Reprinted from THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA VOL. 60, NO. 1 JULY 1976

Consonant environment specifies vowel identity*

Winifred Strange and Robert R. Verbrugge†

Psychology Department, University of Minnesota, Minneapolis, Minnesota 55455

Donald P. Shankweiler

Haskins Laboratories, New Haven, Connecticut 06510
and Psychology Department, University of Connecticut, Storrs, Connecticut 06268

Thomas R. Edman

Psychology Department, University of Minnesota, Minneapolis, Minnesota 55455
(Received 26 December 1975; revised 12 March 1976)

Past studies have shown that, while vowels can be produced with static vocal tract configurations, the resulting steady-state tokens are misidentified frequently by naive listeners. The first experiment compared the perception of isolated vowels with vowels spoken in a fixed consonantal frame by the same set of 15 talkers. Vowels in /p–p/ syllables were identified with far greater accuracy than were comparable isolated vowels in both single and multiple-talker conditions. Acoustical analyses of the test tokens showed that the poor intelligibility of isolated vowels could *not* be attributed to talkers' failure to produce these vowels correctly. In a second experiment, vowels in syllables in which the initial and final stop consonant varied unpredictably from item to item were still identified with greater accuracy than were isolated vowels. These results offer strong evidence that dynamic acoustic information distributed over the temporal course of the syllable is utilized regularly by the listener to identify vowels.

Subject Classifications: [43]70.30, [43]70.40, [43]70.70.

INTRODUCTION

Vowels, unlike consonants, can be produced and identified in isolation. This possibility was exploited early in the investigation of vowel quality, as witnessed by studies of the cardinal vowels (Jones, 1956). Sustained, "steady-state" vowels can be classified by frequencies of the first two or three formants (Potter and Steinberg, 1950). So successful were the efforts to locate the acoustic information sufficient for the perception of sustained vowels that the main focus of research on speech perception shifted to the search for the consonantal cues. But the supposition that the sound pattern is simpler in the case of the vowels than the consonants is unsupportable if a distinction is made between the sustained, isolated vowel and the vowel as it occurs in natural speech.

Although they *can* be produced in a quasi-steady-state manner and in isolation, vowels so produced must be regarded as laboratory artifacts. Ordinarily, vowels occur in coarticulation with consonants in the context of the syllable. The acoustic information in coarticulated vowels is fused and carried in parallel with the consonantal information (see Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Liberman, 1970). It was discovered long ago in tape-cutting experiments of Schatz (1954) and Harris (1953) that vowel quality cannot be discretely localized in any single portion of the syllable, but is distributed throughout the period during which voicing is present.

Studies of perturbations of formant frequencies brought about by uttering vowels in the context of syllables were carried out by Stevens and House (1963), Lindblom (1963), Öhman (1966), and Shearme and Holmes (1962). These investigations demonstrated that steady-state values of the formants are rarely attained because articulatory movement is more or less continuous. Thus, the acoustic description of vowels in ordinary speech is a good deal more complex and problematic than is revealed by the classic studies of the acoustic basis of vowel quality.

If the acoustic structure of the isolated vowel often differs greatly from the "same vowel" in context, it might be inferred that different cues are employed in vowel perception when the vowel is in consonantal context and when it occurs in isolation. It is all the more interesting, therefore, to find indications in the phonetic literature that isolated vowels are difficult to perceive. For example, Fairbanks and Grubb (1961) presented nine isolated vowels produced by phonetically trained talkers to experienced listeners. The overall identification rate was only 74%, which contrasts strikingly with a rate of 94% obtained by Peterson and Barney (1952) for perception of vowels in /h–d/ context. Somewhat better identification of isolated vowels was obtained by Lehiste and Meltzer (1973), with only three talkers producing the tokens. Fujimura and Ochiai (1963) directly compared the identifiability of vowels in consonantal context and in isolation. They found that the center portions of vowels, which had been gated out of CVC syllables, were less intelligible in isolation than in syllabic context. These findings suggest that isolated vowels are misidentified with significantly higher frequency than vowels spoken in at least some consonantal environments. Could it be that the acoustic complexities introduced by syllabic structure better serve the requirements of the perceptual apparatus than do quasi-steady-state formants? If so, then it is surely inappropriate to characterize the cues for vowel identity in terms of static points in a space defined by the first two formants.

It seemed important, therefore, to attempt to demonstrate under carefully controlled experimental condi-

tions that vowels in consonantal contexts are perceived with fewer errors than "the same vowels" presented in isolation. A further purpose of the research reported here was to investigate the sources of information within the CVC syllable that specify the vowel and to explore how that information is utilized by the perceiver in the process of perception. If it is true that consonantal environment generally aids in identification of a vowel, we recognize that there is more than one way the environment might play a facilitating role. One possibility is that portions of the signal that are commonly regarded as consonantal, such as transitions, might aid in normalization for vocal tract differences. Experiments by Fourcin (1968) and Rand (1971) have found that perceptual boundaries between stop consonants vary depending on the vocal tract presumed to have produced a syllable. The phonemic identity of the consonants was fixed and known in advance in the Peterson and Barney (1952) study and in our own investigations (Verbrugge, Strange, Shankweiler, and Edman, 1976). In these cases, the transitions may have allowed listeners to scale the formant frequencies of the medial vowel according to the vocal tract characteristics of the talker and thus reduce vowel ambiguity.

On the other hand, isolated vowels may be difficult to perceive for a more fundamental reason. It is possible that listeners ordinarily rely upon information distributed throughout the whole syllable for identification of the vowel. This seems likely in view of parallel transmission of the consonants and the vowel. If it is the case that syllable-initial and syllable-final transitions specify the vowel as well as the consonants, we could assert that the vowel is inseparable from the syllable, that it is not specified by formant frequencies at any particular cross section in time, but rather is carried in the dynamic configuration of the whole syllable. In this case the presence of transitions should aid identification of the intended vowel whatever additional difficulties may be posed by confronting the listener with multiple vocal tracts.

I. EXPERIMENT I: PERCEPTION OF ISOLATED AND MEDIAL VOWELS

If consonantal environment aids in specifying vowel identity in either of the two ways postulated above, we would expect that the perception of isolated vowels would be less accurate than the perception of medial vowels in listening tests where the tokens on a test were produced by different talkers. Previous studies on the identification of steady-state vowel stimuli support this hypothesis (Fairbanks and Grubb, 1961; Lehiste and Meltzer, 1973). However, these investigations do not directly compare isolated vowels with vowels in syllable frames with the number and type of talkers, number of response alternatives, and other factors held constant. Millar and Ainsworth (1972) report that listeners were able to identify synthetically generated vowels more reliably and uniformly when the vowels were embedded in /h–d/ words than when the acoustically identical segments were presented in isolation. We are not aware of any studies which directly compare the perception of naturally produced isolated vowels with vowels in context.

The present study compares the identifiability of vowels produced in a fixed consonantal frame with isolated vowels when (a) a single talker produced all tokens on a particular listening test ("segregated talker" condition) and (b) when tokens produced by several different talkers are presented in random order ("mixed talker" condition). By independently varying these two factors (consonantal context and talker variation) we can assess the relative contribution of each to the accuracy of vowel identification. Further, the design allows us to test the two hypotheses regarding the way in which consonantal information may be utilized. If consonantal environment aids in vowel identification by serving as a calibration signal for vocal tract normalization, we expect an interaction between the two major variables. That is, we expect that the loss in identifiability of vowels due to the absence of consonantal transitions will be more severe on those tests where talker identity changes, since recalibration is necessary on each trial. We expect no significant disadvantage of the absence of consonantal transitions for those tests in which talker identity is unchanged. Alternatively, if consonantal transitions provide information that specifies vowel identity independent of talker normalization, we expect no such interaction. The identification of isolated vowels should be less accurate than of vowels in consonantal context both for tests on which the talker remains constant and for tests on which talkers are mixed.

This study compares listeners' performance on isolated vowel tests with the results reported previously for medial vowels spoken in /p–p/ environment (Verbrugge *et al.*, 1976). The tests were directly comparable on all factors, such as identity of talkers, order of presentation of alternatives, response alternatives, and recording and reproduction conditions.

A. Method

1. Stimulus materials

The panel of talkers described in our previous research was also utilized for this study. Five men, five women, and five children, none of whom were trained speakers, were selected to represent a wide variety of vocal tract sizes and characteristic fundamental frequencies. According to the judgment of the experimenters, the talkers represented a fairly homogeneous dialect group, that of the upper midwest region from which the listeners were also drawn.

The materials for the /p–p/ tests (mixed and segregated talker) were those described in Verbrugge *et al.* (1976, Expt. II). Talkers read the test syllables, which were printed individually on cards. The /p–p/ words were also used to represent the isolated vowels; talkers were instructed to pronounce the vowels as they would be pronounced in these key words. They were given one practice trial and were instructed to produce the tokens quite rapidly. Each talker produced one token of each of nine isolated vowels: /i/, /ɪ/, /ɛ/, /æ/, /ɑ/, /ɔ/, /ʌ/, /ʊ/, /u/.

For the mixed-talker isolated-vowel test (mixed #–#) three of the nine vowels were selected for each talker, corresponding to the three vowels he or she produced for the /p–p/ test. As in the earlier test, vowels were assigned to talkers randomly with the constraint that each talker contributed only one of the point vowels. Thus, the mixed #–# test consisted of five tokens of each of nine vowels; each of the five tokens was spoken by a different talker.

The segregated-talker isolated-vowel tests (segregated #–#) were comparable to the segregated-talker /p–p/ tests described in Verbrugge *et al.* (1976, Expt. II). One man, one woman, and one child each produced a 45-item test that contained five different tokens of each of the nine vowels.

All test stimuli were recorded in a sound-attenuated experimental room with a ReVox A77 stereo tape recorder and Spher-o-dyne microphone. The 45 tokens on a test were arranged in a random presentation order with the restrictions that the same intended vowel did not appear more than twice consecutively, and tokens produced by the same talker were separated by not less than eight tokens (in the mixed tests). Identical procedures were used to construct each of the four tests so that presentation order, timing, and peak intensity of test tokens were identical for all tests.

2. *Procedure*

Listening tests were presented to small groups of subjects in a quiet experimental room via a Crown CX 822 tape recorder, MacIntosh MC40 amplifier, and AR acoustic suspension loudspeaker. Listeners responded on score sheets which contained nine response alternatives written out in full in each row: "pip, pup, pap, peep, pop, pep, poop, pawp, puup." Before the tests, the experimenter pronounced each of the nine key words, drawing special attention to the last word, "puup," which stood for the syllable /pʊp/. For the #–# tests, the experimenter pronounced each key word followed by the vowel in isolation, again with special attention to the /ʊ/ alternative. Subjects in the mixed-talker conditions were told they would hear "several different talkers"; subjects in the segregated-talker conditions knew they would hear only one voice on each 45-token test.

Independent groups of subjects responded to the /p–p/ and the #–# mixed-talker tests. Each group of subjects completed two repetitions of the 45-token test for a total of 90 judgments per subject, 10 on each intended vowel. In the segregated-talker conditions, three groups of subjects heard the /p–p/ tests and another three groups heard the #–# tests. The order of presentation of the man (M), woman (W), and child (C) tests was counterbalanced across the groups in the orders: MWC, WCM, CMW. Data for only the first two tests were analyzed (i.e., MW, WC, and CM, respectively). Thus, the total number of judgments by the segregated test subjects was equivalent to that for the mixed test subjects (90 judgments) and any effects of fatigue or familiarity were equally distributed across the three talkers for the segregated tests.

3. *Subjects*

The data presented here for the /p–p/ conditions are those obtained in the previous study (Verbrugge *et al.*, 1976, Expt. II). Thirty-three subjects served in the segregated /p–p/ tests (11 in each condition) and 19 subjects were tested on the mixed /p–p/ test. For the tests on isolated vowels, 30 subjects were tested in the segregated #–# test (10 per condition) and 16 subjects heard the mixed #–# test. All subjects were paid volunteers from undergraduate psychology classes at the University of Minnesota. All were native speakers of English and most were natives of the upper midwest region.

B. Results

Errors in vowel identification were tabulated for each condition; an error was defined as the selection of a response other than that intended by the talker. The overall error rate for the four experimental conditions is shown in Fig. 1. The main comparison of interest is between performance reported earlier for vowels in /p–p/ environment and performance on the isolated vowels. On the average, there were 17.0% errors on the mixed /p–p/ test and 9.5% errors on the segregated /p–p/ test. For the isolated vowels, on the other hand, there were 42.6% errors on the mixed test and 31.2% errors on the segregated test. Errors summed over all nine vowels for each subject were submitted to a 2×2 analysis of variance for unequal cell frequencies. The main effects for talker variation (mixed vs segregated) and consonantal context (/p–p/ vs #–#) were both significant, $F(1, 94) = 21.18$ and 125.17, respectively, $p < 0.01$. However, no significant interaction between the two variables was found, $F(1, 94) = 0.93$.

These results indicate that while talker variation does contribute significantly to vowel identification errors for both medial vowels and isolated vowels, the presence or absence of consonantal context is by far the more

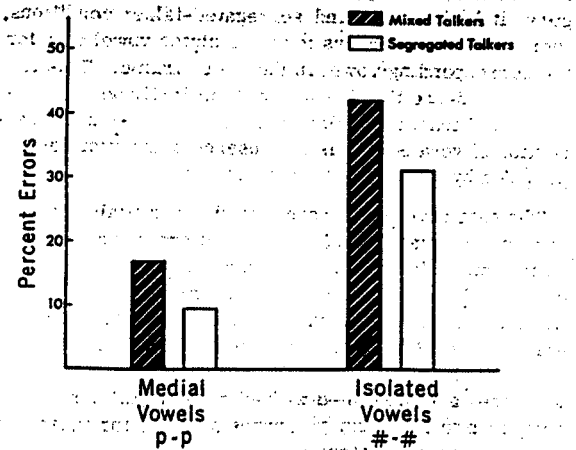

FIG. 1. Overall percent errors for vowels in /p-p/ syllables and isolated vowels. Open bars show errors for segregated talker conditions; shaded bars show errors for mixed talker conditions.

TABLE I. Experiment I: Identification errors (percent) for each intended vowel in four experimental conditions. Error rates excluding /ɑ/–/ɔ/ confusions are given in parentheses. (See Ref. 1.)

Intended vowel	Segregated talkers #–#	Segregated talkers /p–p/	Mixed talkers #–#	Mixed talkers /p–p/
i	16	<1	26	1
ɪ	14	4	23	2
ɛ	46	12	62	27
æ	26	2	48	19
ɑ	64 (19)	23 (4)	61 (32)	20 (10)
ɔ	29 (14)	18 (2)	30 (10)	27 (3)
ʌ	42	8	63	15
ʊ	29	18	49	39
u	14	<1	23	3
Overall errors	31% (25)	9% (6)	43% (38)	17% (13)

important variable. Listeners misidentified approximately *three* times as many isolated vowel tokens as they did the corresponding medial vowels. Thus, it appears that the presence of a consonantal environment is much more critical for accurate vowel indentification than is familiarity with the characteristics of the talkers' vocal tracts.

The hypothesis that consonantal environment contributes to perception of the vowel by providing cues for talker normalization was *not* supported. There was no interaction between the two major variables; the increased error rate due to the absence of consonantal context was almost as great when the talker was constant (an increase of 22%) as it was when talkers varied from token to token (an increase of 26%). We can conclude that the efficacy of the /p–p/ context in aiding vowel identification is directly involved with specification of vowel identity.

A vowel-by-vowel analysis of the identification errors for the four experimental conditions is presented in Table I. (Confusion matrices for the /p–p/ and #–# tests are presented in Tables A-I, A-II, A-III and A-IV.) It is readily apparent that for *every* vowel category, in both mixed- and segregated-talker conditions, there were more errors for the isolated vowel than for the corresponding vowel in the /p–p/ frame. This is strong evidence that the lack of familiarity with a talker's vocal tract is far less detrimental to accurate perception of vowels than is the absence of information provided by a consonantal environment.

The data reveal differences in the identifiability of particular isolated vowels. The pattern of errors is quite similar to that found for medial vowels; the vowels /i/, /ɪ/, and /u/ are most accurately identified while the more central vowels yield relatively more errors in identification. It should be noted, however, that even the former show error rates from 14% to 26% when they are presented without consonantal context, compared to less than 4% errors obtained for these vowels in the /p–p/ context.[1]

A more detailed analysis was undertaken to evaluate the consistency of these results. The percent errors obtained for each of the 45 tokens on the mixed #–# test was compared to the percent errors obtained for the comparable token on the mixed /p–p/ test. Isolated vowel tokens were misidentified more often than medial vowels in 39 out of 45 cases while two pairs produced an equal proportion of errors. In only 4 cases did the /p–p/ token produce more errors than the comparable isolated vowel. Thus, we can conclude that the difference in error rates found between performance on medial and isolated vowels is consistent across individual tokens of the vowels as well as across vowel categories.

The overall results of the segregated tests show that isolated vowels were identified far less accurately than were medial vowels, even when talker variation was absent. Error rates for the man, woman, and child on the segregated #–# tests were 33%, 26%, and 32%, respectively. Comparable error rates for the segregated /p–p/ tests, reported in Verbrugge *et al.* (1976), were 9%, 6%, and 11%, respectively. The differences show a relatively constant advantage of consonantal environment for all three talkers, despite some variability in overall intelligibility of the talkers.

In summary, it is clear that consonantal environment contributes in a major way to the identification of vowels. We reach this conclusion whether we regard the data in terms of overall results, the results for particular vowel categories, for individual tokens or for individual talkers. Isolated vowels are much more poorly identified than vowels embedded in the /p–p/ context.

C. Acoustical analysis

The results of this experiment indicate that isolated, steady-state vowels are poor stimuli from the standpoint of the perceiver. The possibility remains, however, that the perceptual problem in identifying isolated vowels is a result of the way the talkers produced them. Phonetically untrained talkers may be unable to produce specified tokens of vowels reliably in isolation. Acoustical analysis of the vowel utterances by our panel of talkers was undertaken to investigate this possibility.

Center frequencies of the first three speech formants and the duration of the vocalic portion of each syllable were determined from spectrograms and spectral sections produced on a Voiceprint Sound Spectrograph. Recordings of tokens produced by women and children were reproduced at half speed for spectrographic analysis; obtained frequency values were doubled to determine the actual formant frequencies of these tokens. Spectral sections were made at the point of nearest approach to the steady state. (If the vowel was diphthongized by the talker, measurements were obtained from the initial part of the vocalic portion of the syllables.) Two judges, working independently, determined the center frequency values for the speech formants to the nearest 25 Hz. Frequencies reported represent an average of the values obtained by the two judges. In addition, measurements of the duration of the first-formant periodic energy were made.[2]

Measurements were obtained for the 45 tokens of

TABLE II. Average frequency values (Hz) for the first three speech formants of the nine isolated vowels, averaged over five talkers in each group.

		i	I	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u
F1	M	355	447	635	737	757	672	685	497	337
	W	385	482	747	820	843	692	815	577	435
	C	357	580	755	885	1030	770	895	557	500
F2	M	2245	1960	1790	1697	1220	942	1167	1092	1042
	W	2792	2325	2157	2110	1372	1312	1525	1399	1175
	C	3335	2710	2485	2685	1565	1350	1630	1340	1150
F3	M	2937	2575	2510	2445	2347	2453	2307	2352	2165
	W	3482	3060	2960	2900	2915	2875	2847	2815	2735
	C	3680	3630	3765	3680	3700	3540	3725	3613	3150

the mixed-talker /p–p/ test and the 45 isolated vowel tokens in the mixed #–# test. In addition, measurements were obtained for the remaining six isolated vowels spoken by each talker that were not incorporated in the mixed #–# test. Thus, one token of each of nine isolated vowels was measured for each of fifteen talkers. For the segregated tests, one token of each of the nine isolated vowels was selected randomly from each of the three talkers' tests. For comparison, the /p–p/ token which corresponded to each selected isolated vowel was also analyzed.

Looking first at the analysis of the isolated vowels spoken by the full panel of talkers, we can ask whether the poor identification (43% errors) was due to the talkers' inability to produce isolated vowels reliably. Table II presents the average values of the first three speech formants for the men, women, and children. In Fig. 2 the average values for the first and second speech formants are plotted in a two-dimensional "vowel space." On the average, our talkers' productions of the vowels in isolation were systematic in distribution and corresponded closely in formant values to vowels sampled by other investigators (Peterson and Barney, 1952; Tiffany, 1959; Stevens and House, 1963). The formant frequencies showed systematic elevations

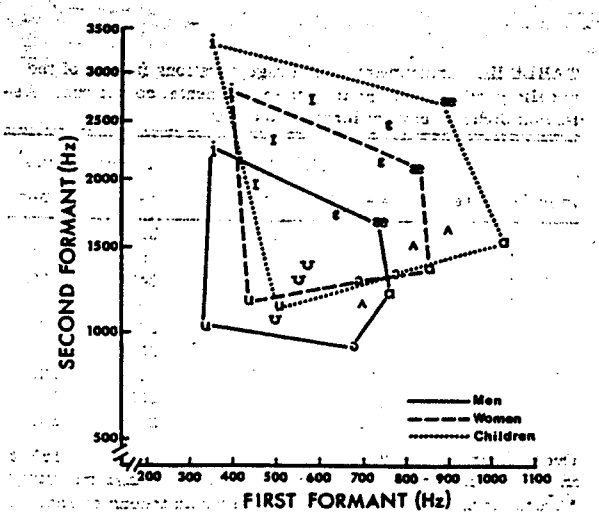

FIG. 2. Average formant 1/formant 2 values for isolated vowels spoken by men, women, and children. (Five talkers in each group.)

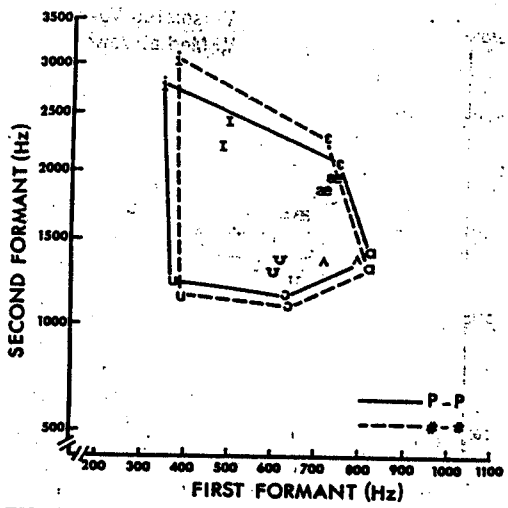

FIG. 3. Average formant 1/formant 2 values for vowels in /p-p/ syllables (solid lines) and vowels in isolation (dashed lines). Values were computed over the five tokens of each vowel in each mixed talker test.

from men to women to children, reflecting a general decrease in the size of these talkers' vocal tracts.

Individual tokens of isolated vowels corresponded closely to values reported in previous studies except for tokens of the vowel /ɔ/ by all talkers, tokens of /ɛ/ spoken by the men and women, three tokens of /æ/ spoken by women, and one token of /ʊ/ spoken by a woman. The deviation in /ɔ/ tokens represents a dialectal difference between our talkers and those recorded by Peterson and Barney (1952). Stevens and House (1963) did not report data for this vowel.

The next question of interest is whether the panel's productions of isolated vowels differed greatly from their corresponding productions of vowels in the /p–p/ consonantal frame. To answer this question we compared the tokens actually used in the two mixed-talker tests. Figure 3 presents the average values of F1 and F2 for the medial vowels and isolated vowels, pooled across men, women, and children. The vowels on the two tests occupied almost the same area in F1/F2 space. The second formant of the medial vowels showed a slight migration toward the center of the space. This is an expected result of coarticulation (where formants fail to reach a steady-state target) and is in accord with results reported by Stevens and House (1963) for vowels produced between consonants with labial and labiodental place of articulation. As Tiffany (1959) noted, this *reduces* the acoustic contrast among vowels spoken in a consonantal frame in comparison to isolated vowels. However, the perceptual data demonstrate that identifiability cannot be predicted from the spread of steady-state formant measurements; medial vowels were perceptually much *more* distinct than vowels in isolation (83% vs 57% correct identifications).

The two sets of vowels were very similar in formant frequencies, in both the central tendency and the variability of values for each vowel. Even so, there were

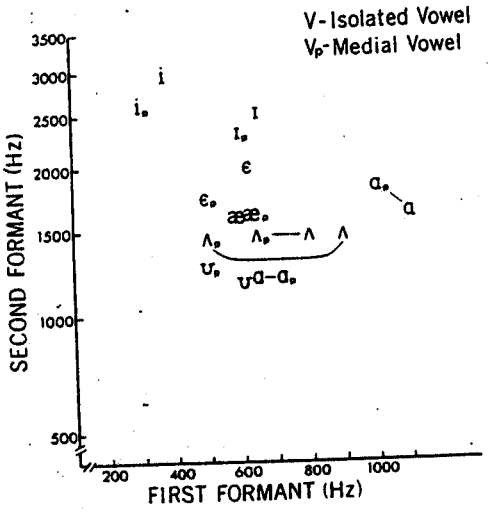

FIG. 4. Formant 1/formant 2 values for the nine pairs of vowels on the mixed talker tests that contributed most to the difference in identification errors. Vowels in /p–p/ syllables are indicated by the subscript *p*.

a few individual tokens that deviated markedly from the central tendencies. It is of interest whether the considerably greater error rate for isolated vowels over that obtained for medial vowels can be attributed primarily to misidentification of tokens which were produced in a deviant manner.

One way to answer this question is to look at those pairs of tokens which contributed most to the difference obtained in the perceptual tests. For nine comparison pairs, errors for the isolated vowel exceeded those for the medial vowel by more than 50% of the opportunities for error. It might be supposed that the formant frequencies of these isolated vowel tokens would show the greatest deviation from the average values and from values for the comparable medial vowel. This is not the case, however, as may be seen from Fig. 4, which shows the nine vowel pairs. For some of these pairs, the first- and second-formant values for both the isolated and medial vowels fell within the range of variation for the appropriate vowel category. For the vowels /æ/, /ɑ/ and /ɪ/ both isolated and medial vowels were displaced from their typical positions. Finally, for the vowels /ʊ/, /ɛ/, and one pair of /ʌ/, the isolated vowel might be considered *less* confusable acoustically than its counterpart in medial position. Thus, there seems to be no close correspondence between perceptual confusability and acoustic deviation from some expected (target) value.

This does not mean, of course, that variations in formant frequency positions have no effect on perception. There were a few pairs of tokens which were "misarticulated" on both the /p–p/ and #–# tests, and which contributed relatively greater numbers of errors in identification. (For instance, one woman's production of /ʊ/ was quite deviant on the medial vowel test, as well as on the isolated vowel test. Listeners made 38% and 100% errors on the isolated and medial tokens, respectively.) However, with respect to the present comparison, the salient point is that deviation in formant structure cannot account for the large and consistent differences between perceptual tests of isolated vowels and vowels in a fixed consonantal frame.

Measurements of formant frequencies of tokens from the segregated talker tests corroborate the results for the mixed-talker tests. Since measurements were made for only a sample of the total set of items, we cannot be sure that deviations in the production of isolated vowels were not responsible for their inferiority as perceptual targets. However, the tokens that were measured gave no indication that the three talkers produced the isolated vowels less consistently than they did the medial vowels. A comparison of pairs of tokens showed that isolated and medial vowels were similar in all but a few cases. Deviations from the normal range of formant values were as likely to be obtained for a randomly selected medial vowel as they were for a randomly selected isolated vowel. Thus, the consistent advantage found in perceptual tests for medial vowels over isolated vowels, for all three talkers and all nine vowel categories, cannot be attributed to deviant formant frequencies of isolated vowels.

While there was no indication of large differences in the formant structure of the vowels in isolation and those in syllables spoken in citation form, these two sets of tokens did differ considerably in terms of overall duration. Table III gives the average duration of the voiced first formants of isolated and medial vowels in segregated- and mixed-talker tests. The isolated vowels were much longer on the average than were the medial vowels. However, a more important consideration is the *relative* durations of the vowels in the two sets. More specifically, are the relative durations of isolated vowels different from those typically found for vowels in consonantal context?

The relative durations of vowels in /p–p/ frames were similar to the values reported by Peterson and

TABLE III. Experiment I: Average durations (msec) of the vocalic portion of tokens in four experimental conditions. Asterisks indicate deviant lengths (see text).

	Segregated Talkers[a]		Mixed talkers[b]	
Intended vowel	#–#	/p–p/	#–#	/p–p/
i	315	128	326*	148
ɪ	228	108	198	138
ɛ	226	111	245*	136
æ	328	194	256	204
ɑ	313*	179	237	177
ɔ	303*	186	251	186
ʌ	246	116	184	138
ʊ	242	124	259*	131
u	311	109*	237	159
Overall errors	279.1	139.4	243.7	157.4

[a]Averages based on three randomly selected tokens of each vowel, one from each of the three talkers.
[b]Averages based on five tokens, each spoken by a different talker.

Lehiste (1960) and House and Fairbanks (1953). The vowels /I/, /ɛ/, /ʌ/, and /ʊ/ were the shortest in duration, /i/ and /u/ were intermediate, and /ɑ/, /ɔ/, and /æ/ were the longest vowels. The only exception to this in our data was the vowel /u/ in the segregated /p–p/ test, for which the average duration was considerably shorter than that reported by other researchers.

As Table III indicates, relative durations for the isolated vowels were similar to those for medial vowels with the following exceptions: for the mixed #–# test, the vowels /i/, /ɛ/, and /ʊ/ showed longer relative durations than they did in consonantal context. For the segregated test, the vowels /ɑ/ and /ɔ/ showed shorter relative durations than their counterparts in consonantal frames.

The atypical durations of these isolated vowels cannot account for the consistent advantage of medial vowels over isolated vowels for every vowel category in the perceptual tests. Even for the deviant vowels, the confusion patterns showed no consistent trend toward responses that would be predicted on the basis of the deviant durations. (See Tables A-III and A-IV for confusion matrices.)

D. Discussion

In this study we found that vowels produced in a fixed consonantal environment were identified with much greater accuracy than were comparable steady-state vowels produced in isolation. This was true both when variation due to talker differences was present and when it was not. Thus, the experiment provides no evidence that coarticulated consonants facilitate identification by enabling the listener to recalibrate for each new talker. Coarticulated consonants are integral to the specification of vowels whether a talker is familiar or not.[3]

Acoustical analyses were undertaken to investigate the possibility that untrained talkers fail to adopt consistent targets for vowels in isolation resulting in a highly unreliable signal for perception. Although there were systematic acoustic differences between vowels produced in consonantal environment and those produced in isolation, the large and consistent increases in confusability among isolated vowels over those obtained for medial vowels could not be explained by increases in the acoustic similarity of vowel categories when defined by formant frequencies. Nor could these differences be attributed to differences in the relative durations of the vowels in isolation and in context. It is interesting to note that medial vowels tend to be *more* similar to each other than comparable isolated vowels in terms of the cross-sectional acoustic parameters that have traditionally been used to differentiate vowel classes. This is additional support for the view that static descriptions of vowels are inadequate for capturing perceptually relevant aspects of the acoustic signals. Our results lead us to conclude that the acoustic information for vowel identity, like that for consonants, is specified in the dynamic configuration of the syllabic pattern as a whole.

In this study, the consonantal environment in which the vowels were produced was constant across all tokens. Thus, the listeners knew beforehand the identity of two of the three phonemes in each test token. It is possible that this knowledge (rather than the presence of formant contours) was the source of superior identification for medial vowels. It would be of limited interest if consonantal environment aided in vowel identification only in this circumstance, since it is not generally the case that listeners have advance knowledge of consonantal identity in natural listening conditions. We therefore undertook an additional experiment to test the effects of a varying consonantal environment on the identification of medial vowels.

II. EXPERIMENT II: PERCEPTION OF VOWELS IN CVC SYLLABLES

We wanted to determine whether a consonantal context which varies from trial to trial (and is therefore unpredictable by the listener) provides important information for vowel identification. We again included conditions where the talkers varied from trial to trial (mixed) and where the same talker produced all tokens on a particular test (segregated), in order to investigate the possible interaction between talker variation and knowledge of consonantal context.

A. Method

1. Stimulus materials

The C–C test syllables were composed from six stop consonants, /p, t, k, b, d, g/, and the nine vowels used in Experiment I. A panel of four adult males, four adult females, and four children (a subset of the 15 talkers used in Experiment I) each produced six tokens for the mixed-talker condition, resulting in a test series of 72 syllables. Within this series, each vowel occurred eight times and each initial and final consonant occurred twelve times. Consonants and vowels were paired such that each vowel was preceded and followed by each consonant at least once. (Both symmetrical and nonsymmetrical pairings were used; for example, syllables such as /t–t/ and /d–t/ both appeared in the test series). The assignment of syllables to talkers was random with the constraint that a talker did not produce the same vowel more than once, nor the same initial consonant more than twice.

The talkers read the test syllables from cards on which they were printed in standard English orthography, except in cases where no unambiguous English spelling existed. For these items, key words were provided beneath the test syllables to indicate the pronunciation of the vowel. All test stimuli were recorded using the equipment and procedures described in Experiment I.

The 72 test syllables were arranged in an order of presentation with the following restrictions: (a) the same intended vowel did not occur more than twice consecutively, (b) there was an equal number of tokens of each intended vowel in the first and second half of the test, (c) the same initial consonant did not occur more

than twice consecutively, (d) tokens produced by the same talker were separated by not less than six tokens, and (e) each talker occurred equally often in the first and second half of the series. For the segregated talker tests, the same three talkers were recorded as in the segregated tests in Experiment I. Each talker recorded the entire list of 72 syllables in the same order as for the mixed-talker test.

2. Procedure

Listening tests were administered to small groups of subjects using the equipment and procedures described in Experiment I. Listeners responded on score sheets printed with columns of key letters representing each of the nine vowels. Above each column, key words containing these letters were printed as follows: "*s*in *s*um *s*and *s*een sh*o*p s*e*nt s*oo*n s*aw* sh*ou*ld." The key letters in the columns were preceded and followed by blank lines. Before the listening test, the experimenter pronounced each key word followed by its vowel in isolation. Special attention was drawn to the key letters that represented the vowel /ʊ/.

Subjects in the mixed-talker condition were required to identify only the vowel in each syllable. They did this by circling, for each syllable, the key letter(s) that symbolized the perceived vowel. Listeners heard the entire test series twice for a total of 144 judgments per subject.

Three groups of subjects were tested in the segregated talker condition. All three groups were required to identify only the vowel in the syllables, and they did so in the same way as the subjects in the mixed-talker condition. As in Experiment I, each group of subjects heard the three talkers in one of three orders: MWC, WCM, or CMW. Again, data for only the first two tests were analyzed, making the number of judgments per subject equal to that for the mixed-talker tests (i.e., 144 judgments per subject). Subjects in all conditions were told that some of the test syllables were real words and that some were nonsense syllables, but that they were to ignore meaning and respond only on the basis of the sound of the syllables.

3. Subjects

All subjects were paid volunteers obtained from undergraduate psychology courses at the University of Minnesota. All were native speakers of English and most were natives of the upper midwest region. Twenty-two subjects served in the mixed-talker condition. Twenty-four subjects were tested in the segregated-talker condition, eight with each of the three counterbalanced orders.

TABLE IV. Overall identification errors (percent) for Experiments I and II.

	Test	Segregated talkers	Mixed talkers
Experiment I	/p–p/	9.5	17.0
	#–#	31.2	42.6
Experiment II	C–C	22.9	21.7

TABLE V. Experiment II: Identification errors (percent) for each intended vowel in two experimental conditions.

Intended vowel	Segregated talkers	Mixed talkers
i	8	6
ɪ	12	17
ɛ	14	24
æ	13	15
ɑ	41 (15)	31 (7)
ɔ	44 (10)	37 (11)
ʌ	11	18
ʊ	46	39
u	17	8
Overall errors	23% (17)	22% (16)

B. Results and discussion

Table IV presents the overall error rates for the two conditions of this experiment along with the results of Experiment I for comparison. There was no significant difference between the error rates for the segregated-talker condition (22.9%) and the mixed-talker condition (21.7%), $t(44\ df) = 0.43$.

The major question of interest was whether consonantal context aids vowel identification even when the context is unpredictable. The results for the C–C test syllables may be compared with those found in Experiment I for /p–p/ syllables and isolated vowels (cf. Table IV). For the mixed-talker condition, vowels in C–C syllables were identified with significantly greater accuracy than were comparable isolated vowels, as tested by a median test: $\chi^2(1\ df) = 18.24$, $p < 0.01$. The overall error rate of 21.7% for C–C syllables was not significantly greater than the 17% errors found for vowels in /p–p/ syllables, $\chi^2(1\ df) = 0.23$. Thus, the results for the mixed-talker condition are clear; both fixed and variable consonantal frames produced a dramatic improvement in vowel identifiability in contrast to isolated vowels. The advantage of a consonantal environment obtains even when the identity of the consonants is not known in advance by the listeners.[4]

The overall results for the segregated-talker condition were less conclusive. Vowels in C–C syllables were, on the average, better identified than isolated vowels: $\chi^2(1\ df) = 6.08$, $p < 0.02$. However, unlike the mixed-talker results, listeners did not identify vowels in C–C syllables as accurately as vowels in /p–p/ syllables, $\chi^2(1\ df) = 25.6$, $p < 0.01$. The error rate for the segregated C–C test appears to be idiosyncratic in that there was *no* advantage over the comparable mixed-talker condition. (For the /p–p/ and #–# tests, the advantage of segregated test over mixed test was 8% and 12%, respectively.)

Table V presents the errors for each vowel category in the two C–C conditions. (Confusion matrices are given in Appendices A-V and A-VI.) Results for individual vowel categories in the mixed-talker condition (right-hand column) verified the pattern found for overall errors. In comparison with the data for the mixed

#–# test (Table I), vowels of each category, with the exception of /ɔ/, were identified with greater accuracy when they were spoken in a variable consonantal frame than when they were spoken in isolation.

Results for individual vowel categories in the segregated talker tests (left-hand column) showed an unexpectedly high error rate for back vowels, /ɑ/, /ɔ/, /ʊ/, and /u/, for all three talkers. Errors on these vowels account for the lack of an overall advantage in the segregated condition over the mixed condition with C–C syllables. We currently have no explanation for this result.

The results of this experiment support the claim that consonantal context aids in the specification of vowel identity by providing important acoustic information to the listener. Even when the consonants are not known in advance, listeners are much more accurate in identifying medial vowels in CVC syllables than they are in identifying isolated steady-state vowels.[5] The acoustic effects of coarticulation carry substantial information about a medial vowel, which aids in vowel identification whether or not the listener has prior knowledge of the consonants' identity.[6]

III. SUMMARY AND CONCLUSIONS

In Experiment I, perceptual tests of vowels produced in isolation and in a fixed CVC context by the same talkers demonstrated that providing a consonantal environment increases the likelihood of correct identification of the intended vowel. This was true both when talker variation was present and when it was not; the advantage of consonantal context was independent of talker variation. Of the two factors investigated, consonantal context was much more important than talker variation in determining listeners' identification of vowels. The increment in error for isolated vowels in comparison to the medial vowels was more than three times greater than the increment attributable to unpredictability of talker.

We considered what might account for the difference in intelligibility between vowels in /p–p/ environment and in isolation. We concluded that the poor intelligibility of isolated vowels could not be attributed to the talkers' failure to produce these vowels in a consistent manner or to their adoption of aberrant formant frequencies. Measurements showed that formant frequency values and relative durations of isolated vowels were generally quite similar to those of vowels in the consonantal frame. The relative intelligibility of a token cannot be estimated very precisely from its position in the space defined by the two formants, a fact also noted by Peterson and Barney (1952).

The second experiment showed that consonantal context aids vowel identification even when the consonant frame varies unpredictably. Vowels produced in randomly varying stop consonant environments were identified more accurately than were isolated vowels both when the talker was fixed within a test block and when talkers, as well as context, varied unpredictably.

These results are surely puzzling if one makes the assumption that target frequencies of the formants alone could fully specify the vowels. If that were so, an isolated quasi-steady-state utterance ought to be an *optimal* signal for perception. It is true that synthetic steady-state vowels based on these formant parameters are fairly intelligible to naive listeners and may be identified quite consistently by experienced listeners (Delattre, 1951). Moreover, in the domain of automatic speech recognition, some success has been achieved with a static model of the vowel. Gerstman (1968) devised an algorithm based on frequencies of the first and second formants of /h–d/ syllables recorded from 76 talkers by Peterson and Barney (1952). Gerstman's algorithm sorted nine vowels in this set with only 2.5% error, less than was made by human listeners. From such a result, one might infer that target formant frequencies can unambiguously specify the vowels of English as produced by a variety of talkers.

However, as we have seen, this conception of the vowel cannot be reconciled easily with certain facts of perception. Vowels in isolation were poor signals from the perceiver's standpoint, even though talkers adopted targets that differed little from those attained in citation-form /p–p/ syllables. Thus, we may suspect that no single cross section through the syllable can fully specify the vowel. This inference is consistent with previous studies in the phonetic literature, to which we have referred. It is also relevant, in this context, to mention the results of an experiment by Bond (1975) on perception of vowels created by iteration of a single cycle from steady-state vowel tokens. Perception of such vowels by naive listeners was even less reliable than the results we obtained for unedited isolated vowels. If target frequencies alone were fully adequate to specify the vowels, it is difficult to understand these results.[7]

We are led to conclude that cues that are ordinarily regarded as consonantal contribute regularly to the perception of the vowel. We suspect that much vowel information is contained in formant transitions, as Lindblom and Studdert-Kennedy (1967) suggested some time ago. Whatever the nature of the contribution consonantal environment makes to the identification of a vowel, the data we have reviewed point to the general conclusion that no single, temporal cross section of a syllable conveys as much vowel information to a perceiver as is given in the dyanmic contour of the formants. From the standpoint of perception, it would seem that the definition of a vowel ought to include a specification of how the relevant acoustic parameters change over time. While listeners may be trained to identify steady-state tokens accurately (Lehiste and Meltzer, 1973), there is no reason to believe that the processes involved in this activity are the same as those typically used for understanding speech in natural situations.

Finally, these results may have implications for understanding the vocal tract normalization problem. Attempts to specify vowels across talkers have usually taken as their basic data, the formant frequency values

of a single cross section of a syllable. Our research indicates that the human perceptual system is ill-equipped to deal with such data. It would seem fruitful to renew the search for invariants across talkers utilizing information defined over the time course of at least a syllable.

ACKNOWLEDGMENTS

This paper reports research begun during the academic year 1972–73 while D. Shankweiler was a guest investigator at the Center for Research in Human Learning, University of Minnesota at Minneapolis. The work was supported by grants to the Center and to Haskins Laboratories from the National Institute of Child Health and Human Development, by grants awarded to D. Shankweiler and J. J. Jenkins by the National Institute of Mental Health, and by a fellowship to R. Verbrugge from the University of Michigan Society of Fellows. We wish to thank Kevin Jones, Kathleen Briggs, and Robert Jenkins for their assistance in the experimental work, and James Jenkins for his advice and encouragement throughout this research.

APPENDIX: CONFUSION MATRICES

Tables report the frequency with which each intended vowel x was identified as response alternative y. In addition, summary statistics for each condition are provided: the percent error for each intended vowel, the overall percent error, and the number of listeners (N).

TABLE A-I. Vowels in /p–p/ syllables: Mixed-talker condition. Overall percent error = 17.0%; N = 19.

Intended vowel	Response i	I	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	Percent error
i	188		1						1		1.1
I		187	1			2					1.6
ɛ			139	47	3			1			26.8
æ			33	154		2				1	18.9
ɑ					152	19	17	2			20.0
ɔ				1	46	138	1	4			27.4
ʌ					18	5	161	6			15.3
ʊ		8			2		47	116	16	1	38.9
u							2	3	185		2.6

TABLE A-II. Vowels in /p–p/ syllables: Segregated-talker condition. Overall percent error = 9.5; N = 33.

Intended vowel	Response i	I	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	Percent error
i	329	1									0.3
I	3	318	4				2	2		1	3.6
ɛ	1		290	20	4	7	5			3	12.1
æ			5	324		1					1.8
ɑ				7	255	62	4	2			22.7
ɔ					55	269	2	4			18.5
ʌ					11	9	305	4		1	7.6
ʊ						29	19	272	10		17.6
u							1	2	327		0.9

TABLE A-III. Isolated vowels: Mixed-talker condition. Overall percent error = 42.6; N = 16.

Intended vowel	Response i	I	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	Percent error
i	119	30	6					1	4		25.6
I	2	124	19			3	6	1	1	4	22.5
ɛ	1	2	61	64	2	6	10	5	3	6	61.9
æ		2	51	84	3	10	1	6	2	1	47.5
ɑ	1		1	20	62	47	21	2		6	61.3
ɔ		1	2	2	18	112	17	6	1	1	30.0
ʌ		1		6	32	31	60	22	4	4	62.5
ʊ		1	5	3	1	15	48	81	1	5	49.4
u	2		1	1		7	6	16	124	3	22.5

TABLE A-IV. Isolated vowels: Segreated-talker condition. Overall percent error = 31.2; N = 30.

Intended vowel	Response i	I	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	Percent error
i	251	3	1	1		1	1	6	33	3	16.3
I	5	259	21		1	3	3	1	4	3	13.7
ɛ	4	7	161	92	9	6	9	7		5	46.3
æ			48	221	3	18	3	2	3	2	26.3
ɑ			2	37	107	135	17	1	1		64.3
ɔ		1	1	12	43	214	19	6		4	28.7
ʌ		1	6	30	47	31	174	9		2	42.0
ʊ			3	4	3	10	51	214	12	3	28.7
u	8	1	1	3	1	2	3	22	258	1	14.0

TABLE A-V. Vowels in C–C syllables: Mixed-talker condition. Overall percent error = 21.7; N = 22.

Intended vowel	Response i	I	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	Percent error
i	331	7	5	1			1		5	2	6.0
I	2	292	53	1			2	2			17.1
ɛ	3	20	269	31	2		21	3		3	23.6
æ			47	298		7					15.3
ɑ			4	2	6	242	85	6	4	1	2	31.3
ɔ	2	3	1	2	91	222	18	6	4	3	36.9
ʌ			21	5	14	4	289	17	1	1	17.9
ʊ	1	6	1		8	10	70	214	41	1	39.2
u	5				2		6	16	323		8.2

TABLE A-VI. Vowels in C–C syllables: Segregated-talker condition. Overall percent error = 22.9; N = 24.

Intended vowel	Response i	I	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	Percent error
i	354	2	17		1			1	5	4	7.8
I	4	339	35					1	1	4	11.7
ɛ	10	21	329	13	1		1		1	8	14.3
æ	2	1	28	333	2	7		1		10	13.3
ɑ		1	1	23	225	100	15	4	6	9	41.4
ɔ		1		11	130	217	4	10	4	7	43.5
ʌ			3	3	16	8	342	8		4	10.9
ʊ	2	4	2		10	1	53	209	91	12	45.6
u	1	1	1		5	2	5	48	318	3	17.2

*Requests for reprints should be addressed to Winifred Strange, Center for Research in Human Learning, 205 Elliott Hall, University of Minnesota, Minneapolis, MN 55455. A partial summary of these results was presented at the 87th Meeting of the Acoustical Society of America, New York, 25 April 1974, and published in W. Strange, R. R. Verbrugge, and D. Shankweiler, "Consonant environment specifies vowel

identity," Haskins Lab. Status Report Speech Res. SR-37/38 (1974). A more complete exposition of the problem of perceptual constancy in speech perception may be found in Shankweiler, Strange, and Verbrugge (in press).

[†]Present address: Department of Psychology, University of Michigan, Ann Arbor, MI 48104.

[1]The extremely high error rate for the vowel /ɑ/ is, in part, due to the considerable confusion bertween /ɑ/ and /ɔ/ in the dialect of the talkers. In Table I, the percentages shown in parentheses for these two vowels represent the error rates excluding /ɑ/–/ɔ/ confusions; that is, a response was counted correct if the subject identified an intended /ɑ/ either as /ɑ/ or as /ɔ/, and likewise for an intended /ɔ/. Adjusted overall error rates also presented in Table I show that subtracting /ɑ/–/ɔ/ confusions has little effect on the relative differences among the four conditions.

[2]For many isolated vowels and some vowels in /p–p/ frames, the offset of periodic energy preceded offset of higher formant energy considerably. However, the rank order of vowels within each listening condition was the same even when the duration of higher formant energy was considered. Thus, the conclusions discussed in the text are valid for both measures of duration.

[3]It has been suggested that the relatively poor performance on the isolated vowels might be due to the lack of correspondence between the stimuli and the orthographic representation of the alternatives provided on the response forms. For both /p–p/ and #–# conditions, subjects were required to respond by selecting the appropriate /p–p/ syllable, (peep, pip, etc.) Thus, subjects in the #–# condition had to "decode" the orthography to match the isolated vowel, whereas subjects who heard medial vowels had only to match the orthographic syllable to the perceived syllable. Since the preparation of this manuscript, we have used different response forms for both /p–p/ and #–# tests. The symbols on the response forms corresponded to vowels in isolation, (EE, IH, EH, etc.), and subjects were given practice to make sure they could use the symbols appropriately. Results of these studies, when compared to those from conditions using the syllable response alternatives, showed *no* difference in performance for the isolated vowels. On the other hand, errors for vowels in /p–p/ syllables were somewhat greater when we used the isolated vowel symbols. However, identification of medial vowels was still significantly better than for isolated vowels. Further studies of the effects of different response forms are underway and will be reported in a subsequent article. We feel quite confident that the large and consistent differences found in the present study were due *primarily* to perceptual effects.

[4]It is worth noting that tokens by the subset of 12 talkers used in the C–C test yielded 20% errors on the /p–p/ test. Thus, if anything, errors in the C–C study are probably overestimated relative to the results one might expect for a test including all 15 talkers.

[5]In a separate study, similar results were found when subjects were asked to identify both the consonants and the vowel in each test syllable. Errors in vowel identification averaged 29%. Thus, even with the additional task of identifying the consonants, error rates were substantially lower than when listeners were required to identify vowels in isolation.

[6]Two aspects of the design of the C–C tests make further interpretation of the results problematic. First, although each consonant appeared equally often, the occurrences of consonants in initial and final position were not balanced across vowels, nor were equal numbers of consonants contributed by different talkers in the mixed test. As a result, we cannot make precise statements about the relative advantages of fixed and variable contexts, about the interaction of context with talker variation, or about the relative effects of different consonants on the identifiability of coarticulated vowels. A second problem concerns a possible interaction between vowel categories and prior familiarity with particular test items. Many of the C–C syllables are words which are familiar to the listeners. If this factor has a major effect on the perception of vowels in tasks like ours (in spite of the closed response set and the instructions to ignore meaning), the superior recognition of C–C syllables might have little to do with the type of acoustic information made available. If so, one might expect that listeners would do far better on syllables that formed words than on those which were nonsense syllables. Of the 72 C–C syllables included in the present experiment, 38 were English words. The overall error rate for these tokens in the mixed-talker test was 15%, compared to a 25% error rate for the 34 remaining C–C syllables. While this suggests that linguistic experience is a factor in vowel identification under these conditions, two further observations should be made. First, both error rates are well below that obtained for isolated vowels. Thus, if experience is a factor at all, it is probably secondary to the presence of phonetic context. Second, the error rates for the real words and nonsense syllables are difficult to interpret, since the fraction of C–C syllables which are real words varies with different vowel categories. The analysis is further complicated by intrinsic differences in perceptual difficulty among the nine vowels and by differences among the C–C syllables in orthographic representation.

[7]The implications of the specification of vowels in terms of idealized "targets" is explored further in Shankweiler, Strange, and Verbrugge (in press).

Abramson, A. S., and Cooper, F. S. (1959). "Perception of American English vowels in terms of a reference system," Haskins Lab. Q. Prog. Report QPR-32, Appendix 1.

Bond, Z. S. (1975). "Identification of vowels excerpted from context," J. Acoust. Soc. Am. 57, S24(A).

Delattre, P. C. (1951). "The physiological interpretation of sound spectrograms," Publ. Mod. Lang. Assoc. Am. 66, 864–875.

Fairbanks, G., and Grubb, P. (1961). "A psychophysical investigation of vowel formants," J. Speech Hearing Res. 4, 203–219.

Fourcin, A. J. (1968). "Speech source inference," IEEE Trans. Audio Electroacoust. AU-16, 65–67.

Fujimura, O., and Ochiai, K. (1963). "Vowel identification and phonetic contexts," J. Acoust. Soc. Am. 35, 1889 (A).

Gerstman, L. H. (1968). "Classification of self-normalized vowels," IEEE Trans. Audio Electroacoust. AU-16, 78–80.

Harris, C. M. (1953). "A study of the building blocks in speech," J. Acoust. Soc. Am. 25, 962–969.

House, A. S., and Fairbanks, G. (1953). "The influence of consonant environment upon the secondary acoustical characteristics of vowels," J. Acoust. Soc. Am. 25, 105–113.

Jones, D. (1956). *An Outline of English Phonetics* (Heffer, Cambridge, England).

Lehiste, I., and Meitzer, D. (1973). "Vowel and speaker identification in natural and synthetic speech," Lang. Speech 16, 356–364.

Liberman, A. M. (1970). "The grammers of speech and language," Cognitive Psychol. 1, 301–323.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). "Perception of the speech code," Psychol. Rev. 74, 431–461.

Lindblom, B. E. F. (1963). "Spectrographic study of vowel reduction," J. Acoust. Soc. Am. 35, 1773–1781.

Lindblom, B. E. F., and Studdert-Kennedy, M. (1967). "On the role of formant transitions in vowel recognition," J. Acoust. Soc. Am. 42, 830–843.

Millar, J. B., and Ainsworth, W. A. (1972). "Identification of synthetic isolated vowels and vowels in h-d context," Acustica 27, 278–282.

Öhman, S. E. G. (1966). "Coarticulation of VCV utterances:

Spectrographic measurements," J. Acoust. Soc. Am. 39, 151–168.
Peterson, G. E., and Barney, H. L. (1952). "Control method used in a study of the vowels," J. Acoust. Soc. Am. 24, 175–184.
Peterson, G. E., and Lehiste, I. (1960). "Duration of syllable nuclei in English," J. Acoust. Soc. Am. 32, 693–703.
Potter, R. K., and Steinberg, J. C. (1950). "Toward the specification of speech," J. Acoust. Soc. Am. 22, 807–823.
Rand, T. C. (1971). "Vocal tract size normalization in the perception of stop consonants," Haskins Lab. Stat. Report Speech Res. SR-25/26, 141–146 (unpublished).
Schatz, C. (1954). "The role of context in the perception of stops," Language 30, 47–56.
Shearme, J. N., and Holmes, J. N. (1962). "An experimental study of the classification of sounds in continuous speech according to their distribution in the formant 1–formant 2 plane," in *Proceedings of The 4th International Congress of Phonetic Sciences* (Mouton, Hague), pp. 234–240.
Shankweiler, D. P., Strange, W., and Verbrugge, R. R. (in press) "Speech and the problem of perceptual constancy," in *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, edited by R. Shaw and J. Bransford (Lawrence Erlbaum Associates, Hillsdale, NJ).
Stevens, K. N., and House, A. S. (1963). "Perturbations of vowel articulations by consonantal context: An acoustical study," J. Speech Hearing Res. 6, 111–128.
Tiffany, W. R. (1959). "Nonrandom sources of variation in vowel quality," J. Speech Hearing Res. 2, 305–317.
Verbrugge, R. R., Strange, W., Shankweiler, D. P., and Edman, T. R. (1976). "What information enables a listener to map a talker's vowel space," J. Acoust. Soc. Am. 60, 198–212.