Perception & Psychophysics
1985, 37 (2), 93-102

Perceptual classification of information in vowel-consonant syllables

RICHARD B. IVRY and PETER W. JUSCZYK
University of Oregon, Eugene, Oregon

Two experiments are reported which examined whether information specifying consonant-identity was available in brief segments at the offsets of vowel-consonant syllables. The first experiment employed a classification task in which the subjects were required to sort eight synthetic stimuli composed of two stop consonants, /b/ and /d/, in four vowel environments. It was found that the subjects' responses were best described by a classification strategy based on overall acoustic similarities between the stimuli. It was hypothesized that these acoustic similarities could be predicted by averaging the frequencies of the second and third formants at offset. A perceptual learning task was used in Experiment 2. Although the salience of the acoustic similarities was again evident, the results also indicated that the subjects were able to learn classification schemes based on acoustic-phonetic similarities. Subjects made fewer errors in learning to sort the stimuli by both consonant-identity and vowel-similarity rules in comparison to an arbitrary division when all of the formants were left intact. These data are interpreted as an indication that brief segments of speech contain invariant cues to phonetic identity and that the salience of phonetic classifications increases as the sounds retain more of the information found in speech.

For many years there has been much debate concerning whether speech contains invariant cues which allow the listener to abstract the phones that compose a particular utterance. On the basis of evidence drawn from analyses of sound spectrograms, Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967) argued that there were no invariant acoustic properties of each phone which signaled its presence in all contexts. For example, spectrographic analyses revealed that the formants for a phone such as /d/ varied greatly across different vowel contexts. Thus, Liberman et al. suggested that the coarticulation of consonants with adjacent vowels makes it impossible to isolate consonants in the speech stream. However, recent advances in understanding the physiology and psychophysics of the auditory system have led to the development of new ways of analyzing the speech signal. These alternative ways of analyzing the speech signal have rekindled interest in the existence of possible acoustic invariants for phones (e.g., Kewley-Port, 1980, 1983; Searle, Jacobsen, & Rayment, 1979; Stevens & Blumstein, 1978, 1981).

Stevens and Blumstein (1978, 1981; Blumstein & Stevens, 1979) presented an approach based on constraints on the acoustic signal imposed during speech production by the articulatory system. For instance, since the burst and formant transitions of a prevocalic stop consonant are produced by the same articulatory gesture, Stevens and Blumstein chose to look for an acoustic invariant of the consonant in a 10-20-msec time window that integrated information from this section of the speech signal, which they called "integrated" cues. Through this method of analysis, a set of templates were derived to capture the essential and invariant characteristics of particular phones. Although promising as an initial approximation (85% correct classification for the prevocalic consonants), the templates failed to approach the near-perfect identification rates achieved by humans in everyday perceptual experience. In addition, the templates were not particularly successful with postvocalic consonants (76% correct).

A different approach has been employed by Kewley-Port (1980, 1983), Searle et al. (1979), and Zwicker, Terhardt, and Paulus (1979). Rather than integrating the acoustic information across the whole consonantal portion of the sound, these researchers have emphasized the continuously changing energy distribution by sampling the signal at shorter intervals (e.g., every 5 msec). Through this method, Kewley-Port and her collaborators (1983; Kewley-Port, Pisoni, & Studdert-Kennedy, 1983) have identified a number of primary and secondary features which appear to be invariant as to place of articulation for stop consonants.

Regardless of the success of these latest attempts to identify acoustic invariants, it remains to be demonstrated that the human auditory system makes use of such cues in the natural processing of speech. As Jusczyk, Smith, and Murphy (1981) point out, "A description of the speech signal in terms of some invariant physical properties will provide a successful account of speech perception only insofar as it isolates those properties on which the per-

We wish to acknowledge the assistance provided to the second author through a grant from N.I.C.H.D. (HD 15795). We would also like to thank Steven Keele, Alvin Liberman, Michael Posner, James Sawusch, Linda B. Smith, and especially Deborah Kemler Nelson for comments they made on earlier versions of the present manuscript. Requests for reprints should be sent to Peter W. Jusczyk, Department of Psychology, University of Oregon, Eugene, OR 97403.

ceiver operates" (p. 11). Thus, any attempts to explain speech perception in terms of the detection of invariant acoustic properties must not only demonstrate the existence of such properties, but also show that these properties are the relevant ones for the perceiver (see also Dorman, Studdert-Kennedy, & Raphael, 1977).

One empirical link between the search for invariants in the acoustic signal and the psychological process of speech perception is provided by research that addresses the success of perception when the information available in the acoustic stimulus is limited to durations comparable to those for the proposed spectral templates. To the extent that the perceiver is able to assign such limited-duration segments to the appropriate phonemic categories, one can claim that the invariant acoustic cues in such segments are indeed psychologically relevant. Several investigators have examined how well subjects can identify the place of articulation when presented with only brief segments of monosyllables. For example, Stevens and Blumstein (1978) found that subjects were successful in identifying the consonant for 90% of consonant-vowel (CV) stimuli when both the burst and formant transitions were left intact. Removal of the bursts only reduced the accuracy rates to 81%. Subsequently, Blumstein and Stevens (1980) confirmed that stimulus duration had little effect on subjects' performance. High identification rates were obtained even with stimuli as short as 10 msec. Similarly, in another study using truncated speech stimuli, Kewley-Port (1980) found that accuracy rates reached 95% for the identification of 20-msec stimuli derived from natural speech tokens.[1]

Although these results are consistent with the view that there are invariant acoustic cues for stop consonants available to the perceiver, the labeling tasks used in these experiments were highly constrained with regard to possible response categories.[2] A more conservative measure was employed by Jusczyk et al. (1981), who used a mixed set of consonant-vowel (CV) syllables and a free classification task. They generated eight synthetic syllables by combining /b/ and /d/ with four vowels (/ɛ/, /i/, /o/, /ɔ/). In addition, two sets of 30-msec-duration truncated speech stimuli were derived from these syllables: A full formant set produced by truncating the syllables at the point at which the formant trajectories attained their steady state values, and a two-formant set that included only the second and third formants. The latter set was of interest because the acoustic properties of the other formants are the same for /b/ and /d/. Hence, any differential acoustic information regarding the identity of the stop consonants might be expected to lie within the region of the second and third formants.

Jusczyk et al. (1981) found that some subjects did spontaneously classify both the syllables and the truncated full-formant stimuli into two categories corresponding to the phonemic labels /b/ and /d/. This result suggested that, within the truncated full-formant stimuli, there are psychologically relevant cues sufficient to identify stop consonants.

Additional support for this view came from another experiment reported by Jusczyk et al. (1981), which was designed to investigate whether subjects could learn various rules for grouping each set of stimuli. Two of these rules required phonetic groupings: consonant-identity (/b/ vs. /d/) and vowel-similarity (/i/, /ɛ/ vs. /ɔ/, /o/). This grouping, which was one that emerged in subjects' spontaneous classification of the stimuli in earlier experiments, corresponds to a front versus back vowel distinction. By contrast, the third rule imposed an arbitrary grouping of the stimuli (i.e., one with no phonetic basis). For present purposes, the critical result was that subjects learned the consonant-identity rule significantly better than the arbitrary one for the syllables and the full-formant truncated stimuli, although not for the two-formant versions of these stimuli. Jusczyk et al. interpreted these findings as an indication that there are sufficient cues to consonant-identity in the truncated full-formant versions of the syllables but not in the two-formant ones despite the fact that the latter, nominally, includes the same acoustic information as do syllables. Accordingly, they argued that it was likely that the relationship between the formants, rather than the absolute values of formants, was crucial to the listener's perception of stop consonants.

Although recent efforts to uncover invariant acoustic properties for stop-consonant segments are encouraging, whatever successes have been achieved are almost exclusively limited to stops in syllable-initial positions. Attempts to provide templates for stops in syllable-final position have been considerably less successful (e.g., Blumstein & Stevens, 1979). In this respect, the lack of success up to now may be attributable to a variety of factors, including inadequate templates or even the possibility that there are no such acoustic invariants usable by the perceiver. One way of exploring this issue is to determine whether perceivers are able to employ information from brief segments of speech at the ends of VC syllables in order to group the segments according to consonant identity (e.g., /b/ vs. /d/). Although an investigation of this sort would not provide a description of the specific acoustic cues that the perceiver was operating on, it would at least indicate whether there was sufficient information in such brief segments to specify consonant identity. Accordingly, the primary impetus for the present study was to employ the methods of Jusczyk et al. (1981) to examine whether listeners could utilize information in the formant transitions of VC syllables to determine consonant identity.

In addition, we also wished to clarify the basis for the predominant classification pattern that Jusczyk et al. (1981) found in their study. As noted above, Jusczyk et al. observed that the subjects' preferred grouping scheme, with both the syllables and truncated stimuli, corresponded to a front-back vowel distinction. For this reason, they suggested that subjects were responding to phonetic qualities of the stimuli such as vowel similarity. However, an alternative possibility is that subjects responded to some more general acoustic property, such as overall pitch at

stimulus offset. By employing VC stimuli in the present study, it was possible to observe whether subjects were more prone to group the stimuli according to perceived vowel qualities or to overall pitch at offset.

EXPERIMENT 1

One test of the psychological relevance of any acoustic invariants in brief speech segments is to present the listener with a variety of different stimuli and have him or her group the stimuli into categories. If invariant information concerning phonetic identity is particularly salient, then the listener might be expected to form groups on this basis. Hence, the first experiment employed a classification task in which subjects heard a variety of different stimuli and were asked to assign them to two groups.

Following Jusczyk et al. (1981), we focused on the stop consonant pair /b/ and /d/ in four different vowel contexts (/i/, /ɛ/, /ɔ˞/, /o/). The vowel contexts were chosen to maximize differences in the relationships among the first, second, and third formants in order to provide the strongest possible test of potential invariant cues to the final consonants. Three types of stimuli were employed: VC syllables without release bursts, truncated full-formant versions of these syllables (containing the last 30 msec of each formant), and truncated two-formant versions of the syllables (containing the last 30 msec of the second and third formants only). Full VC syllables were included, since it was expected that subjects might easily sort these stimuli into categories based on the identity of their final consonants. The truncated full-formant stimuli were employed to examine whether there was invariant information in the final formant transitions which specifies consonant identity for the perceiver. The truncated two-formant stimuli were chosen as a further test of Jusczyk et al.'s claim that it is the relationship among the formants, and not merely the spectral frequency values of the second and third formants, that is critical for determining stop consonant identity.

Finally, as noted above, the predominant grouping strategy followed by subjects in the study by Jusczyk et al. could be based on either vowel similarity or the overall pitch at offset. In the present experiment, these two bases for classification were unconfounded. As before, a tendency to employ a vowel-similarity strategy would lead subjects to put /ib/, /id/, /ɛb/, /ɛd/ into one group and /od/, /ob/, /ɔ˞d/, /ɔ˞b/ into the other.[3] However, a grouping according to overall pitch at offset would result in the groups /ɛb/, /ob/, /od/, /ɔ˞b/ (low pitch) and /ib/, /id/, /ɛd/, /ɔ˞d/ (high pitch).[4]

Method

Stimuli. The stimuli consisted of eight synthetic syllables (/ib/, /id/, /ɛb/, /ɛd/, /ob/, /od/, /ɔ˞b/, /ɔ˞d/) plus two truncated versions of each. All stimuli were prepared on a LSI 11/23 computer in the Speech Perception Laboratory at the University of Oregon, and were generated with the cascade-parallel synthesizer designed by Klatt (1980) and modified by Kewley-Port (1978). Eight natural speech tokens spoken by P.W.J. served as models for constructing the synthetic syllables. The natural speech tokens were analyzed by the VOCODE program developed by Mertus (1982) which computes the frequency, bandwidth, and amplitude of the first four formants at 5-sec intervals by using a 26-msec time window.

The syllable stimuli were all generated without final release bursts and were equated for overall duration (295 msec) and pitch contour. The latter had an initial value of 121 Hz, rose to a peak of 125 Hz after 45 msec, and then fell linearly to a terminal value of 100 Hz. The amplitude of voicing had an initial value of 50 dB, rose to a peak value of 66 dB after 20 msec, and then dropped only slightly to 65 dB across the duration of the stimulus. This control of amplitude contours was done to avoid the possibility that the final formant transitions would be obscured by reductions in voicing amplification.

Syllables sharing a common vowel (e.g., /ib/ and /id/) were equated in all respects except for their second- and third-formant transition values. Table 1 presents the values of the first-, second-, and third-formant values sampled at four points in the duration of each synthetic syllable. To insure that the synthetic syllables were accurate representations of real speech sounds, an identification test was conducted. The eight synthetic syllables were each presented 10 times in a random order to 10 subjects. The subjects were given eight labels ("eb," "ed," "eeb," "eed," "ob," "od," "erb," "erd") and asked to identify each stimulus. The overall correct identification rate was 95.9% for the eight synthetic stimuli, ranging from a low of 85% for /ib/ to a high of 100% for /ɔ˞b/ and /ɔ˞d/.

The truncated full-formant stimuli were produced by removing the first 265 msec of each syllable, at which point the transitions of first, second, and third formants began. Thus, the truncated stimuli were 30 msec in duration. The relevant formant trajectories are, of course, identical to those of the full syllables and are displayed

Table 1
First, Second, and Third Steady-State Formant Frequencies (0-265 msec) and Transitions (265-295 msec) in Hertz for the Eight Vowel-Consonant Syllables Sampled at Four Points

Syllables	Formant Transitions	0 msec	150 msec	265 msec	295 msec
	1	220	265	300	200
(ib)	2	2250	2335	2400	1600
	3	3200	3200	3200	2400
	1	220	265	300	200
(id)	2	2250	2335	2400	2000
	3	3200	3200	3200	3000
	1	600	572	550	200
(ɛb)	2	1750	1807	1850	1100
	3	2500	2500	2500	2100
	1	600	572	550	200
(ɛd)	2	1750	1807	1850	1700
	3	2500	2500	2500	2700
	1	500	462	400	250
(ob)	2	1050	908	800	550
	3	2200	2252	2400	2500
	1	500	462	400	250
(od)	2	1050	908	800	550
	3	2200	2252	2400	2100
	1	600	600	600	200
(ɔ˞b)	2	1200	1200	1200	800
	3	1600	1600	1600	1100
	1	600	600	600	200
(ɔ˞d)	2	1200	1200	1200	1800
	3	1600	1600	1600	2760

Note – The relationship between the frequencies of all adjacent samples is linear.

in the last two columns of Table 1. Moreover, since the full-formant stimuli are merely abbreviated versions of the complete syllables, the spectrum for a given truncated full-formant stimuli is identical to that of the offset spectrum of the syllable from which it is derived.

The truncated two-formant stimuli were generated by removing the first-, fourth-, fifth-, and sixth-formant information. Since the removal of this information can result in a drastic change in the amplitude relations between the second and third formants, measurements of the amplitudes of the transition portions of these formants were made from each syllable using the VOCODE program devised by Mertus. The two-formant patterns were then generated on the parallel branch of the Klatt synthesizer, taking care to maintain the appropriate amplitude relations of the formants throughout the duration of the stimuli. Owing to the lack of acoustic energy in the regions of the first, fourth, fifth, and sixth formants, the spectra for the two-formant patterns differ considerably from those of the syllables and truncated full-formant patterns.

The stimuli were converted to analog form in real time via a 12-bit digital-to-analog converter, low pass filtered at 4.8 kHz.

Subjects. Thirty-six undergraduates at the University of Oregon served as subjects in the experiment. All were native speakers of English and reported no history of either speech or hearing disorder. The subjects received either course credit or $3 for participating in the experiment.

Procedure. The subjects were tested in groups ranging in size from two to six subjects. Each individual was seated at a partially enclosed booth equipped with a set of TDH-39 headphones and a response box. All of the sounds were presented on line by an LSI 11/23 computer. The order of presentation was always randomly determined within a series of the eight stimuli for a given condition. The sounds were separated by a 4-sec response period. The volume was adjusted with reference to a sound-level meter (Quest Electronics Model 215) so that the stimuli were played at a level of approximately 72 dB (A) SPL. Responses were recorded on line by the registration of which of two response buttons each subject pressed during the response period.

An equal number of subjects were assigned randomly to each of six experimental conditions. These conditions constituted a 3 × 2 factorial design in which one factor was stimulus type (syllables, full-formant stimuli, or two-formant stimuli) and the other factor was instruction ("form two groups" or "form two groups based on the final position similarities"). Depending on their test condition, subjects were instructed that they would be hearing syllables or brief segments of eight different speech sounds and that they would have to sort the sounds into two groups. The subjects in the final-position conditions were told to form two groups by focusing on similarities in the final portion of each sound. The subjects in the other instruction conditions were simply directed to form two groups by "putting together the stimuli which sound the most alike." For subjects presented with the truncated stimuli, it was reiterated that the stimuli were shortened versions of speech sounds. Following Nusbaum, Schwab, and Sawusch (1983), we hoped to encourage the subjects to use whatever linguistic information was available in the truncated stimuli.

The subjects then heard each of the eight sounds of the stimulus set to which they were assigned so that they might familiarize themselves with the stimuli. Following this, the subjects were directed to listen carefully to two more series of the eight test items and begin sorting the stimuli into two groups by pressing response keys labeled "1" and "2." These responses were recorded and scored by the computer. The subjects were instructed to make a response following each sound even if they were uncertain as to which group that sound belonged to. After the practice trials, the subjects were administered an 80-item test sequence. A 5-min break followed this first sequence, after which the subjects were run through two more practice series of the same stimulus set and then a second test sequence of 80 items. At the conclusion of this final phase of the experiment, the participants were asked to write down the strategies and criteria they had used in forming their groups.

A complete experimental session lasted approximately 40 min.

Results and Discussion

Following Jusczyk et al. (1981), only the data from the second 80-item test series were analyzed, since response patterns showed little difference between the two test sequences. The first phase of the analysis focused on the question of how consistent subjects were in assigning a given stimulus to a particular group. Inconsistent classifications would imply that the subjects had difficulty either in classifying certain sounds or remembering the groups they had formed. Following Jusczyk et al. (1981), the relative H statistic was used to measure the amount of uncertainty present in subjects' categories (Attneave, 1959; Garner, 1962). In situations that involve equiprobable alternatives, H represents the minimum number of binary digits into which an event may be encoded. The consistency with which a subject classified a given stimulus is equivalent to 1-Rel H, where Rel H = H/H max.[5] Single consistency scores for each of the eight stimuli within the four conditions were computed. These scores were then submitted to an ANOVA of a 3 (stimulus type) × 2 (rule) × 8 (sounds) mixed design, which revealed significant main effects for stimulus type [$F(2,30) = 10.12$, $p < .001$], and sounds [$F(7,210) = 11.58$, $p < .001$], and the interaction of these factors [$F(14,210) = 3.21$, $p < .001$]. None of the remaining main effects or interactions approached significance. Post hoc analyses conducted according to the Tukey method (overall $p < .05$) revealed that subjects were more consistent in classifying the syllables than either of the truncated stimulus sets, which suggests that the former were more discriminable and/or easier to remember. The significant interaction was the result of the fact that subjects in the syllable condition were more accurate in maintaining their groups with only four of the sounds (viz. /ɛb/, /id/, /od/, /ɔ˞d/). More important, the overall consistency score of .70 (SD = .17) is acceptable. (Note that had all six subjects in a condition assigned a particular sound eight times to one group and two times to another, the consistency score for that sound would be only .28, a score well below the observed scores for all eight sounds.) Hence, it can be concluded that the subjects were able to consistently sort the stimuli into two groups.

Naturally, the data of greatest interest are those relevant to the kinds of groupings the subjects formed. For this purpose, we calculated the mean proportion of trials that fit the vowel-similarity and consonant-identity groupings for each condition. With respect to the syllable stimuli, none of the subjects employed a consonant-identity grouping, two subjects followed the vowel-similarity pattern (i.e., [i], [ɛ], vs. [o], [ɔ˞]), and the remaining four subjects adopted different vowel-based groupings (e.g., [i], [o] vs. [ɛ], [ɔ˞]). A different pattern of results emerged for those subjects instructed to sort the

syllables according to similarities in their final position. Four of the subjects tended to group the stimuli according to consonant-identity, although two of these subjects generally placed /ib/ in the /d/ group. It should be noted that subjects had the most difficulty in labeling this sound in our preliminary identification test. The other two subjects classified the sounds according to vowel similarities. The mean proportion of the responses for this condition that conform to a consonant-identity grouping is .78, whereas the mean proportion that follow a vowel-similarity grouping is .68.[6] These means do not differ significantly [$t(5) < 1.0$]. In view of the instructions to sort the stimuli according to similarities in their final position, it was somewhat surprising that two of the subjects did not employ consonant-identity groupings. This is probably due to the relatively brief portion of the sound containing formant transitions and also a result of the burstless construction of these stimuli. Malecot (1956) has reported a similar finding. Nevertheless, it seems clear not only that the vowels are salient in the syllables, but that the consonant information can also be abstracted.

Since our principal question concerned the manner in which listeners process speech sounds containing only the transitional portion of the signal, classifications of the two sets of truncated stimuli were analyzed by examining responses with respect to three different classification strategies. Two of these correspond to phonetic rules—one based on consonant-identity in which a perfect grouping would distinguish between the /b/ and /d/ sounds and the other to vowel similarities, a front-back vowel distinction (i.e., /i/, /ɛ/, vs. /o/, /ɔ/). An acoustic rule was also tested which split the stimuli into two equal groups according to the mean frequency of the second and third formants at offset. As noted previously, it was expected that if subjects were to apply an acoustic rule of this sort, then one group would be composed of /ɛb/, /ob/, /od/, and /ɔb/ and the other group would contain /ɛd/, /ib/, /id/, and /ɔd/. As is evident in Figure 1, the acoustic rule accounts for a greater proportion of classifications in all four conditions. Neither the instructions [$t(5) < 1.0$] nor the stimulus sets [$t(10) < 1.0$] significantly altered the classification patterns. Across all conditions, the mean proportions of responses sorted by consonant-identity, vowel similarity, and acoustic-similarity are .65, .78, and .83, respectively. Paired t tests were performed to test these differences. Both the acoustic and vowel rules were significantly better than the consonant-identity rule [vowel vs. consonant, $t(23) = 5.31$, $p < .001$; acoustic vs. consonant, $t(23) = 12.37$, $p < .001$]. Moreover, the acoustic rule was found to be significantly better in describing subjects' classifications than the vowel-similarity rule [$t(23) = 3.39$, $p < .01$]. Note that the high score observed for the vowel-similarity rule is probably attributable to its great overlap with the acoustic-similarity rule in the present case. If subjects consistently employed the acoustic-similarity rule, they would be scored correct on the vowel-similarity rule 75% of the trials. Thus, it appears that subjects in the present experiment found that the acoustic similarities at offset provided the most salient means of grouping the truncated stimuli.

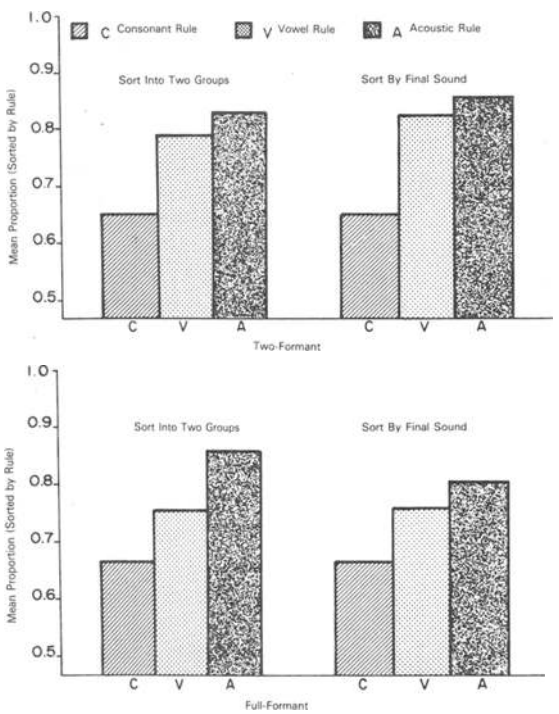

Figure 1. The mean proportion of trials correctly sorted by the consonant-identity, vowel-similarity, and acoustic rules for the two types of vowel-consonant chirp stimuli.

The present results, then, suggest that subjects in the earlier experiment by Jusczyk et al. (1981) might be more appropriately described as following an acoustic-similarity rule relating to overall pitch at offset, as opposed to a vowel-similarity rule. This tendency to group stimuli undergoing rapid spectral changes in terms of their offset frequencies has been observed in other contexts (e.g., Brady, House, & Stevens, 1961; Grunke & Pisoni, 1982; Shattuck & Klatt, 1976).

Thus, it appears that acoustic similarities rather than phonetic similarities are the most salient factors in adults' classifications of truncated speech stimuli. Nevertheless, it need not follow that phonetic information is unavailable in these truncated stimuli, although the rather poor fit of the consonant-identity rule to the data raises questions as to whether information about consonant identity really is available in the truncated VC stimuli. Only one of the 24 subjects in the present study could be classified as using a consonant-identity rule on at least 80% of the trials.[7] However, it is possible that phonetic information, though less salient than acoustic information in such stimuli, is accessible to the listener under certain conditions. For example, a perceptual learning task may prove to be a more sensitive measure of the degree to which listeners can abstract certain kinds of structural relations (e.g., Grunke & Pisoni, 1982; Jusczyk et al., 1981). Hence, the following experiment employed a perceptual

learning task to evaluate whether invariant cues to phonetic identity were present in truncated segments of VC syllables.

EXPERIMENT 2

Although subjects do not spontaneously group the truncated stimuli according to consonant identity, they might still be able to abstract consonant information from these stimuli if required to do so explicitly. Specifically, if there is information available to the perceiver for a partitioning of the stimuli according to consonant identity, subjects should be able to learn to sort by this rule. In particular, if such stimuli contain invariant information about consonants that is psychologically meaningful, it should be easier to learn to sort by a consonant-identity rule than by an arbitrary one.

In Experiment 2, subjects were trained to group the eight vowel-consonant syllables, or truncated versions of these, according to four different classification schemes. Two of these were based on the phonetic properties of the sounds, that is, vowel similarity or consonant identity. A third rule, acoustic similarity, required the subjects to split the stimuli into two groups by distinguishing the high-frequency sounds at offset from the low ones. A fourth rule was devised which imposed an arbitrary organization on the stimuli and, therefore, could not be characterized by either phonetic or acoustic properties. This last rule served as a baseline condition, since the only way it could be learned was by memorizing the individual items belonging to a group. Thus, differences in the number of trials required to learn these phonetic and acoustic rules relative to the arbitrary rule should provide an index of the psychological status of these different classes of information. In addition, a second phase of this experiment looked at the speed with which subjects were able to employ these rules once learning had been achieved. It was expected that the reaction time data would provide converging evidence regarding the psychological status of the different groupings.

Method

Stimuli. The stimuli were identical to those used in Experiment 1.

Subjects. Twenty-four undergraduates at the University of Oregon served as subjects in the experiment. All were native speakers of English and reported no history of either speech or hearing disorder. The subjects received either course credit or $6 for participating in the experiment.

Procedure. Each subject was tested individually in a subject station that allowed on-line presentation of the stimuli and recording of responses (see description in Experiment 1). An equal number of subjects (eight) were assigned randomly to each of the three stimulus conditions (i.e., full VC syllables, truncated full-formant stimuli, or truncated two-formant stimuli). Within a given stimulus condition, each subject was trained to sort the stimuli according to *all four of the classification rules*. These rules are presented in Table 2.

The order of learning the four rules was counterbalanced within each condition. For each of the rules, the following procedure was employed. A subject was instructed that he or she would be hearing eight different sounds. Subjects in the syllable condition were told that the stimuli were synthetic speech sounds, whereas those in the truncated stimulus conditions were told that the stimuli were shortened versions of speech sounds. The subjects were informed that four of these sounds constituted Group 1 and that the other four were the members of Group 2. The subject's task was to learn to assign each stimulus to its designated group. The experimenter then presented the four sounds that belonged to Group 1 at a rate of one sound every 2 sec. Following a 5-sec pause, the four members of Group 2 were played. The instructions were then repeated, and the groups demonstrated a second time. Following this, a training period was conducted in which each of the eight stimuli were played three times in a random order at a rate of one stimulus every 4 sec. The subjects were instructed to press one of two response buttons, depending on which group each stimulus was perceived to belong to. The ordering of the buttons was counterbalanced across subjects. Feedback was provided by a light that would come on above the correct group 2.5 sec after the stimulus had been presented. Any response made after the feedback light had gone on was counted as incorrect. The light would go off after 1.0 sec, thus leaving .5 sec in which the subject could prepare for the next sound.

A subject was deemed to have successfully learned a given rule if he or she responded correctly on at least 20 of the 24 training trials. If the subject fell below this criterion, the procedure was repeated. Once again, the two demonstration sets were played, followed by another block of 24 trials. Testing continued in this manner for each rule until either a subject learned the rule or four unsuccessful training blocks had been completed. In the latter circumstance, testing on the rule was terminated. Whenever a subject did succeed in learning a particular rule, he or she was immediately tested on the corresponding speeded classification task. The subject was instructed to continue assigning the sounds to their appropriate groups and told that the response times would also be measured. Thus, the subject was encouraged to respond as quickly as possible while maintaining accuracy. No feedback was provided regarding the correctness of responses during the speeded classification tasks. An 80-item test block was composed of 10 series of the eight sounds. The stimuli were randomized within a series and were spaced at 4-sec intervals. No response was recorded if the subject failed to respond within 3 sec.

Following completion of testing with the first rule (after either the speeded classification task or four unsuccessful training blocks), the subjects were given a 5-min break before the entire process was repeated for a second rule. To reduce fatigue effects, the subjects were required to return the following day for testing with the third and fourth rules. An entire experimental session took approximately 1½ h—45 min each day.

Table 2
Four Sorting Rules for Learning and Speeded Classification Tasks

Consonant			Vowel Similarity			Acoustic			Arbitrary		
(ib)		(id)	(ib)		(ob)	(ɔ·b) (950)		(ib) (2000)	(ib)		(id)
(ɛb)		(ɛd)	(id)		(od)	(ob) (1525)		(ɛd) (2200)	(ob)		(od)
(ob)	vs.	(od)	(ɛb)	vs.	(ɔ·b)	(ɛb) (1600)	vs.	(ɔ·d) (2280)	(ɛd)	vs.	(ɛb)
(ɔ·b)		(ɔ·d)	(ɛd)		(ɔ·d)	(od) (1800)		(id) (2500)	(ɔ·d)		(ɔ·b)

Note—Mean frequency of second and third formants at offset is listed after each stimulus under acoustic rule.

Results

Table 3 presents the number of subjects who learned the four different sorting rules for each condition. All of the subjects who heard the syllables were able to learn the vowel rule; 75% were successful in mastering the acoustic and consonant rules. Only two subjects learned the arbitrary rule to criterion. For both types of truncated stimuli, subjects were most successful with the vowel and acoustic rules. The number who reached criterion with the consonant rule was considerably less—three subjects in the full-formant condition and only two in the two-formant condition. One of these subjects in the latter condition also successfully completed training with the arbitrary rule.

As displayed in Figure 2, the number of errors made during training serves as an index of perceptual learning. It is obvious that subjects tended to experience less difficulty with the vowel and acoustic rules. To verify this, the error data for individual subjects were submitted to an ANOVA of a 3 (stimulus type) × 4 (classification rule) mixed design. The main effect for stimulus type approached significance [$F(2,21) = 3.07$, $p < .10$], and there was a highly significant main effect for rule [$F(3,63) = 45.96$, $p < .001$]. Furthermore, there was also a significant interaction between these two factors [$F(6,63) = 3.39$, $p < .01$]. Post hoc comparisons based on the Tukey method (overall $p < .05$) establish that a number of factors contribute to this interaction. First, consider the ease with which the various rules were learned. The vowel-similarity rule proved to be easier to learn than either the consonant-identity or arbitrary rules for all three stimulus types. In addition, the vowel-similarity rule was superior to the acoustic rule for the syllable stimuli, whereas there were no significant differences between these two rules for either of the truncated stimulus types. Hence, the acoustic rule was significantly better than either the consonant-identity or arbitrary rules for the truncated stimuli. Finally, the consonant-identity rule was superior to the arbitrary rule with both the syllables and truncated full-formant stimuli, but there was no difference between the rules with the truncated two-formant stimuli.

Comparisons across the three stimulus types revealed an interesting tendency. The consonant-identity rule was learned more easily in the syllable condition than in either truncated speech condition, whereas for the acoustic rule nearly the opposite was true (viz, it was easier in the two-formant condition than in the syllable condition). It was also observed that both the vowel-similarity and arbitrary rules were learned more successfully in the syllable condition than in the full-formant condition.

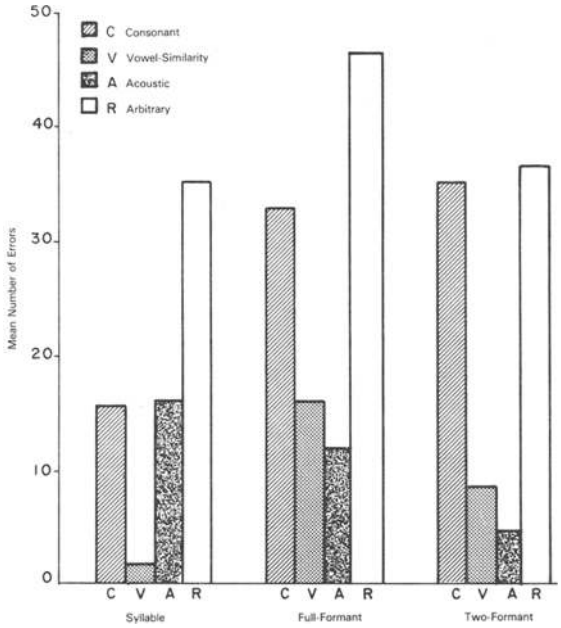

Figure 2. Mean number of errors for each type of stimulus when subjects were required to sort the stimuli according to the four classification rules.

The speeded classification results exhibited much the same pattern as the learning results. The mean reaction times for those rules which were learned to criterion by at least two subjects are given in parentheses in Table 3. One-way ANOVAs with repeated measures were conducted separately for each stimulus type. All three analyses attained significance [syllable, $F(3,21) = 10.45$, $p < .001$; full-formant, $F(2,12) = 4.99$, $p < .05$; two-formant, $F(2,14) = 8.86$. $p < .01$]. Post hoc comparisons were again conducted with the Tukey method (overall $p < .05$ for each condition). Paralleling the learning results, the vowel-similarity and acoustic rules produced faster reaction times than the consonant-identity rule and did not differ from each other for either of the truncated stimulus types. For the syllables, comparisons between the vowel-similarity rule and the other rules are not particularly informative because the vocalic portion of the stimulus is available 260 msec sooner than the formant transitions. Indeed, the mean RTs for two of the subjects in this condition were less than 265 msec. However, the other comparisons in the syllable condition are appropriate and revealed, consistent with the learning results, that the arbitrary grouping was significantly slower than either the consonant-identity rule or the acoustic rule, which did not differ from each other.

Table 3
Number of Subjects in Each Stimulus Condition Who Learned the Four Sorting Rules to Criterion

Stimulus Condition	Sorting Rules			
	Consonant	Vowel Similarity	Acoustic	Arbitrary
Syllables	6 (673.2)	8 (486.5)	6 (786.3)	2 (1098.0)
Full-Formant	3 (798.0)	5 (620.6)	6 (639.1)	0
Two-Formant	2 (930.7)	7 (734.4)	8 (716.5)	1

Note—Mean reaction times in speeded classification task given in parentheses for any condition in which at least two subjects reached criterion.

Discussion

The present experiment sought to determine whether truncated portions of VC syllables contain sufficient information to specify phonetic (and, in particular, consonant) identity to the perceiver. To the extent that subjects were better able to learn to group the syllables and truncated full formants more easily with the consonant-identity and vowel-similarity rules than with the arbitrary rule, it would seem reasonable to conclude that there is some psychologically relevant invariant phonetic information available in these stimuli. The organization of the stimuli into groups according to shared vowel-similarity or consonant-identity provided subjects with an advantage over a purely arbitrary grouping. Note that the superior performance of subjects with the consonant-identity rule, as compared with their performance with the arbitrary rule, cannot be ascribed to greater overlap with the acoustic-similarity rule. Both the consonant-identity and arbitrary rules overlapped with the acoustic-similarity rule to the same extent. Thus, it seems likely that the organization provided by the consonant-identity rule does convey some psychologically relevant advantage to the perceiver.

In contrast to Jusczyk et al.'s (1981) results with CV stimuli, the consonant-identity rule proved more difficult to learn than the vowel-similarity rule for all three types of stimuli. One likely explanation for these results is that the VC syllables in the present study were all burstless stops. It is a well-known finding that stops without bursts in syllable-final positions are less well perceived than those with release bursts (Malecot, 1956). Similarly, in their attempt to isolate acoustic invariants for stops, Blumstein and Stevens (1979) found that sampling across the burst portion of the sound greatly increased the percentage of final position /d/s, which were matched to their /d/-template. Nevertheless, the present results with burstless stops do provide some encouragement to those looking for invariant cues to identity of final stops.

Consistent with the earlier results of Jusczyk et al. (1981) with two-formant patterns, there was no indication that subjects in the present study learned the consonant-identity rule appreciably better than they did the arbitrary one. Hence, the presence of energy in the first formant region seems to be a necessary part of any invariant cues to consonant identity.

Although there may be perceptually accessible information about consonant-identity in the syllables and truncated full-formant stimuli, it certainly is not the most salient type of information. Across all three types of stimuli, the vowel-identity rule proved easier to learn than the consonant-identity rule. Similarly, for both sets of truncated stimuli, a rule based on acoustic-similarity at offset proved to be significantly easier to learn than the consonant-identity rule. There are a number of factors which may have contributed to the greater salience of the vowel-similarity and acoustic-similarity rules. Certainly with respect to the VC syllables, the vocalic portion is longer and louder than the consonantal portion. However, there are, clearly, other important factors responsible for the saliency of the vowel-similarity grouping, since it is found even with the truncated stimulus sets where there is no durational advantage for the vowel over the consonant. One possible explanation for the ease with which the vowel-similarity rule was learned across all stimulus types is that there are two different types of structural properties that could be used to form stimulus groupings that conform to this rule. Thus, in addition to a phonetic division of the stimuli into front and back vowels, there is a potential acoustic basis for the same groupings. In particular, a division of the stimuli according to the mean of the second and third formant frequencies at *onset* would produce the same kinds of groups as a phonetic division according to front versus back vowels. Hence, it is possible that, across the different stimulus types, different subjects may have been using different information to learn the same rule.

The fact that the acoustic-similarity rule was so readily learned, especially with the truncated stimuli, is consistent with findings from a number of other studies dealing with nonspeech stimuli. In particular, this rule grouped the stimuli according to acoustic-similarity at offset. Thus, Brady et al. (1961) found that subjects were most likely to match a comparison tone to the offset frequencies of stimuli undergoing rapid spectral changes. Similarly, in a study employing nonspeech sine-wave stimuli, Grunke and Pisoni (1982) found that subjects were considerably more adept at learning to group these stimuli according to their offset characteristics than according to their onset characteristics. In this respect, it is worth noting that subjects seemed to encounter the most difficulty in learning the acoustic-similarity rule when the stimuli were the most speechlike, that is, with the syllables.

In the overall pattern of results, there was some suggestion that the phonetic rules were easier to learn, the more speechlike the stimuli were, and conversely, the acoustic rule was easier, the more nonspeechlike the stimuli were. Closer inspection of the error data provided some additional support for this contention. Most subjects who listened to the full-formant stimuli had difficulty reaching criterion for the acoustic rule because they kept producing groups that followed a vowel-similarity rule. The opposite was true for the two-formant stimuli, namely, these subjects often erred in learning the vowel-similarity rule because they kept splitting the stimuli according to the acoustic similarities of offset. This tendency was verified in an ANOVA on the error data, which showed the expected three-way interaction between stimulus set, rule, and sounds [$F(7,98) = 10.50$, $p < .01$].

GENERAL DISCUSSION

The present study provides support for the notion that brief segments at the offsets of VC syllables contain psychologically relevant invariant cues to stop-consonant identity. The fact that subjects were able to master a classification rule based on consonant identity more readily than an arbitrary grouping rule suggests that there is some

special psychological status to an organization that partitions the stimuli according to consonant identity. However, it need not follow that the perceiver directly extracts the relevant phonetic categories in performing the task. Rather, it is sufficient that the perceiver focus on acoustic properties of the signal which are highly correlated with phonetic categories. It is a description of these properties which might be forthcoming in the new approaches recently employed in analyzing the speech signal (e.g., Blumstein & Stevens, 1980; Kewley-Port, 1983; Kewley-Port & Luce, 1984; Searle et al., 1979; Zwicker et al., 1979).

It is possible that performance on the truncated stimuli in the present study would have improved had release bursts been included. We elected not to include release bursts for several reasons. First, release bursts are often absent in fluent conversational speech and the listener still must detect the cues for consonant identity for such utterances. Second, the short time windows employed in some of the acoustic analyses to date (e.g., 26 msec for Blumstein & Stevens, 1980) would make it virtually impossible to include information about both final formant transitions and bursts, since these are separated by 30-50 msec of silence (corresponding to vocal tract closure) in natural speech. Thus, the only alternative would be to employ truncated segments of considerably longer duration (on the order of 80-90 msec) than some of the proposed templates. Ultimately, even if longer duration templates proved more successful, it would still be necessary to explain how consonant-identity is extracted from unreleased segments.

Although the present study offers some encouragement to those searching for psychologically relevant invariant cues for stop consonants in a syllable-final position, any optimism here must be tempered by the relatively low salience of classification according to consonant identity. What tendency there was for subjects to employ an organization consistent with consonant identity emerged only under conditions in which they were explicitly instructed to do so. Even here, subjects' performance levels were considerably below those observed by Jusczyk et al. (1981) for CV stimuli. In particular, Jusczyk et al. found that for truncated full-formant stimuli performance with the consonant-identity rule was equivalent to that with the vowel-similarity rule, whereas in the present case performance with the consonant-identity rule was significantly worse than that with either the vowel-similarity rule or the acoustic-similarity rule. Whether the lower salience of the consonant-identity rule for VC stimuli is a consequence of weaker invariant cues in syllable-final position overall or only in the formant transition interval that we examined is difficult to say.

Lastly, there is an interesting pattern to the classification that subjects found most salient. With respect to the truncated stimuli, a classification according to acoustic similarity was the most prevalent. Given the confounding of this classification scheme with a vowel-similarity one in the study by Jusczyk et al. (1981), it seems likely that subjects in that study may also have been utilizing an acoustic-similarity organization, at least for the truncated two-formant stimuli. However, as the present study also indicates, there is a tendency for subjects to prefer, and employ more readily, classifications corresponding to phonetic groupings as the stimuli become more speech-like. Thus, the ability of subjects to learn the consonant-identity rule showed a marked improvement as the stimuli progressed from truncated two-formant representations to truncated full-formant representations to VC syllables. In this regard, the addition of information presumably redundant to stimulus pairs, such as identical first-formant transitions, evidently plays an important role in determining the favored perceptual classification. Sawusch and Nochajski (1985; also Sawusch, in press) have observed a similar phenomenon in the perception of glissandos, whereby the addition of a redundant glissando decreased reaction times in a variety of sorting tasks. They have hypothesized that the addition of redundant information yields patterns of "emergent features" which make certain stimulus distinctions more discriminable. The inclusion of the additional formant transitions with the truncated full-formant stimuli may have had a similar effect in the present study. In any event, it is obvious that the context in which the critical stimulus differences (in this case the second- and third-formant transitions) are set dramatically affects the preferred perceptual organization of the stimuli (see Foard & Kemler Nelson, 1984, for a general argument along these lines).

REFERENCES

ATTNEAVE, F. (1959). *Applications of information theory to psychology*. New York: Holt, Rinehart & Winston.

BLUMSTEIN, S. E., & STEVENS, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America*, **66**, 1001-1017.

BLUMSTEIN, S. E., & STEVENS, K. N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America*, **67**, 648-662.

BRADY, P. T., HOUSE, A. S., & STEVENS, K. N. (1961). Perception of sounds characterized by rapidly changing resonant frequency. *Journal of the Acoustical Society of America*, **33**, 1357-1362.

DORMAN, M. F., STUDDERT-KENNEDY, M., & RAPHAEL, L. J. (1977). Stop consonant recognition: Release bursts and formant transitions as functionally equivalent context-dependent cues. *Perception & Psychophysics*, **22**, 109-122.

FOARD, C. F., & KEMLER NELSON, D. G. (1984). Holistic and analytic modes of processing: The multiple determinants of perceptual analysis. *Journal of Experimental Psychology: General*, **113**, 94-111.

GARNER, W. R. (1962). *Uncertainty and structure as psychological concepts*. New York: Wiley.

GRUNKE, M. E., & PISONI, D. B. (1982). Some experiments on perceptual learning of mirror-image acoustic patterns. *Perception & Psychophysics*, **31**, 210-218.

JUSCZYK, P. W., SMITH, L. B., & MURPHY, C. (1981). The perceptual classification of speech. *Perception & Psychophysics*, **30**, 10-23.

KEWLEY-PORT, D. (1978). *KLTEXC: Executive program to implement the Klatt software synthesizer* (Research on Speech Perception, Progress Report 4). Bloomington: Indiana University.

KEWLEY-PORT, D. (1980). *Representations of spectral change as cues to place of articulation in stop consonants* (Research on Speech Perception, Technical Report No. 3). Bloomington: Indiana University.

KEWLEY-PORT, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America*, **73**, 322-335.

KEWLEY-PORT, D., & LUCE, P. A. (1984). Time-varying features of initial stop consonants in auditory running spectra: A first report. *Perception & Psychophysics*, **35**, 353-360.

KEWLEY-PORT, D., PISONI, D. B., & STUDDERT-KENNEDY, M. (1983). Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, **73**, 1779-1793.

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

LIBERMAN, A. M., COOPER, F. S., SHANKWEILER, D. P., & STUDDERT-KENNEDY, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 431-461.

MALECOT, A. (1956). The role of releases in the identification of released final stops. *Language*, **34**, 370-380.

MERTUS, J. (1982). *VOCODE* [Computer program]. Providence, RI: Brown University, Department of Linguistics.

NUSBAUM, H. C., SCHWAB, E. C., & SAWUSCH, J. R. (1983). The role of "chirp" identification in duplex perception. *Perception & Psychophysics*, **33**, 323-332.

SAWUSCH, J. R. (in press). Auditory and phonetic coding of speech. In E. C. Schwab & H. C. Nusbaum (Eds.), *Perception of speech and visual form: Theoretical issues, models, and research.* New York: Academic Press.

SAWUSCH, J. R., & NOCHAJSKI, T. H. (1985). *Auditory pattern processes and emergent features in the perception of speech based stimuli.* Manuscript in preparation.

SEARLE, C. L., JACOBSON, J. Z., & RAYMENT, S. G. (1979). Phoneme recognition based on human audition. *Journal of the Acoustical Society of America*, **65**, 799-809.

SHATTUCK, S. R., & KLATT, D. H. (1976). The perceptual similarity of mirror-image acoustic patterns in speech. *Perception & Psychophysics*, **20**, 470-474.

SINGH, S., & WOODS, D. R. (1971). Perceptual structure of 12 American English vowels. *Journal of the Acoustical Society of America*, **49**, 1861-1865.

STEVENS, K. N., & BLUMSTEIN, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, **64**, 1358-1368.

STEVENS, K. N., & BLUMSTEIN, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech.* Hillsdale, NJ: Erlbaum.

ZWICKER, E., TERHARDT, E., & PAULUS, E. (1979). Automatic speech recognition using psychoacoustic models. *Journal of the Acoustical Society of America*, **65**, 487-498.

NOTES

1. Interestingly, both Blumstein and Stevens (1980) and Kewley-Port (1980) found that subjects were also usually able to identify the vowel of the syllable from which the stimulus had been excerpted. However, success in identifying vowels did show an improvement with increased stimulus duration.

2. In particular, Blumstein and Stevens (1980) constrained subjects' responses by limiting their choice of perceptual categories to B, D, and G. In addition, they used a block design in which each vowel context was tested separately. Similarly, although Kewley-Port (1980) provided her subjects with six different categories (B, D, and G and thier voiceless counterparts, P, T, and K), her results are collapsed across the voicing dimension. Given such constraints, the results reported by Blumstein and Stevens and Kewley-Port cannot be unambiguously interpreted as evidence that phonetic categories are being abstracted from their stimuli. At best, Kewley-Port's results suggest that place categories are abstracted. An alternative hypothesis is that subjects judged each stimulus in terms of its resemblance to each of the available response choices. Appropriately designed single-tone glissandos might yield similar accuracy rates but certainly would not be perceived as speech.

3. Jusczyk et al. (1981) chose the designation "vowel similarity" to describe this partitioning of the stimuli because it corresponds to the phonetic front-back distinction. However, as one of the present reviewers observed, according to the ratings collected by Singh and Woods (1971), while [i] and [ɛ] are highly similar, [o] and [ɔ] are actually quite dissimilar to each other. Hence, it is possible that subjects may respond by putting the similar vowels together into one group and the remaining vowels into an "other" category. Although it is possible that the selection of other vowels might have resulted in a stronger tendency to employ a vowel-similarity strategy, the present vowel set was chosen to provide a diverse set of formant transition cues for [b] and [d] in different vowel contexts.

4. Our estimate of overall pitch at offset is based upon the mean of the second and third formants at offset. We considered other alternatives, such as differentially weighting the two formants—for example, decreasing the weighting of the third formant due to the decreasing amplification of higher formants—as suggested by Shattuck and Klatt (1976). However, a partitioning according to the value of the second formant alone produces exactly the same high/low-pitch classification of the stimuli as the one we employed. Thus, in the present instance, a weighted function does not seem necessary.

5. The computational formula for calculating Rel H = $\Sigma p \log_2 p/\#$ bits, where p refers to the proportion of trials that a given stimulus was assigned to a particular category by each subject and the number of bits is determined from the number of responses categories employed. The number of bits equals the power to which 2 must be raised to equal the number of response alternatives. In the present case, with only two response alternatives, the number of bits equals 1.

6. Note that these two groupings are not orthogonal. Thus, a subject employing a consonant-identity grouping 100% of the time would still receive a score of 50% on the vowel-similarity grouping.

7. In comparison, Jusczyk et al. (1981) found evidence for consonant-identity groupings in 2 out of 15 subjects under comparable experimental conditions, but with truncated CV stimuli.

(Manuscript received August 14, 1984;
revision accepted for publication January 27, 1985.)