Perception & Psychophysics
1978, Vol. 24 (4), 327-336

Acoustic cues and psychological processes in the perception of natural stop consonants

MARCEL ADAM JUST, RICHARD LEE SUSLICK, STEVE MICHAELS
and LINDA SHOCKEY
Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213

These studies examined the perceptual role of various components of naturally produced stop consonants (/b, d, g, p, t, k/) in CV syllables. In the first experiment, the context-sensitive voiced formant transitions were removed with a computer-splicing technique. Identification accuracy was 84% when the consonant was presented with the same vowel as had been used to produce it. Performance fell to 66% when the consonant was juxtaposed with a different vowel. The second experiment not only deleted the voiced formant transition, but also replaced the aspiration with silence. Here, identification accuracy dropped substantially, especially for voiceless stops, which had contained devoiced formant transitions in the replaced interval. The pattern of errors suggested that listeners try to extract the missing locus of the consonant from the vowel transition, and in the absence of a vowel transition, they try to extrapolate it from the second formant of the steady-state vowel.

Certain findings in the area of speech perception have had an impact far beyond the boundaries of their discipline. In particular, it has been shown that the stop consonants /b, d, g, p, t, k/ have important acoustic properties that are context-sensitive; that is, some of the physical energy associated with a stop consonant varies as a function of the phonemes that precede and follow the stop. This fact suggested that human speech recognition mechanisms must be context-sensitive. The present research examined some acoustic cues that vary in context sensitivity in order to assess their role in the perception of natural stop consonants.

The argument that speech perception is context-sensitive has been extremely influential. Perhaps one of the best demonstrations of its influence is the frequency with which introductory textbooks publish the synthetic speech spectrograms of /di/ and /du/ (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). These spectrograms show that /di/ and /du/ have different second-formant vowel transitions. For /di/, the transition rises from about 2,200 Hz to about 2,600 Hz. For /du/, the transition falls from about 1,200 Hz to about 700 Hz. Although the transitions associated with any one consonant vary from vowel context to vowel context, they provide a sufficient cue for identifying the consonants. In a landmark study in this area, stop consonants (/b, d, g/) were synthesized with identical first formants and the same steady-state second formant (Delattre, Liberman, & Cooper, 1955). The only distinguishing feature was the second-formant transition. Different second-formant transitions resulted in the perception of different stop consonants; for example, when placed before a steady-state /a/, a rising second-formant transition was identified as /b/; a slightly falling second-formant transition was identified as /d/, and a sharply falling transition was identified as /g/. These studies demonstrated that the transition can be a sufficient cue for perceiving the place of articulation of the voiced stop consonant. The sufficiency of this cue in identifying phonemes was significant because the transition varied from vowel to vowel, and so the perception of stop consonants appeared to be context-sensitive (Liberman et al., 1967).

Given the theoretical significance of formant transitions as cues to phoneme identification, it has been important to determine to what extent these cues are necessary for recognition. The study of Delattre et al. (1955), among many others, demonstrated sufficiency. However, their necessity has not been established. In fact, listeners can identify stop consonants in certain vowel environments after · the vowel transitions have been removed (Cole & Scott, 1974; Fischer-Jørgensen, Note 1). In both of these studies, listeners correctly identified /bi/ and /di/ at least 70% and 80% of the time, respectively, after the voiced formant transitions had been removed from the syllable. These experiments, unlike those of Liberman et al. (1967), involved natural speech and

This research was supported in part by the National Institute of Mental Health, Grants MH-07722 and MH-29617. We thank Patricia Carpenter and Ron Cole for their help on previous drafts of this manuscript. Requests for reprints should be sent to Marcel Adam Just, Psychology Department, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

the vowel transitions were removed with tape-splicing procedures. Moreover, in the Cole and Scott experiment, listeners could also recognize stops presented in a vowel context other than the one in which they were originally produced (again, without the vowel transitions). Recognition performance was extremely good. Thus, formant transitions seem not to be necessary cues in certain circumstances.

The present research examined how subjects identify stop consonants in the absence of second-formant transition information. One of the main features of the present work is that computer digitizing and editing allowed very precise control over the stimulus splicing. The study focused on the perception of natural speech CV syllables composed of a stop consonant /b, d, g, p, t, k/ followed by one of three vowels (/a, i, u/). The first experiment examined the accuracy of untrained observers in identifying the six stops without voiced formant transitions. The second experiment examined the role of the aspirated interval between the burst and voiced formant transition to determine how this interval contributes to recognition.

EXPERIMENT 1

Method

Stimuli. The source utterances were CV syllables consisting of one of the stops /b, d, g, p, t, k/ followed by one of three vowels /a, i, u/. Each syllable was uttered 21 times for a total of 378 CV source utterances. These utterances were produced in a random order by a phonetically untrained male English speaker in a sound-attenuated room, and recorded on a Scully ½-in. master tape recorder. The tape was then filtered for frequency cutoff at 12 kHz, digitized using a 9-bit analog-to-digital converter, and stored on a PDP-10 computer. The syllables were highly intelligible at the conclusion of this processing.

Each of the 378 utterances was segmented into four consecutive parts defined by the following boundaries. The first and second boundaries were at the beginning and end of the stop explosion, defining the burst. The third boundary was at the beginning of the voiced formant transition. Therefore, for voiceless stops, the interval demarcated by the second and third boundaries contained aspiration and low-energy voiceless formant transitions. For voiced stops, the interval also contained some aspiration, and perhaps some formant-transition information, but at still lower intensities and durations, so much so that it often resembled silence or low-energy noise. For convenience,the interval will be referred to as aspiration for both voiceless and voiced stops. The third and fourth boundaries demarcated the voiced formant transition.

Figure 1 shows a speech spectrogram of one stimulus syllable, /da/. The figure shows the first and second boundaries located at the beginning and end of the burst. The third boundary is at the beginning of the voiced formant transition. The portion between the second and third boundary contains some aspiration. The fourth boundary is between the end of the transition and the beginning of the steady-state vowel. It was the interval between the third and fourth boundaries that was eliminated in Experiment 1.

The segmentation was done by a combination of human and machine processes. The machine processes were speech segmentation and editing programs used by the speech recognition system (Hearsay) at Carnegie-Mellon (cf. Reddy & Newell, 1974). Each utterance was displayed in both spectrogram and waveform format. We examined these displays and judged where the boundaries between the four parts of the syllable should be placed. The time scale on the two displays allowed us to specify the time at which a particular boundary, such as the end of a burst, should be placed. After the boundary locations had been input to the computer, the spectrogram and waveform plot were displayed again, this time with the boundaries marked in, so that we could see that they had been placed appropriately. The markers were placed to within 1 msec accuracy. In almost all cases, it was very easy to judge where the boundaries should be placed. The first and second boundaries, at the beginning and end of the burst, respectively, were easy to specify in the waveform plot. The stop /k/ sometimes had two bursts, and in those cases both bursts were included. The third boundary, at the beginning of the voiced formant transition, was also easy to specify in the waveform plot. The fourth boundary, between the transition and the steady-state vowel, was determined by a collaborative effort of man and machine. A computer program that detected transitions (on the basis of the formant's rate of change) indicated where it considered the transition to end (Baker, Note 2). The human judge referred to the spectrogram, indicating where he thought the boundary should be. When the two methods were within 20 msec of each other, the boundary was set where the program had decided. When there was a substantial difference (this only occurred about 15% of the time), the human resolved the conflict using his best judgment.

Table 1 shows the durations of the three segments averaged over 21 repetitions and 3 vowels. The obtained durations in-

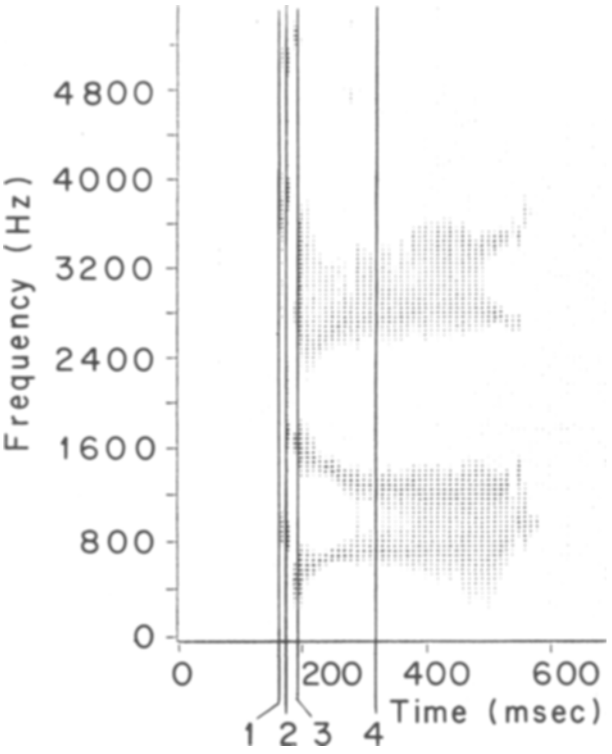

Figure 1. Spectrogram of a /da/ with the four splicing boundaries indicated by the vertical lines. See text for criteria of boundary placement.

Table 1
Mean Duration of CV Intervals (in Milliseconds)

	Burst	Aspiration	Voiced Formant Transition	VOT
/b/	6	9	112	15
/d/	7	16	116	23
/g/	9	22	112	31
/p/	9	64	90	73
/t/	10	71	93	81
/k/	14	74	85	88

Note–VOT is the voice onset time, the duration of the burst plus the aspiration. In Figure 1, VOT is the interval between boundaries 1 and 3.

dicate that the temporal characteristics of the sample of syllables are similar to normative values (cf. Lisker & Abramson, 1967). In particular, the VOTs (voice onset times) for the three voiced stops, /b, d, g/, increased by an average of 8 msec as the place of articulation progressed from front (/b/) to middle (/d/) to back (/g/). Similarly, the VOTs for the three voiceless stops increased by an average of 8 msec as the place of articulation progressed from front (/p/) to middle (/t/) to back (/k/). Thus, place and voicing have perfectly additive effects on VOT in the sample.

Table 2 indicates the duration of the voiced formant transition for the 18 syllables. The transition is generally longer for /a/ than for /i/ (the exceptions being the bilabial stops) and always longer for /a/ than for /u/. The transition durations also vary as a function of the consonant.

A splicing program was used to combine segments of the syllables to form two types of stimuli, control syllables and experimental syllables. The control syllables were constructed by deleting the voiced formant transition from the syllable and joining the burst and aspiration of a stop to a steady-state vowel. In half the cases, the burst and aspiration were recombined with the original steady-state vowel they had been produced with. In the other half, they were combined with the same vowel type but not the original vowel token. Thus, half of the control /pa/ syllables consisted of original /pa/s with the voiced formant transition deleted and the other half consisted of new combinations of /p/ and /a/ that had been produced in separate /pa/ syllables, also with the voiced formant transition deleted. (These two ways of producing control syllables produced similar results, so the distinction will not be discussed further.) Since there were three consonants and six vowels, there were 18 types of control syllables. These syllables generally sounded "correct" to the authors, and the results will bear out this observation.

The experimental syllables were constructed by exchanging consonants and vowels across syllables, and deleting voiced formant transitions. Each of the three consonants (burst plus aspiration) was placed before two steady-state vowels (/a/, /i/, or /u/), different from the one used in producing the original syllable. For example, /p/s that had been produced in the context of /a/s were combined with /i/s and /u/s. None of the stimulus syllables contained voiced formant transitions. Since there were 6 consonants and 3 vowel contexts (allowing for 6 kinds of exchanges), there were 36 types of experimental syllables.

Table 2
Mean Duration of Voiced Formant Transition (in Milliseconds)

	/a/	/i/	/u/
/b/	113	129	96
/d/	137	117	95
/g/	151	87	99
/p/	91	103	77
/t/	104	82	91
/k/	102	71	82
Mean	116	98	90

Using the segmentation boundaries discussed above, a digital splicing program was used to combine various parts of each CV source utterance to produce the experimental stimuli. To ensure a smooth transition between the spliced portions and to avoid telltale clicks, splices were made only at a down zero-crossing of the speech signal for trailing splice sections and only at up zero-crossing for the leading splice. Such splices are possible only if there exists a zero-crossing within some given distance (say .1 msec) of the designated boundary. By rejecting all of the stimuli that did not meet this criterion, a set of stimuli was obtained consisting of 3 tokens of each type of control syllable (of which there were 18 types) and 7 tokens of each type of experimental syllable (of which there were 36 types) for a total of 306 stimulus syllables.

The stimuli were then output in a random order using a 9-bit digital-to-analog converter passed through a 14-kHz filter and recorded on a Sony TC-650 tape recorder, such that a new syllable started every 4 sec.

Procedure. Forty phonetically untrained, Carnegie-Mellon undergraduates listened to the CV syllables presented over loudspeakers at a comfortable listening level, and wrote down what they heard immediately after each syllable was presented. The subjects were told that they would hear 1 of 18 possible syllables, bi, ba, bu, di, da, du, etc. There were 15 practice trials at the beginning of the session. Then the 306 syllables were presented in a random order. The testing session lasted approximately 30 min.

Results

The results are summarized in Table 3. The rows indicate the six stimulus consonants. The column headings indicate the original vowel used to produce the syllable and the vowel that was substituted in producing the final stimulus syllable. For example, /i/ → /a/, the second column, indicates that the consonant was originally produced before an /i/ but that the steady state /a/ was substituted for the /i/ in the stimulus that was presented. Columns 1, 4, and 7, /a/ → /a/, /i/ → /i/, /u/ → /u/, are the control conditions.

Table 3 shows the percentage of correct identifications of the consonant. The second (and third) entries in each cell indicate the erroneous responses that occurred at least 12% of the time, while an asterisk indicates that there was no frequent erroneous response. Subjects seldom misidentified the vowel; misidentification of the vowel or null responses account for only 3% of the responses, as indicated in the last row and column of Table 3. Such responses were somewhat more frequent for voiced than for voiceless stops, and more frequent for /a/ than for the other two vowels.

To compare the performance in this task to chance performance, an appropriate chance level of responding must be selected from two possibilities. One possible approach is to examine how well listeners can identify stop consonants in the absence of voiced formant transitions. Since there are six

Table 3
Percentage of Correct Responses (and the Most Common Error) for Experiment 1

	/a/→/a/	/i/→/a/	/u/→/a/	/i/→/i/	/a/→/i/	/u/→/i/	/u/→/u/	/a/→/u/	/i/→/u/	Mean	Errors
/b/	74 *	81 *	81 *	87 *	59 (d 20)	65 (d 14)	96 *	92 *	95 *	81	6
/d/	38 (b 38)	43 (b 23)	50 (b 18)	97 *	83 (g 11)	66 (g 28)	93 *	55 (b 34)	74 (b 21)	67	4
/g/	13 (d 48) (b 18)	12 (d 32) (b 22)	38 (b 18)	79 (d 18)	27 (d 56)	51 (d 17) (b 15)	75 (b 14)	32 (b 44) (d 16)	11 (d 46) (b 29)	38	6
/p/	92 *	84 *	86 *	98 *	88 *	83 *	97 *	96 *	91 *	90	1
/t/	86 *	72 *	88 *	100 *	95 *	88 *	100 *	61 (p 25)	87 *	86	1
/k/	97 *	40 (t 22) (p 21)	98 *	100 *	47 (t 43)	82 (p 12)	99 *	76 *	18 (t 34) (p 31)	73	2
Mean	66	55	73	93	66	72	93	69	63	72	
Errors	4	7	6	1	2	3	1	3	2	3	

Note—Data are based on 280 observations per cell for the experimental syllables and 120 observations per cell for the control syllables. "Errors" are the percentage of times no response was given or the vowel was incorrect.
**Confusions were too few or no one kind received more than 11% of the responses.*

stop consonants, the probability of responding correctly by chance is .167. Using a binomial distribution with $p = .167$, and assuming independence of trials, the boundary of the 95% confidence interval for control syllables (N = 120) is 25% correct. For the experimental syllables (N = 280), the boundary is 21% correct.

An alternative approach is to hypothesize that voiced formant transitions are necessary cues to place of articulation, and then determine whether listeners can correctly identify the place in the absence of voiced formant transitions. In this view, the probability of responding correctly by chance is 33% (since there are three possible places of articulation in this situation). Here the boundary of the 95% interval is 42% for control syllables and 39% for experimental syllables. Thus, these two approaches yield different confidence intervals for "chance." For simplicity, it may be useful to note that anything above 42% is better than chance, and anything under 21% is no better than chance for both control and experimental syllables, no matter which of the two approaches is adopted.

The mean percentage of correct responses was 72%. In general, performance was best for bilabial stops (/b/, /p/), slightly poorer for alveolar stops (/d/, /t/), and noticeably poorer for the velar stops (/g/, /k/). The /g/ was correctly identified only 38% of the time, much less frequently than the other stops. The /g/ and /k/ account for 9 of the 11 cells where correct performance was less than 50% (the other two are /da/ → /da/ and /di/ → /da/). It is not surprising that the /g/ (and to some extent /k/) produces relatively poor performance. It has been shown that the burst frequencies associated with velar consonants vary much more from vowel context to vowel context than do the burst frequencies for /p/ and /t/ (Liberman, Delattre, & Cooper, 1952; Delattre et al., 1955). Consequently, the burst and aspiration would be expected to be poorer cues for /g/ and /k/ than for the other consonants.

The control syllables establish a baseline of recognition performance when the voiced formant transitions are removed. Performance in these conditions was fairly good, 84% overall; however, there were considerable differences among the three vowels. Recognition was excellent for /i/ → /i/ and /u/ → /u/; 93% of the responses were correct for these vowels. The major decrement occurred for /a/ → /a/, Column 1, where performance was only 66% correct. The differences in performance may be related to the fact that the excised vowel transitions were generally longer for /a/ than for /i/ and /u/, as Table 2 indicates. When more voiced formant transition is removed, performance is poorer. In fact, among the 18 control syllables, the correlation between the duration of the excised voiced formant transition and correct recognition was substantial, $r(17) = -.78$, $p < .01$. For example, for the syllables /da/ and /ga/, the excised voiced formant transitions were very long (137 and 151 msec, respectively) and recognition performance was extremely poor (38% and 13%, respectively). A longer formant transition seems to be associated with greater perceptual importance as a cue to consonant identification. One possibility is that a longer transition may play a more important role in the speech envelope, the overall configuration of acoustic energy. Consequently, removing a longer transition would alter the shape of the speech wave more.

The mean identification rate on the experimental syllables was 66%, and the percentage in individual conditions was also correlated with the duration of the excised voiced formant transition, $r(35) = .34$, $p < .05$. The durations involved in this computation were the mean durations of the transitions of the vowel that was ultimately presented with the consonant. For example, the covariate for the percent correct on /bi/ → /ba/ and for /bu/ → /ba/ was the mean transition duration for the /ba/s. A similar correlation, computed between the correct identification of a consonant and the duration of the transition originally used to produce the consonant, was much lower, $r(35) = -.08$, n.s. This asymmetry indicates that the vowel that was presented with a consonant was more important in determining performance than the vowel used to originally produce the consonant. The longer the voiced formant transition excised from the presented vowel, the less probable was the correct identification of the consonant.

Analysis of errors. Of the 12,240 responses in this study, 72% were correct. An additional 3% of the responses were null responses and misidentifications of the vowel. Thus, in 25% or 3,060 cases, the consonant was incorrectly identified. These errors often fall into systematic patterns that yield insights into the cues that are used in consonant identification. The distribution of these errors (i.e., which conditions they occurred in) indicates where the cues were lacking or misleading for correct consonant identification. The type of error that was made indicates how the presented cues were interpreted.

The erroneous responses were fairly systematic; often one or two types of errors predominated in each condition. Table 3 lists the erroneous responses that accounted for more than 11% of the responses in each condition. These results show that there was little confusion of the voicing features, consistent with the fact that VOT relations were preserved in the stimuli; voiced stops were seldom perceived as voiceless stops and vice versa (cf. Miller & Nicely, 1955; Fischer-Jørgensen, Note 1). Experiment 2 will demonstrate that the VOT duration is not the only cue for the voiceless-voiced distinction. Another cue is whether this interval is silent or primarily filled with aspiration.

The pattern of the errors (i.e., which erroneous responses tended to be produced) can often be accounted for by a locus-extrapolation process. The locus of a particular nonvelar stop is the frequency "pointed to" by the second formant vowel transition associated with that stop (Delattre et al., 1955). The second formant vowel transition of a nonvelar stop always points (in a backward, right-to-left direction on a spectrogram) to the same frequency, regardless of what the vowel is. Presumably, the locus frequency reflects the resonance characteristics that are fairly constant when that consonant is being produced. For example, 1,800 Hz is the locus for /d/, and the second formant vowel transitions into all vowels point to this locus, regardless of whether the vowel's second formant is above or below this locus. If listeners can extrapolate this invariant locus from the vowel transition, they could determine what the preceding stop consonant was. In fact, the sufficiency of second formant vowel transitions indicates that there must be some mental process that can compute the identity of a stop consonant with only the transition information as input. The locus-extrapolation process is one possible candidate, at least for nonvelar stops.

In the Delattre et al. experiments, it was hypothesized that the listener extrapolated the locus from the second-formant transitions. In the Delattre et al. spectrograms, /b/ has a locus around 720 Hz; /d/ has a locus around 1,800 Hz; and /g/ has a locus around 3,000 Hz, but only when the vowel has a second formant above 1,200 Hz. For convenience, these will be called the canonical loci of the respective consonants. In the current experiment, listeners would have to extrapolate the locus from syllables that contained no voiced formant transition but only a steady-state F2. To the extent that the locus extrapolated from the steady-state F2 is discrepant with the canonical locus, listeners should make identification errors.

To illustrate this error-generation process, consider the /du/ syllable which has a steady state F2 around 700 Hz. If the listener extrapolated the locus from this steady-state formant, he would infer that the locus of the consonant was around 700 Hz. In fact, this is close to the canonical locus for /bu/. Consequently, to the extent that listeners extrapolated the locus from the steady-state formants, they should mistake /du/ for /bu/. This did occur; as Table 3 shows, /da/ → /du/ and /di/ → /du/ were erroneously identified as /bu/ 34% and 21% of the time, respectively.

This process manifests itself very regularly in the data. For example, /bi/ might be erroneously identified as /di/ because the steady-state F2 for /bi/ has a high frequency. Consequently, the locus of /bi/ extrapolated from the steady-state F2 would appear to be closer to the canonical locus of /d/ (1,800 Hz) than to the canonical locus of /b/ (700 Hz). Similarly, /di/ might be erroneously identified as /gi/ because the steady-state F2 for /di/ has a higher frequency than the canonical locus of the /d/. Consequently, the extrapolated locus would be closer to the canonical locus of /gi/ (3,000 Hz) than the canonical locus of /d/ (1,800 Hz). A similar explanation might explain the pattern of responses for /g/; however, /g/ is more complex, since the 3,000 Hz locus is effective only for certain vowel contexts. In summary, the error patterns are generally systematic. However, it is important to remember that the responses were correct 72% of the time, so that the locus extrapolated from F2 was not necessarily the major cue listeners used, otherwise the overall level of performance would not have been so high. Still, it was a cue that listeners sometimes used.

Performance was better for the voiceless stops than for the voiced stops. One reason may be that the aspiration following the burst of a voiceless stop carries voiceless formant transition information, and this voiceless formant transition can cue the perception of the stop consonant. This possibility was explored further in Experiment 2.

Discussion

Experiment 1 generally indicated that even in the absence of voiced formant transitions, natural stop consonants can be correctly identified fairly well, 84% of the time when the stop is presented with the original vowel. In some cases, they are even identified well when placed in a vowel context that is different from the production context. The probability of correct identification varies considerably from stop to stop and depends on the source and destination vowel contexts. For /b, d, p, t, k/, the consonants are identified correctly in foreign contexts 74% of the time. It would suggest that the voiced formant transitions are not a necessary cue and that in natural speech, there are other cues that are sufficient to attain fairly high identification. However, for /g/, the identification in foreign vowel contexts was only 28%. Thus, for /g/, the voiced formant transition seems to be an important cue.

The results in Table 3 can be compared to those of Cole and Scott, who used manual splicing to produce their stimuli (Cole & Scott, 1974, Experiments I and II). They examined 36 of the 54 conditions reported in Table 3. In those 36 conditions, Cole and Scott reported a mean identification rate of 91%, while the current experiment found 80% correct identification. In the remaining 18 conditions not examined by Cole and Scott, the current study found a mean identification rate of 58%.

The results may also be compared to those of a recent study which examined the perception of /b, d, g/ in eight vowel contexts, including /-id, -ad, -ud/ (Dorman, Studdert-Kennedy, & Raphael, 1977).[1] Two different speakers produced the stimuli, and their utterances produced dissimilar results. For the conditions corresponding to the nine control syllables Dorman et al. obtained 30% correct identification with Speaker 1 and 60% with Speaker 2, while the current study obtained 72% correct identification. The results for Speakers 1 and 2 were moderately correlated with each other across these nine conditions, r(8) = .63. An equally good correlation was obtained between Speaker 1 and the current results, r(8) = .69; but the correlation was low with Speaker 2, r(8) = .14. Thus, the current results are similar to Dorman et al.'s Speaker 2 in absolute performance level and similar to Speaker 1 in the performance pattern.

The analysis of the errors in the present study indicated the role that voiced formant transitions might play in the identification of stops. The identification of a stop depended on what vowel was presented with the stop. For example, the type of errors that occurred for /d/ depended on whether the /d/ was presented before an /a/ or an /i/. The voiced formant transition (or in its absence, the steady-state second formant) was used to extract the locus of the consonant. Thus, listeners computed the identity of the stop partially on the basis of the transition and vowel that followed it.

In summary, recognition rates for syllables lacking voiced formant transitions are fairly high. The role of voiceless formant transitions was examined in Experiment 2.

EXPERIMENT 2

The aspiration following the burst of a voiceless stop may contain sufficient information for consonant identification (Fischer-Jørgensen, Note 1). In fact, when the aspiration and burst provide conflicting cues to the identity of a voiceless stop consonant, the aspiration often determines the resulting perception (Fischer-Jørgensen, Note 1). The aspiration contains some low-energy formant-transition information. Consequently, it is important to determine the role of the aspiration in perceiving natural stop consonants.

To examine the contribution of the aspiration, the same design was used as in Experiment 1, except that the aspiration was replaced by a silent interval of the same duration. This procedure should be particularly detrimental to the identification of the voiceless stops for three reasons. First, the procedure removes the aspiration that carries formant transition information useful in identifying a stop consonant's place of articulation. Second, the interval is fairly long (64-74 msec) in the case of voiceless stops. Third, the inserted silent interval may provide a misleading cue to the voicelessness. In fact, silence followed by a fairly abrupt onset of voicing is often interpreted as a voiced stop. In many synthetic speech perception experiments on voiced stops, the stimuli consist of only voiced formant transitions and steady-state vowels. The listeners hear silence, then a vowel transition, then a steady-state vowel, and under these circumstances they almost always report hearing a voiced stop rather than an unvoiced one. In other words, the silence that precedes the formant transition is part of a cue complex that is often interpreted as a voiced stop. Consequently, one might predict that voiceless stops whose voicing onsets are preceded by complete silence may be sometimes be mistaken for voiced stops.

The effect of removing the aspiration following a voiced stop is harder to predict. The replaced interval in the case of voice stops is brief (9-22 msec) and contains only silence and very low-energy voiceless formant transitions, if any. The current procedure will be detrimental to the extent that the energy in this interval is used in the identification process. As in Experiment 1, in both the voiceless and voiced cases, the voiced formant transitions were also eliminated.

Procedure

After completing Experiment 1, the subject was given a short rest and then began Experiment 2. The same procedure was used. The 306 stimuli were presented in random order, and the subject wrote down what he heard as soon as it was presented.

Results

Recognition deteriorated markedly when the aspiration was replaced by silence and the voiced formant transition was removed, as shown in Table 4. The mean correct identification, 38%, was approximately half the rate in Experiment 1. Again, performance was better for the bilabial and alveolar consonants than for the velar consonants. Also, correct recognition was higher for two of the control conditions, /i/ → /i/ and /u/ → /u/, and lower for /a/ → /a/ and the experimental conditions in which the vowels were exchanged. The number of unclassifiable responses was higher than in Experiment 1; on 14% of the trials, the subjects failed to respond or misidentified the vowel.

Voiceless stops. Performance for the voiceless stops, /p/, /t/, /k/, was greatly impaired; mean correct identification was only 29% in this experiment, whereas it was 83% in Experiment 1. For the control syllables, the average performance was only 31% correct, although 7 of the 9 control cells showed above-chance performance on correct phoneme identification. For the experimental syllables, in which the vowels were exchanged, performance was also poor, only 28% overall. Still, performance was clearly above chance in 11 of the 18 cells; 3 more cells were borderline and 4 were clearly no better than chance. This analysis indicates that listeners could correctly identify voiceless stops at better than chance level (16.7%) in 18 of the 27 cells in the absence of voiced or voiceless formant transitions.

While the identification of voiceless stops was low

Table 4
Percentage of Correct Responses (and the Most Common Error) for Experiment 2

	/a/→/a/	/i/→/a/	/u/→/a/	/i/→/i/	/a/→/i/	/u/→/i/	/u/→/u/	/a/→/u/	/i/→/u/	Mean	Errors
/b/	59 *	76 *	66 *	61 (d 13)	46 (d 19)	54 (d 16)	87 *	84 *	83 *	68	12
/d/	21 (b 55)	32 (b 31)	40 (b 20)	68 (g 12)	57 (g 15)	56 (g 25)	66 (b 13)	32 (b 46)	45 (b 32)	46	10
/g/	9 (d 31) (b 27)	11 (d 34) (b 21)	21 (b 16) (k 14) (t 12)	53 (d 20)	28 (d 28) (b 21)	24 (b 26)	68 (b 13)	28 (b 42)	7 (b 34) (d 34)	28	15
/p/	26 (b 41)	26 (b 18) (d 14)	30 (b 26)	30 (b 20) (g 12)	34 (b 14)	33 (b 25)	41 (b 26)	32 (b 28)	33 (b 25)	32	17
/t/	8 (p 23) (b 19) (g 16)	30 (p 16) (b 14) (d 14)	29 (d 15) (g 13)	20 (g 21) (b 13)	28 (d 16) (p 13)	15 (k 20) (g 18) (d 14) (p 12)	36 (p 14) (d 14)	18 (d 20) (p 14)	23 (d 33) (p 13)	23	16
/k/	32 (g 28)	15 (t 15) (b 15) (g 15)	38 (g 29)	42 (g 31)	23 (p 21) (g 17)	57 *	45 (g 42)	23 (g 25) (p 13)	11 (t 20) (p 16) (d 15)	32	13
Mean	26	32	37	46	36	40	57	36	34	38	
Errors	17	14	15	13	14	13	12	15	13	14	

Note–Data are based on 280 observations per cell for the experimental syllables and 120 observations per cell for the control syllables. "Errors" are the percentage of times no response was given or the vowel was incorrect.
**Confusions were too few or no one kind received more than 11% of the responses.*

overall, Table 4 indicates that subjects could correctly identify the place of articulation about half the time. For example, the mean identification accuracy of /p/ was 32% overall; however, if all responses that preserved place of articulation were counted as correct, then the /b/ response is no longer an error and performance is at the 56% level. This computation requires that responses that preserved place be added to the correct responses. (These erroneous response rates appear in Table 4 only if they occurred more than 11% of the time.) With this scoring procedure, /p/ produced responses that preserved place 56% of the time; /t/ and /k/ produced correct place responses 39% and 56% of the time, respectively. Chance performance would be 33% rather than 16.7%; nevertheless, place is perceived significantly better than chance in all nine cases for /p/, for seven of nine cases involving /t/, and seven of nine for /k/. Thus, place is very often correctly perceived at better than chance level for voiceless stops whose vowels contain no voiced or voiceless formant transitions; overall, the place of articulation is correctly identified roughly 50% of the time.

A comparison between Experiments 1 and 2 indicates the extent to which the aspiration following a voiceless stop is used to identify the stop. The aspiration carries formant transition information that helps identify the place of articulation. When this cue was removed in Experiment 2, the place of articulation was perceived more poorly than in Experiment 1, with a 33% decrement in performance.

The aspiration following a voiceless stop is also part of the cue complex that identifies the stop as voiceless. The aspiration is a filled interval (lasting 64-74 msec in this sample) that may provide continuity between the burst and the vowel. This continuity makes the voicing onset more gradual and helps identify the feature of voicelessness. When the aspiration was replaced by silence, the voiceless stops were sometimes identified as the voiced stops of the corresponding place of articulation. In fact, /p/ was incorrectly perceived as /b/ 24% of the time, /k/ was incorrectly perceived as /g/ 23% of the time, and /t/ was incorrectly perceived as /d/ 16% of the time. Such errors did not occur in Experiment 1 where the aspiration was present. Thus, the aspiration provides an important cue to perceiving a stop as voiceless.

Voiced stops. Replacing the aspiration did not impair the identification of the voiced stops as much as the identification of the voiceless stops. The recognition accuracy for voiced stops was 47% in Experiment 2 as compared to 62% in Experiment 1. Clearly, listeners could still identify /b/ and /d/ at better than chance level. Nevertheless, the 15% decrement suggests that the interval between the burst and the vowel transition does carry some information, even in voiced stops. It is interesting to note that in Experiment 2, where there were no formant transitions, place of articulation was perceived approximately equally well for voiced (47%) and voiceless (50%) stops. In addition, the voicing feature was almost always correctly perceived. As mentioned above, the inserted silent interval preceding the steady-state vowel is a cue that leads to a correct response for voiced stops.

In the nine voiced control syllables (e.g., /ba/ → /ba/), identification accuracy was on average 55%, with /a/ → /a/ producing worse performance than /i/ → /i/ and /u/ → /u/. In the corresponding nine conditions of the Dorman et al. (1977) study, their Speaker 1 produced 16% accuracy while Speaker 2 produced 58% accuracy. The correlations between Speakers 1 and 2, and between each speaker and the current study were all around .3.

In the 27 voiced experimental syllables (e.g. /ba/ → /bi/), correct identification was at 47%, with performance generally poorer if the stop was /g/ or if the presented vowel was /a/. The corresponding performance level in the Dorman et al. (1977) study (involving only Speaker 2) was similar, 45%, and the correlation between the two experiments was r(26) = .48.

The analysis of errors. The confusion errors for the voiced stops were quite systematic and similar in pattern to those of Experiment 1. The most common misidentification for /bi/ was /di/, for /da/ was /ba/, for /di/ was /gi/, and for /du/ was /bu/. Again, these may be explained by assuming that the listener extrapolates the consonant locus from the second formant of the steady-state vowel and uses this extrapolated locus as one cue in identifying the consonant. Since /g/ does not have a fixed locus, the locus extrapolation process may yield a poorer cue for recognition, leading to a less systematic pattern of error responses.

GENERAL DISCUSSION

This discussion will first evaluate the relative contributions of various components of a stop consonant to its recognition, and second, evaluate the theoretical implications of these results.

The burst

In Experiment 2, subjects were presented with a burst, followed by silence, followed by a steady-state vowel. Under these circumstances, the consonant was correctly identified 38% of the time, and the place of articulation was correctly identified 49% of the time as compared to chance levels of 16.7% and 33%, respectively. The burst must be providing the information that makes performance above chance. Specifically, the burst frequency is likely to be an important cue to place of articulation. Furthermore, the duration of elapsed time between the burst and the onset of voicing (VOT) may provide another cue to place. This is possible since VOT increased by 8 msec as the place progressed from front to middle to back. Thus, the burst frequency and its temporal separation from the onset of voicing are two aspects of the burst that may provide cues to the place of articulation. Finally, the temporal separation between the burst and the onset of voicing is the major cue to indicate whether the consonant is voiced or voiceless. Specifically, a burst that is followed by a short VOT (say less than 40 or 50 msec) in which the voicing onset is fairly abrupt and preceded by silence is interpreted as a voiced stop. A burst followed by a much longer VOT (say over 50 or 60 msec) and a more gradual onset of voicing is a cue to a voiceless stop. Thus, a burst may be a cue to a stop consonant in terms of its own characteristics and in terms of its temporal relation to the vowel.

Aspiration

The aspiration between the burst and the voiced formant transition appears to be a cue to two features, place and voicing. Since Experiments 1 and 2 differed only in whether there was aspiration or silence, the contribution of the aspiration can be directly evaluated. First, the accuracy of identifying the place of articulation dropped by 24% when the aspiration was replaced by silence. Presumably, the cue to place of articulation in the aspiration was the voiceless formant transition.

The replacement of the aspiration by silence also caused a 21% incidence of incorrectly identifying voiceless stops as voiced in Experiment 2; this did not occur in Experiment 1. Thus, the aspiration is part of the long VOT cue complex that signals a voiceless stop.

Voiced Formant Transitions

These important components of stop consonants were excluded from this study to evaluate recognizability of stops in their absence. Experiment 1 showed that identification was fairly high overall (72%) and above 85% about half the time when there were no voiced formant transitions. However, the absence of voiced formant transitions was very damaging to the recognition of /g/.

Theoretical Implications

Sufficiency of cues. There is no obvious criterion to determine whether a cue is sufficient for perception. A stringent definition is that a cue is sufficient if and only if recognition of the stimulus with only that cue present is no worse than when all cues are present. A very lenient definition would require a sufficient cue to produce identification accuracy that is better than chance. By the stringent criterion, release bursts are not sufficient cues for the recognition of stop consonants, but they are sufficient cues by the weak definition. The sufficiency of aspiration and voiced formant transitions could be evaluated by similar experiments, but cannot be evaluated on the basis of the studies reported here.

Invariance of cues. To some extent, the issue of invariance is a question about the physical structure of the stimulus more than a question about psychological representations or processes. Either a cue is invariant across contexts or it is not. This issue could be left entirely to the field of acoustic phonetics provided that the available instruments and methodologies can isolate and distinguish all possible physical acoustic cues. To the extent that human perceptual mechanisms might be more accurate or relevant sensors of cues (invariant or otherwise), it is useful to study the human ability to detect invariants.

In the case of release bursts associated with stop consonants, acoustic phonetic examination indicates that some components are invariant, while others are not. The original discovery of the variability of voiced formant transitions across contexts stimulated a great deal of research, partially because the result contradicted our intuitions that we utter the same sound in different contexts. More current research may direct the focus from the variability per se to the psychological processes that interpret the cues. In this regard, the current study indicates that listeners seemed to extrapolate the consonant locus, attempting to use the vowel transition information. The striking pattern of errors suggests that listeners extrapolated from the steady-state F2 when the transition was removed. Consequently, when the extrapolated locus differed from the canonical locus, errors were more likely. Moreover, the precise nature of the error was often explicable in terms of the extrapolation process. Extrapolating a stop consonant locus is just one mental process that might be used in speech perception. A theory of speech perception will have to specify other such processes, instead of focusing predominantly on the characteristics of speech itself.

REFERENCE NOTES

1. Fischer-Jørgensen, E. *Tape cutting experiments with Danish stop consonants in initial position.* Annual Report VII, Institute of Phonetics, University of Copenhagen, Copenhagen, Denmark, 1972.

2. Baker, J. K. Machine-aided labelling of connected speech. In *Working papers in speech recognition II.* Computer Science Department, Carnegie-Mellon University, Pittsburgh, 1973.

REFERENCES

Cole, R. A., & Scott, B. The phantom in the phoneme: Invariant cues for stop consonants. *Perception & Psychophysics*, 1974, **15**, 101-107.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 1955, **27**, 769-773.

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics*, 1977, **22**, 109-122.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. Perception of the speech code. *Psychological Review*, 1967, **74**, 431-461.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. The role of selected stimulus variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, 1952, **65**, 497-516.

Lisker, L., & Abramson, A. S. Some effects of context on voice onset time in English stops. *Language and Speech*, 1967, **10**, 1-28.

Miller, G., & Nicely, P. An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, 1955, **27**, 338-352.

Reddy, R., & Newell, A. Knowledge and its representation in a speech understanding system. In L. W. Gregg (Ed.), *Knowledge and cognition*. Hillsdale, N.J: Erlbaum, 1974.

NOTE

[1] The article of Dorman et al. (1977) appeared in print while the current manuscript was under editorial review, and the consideration of the Dorman data was included in the revision of the manuscript. The Dorman data points have been estimated from the published graphs. The responses in the Dorman study were scored only in terms of place of articulation, ignoring voicing errors. However, since our data showed almost no voicing errors for voiced stops, we assume that Dorman's didn't either, so the results of the two different scoring procedures may be comparable.

(Received for publication August 15, 1977;
revision accepted June 18, 1978.)