisely what closure duration is needed between the VC₁ and C₂V waveforms to perceive both stop consonants on 50 percent of the trials; a typical value for this perceptual boundary on a continuum of varying silent closure durations is 70 msec. However, the explanation of the phenomenon is still far from clear.

Two basic possibilities may be distinguished. One is that the effect in question is entirely auditory; e.g., it might be due to interference of the cues for the second stop (the formant transitions out of the closure) with the processing of the cues for the first stop (the formant transitions into the

Acknowledgment: This research was supported by NICHD Grant HD01994 and BRS Grant RR05596 to the Haskins Laboratories. Most of the work was done by my able assistant, Patti Price. I thank Gary Kuhn for permission to draw on his collection of VCV stimuli. Leigh Lisker, Virginia Mann, and Ignatius Mattingly contributed helpful comments.

[HASKINS LABORATORIES: Status Report on Speech Research SR-63/64 (1980)]
different stimuli of both about relevant auditory other appears to be so, then variations in the closure duration needed to perceive two stop consonants should be correlated with similar variations in the average (or, perhaps, the minimum) closure duration in naturally produced VC1C2V sequences. Neither of these alternatives has been unequivocally supported or rejected in recent studies of the influence of three primary auditory stimulus parameters (spectrum, duration, and amplitude of the two signal portions) on the location of the perceptual boundary (Repp, 1979a, 1979b). In part, this is due to an absence of systematic acoustic data based on natural productions, and to the consequent uncertainty as to the predictions of the "articulatory hypothesis".

The present paper remedies this situation by directly comparing perception and production of a set of utterances selected to be particularly relevant to the articulatory hypothesis. The set consists of the six possible sequences of the three voiced stop consonants of English, in vocalic context: /VbdV/, /VbgV/, /VdgV/, /VdbV/, /VgbV/, and /VgdV/. A preliminary study comparing perceptual boundary values (the closure duration needed to hear both stops, rather than only the second) for these six stimulus types was reported briefly by Liberman (1975). The stimuli in that experiment were synthetic and of the form /bâC1C2a/; the silent closure interval was varied from 0 to 125 msec in a number of steps. The results were quite clear: On one hand, stimuli in which place of stop articulation moved from front to back (/bd/, /bg/, /dg/) had boundary values of 75-90 msec; on the other hand, stimuli in which place of stop articulation moved from back to front (/db/, /gb/, /gd/) had boundaries between 0 and 25 msec of silence. These data pointed towards a possible articulatory basis: perhaps, back-front sequences are easier to articulate (and, hence, have shorter closures) than front-back sequences. However, no articularatory or acoustic observations were available that spoke to this suggestion.

Recently, Raphael and Dorman (in press) replicated the Liberman study using natural speech. In view of the fact that they used single tokens produced by a single speaker (stimuli nearly as unrepresentative as the synthetic tokens used by Liberman), the agreement with the results of the earlier study was striking. Front-back sequences again required 75-90 msec of closure for both stops to be heard; back-front sequences, on the other hand, had perceptual boundaries between 0 and 50 msec. Curiously, Raphael and Dorman did not raise the possibility of an articulatory basis for their results; instead, they briefly considered two psychoacoustic hypotheses, neither of which was well supported by their data. However, they acknowledged—as did Liberman (1975)—the need to replicate this pattern of results in vocalic contexts other than /a−a/.

This is one purpose of the present studies. It seems likely that any articulatory constraint relating to front−back vs. back−front movement in place of stop articulation would be essentially constant across different vocalic environments; therefore, if perception follows production—as the articulatory hypothesis asserts—the pattern of perceptual results, too,
should be invariant across different vocalic contexts. In the less likely case that the articulatory dynamics of two-stop sequences strongly depend on the vocalic environment, the question becomes whether changing articulatory patterns correspond in any way to changing perceptual requirements as a function of vocalic context. If psychoacoustic factors are at work in the perceptual suppression of the first stop, considerable variability in the pattern of results might be expected across different vocalic contexts because the acoustic properties of the stimuli change radically with changes in the surrounding vowels; in particular, the formant transitions conveying the places of articulation of the two stops may change in extent, shape, and direction. According to the auditory hypothesis, however, the pattern of variability observed in perception should have little relation to what occurs in speech production.

Thus, the present studies address three issues: (1) Does the perceptual boundary indeed vary across different combinations of stops, as earlier studies suggest, and if so, is this pattern of results stable across different vocalic contexts? (2) Do closure durations in corresponding natural utterances vary across different combinations of stops, and if so, is this pattern stable across different vocalic contexts? (3) Is there any consistent relationship between the patterns observed in perception and in production?

**EXPERIMENT 1: PERCEPTION—SYNTHETIC STIMULI**

**Method**

**Subjects.** Eleven subjects participated. They included nine paid volunteers (mostly Yale undergraduates), one research assistant, and the author. All were native speakers of American English except for the author whose native language is German. Earlier studies indicated no systematic differences between his perception of VC\(_1\)C\(_2\)V stimuli and that of native speakers of English.

**Stimuli.** Because convincing unreleased syllable-final stops at all three places of articulation are difficult to synthesize following vowels other than /a/, the vowel in the first syllable was always /a/, and only the vowel in the second syllable was varied. The basic stimulus components were three VC syllables—/ab/, /ad/, and /ag/—and nine CV syllables: /ba/, /da/, /ga/, /bi/, /di/, /gi/, /bu/, /du/, /gu/. All syllables were produced by the OVE IIIc serial resonance synthesizer at Haskins Laboratories. Out of convenience, the parameters were taken from a set of VCV utterances previously synthesized by a colleague using a computer procedure (CONVERT) which permits the conversion of parameters of natural-speech spectrograms into synthesizer parameter values. Thus, the synthetic syllables were simplified recreations of natural speech; the fact that they were derived from VCV (rather than VC\(_1\)C\(_2\)V) utterances seemed unimportant, especially since there were no obvious coarticulatory effects across the closure period (cf. Öhman, 1966) in the original utterances. Only periodic excitation was used in the synthetic stimuli.

The stimuli were regularized with respect to duration and fundamental frequency. All VC syllables were 180 msec long and had a constant fundamental
of 120 Hz. All CV syllables were 290 msec long and had a fundamental frequency contour that began at 120 Hz, remained steady for 40–140 msec (depending upon the individual stimulus, as copied from natural speech), and then fell steadily to a value between 94 and 105 Hz. All amplitudes and formant trajectories remained as traced from natural speech. This implied lower output amplitudes for /Cu/ than for /Ca/ and /aC/ syllables, with /Ci/ amplitudes in between. (Repp, 1979b, showed that stimulus amplitude plays only a minor role in the paradigm used here.)

All synthetic stimuli were digitized at 10 kHz using the Haskins Laboratories PCM system. Three test tapes were then created, identical except for the vowel of the CV syllables (/a/, /i/, /u/), which varied across tapes. Each tape contained first a randomized sequence of the six component syllables (/ab/, /ad/, /ag/, /bV/, /dV/, /gV/) in which each stimulus occurred 10 times, with interstimulus intervals (ISIs) of 3 sec. The stimuli in the main portion of the test consisted of the six possible /aC1C2V/ disyllables (C1 ≠ C2), with silent closure intervals varying in ten 10-msec steps from 15 to 115 msec. The resulting 66 disyllabic stimuli were recorded in five different randomizations, with ISIs of 3 sec.

Procedure. The subjects listened in a quiet room over TDH-39 earphones. The tapes were played back at a comfortable intensity on an Ampex AG-500 tape deck. Each subject participated in two sessions. In each session, all three tapes were presented in counterbalanced order. Thus, each subject gave a total of 10 responses to each individual VC-CV stimulus combination, 20 responses to each isolated CV syllable, and 60 responses to each isolated VC syllable (since the same VC syllables occurred on each tape). The task was to identify by forced choice (in writing) all stop consonants heard. In the monosyllabic series, the response choices were "b", "d", "g"; the subjects were told that the stops could occur in either initial or final position. In the VC-CV series, there were nine response choices: "b", "d", "g", "bd", "bg", "dg", "db", "gb", "gd". The subjects were informed about the structure of these stimuli—that they were made up from the monosyllabic components just heard, with varying intervals of silence between them. They were also told that, at short intervals of silence, the first (syllable-final) stop tends to disappear from perception. They were asked to write down only what they heard, not to guess a supposed consonant that was not actually perceived.

Results and Discussion

Two subjects (paid volunteers) unexpectedly failed to hear a sufficient number of single stops in VC-CV combinations—they generally heard two stops, usually the correct ones, even when little or no silence was present. Their data were excluded, so that the following results are based on nine subjects.

Monosyllables. The identifiability of the stops in the isolated VC and CV components was good to excellent, considering the fact that most of the subjects had little experience with synthetic speech. The majority of the confusions was due to a few individual listeners who more or less consistently misidentified an individual stimulus. The /Ci/ set generated more confusions than the /Ca/, /Cu/, and /aC/ sets; the respective percentages of correct responses were 80.4, 98.0, 97.6, and 95.7.
VC-CV combinations: Two-stop vs. one-stop responses. The responses to VC-CV combinations were first scored in terms of two-stop vs. one-stop responses, regardless of whether the responses were correct (i.e., the equivalent of C1, C2, or C1C2) or not. (Exclusion of errors would have distorted the data because of certain systematic misidentifications, which are discussed below.) All VC-CV combinations showed the expected increase in two-stop responses as the silent interval increased in duration. The boundary values (50-percent cross-over points) for all but two of the labeling functions fell between 55 and 80 msec. Two functions, however, stood out—those for /agba/ and /adba/; these stimuli required much less silence for both stops to be heard, and they received a nonnegligible number of two-stop responses even at the shortest silence duration. Note that both stimuli contain back-to-front movements of place of articulation, in agreement with Raphael and Dorman (in press).

Figure 1 summarizes the data in terms of percentage single-stop responses, averaged across all silence durations—a measure that takes into account differences in the lower and upper asymptotes of the response functions. (However, a plot in terms of boundary values yields a very similar pattern.) It can be seen that the deviant results for /db/ and /gb/ in the /Ca/ set have no parallel in the /Ci/ and /Cu/ sets; clearly, they are specific to the /Ca/ stimuli (to /ba/ in particular). The hypothesis that front-back sequences (the first three stimuli on the abscissa in Figure 1) would have lower boundary values (i.e., more single-stop responses) than back-front sequences (the last three stimuli on the abscissa) is not supported in the /Ci/ and /Cu/ sets, and only partially supported in the /Ca/ set, since /ag-da/ did not have a low boundary value.

The deviant results for /adba/ and /agba/ led to highly significant effects in an analysis of variance. However, after exclusion of all /db/ and /gb/ stimuli from the analysis, there was no significant effect of either consonant combinations or vocalic context; the interaction of these two factors was marginally significant, F(6,48) = 3.0, p < .05, but difficult to interpret.

VC-CV combinations in the /Ca/ set tended to have somewhat shorter boundaries than those in the /Ci/ and /Cu/ sets, even if the two extreme cases (/adba/, /agba/) are disregarded. This tendency (though not significant) is interesting since Repp (1979a) found shorter boundaries in stimuli of the type /V1bgV2/ when V1 = V2 than when V1 ≠ V2. The V1 = V2 condition was met by the present /Ca/ set, since all VC stimuli began with /a/. Thus, this difference might reflect a perceptual effect of contextual homogeneity, with a possible basis in articulation.

VC-CV combinations: C1 responses and errors. To the extent that they do not derive from C2 misidentifications, C1 responses violate the principle that, at short silent intervals, C2 is perceptually dominant over C1. A high percentage of these responses occurred in /adbu/ and /agbu/; several subjects had difficulty perceiving the stop in /bu/ even at the longer silent intervals (cf. Repp, 1979a), most likely because this stimulus had only minimal formant transitions that were difficult to detect and therefore were overpowered by more pronounced cues in the preceding signal portion. C1 responses were also frequent in /adbi/, /adbi/, and /agbi/; they could only in part be accounted
Figure 1. Percent single-stop responses (averaged over all silence durations) to the 18 VC₁-C₂V combinations (synthetic speech).
for by $C_2$ confusions between /bi/ and /di/. Many of the remaining $C_1$ responses could be predicted from the way the isolated stimulus components were perceived, except for a small percentage occurring in response to /adba/ and /adga/. Note that nearly all these cases involve labial stops in second position; thus, syllable-initial labial formant transitions seemed to be less effective in competition with conflicting syllable-final transitions than syllable-initial alveolar and velar transitions.

A large proportion of the error responses (responses other than the equivalents of $C_1$, $C_2$, and $C_1C_2$) could be predicted from the misidentifications of the monosyllabic components. There were certain unpredicted errors, however, that showed up with consistency. They included "bg" responses to /adga/ and especially to /adgi/ (rarely to /adgu/), which constituted the large majority of error responses to these stimuli (total: 9.2 percent); and "bd" responses to /adga/, /adgi/, and /adgu/, which made up about two thirds of the errors to these stimuli (total: 11.2 percent). These errors involve alveolar-velar combinations (in either order) in which the first stop was mislabeled as "b". (Neither /ad/ nor /ag/ was misidentified as "ab" in isolation.) We may be dealing here with a form of perceptual contrast (cf. Repp, 1978).

**EXPERIMENT 2: PRODUCTION—ACOUSTIC MEASUREMENTS**

Experiment 2 provided acoustic measurements of natural VC$_1$C$_2$V utterances, in order to see whether there is any relationship between the amount of silence required in perception and the average durations of closure periods in natural speech. While there have been several studies of closure durations associated with single intervocalic stops, the only study of two-stop sequences to date seems to be the unpublished work of Westbury (Note 1). However, he examined only clusters that were heterogeneous with respect to voicing (i.e., clusters of one voiced and one voiceless stop), whereas the present study was concerned with sequences of two voiced stops. Nevertheless, his results are highly relevant. He found that total closure durations were shorter when the first stop was alveolar than when it was labial or velar; they were also shorter when the second stop was velar than when it was alveolar or labial. In addition, he found an effect of vocalic environment, which he interpreted as a tendency towards temporal compensation for intrinsic variations in vowel duration: the longer the duration of the context (/bVC$_1$C$_2$Vt/), the shorter the closure duration. He did not report any changes in the effects of stop place of articulation across different vocalic environments.

The present study not only used somewhat different stimulus materials but also went beyond Westbury's by dividing closure periods into two portions. This was possible since most of the utterances measured contained release bursts of the syllable-final stop ($C_1$). (Westbury's utterances either did not contain such bursts, or he did not take them into account in his measurements.) In perceptual studies using natural speech, $C_1$ release bursts are deleted to produce the perceptual phenomenon of interest (Raphael & Dorman, in press; see Exp. 3 below). However, since the acoustic information for the syllable-final stop really includes the $C_1$ release and the preceding closure, this fact needs to be taken into account in any explanation of perceptual
results: It may be that the amount of silence listeners need in perception is more directly related to the closure preceding the release ("C₁ closure") than to the total closure duration in production.

Method

Subjects. The subjects were two female research assistants, both native speakers of American English, and the author. The author, a native speaker of the Viennese variety of German, has lived in the United States for over 11 years but has retained a foreign accent. However, it was considered unlikely that the pronunciation of voiced stop consonant sequences in meaningless isolated disyllables would show any systematic influence of native language.

Utterances. The utterances were the same as in Experiment 1. The 18 disyllables were arranged into 10 different random lists that were typed onto a sheet of paper in simple spelling (e.g., abdi, adgu, etc.). After listening to sample pronunciations and practicing for a few minutes, the subjects read from the lists at an even pace, pronouncing each utterance at a fairly fast rate, with stress on the second syllable but without neutralizing the initial vowel. The recordings were made in a soundproof booth, using a Shure microphone and an Ampex AG-500 tape recorder.

Measurement procedure. All measurements were performed on a large-scale oscillographic display provided by a GT40 computer. After inputting an utterance from audio tape, critical points in its digitized waveform were located in the continuous, magnified display by means of a cursor, and the distance from one critical point to the next was measured to the nearest tenth of a millisecond using an automatic counter. Seven measurement points were defined:

A. Approximate onset of utterance.
B. Offset of VC portion. (Sometimes, voicing pulses persisted into the closure; in this case, the onset of significant damping—indicating closure of the vocal tract—was taken as the criterion.)
C. Onset of C₁ release burst.
D. Offset of C₁ release burst (approximate within a few msec).
E. Onset of CV portion.
F. Onset of periodicity in CV portion.
G. Approximate end of utterance.

From these measurement points, the following durations were derived:

\[ F - A = \text{Total utterance.} \]
\[ B - A = \text{VC portion.} \]
\[ D - B = \text{Total closure.} \]
\[ C - B = "C₁ closure". \]
\[ D - C = \text{C₁ release burst.} \]
\[ E - D = "C₂ closure". \]
\[ G - E = \text{CV portion.} \]
\[ F - E = \text{C₂ burst and aspiration.} \]
\[ G - F = \text{CV voiced portion.} \]

All measurements were performed by a research assistant (a graduate student in phonetics) after thorough consultation with the author. Analyses
of variance were performed on all measures of interest, with the factors Speakers, Vowels (three final vowels), and Consonants (six combinations). Since C₁ and C₂ were not orthogonal factors, their separate influences were examined in post-hoc (Newman-Keuls) tests comparing those six pairs of utterances that differed in one component only (C₁ effects: /bg/ vs. /dg/, /bd/ vs. /gd/, /db/ vs. /gb/; C₂ effects: /gb/ vs. /gd/, /db/ vs. /dg/, /bd/ vs. /bg/). The pooled within-cell variance (10 observations per cell, i.e., per utterance) was taken as the error term. Missing values, due to rare mispronunciations or acoustic anomalies, were replaced with the cell mean prior to analysis.

Results and Discussion

Total closure duration. The pattern of average closure durations as a function of consonant combinations and final vowels is shown in Figure 2, separately for each speaker. The grand average duration was 168 msec, with an average within-cell standard deviation of 15 msec. Statistical analysis revealed, first of all, a speaker effect, $F(2,486) = 262.3, p << .001$: BHR's closures were longer (188 msec, on the average) than DK's (162 msec) and SP's (154 msec). More interestingly, there was a highly significant vowel effect, $F(2,486) = 36.1, p << .001$: Closure durations were shorter for final /a/ (160 msec) than for final /i/ (172 msec) and /u/ (172 msec). This effect was shown (on the average) by all three speakers and by each individual consonant combination; no statistical interaction involving the vowel effect approached significance. Finally, there was a significant consonant effect, $F(5,486) = 8.5, p < .001$, which did not interact with any other factor, despite (or because of) the considerable variability evident in Figure 2. The six consonant combinations were arranged as follows: /dg/ (161 msec), /bg/ (165 msec), /db/ (168 msec), /gd/ (170 msec), /gb/ (172 msec), /bd/ (173 msec). Newman-Keuls tests revealed one significant effect of the first stop (/d/ shorter than /b/, $p < .05$) and two significant effects of the second stop (/g/ shorter than /b/ and /d/, both $p < .01$), out of three comparisons in each case.

Certainly, these data provide no evidence for closures to be shorter in back-front sequences than in front-back sequences, or to be especially short in /adb/a/ and /gab/a/ (cf. Exp. 1). However, the results are in excellent agreement with Westbury's (Note 1) measurements, which showed closures to be shortest for alveolar stops in first position and for velar stops in second position. Westbury also found, in agreement with the present results, that closure durations were shortest in /a-a/ context, and he related this finding to the relatively long durations of these vocalic portions. We will return to this issue below.

C₁ closure. The C₁ closure measurements are shown in the left half of Figure 3. Since speaker SP did not consistently produce C₁ release bursts, her closure durations could not be broken down into components. Speakers BHR and DK, on the other hand, produced release bursts in all utterances. Their average C₁ closure lasted 74 msec, with an average within-cell standard deviation of 13 msec. C₁ closures were significantly longer in BHR's productions (80 msec) than in DK's (67 msec), $F(1,324) = 90.8, p << .001$, which parallels the difference in total closure durations reported above. Interestingly, there was no significant effect of the final vowel here,
Figure 2. Average total closure durations in 18 VC₁-C₂V combinations produced by three speakers.
although there had been such an effect on total closure duration. However, there was a highly significant effect of consonants, $F(5,324) = 34.7$, $p < .001$, which also interacted with speakers, $F(1,324) = 3.7$, $p < .01$. Averaging over speakers (which seems permissible since the interaction was quite small), the rank order was: /gb/ (63 msec), /db/ (66 msec), /dg/ (70 msec), /gd/ (72 msec), /bd/ (80 msec), /bd/ (88 msec). Newman-Keuls tests showed C1 closures to be clearly longer when C1 was /b/ than when it was /d/ or /g/ ($p < .01$) -- a result that is in striking agreement with measurements of closure durations in single intervocalic stops, which show longer durations for labials (e.g., Kohler, 1979; Umeda, 1977; Westbury, Note 1). However, C2 also affected C1 closure duration: C1 closures were longer preceding /d/ than preceding /b/ ($p < .01$) or /g/ ($p < .01$, but only shown by speaker DK). Thus, while a syllable-final /b/ led to long C1 closures, a following syllable-initial /b/ was associated with rather short C1 closure durations.

C2 closure. The C2 closure measurements for speakers BHR and DK are shown in the right half of Figure 3. The average C2 closure lasted 84 msec, with an average within-cell standard deviation of 16 msec. BHR's C2 closures were significantly longer (90 msec) than DK's (78 msec), $F(1,324) = 46.1$, $p << .001$, as had been his C1 closures. There was a significant vowel effect, $F(2,324) = 6.9$, $p < .001$, C2 closures being shorter preceding /a/ (80 msec) than preceding /i/ (85 msec) or /u/ (87 msec). Since C1 closure had shown no vowel effect, it was C2 closure that was responsible for the variations in total closure duration with final vowel. C2 closure durations varied significantly across different consonant combinations, $F(5,324) = 13.2$, $p < .001$, and the pattern differed somewhat between the two speakers, $F(5,324) = 4.2$, $p < .001$. Overall, however, the rank order was nearly the inverse of that for C1 closure duration: /bd/ (75 msec), /dg/ (78 msec), /bg/ (82 msec), /gd/ (84 msec), /bd/ (90 msec), /gb/ (95 msec). Newman-Keuls tests showed that syllable-initial /b/ (C2) was associated with longer C2 closures than either /d/ or /g/ ($p < .01$), with somewhat longer closures for /g/ than for /d/ ($p < .05$), whereas C2 closures were shorter when the preceding stop was /b/ than when it was /g/ ($p < .01$). Thus, C2 closures, like C1 closures, were longest when the associated stop was labial, but tended to be short when the other stop was labial.

Other signal portions. Since only closure duration measures are directly relevant to the topic of this paper, the other measurements will be summarized only very briefly. C1 release bursts (average duration 17 msec) were markedly shorter for syllable-final /b/ in BHR's utterances, but not in DK's. VC portions (average duration 105 msec) showed no speaker difference (in contrast to the closure measures) but an effect of C1: The vocalic portion was shorter for /b/ than for either /d/ or /g/ ($p < .01$). The C2 burst and aspiration portion—the voice onset time (VOT) of C2—showed the familiar effect of C2 place of articulation, VOTs being shortest for /b/ (11 msec) and longest for /g/ (24 msec), with /d/ (19 msec) in between. Two speakers (BHR and DK) had shorter VOTs before /a/; speaker SP, however, showed the opposite pattern. (SP also had much shorter VOTs than the other two speakers.) The voiced CV portion (average duration 221 msec) was longer for /a/ for two speakers; again speaker SP differed by showing no vowel effect. There was no consonant effect here but a speaker difference, DK being slower than BHR (and both much slower than SP). Since DK had shorter closures than BHR, and since VC portions showed no speaker differences, independent temporal control of the different signal portions is suggested.
Figure 3. Average $C_1$ and $C_2$ closure durations in 18 VC$_1$-C$_2$V combinations produced by two speakers.
Summary. Closure durations were affected by the identity of both consonants as well as by the final vowel. C₁ and C₂ generally had opposite effects; thus, total closure durations ranked /d/ < /b/ < /g/ with respect to C₁ but /g/ < /b/ < /d/ with respect to C₂. When C₁ and C₂ closure segments were considered separately, however, the consonant effects were found to reflect primarily labial articulation: Both C₁ and C₂ closures were longest when the associated consonant was /b/ and tended to be shortened when the other stop was /b/. Total closure durations were shortest in /-a/ context, and this effect was entirely due to variations in C₂ closure.

This pattern of results does not show a close resemblance to the perceptual results of Experiment 1. The abnormal perceptual boundaries for /adba/ and /agba/ have no parallel in production, and systematic effects of C₁ and C₂ across all three vocalic contexts are observed in production only, not in perception. Only the final-vowel effect (shorter closures in /-a/ context) corresponds to a tendency towards shorter perceptual boundaries in that context. However, this effect could easily have an auditory basis: Several studies have shown that silent gaps are easier to detect in spectrally homogeneous than in heterogeneous environments (Collyer, 1974; Perrott & Williams, 1971; Williams & Perrott, 1972). Since the initial vowel in the present stimuli was always /a/, stimuli ending in /-a/ were spectrally more homogeneous than stimuli ending in /-i/ or /-u/, and perhaps this homogeneity facilitated the detection of the silent closure period.

EXPERIMENT 3: PERCEPTION—NATURAL SPEECH STIMULI

So far, our comparison of perception (Exp. 1) and production (Exp. 2) of two-stop sequences has been disappointing. However, the results of Experiment 1 may not have been representative, due to peculiarities of the synthetic stimuli. Although this possibility seems less likely in view of the good agreement between portions of the results of Experiment 1 and the earlier findings of Liberman (1975) and Raphael and Dorman (in press), it seemed desirable to replicate Experiment 1 using natural-speech stimuli. This was the purpose of Experiment 3.

Method

Subjects. Twelve subjects participated. They included ten paid volunteers with little experience in speech perception experiments and two subjects with considerable experience as listeners (a graduate research assistant and the author).

Stimuli. The stimuli were constructed from speaker BHR's utterances, which had been collected and measured in Experiment 2. To avoid token-specific irregularities and to permit an estimate of natural variability, four different tokens of each of the 18 utterances were selected from the 10 originally recorded. Thus, the initial stimulus pool consisted of 4 x 18 = 72 utterances. All utterances were digitized at 10 kHz and edited using the Haskins Laboratories Pulse Code Modulation system. The original closure periods (including the C₁ release bursts) were excised, and various amounts of silence (0-100 msec, in 10-msec steps) were inserted instead. The VC and CV portions were also stored in separate files.
The experimental tapes were analogous to those of Experiment 1. Three parallel sets were recorded, one for each final vowel. In each set, the first stimulus sequence consisted of the isolated VC and CV portions in random order, arranged in 5 blocks of 48, with ISIs of 2.5 sec and 10 sec between blocks. The 48 stimuli resulted from 4 tokens of each of 2 portions (VC and CV) of 6 utterances. The second stimulus sequence contained the VC-CV combinations in random order, arranged in 4 blocks of 66, with ISIs of 2.5 sec and 10 sec between blocks. The 66 stimuli resulted from 11 closure durations for one token of each of the six utterances. Different tokens were used in each of the four blocks; thus, there were in fact 4 x 66 = 264 physically different stimuli.

Procedure. Each subject participated in three sessions, one for each final-vowel condition. The order of final-vowel conditions was counterbalanced across subjects. In each session, the isolated VC and CV portions were presented first. A total of 5 responses for each token of each utterance was obtained, i.e., 20 responses for each utterance when token variation is ignored. Subsequently, the VC-CV combinations were presented three times, separated by appropriate rest periods. That is, each subject gave a total of three responses to each individual stimulus, or 12 responses when ignoring token variation.

Results and Discussion

Monosyllables. The natural-speech CV stimuli were quite intelligible, but the VC stimuli were less well identified than the synthetic stimuli in Experiment 1. The stop in /ag/, in particular, was frequently misidentified, with "b" confusions being about twice as frequent as "d" confusions. This poor identifiability was obviously a consequence of removing the C1 release burst. The percentages of correct responses for the /Ci/, /Ca/, /Cu/, and /aC/ sets were 90.2, 96.3, 99.4, and 82.3, respectively (52.1 for /ag/). The confusion patterns did not seem to reflect in any way the context in which a given stimulus portion had been pronounced; thus, there seemed to be little coarticulation between VC and CV portions.

VC-CV combinations: Two-stop vs. one-stop responses. The results of the main part of the experiment were somewhat startling. Although the two experienced subjects produced what seemed to be typical and orderly results, a number of the naive subjects failed to show the VC-CV interference phenomenon, i.e., the predominance of single-stop percepts at short closure durations. All naive subjects reported two stops at short silence durations for at least some of the stimuli. Moreover, these responses were correct more often than not, and those misperceptions that occurred were typically consistent and stimulus-specific.

This outcome was quite unexpected, even though it will be recalled that two subjects in Experiment 1 had to be excluded for the same reason. To make sure that no problem of instructions was involved, two of the subjects were recalled and carefully instructed by the author. The same result was obtained: There were very few single-stop responses. Inspection of the stimuli did not reveal any reason for this "abnormal" behavior of the majority of listeners. Of course, researchers have known for a long time that speech cues—silence in particular—do not always have a perceptual effect: Their
effect depends on the values of other relevant cues in the signal. In the present case, the formant transitions in and out of the closure and the $C_2$ release burst may have provided stop manner cues strong enough to override the perceptual effect of silence. What is surprising is that this occurred only for the naive listeners, as if they assigned less weight to the silence cue than the two experienced listeners. Interestingly, very similar observations have recently been reported by May, Porter, and Miller (1980).

Five subjects had to be excluded because they either gave no single-stop responses at all or just a few that were fairly randomly distributed. However, the responses of the remaining five naive listeners fell into a fairly orderly pattern that, moreover, resembled the results of the two experienced listeners, BHR and PP. Therefore, the data of all seven subjects were combined. They are plotted in Figure 4, which is analogous to Figure 1.

The figure first shows a pronounced vowel effect, $F(2,12) = 4.7, p < .05$: Considerably more silence was required to hear both stop consonants in /-i/ context than in /-a/ and /-u/ contexts, and slightly less silence was required in /-a/ context than in /-u/ context. While the latter tendency parallels the findings of Experiments 1 and 2, the first, larger difference has no correspondence in the earlier results. This difference was primarily due to the naive subjects since neither BHR nor PP showed any vowel effects that were consistent across all six consonant combinations. Inspection of the test schedule suggested that the effect was not an artifact of test order, which was still nearly balanced across the selected subjects.

The second effect seen in Figure 4 is a pattern of differences across the six VC-CV combinations, $F(5,30) = 8.1, p < .001$, that was quite consistent across the three final-vowel contexts. (The interaction was marginally significant.) In each case, the longest silences were required for /dg/; /bg/ ranked second in two contexts and third in the third. The shortest silence durations were required in /bd/ and /gd/, except in /-u/ context where /db/ had the shortest boundary. Once again, this pattern does not consistently follow the front-back vs. back-front distinction. Rather, it seems to reflect an effect of $C_2$: Longer silences were required when $C_2$ was /g/ than when it was either /b/ or /d/ ($p < .01$ in Newman-Keuls tests). Note that the boundary rank order /g/ > /b/ > /d/ with regard to $C_2$ is precisely the opposite of that obtained in production, indicating that VC-CV combinations with longer total closure durations in production required less silence in perception. This runs counter to the articulatory hypothesis, as conceived at the outset.

VC-CV combinations: $C_1$ responses and errors. Given the high frequency of /ag/ misidentifications, a large number of errors, as well as single-stop responses at long closure durations, might be expected in VC-CV stimuli containing that component. The errors did occur; however, single-stop responses were not as frequent as expected. To /gb/ combinations, subjects frequently responded "db"; and "bd" responses to /gd/ stimuli were extremely common. Thus, listeners tended to prefer that confusion of /ag/ that led to the perception of two stops over the one that led to single-stop responses, perhaps because of the acoustic inappropriateness of the /ag/ transitions for a single "b" or "d" percept. Other common confusions that could not be fully accounted for by misperception of the monosyllabic components were "gb" responses to /db/ and "bg" responses to /dg/. All these errors involved, of
Figure 4. Percent single-stop responses (averaged over all silence durations) to the 18 VC1-C2V combinations (natural speech).
course, the perception of \( C_1 \); \( C_2 \) was very rarely misidentified. What is noteworthy is that a large number of errors occurred at all silence durations, including the longest, and that they were always in the direction of hearing two stops, rather than one. In other words, the listeners seemed to "know" that conflicting VC and CV cues could not be integrated into a single percept; it is not clear, however, what led them to misidentify \( C_1 \) so frequently in VCCV context. Note that the error pattern in the present experiment resembled that found in Experiment 1.

CONCLUSIONS

Systematic variations in the amount of silence required to hear two stops in utterances of the VC\( _1 \)C\( _2 \)V type do not appear to be correlated with variations in closure durations of corresponding natural utterances. They even differ a good deal between perceptual experiments employing synthetic and natural stimuli, respectively. Thus, the cause for the perceptual variability must be sought in auditory properties of the stimuli; it does not seem to be grounded in listeners' knowledge of articulatory dynamics. Presumably, the effective amount of silence perceived, or the effective value of some other relevant stimulus characteristic, is modified by the acoustic environment (in ways not yet understood) before it enters the phonetic decision process.

This conclusion underlines the importance of distinguishing between auditory and phonetic (or articulation-based) phenomena in speech perception. A number of perceptual effects have been reported that seem to require an explanation that makes reference to speech production (for recent examples, see Repp et al., 1978; Mann & Repp, 1980, in press). Indeed, the basic fact that silence plays a role at all in the perception of stop consonants may still belong in that category, although it also invites auditory hypotheses of various sorts. However, the present experiments, in conjunction with earlier data (Repp, 1979a, 1979b), suggest that variations in the amount of silence required for accurate perception arise at an auditory level. Since speech must pass through the auditory system on its way to higher centers of processing, we must expect that the perceptual phenomena we uncover in the laboratory will reflect both auditory and phonetic processes. To distinguish between these two sources of variation in each individual case is perhaps the most pervasive, and the most challenging, problem of speech perception research.

REFERENCE NOTE


REFERENCES


Kohler, K. J. Dimensions in the perception of fortis and lenis plosives. Phonetica, 1979, 36, 332-343.


