Perception & Psychophysics
1983, 34 (5), 441-450

Identification of vowels in "vowelless" syllables

JAMES J. JENKINS, WINIFRED STRANGE, and THOMAS R. EDMAN
University of Minnesota, Minneapolis, Minnesota

Traditionally, it has been held that the primary information for vowel identification is provided by formant frequencies in the quasi-steady-state portion of the spoken syllable. Recent research has advanced an alternative view that emphasizes the role of temporal factors and dynamic (time-varying) spectral information in determining the perception of vowels. Nine vowels spoken in /b/+vowel+/b/ syllables were recorded. The syllables were modified electronically in several ways to suppress various sources of spectral and durational information. Two vowel-perception experiments were performed, testing subjects' ability to identify vowels in these modified syllables. Results of both experiments revealed the importance of dynamic spectral information at syllable onset and offset (in its proper temporal relation) in permitting vowel identification. On the other hand, steady-state spectral information, deprived of its durational variation, was a poor basis for identification. Results constitute a challenge to traditional accounts of vowel perception and point toward important sources of dynamic information.

Traditionally it has been held that the primary information for the perception of vowels is provided by their "target" formant frequencies (Joos, 1948). These targets are taken to be the center frequencies of the vocal tract resonances for each vowel when the vowel is produced as a sustained, isolated token. It is usually the case that the frequency loci of the first two speech formants are sufficient to differentiate the nine American monophthongs when they are spoken in isolation or pronounced in carefully articulated syllables by a single speaker. These so-called steady-state vowels can be represented as static points arrayed in a formant frequency "vowel-space," as in the classic study by Peterson and Barney (1952).

Several early studies (House & Fairbanks, 1953; Lindblom, 1963; Stevens & House, 1963) indicated, however, that vowels spoken in syllables at normal or rapid rates often failed to reach their target frequencies and showed considerable acoustic variation as a function of syllabic context. That is, vowels coarticulated with consonants yield acoustically variant signals. Until recently, little research was devoted to an understanding of the perceptual consequences of these facts (but see Lindblom & Studdert-Kennedy, 1967). When vowel perception was studied, isolated, steady-state vowels (either spoken or synthetically generated) were customarily used. It was commonly assumed that variations due to consonantal context, speaking rate, and speaker characteristics were compensated for by normalizing mechanisms (Gerstman, 1968; Stevens & House, 1963; Summerfield & Haggard, 1975).

Recent research in our laboratories and elsewhere has led us to question the adequacy of a description of vowels as static points in a Formant 1/Formant 2 space. Strange, Verbrugge, Shankweiler, and Edman (1976) found that medial vowels in naturally produced, consonant-vowel-consonant (CVC) syllables were identified by naive listeners more accurately than were vowels spoken as sustained, isolated tokens, even when the syllables were produced by many different speakers.[1] In another study of vowels produced in both citation-form syllables and syllables excised from sentences, Verbrugge, Strange, Shankweiler, and Edman (1976) again found relatively good identification of CVC syllables, despite considerable ambiguity in target information contributed by age and sex differences among the speakers. They also found that information that specified speaking rate played an important role in vowel identification.

Strange, Edman, and Jenkins (1979) examined listeners' identification of vowels in syllables of four different structures: CVC, CV, VC, and isolated vowels. Tests of vowel identification, using /b/ and

This research was supported by grants to James J. Jenkins and Winifred Strange from the National Institute of Mental Health (MH-21153) and to the Center for Research in Human Learning from the National Institute of Child Health and Human Development (HD-0098) and the National Science Foundation (BNS 75-03816). We are happy to acknowledge the technical support of Haskins Laboratories, which made the stimulus preparation possible, and the stimulating interaction with the personnel of Haskins, which made the work pleasant and rewarding. We wish to thank Kathleen Briggs, Christopher Jenkins, Kevin Jones, and Deb Kasma for their assistance in data collection and analysis. Part of this work was reported at the 93rd meeting of the Acoustical Society of America. James J. Jenkins and Winifred Strange are now at the University of South Florida. Thomas R. Edman is now at the Technology Strategy Center, Honeywell Inc., Roseville, Minnesota. Requests for reprints may be sent to Winifred Strange, Department of Communicology, University of South Florida, Tampa, Florida 33620.

/p/ as consonants, again demonstrated that most consonantal contexts improved vowel identification relative to that demonstrated for isolated vowels. In addition, it was shown that closed-syllable (VC) contexts aided identification more than open-syllable (CV) contexts, presumably because they provided better information about "intrinsic vowel duration," which is usually considered a secondary cue to the identity of English vowels. The investigators concluded that both temporal factors and time-varying spectral information played an important role in determining the identifiability of vowels. Gottfried and Strange (1980) repeated these experiments with the velar consonants /k/ and /g/ with similar results, although vowels in all of the /g/ contexts were identified relatively poorly.

The research reviewed above motivated a return to the basic questions: How are vowels specified in the speech stream? What acoustic parameters provide the listener with critical information for unambiguous identification of American English monophthongs as spoken in continuous coarticulated speech? From our earlier research, we have concluded that in normal speech, vowels, like consonants, are specified by time-varying information defined over the entire syllable (and, perhaps, beyond). The research reported here was undertaken to examine more closely the degree to which such time-varying information specifies the vowel and to explore the nature of that information.

The present experiments represent our first attempt to isolate and manipulate three major sources of information for vowel identity in CVC syllables: (1) static "target" information provided in the quasi-steady-state portion of the syllable, (2) time-varying spectral information provided in the formant transitions into and out of the vowel nucleus, which will be referred to as dynamic spectral information, and (3) temporal information that reflects "intrinsic vowel duration." The phonetic feature of vowel length (redundant in English) has been most closely associated in acoustics with the duration of the vocalic nucleus. Such a definition, of course, confounds energy and elapsed time. Syllables containing phonetically "long" vowels have more overall energy than spectrally similar "short" vowels, and the time between initial consonant release and final consonant closure is, of course, longer for such syllables.

We decided to modify syllables of natural speech rather than attempting to manipulate these parameters in synthetic stimuli, for the obvious reason that we did not yet know how to synthesize time-varying vowel information. The experiments, therefore, are explorations of the information for vowel identity available in electronically edited segments of real speech. The operations performed were, for the most part, relatively simple. To isolate the steady-state targets of the syllable, the transitions at the beginning and end of the syllable were attenuated to silence (as in Fujimura & Ochiai, 1963). Conversely, to investigate the role of formant transitions, the quasi-steady-state portion of each syllable was attenuated to silence, creating the "vowelless" syllables referred to in the title. The effects of durational information were explored by manipulating the elapsed time between the initial and final transition segments of the vowelless syllables, and (in Experiment 2) the duration of the vocalic segment that was retained.

The logic of the experiment was simple. Using the above types of edited syllables, we conducted vowel identification tests. If subjects, listening to altered syllables, could accurately identify the vowels intended in the original syllables, that would provide direct evidence for the presence of information sufficient to specify the vowel. (It must be noted, of course, that a failure of accurate identification would not mean that the information in that segment was not involved in normal vowel identification. It would only mean that in such artificial isolation the information was not sufficient to permit accurate identification.) In essence, then, the experiments are a "brute force" effort to provide an evaluation of the likely major sources of information for vowel identification.

EXPERIMENT 1

Method

Stimulus materials. One token of each of nine /b/ + vowel + /b/ syllables spoken "briskly" by an adult male speaker of Upper Midwestern dialect was recorded, low-pass filtered (3860 Hz), and

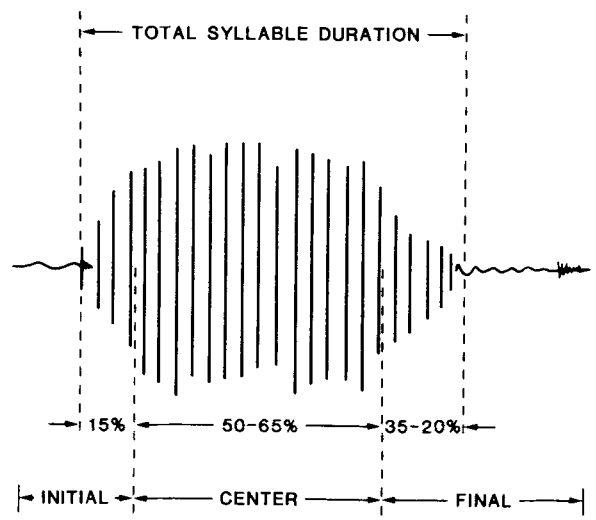

Figure 1. Schematic representation of the acoustic waveform of a syllable. Each syllable was divided into three components that were proportions of the total syllable duration, from initial stop release to final stop closure.

digitized (10 kHz) using the Haskins Laboratories Pulse Code Modulation system. The digitized waveform of each syllable was divided into three components, as illustrated schematically in Figure 1.

The total syllable duration was measured from the initial consonant release (not counting any prevoicing, when present) to the final closure (not counting the period of closure or the following release). The initial component consisted of the first 15% of the total duration for each syllable (plus the prevoicing when present). The center component consisted of the next 50% to 65% of the total duration, depending on the vowel: 50% for the intrinsically short vowels /I/, /ɛ/, /ʌ/, and /ʊ/, 60% for the intermediate vowels /i/ and /u/, and 65% for the long vowels /æ/, /a/, and /ɔ/. These percentages were selected on the basis of data from Lehiste and Peterson (1961) to insure that all of the quasi-steady-state part of each syllable was included in the center. (Later spectrographic analysis showed that the centers as well as initial and final components contained formant movement associated with the initial and final consonants, especially for the short vowels.) Last, the final component was defined as the remaining 35% to 20% of the syllable, plus the closure period and release. These three components are indicated in the bottom of Figure 1. The appendix gives the duration of each component for each of the nine syllables.

Having defined these three components for each of the nine syllables, seven sets of stimuli were generated as shown in Figures 2A and 2B. The control stimuli (shown at the top of each figure) were the nine full syllables, that is, the digitized versions of the original /b/ + vowel + /b/ syllables. Six sets of modified syllables were constructed as follows:

The silent-center syllables (Figure 2a, second row) were generated by retaining the initial and final components in the appropriate temporal relationship, and attenuating the center component until no signal remained. These are the so-called "vowelless" syllables, in that they contained no quasi-steady-state nucleus associated with the vowel targets. Both dynamic spectral information and durational information (elapsed time), available in the original syllables, were retained in these modified syllables.

The variable (length) centers, shown in the bottom row of Figure 2A, were the exact converse of the silent-center syllables. The vowel nucleus remained, and the initial and final components were deleted. Because the center components were defined as a proportion of the total syllable duration, these stimuli varied in duration (hence, the name variable centers) such that differences between intrinsically short, mid, and long vowels were actually enhanced. Thus, these stimuli contained both the target information thought to be the primary cue for vowel identity and the secondary temporal cue of relative vowel duration.

Because the silent-center syllables contained as much as 120 msec of absolute silence, we were concerned that naive listeners might not perceive the two parts of the syllable as an integrated whole. In an attempt to give these vowelless syllables some continuity, we created another set of stimuli which filled the gap in the silent-center syllables with naturally produced speech noise, /ʃ/. (These are illustrated in the second row of Figure 2B). The attempt was to create stimuli that sounded like syllables transmitted over a channel containing intermittent static. (The noise was adjusted to an amplitude considerably lower than the portion of the speech signal that it replaced.) These stimuli were called Hiss-Center Syllables.

In order to partial out the effects of durational information and dynamic spectral information, a further set of stimuli was constructed using the initial and final components. However, these components were juxtaposed in time, by adjoining the two digitized waveforms with no silent interval between them, as illustrated in the bottom row of Figure 2B. These stimuli are referred to as the abutted syllables. They contain the formant transitions into and out of the vocalic nuclei of the original syllables, but elapsed time differences between long and short vowels have been neutralized such that all syllables are short. Furthermore,

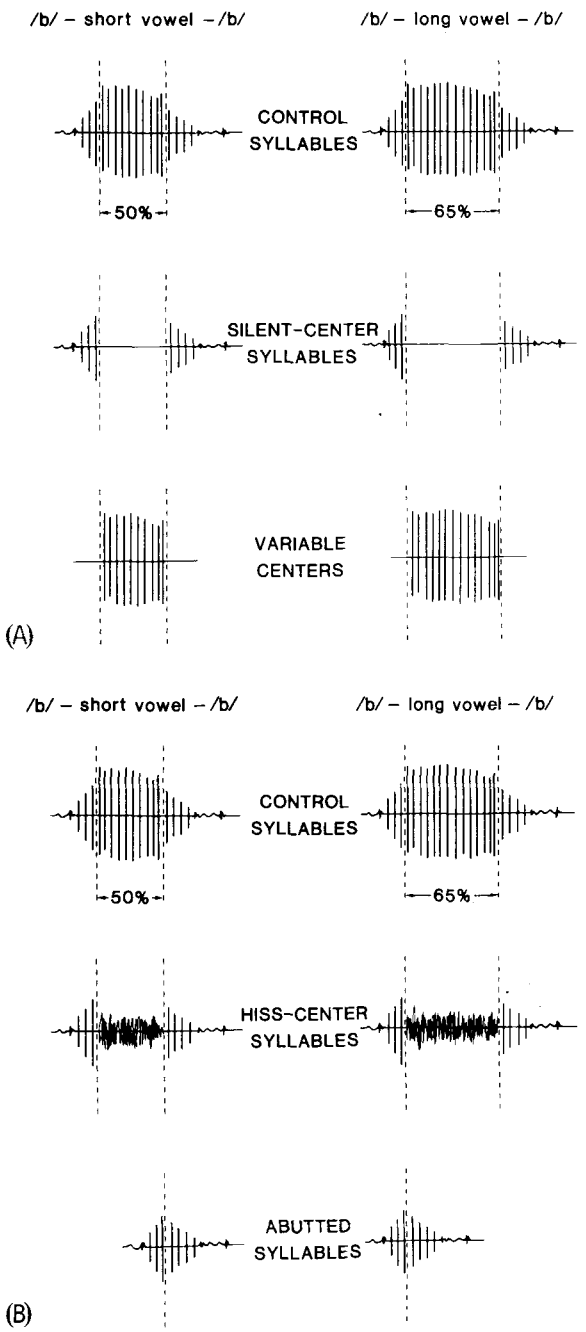

Figure 2. (A) Schematic representations of the acoustic waveforms of control syllables (top row), silent-center syllables (middle row), and variable-centers stimuli (bottom row). (B) Schematic representations of the acoustic waveforms of control syllables (top row), hiss-center syllables (middle row), and abutted syllables (bottom row).

there are major discontinuities in the middle of each stimulus in the formant pattern, the pitch contour, and the energy envelope.

Two final stimulus sets (not shown), called the initials and the finals, were constructed by keeping the appropriate component and deleting the two components not desired. These were included as control conditions to test whether the vowels could be ac-

curately perceived on the basis of either of these components taken by itself. Although spectral analysis showed that vowel targets were not reached in either of these components (that is, one or more formants had not attained steady-state values), we wished to make a perceptual assessment of this as well.

Perceptual tests for these seven sets of stimuli were recorded separately. The nine tokens of each set were repeated 10 times each in random order for a total of 90 items on a test. A between-subjects design was used in the experiment to circumvent confoundings of orders and subject experience with the materials.

Subjects. All subjects were volunteers from undergraduate psychology classes at the University of Minnesota; they received partial course credit or $2 for participating in the experiment. All were native speakers of American English, and most were native to the Upper Midwest region of the United States. All reported having no hearing difficulties. Small groups of subjects were assigned randomly to the seven conditions until there were between 15 and 20 subjects in each condition. A total of 126 subjects were tested.

Procedure. Tests were presented to groups of two to eight subjects in a quiet experimental room via a Revox A77 tape recorder, MacIntosh MV49 amplifier, and an AR acoustic suspension loudspeaker. Amplification levels were necessarily different for the seven test conditions because of differences in length, peak amplitude, and energy envelope of the stimuli. The initials and finals were amplified the greatest amount relative to the controls, but the silent-center syllables were also amplified, relative to the controls. The level for each stimulus series was adjusted so that the stimuli were clearly audible to all listeners. The level was monitored by a Heathkit VTVM placed across the output to the loudspeaker, so that levels were the same for all subjects within a stimulus condition.

The subjects responded on score sheets which contained nine response alternatives, written in English orthography and arrayed in rows. The subjects were told that their task was to identify the vowels in the syllables, or parts of syllables, by circling the key word containing the vowel they had heard. They were informed that the stimuli were speech syllables that had been modified by computer. They were familiarized with the response sheets and with the experimental stimuli at the same time, and given detailed instructions, examples, and feedback for 27 trials, three instances of each vowel. Then they responded to the 90-item test, without feedback. After the subjects completed their experimental series, they were all tested on the 90-item control series. (For the control subjects, this was a replication of the first test.)

Results

The identification score for each subject on the control series (second test for the control-condition subjects) was inspected first, in order to eliminate subjects who were unable to perform the identification task with the original syllables within an acceptable range of accuracy (2 SDs from the overall mean). Records for any subject who made more than 20% errors on the control syllables were discarded from the experiment without regard to the subject's prior performance on the experimental materials. Only 5 of 126 subjects were removed from the experiment by this criterion: 2 from the variable-centers condition, 2 from the finals condition, and 1 from the abutted syllables condition.

The identification results for each of the seven experimental conditions are given in Table 1, which presents the overall error rate summed over all nine vowels within each condition. An error was defined as an omission (there were very few) or a response other than the vowel intended by the speaker in the original syllable, with one exception. Confusions between /*a*/ and /ɔ/ were not counted as errors because this distinction is not made in the dialect of many of our subjects. However, responses other than /*a*/ or /ɔ/ on these stimuli were included as errors.

It can be seen that the error rate varied greatly as a function of stimulus condition. A one-way analysis of variance of subjects' errors yielded a significant difference across conditions [$F(6,114) = 47.85$, $MSe = 83.04$, $p < .001$]. It is apparent that performance in the silent-center syllables condition was best of all the modified conditions; indeed, performance there was not different from performance on the controls. Variable centers and hiss-center syllables also yielded relatively accurate performance, while abutted syllables, initials, and finals produced many more vowel-identification errors.

To assess the statistical reliability of these differences, the critical range was calculated via the least significant difference test ($CR = 8.16$, $p = .01$). By this criterion, the controls, silent-center syllables, and variable centers were different from the other conditions but not different from each other. Thus, it appears that sufficient information for accurate identification of the vowels was provided by these two experimental conditions. The variable centers and the hiss-center syllables were not significantly different from each other, but both were better than the abutted syllables, which were, in turn, markedly better than the initials and finals.

Inspection of the distribution of subjects' scores in

Table 1
Identification Errors Over All Vowels, Experiment 1 (Excluding /*a*/–/ɔ/ Confusions)

Condition	N	Mean	SD	Percent Error	LSD Clusters
Control	18	6.22	5.87	6.9	
Silent Center Syllables	16	6.88	6.99	7.6	
Variable Centers	20	11.85	11.68	13.2	
Hiss-Center Syllables	15	16.80	11.82	18.7	
Abutted Syllables	18	27.33	9.70	30.4	
Initials	15	41.07	7.15	45.6	
Finals	19	41.91	6.43	46.6	

each condition lends support to these clusters. Only 2 of the 54 subjects in the three most accurate groups had more errors than the median of the abutted-syllables condition, and only 1 of 18 subjects in the abutted-syllables condition had fewer errors than the median of the three accurate conditions combined. In parallel fashion, only two subjects in the abutted-syllables condition had more errors than the median of the initials and finals conditions and only one subject in the latter two groups had fewer errors than the median of the abutted-syllables condition.

Discussion

These results convincingly demonstrate that listeners can respond appropriately to the silent-center syllables as integrated speech utterances and that they find sufficient information in these "vowelless" syllables to identify vowels accurately. Surprisingly, there was no significant decrement in the accuracy of identification of these stimuli as compared with performance on the control stimuli. This finding is of both theoretical and experimental importance.

The fact that listeners can identify the vowels accurately in syllables in which only the consonant onsets and offsets are given is strong evidence for the adequacy of dynamic information in specifying the vowel. Both spectral dynamics and durational information were given in these stimuli, of course, and the importance of these sources (as suggested in our earlier work) seems to be amply confirmed in this study. Indeed, the results suggest that such sources of information are sufficient to specify the vowels unambiguously. The practical significance of the finding is the confirmation that modified syllables such as these can be used effectively with naive listeners to explore sources of information involved in the identification of vowels.

The hiss-center syllables offered no advantage over the silent-center syllables as experimental stimuli, and the significant increase in errors for the former suggests that the addition of noise in the silent interval actually worked to some subjects' disadvantage. Identification errors ranged from 0% to 44% across subjects. Perhaps the sudden transition into another speech-like sound in midsyllable was interfering rather than facilitating in perceiving the continuity of the syllable. Alternatively, the level of noise relative to vocalic signal may have been sufficient to mask relevant information in the initial and final segments. Because the silent-center syllables yielded very good performance, and subjects had no difficulty in perceiving their unity, there was no motivation for further experimentation with the hiss-center syllables.

The variable centers provided relatively good information for vowel identity. It must be noted, however, that these stimuli provided not only steady-state spectral information of the sort represented in a "vowel space," but also provided information for phonetic vowel length in the form of relative duration differences. Duration differences were actually enhanced over those specified in the control stimuli because a greater proportion of the original syllable was included for long vowels than for the short vowels. Thus, while the ratio of long to short original syllables was 1.37, the ratio of long to short variable-centers stimuli was 1.75.

Identification of vowels in the abutted-syllables condition was relatively poor, as might have been expected. Information for vowel length in the form of durational differences was unavailable in these stimuli, and the resulting stimuli were quite brief (about 60 msec). Perhaps more importantly, there were discontinuities in formant frequencies, amplitude, and pitch contours in the middle of the re-formed syllables that might well have contributed to identification difficulty. This is borne out by the observation that both long and short vowels were misperceived more often, relative to the silent-center syllables. If the decrement were due to the neutralization of duration information alone, we would expect an increase in errors primarily for long vowels.

As expected, the initials and finals conditions were grossly inadequate in providing effective information for vowel identification. In both of these conditions, vowel targets were not reached, and thus the static spectral information was ambiguous when either segment was considered by itself. Durational information, perhaps available in the rates of transitions, appeared not to be effective when the initials and finals were presented alone. The failure of these isolated components to yield accurate vowel identification is in sharp contrast to the high level of identification accuracy achieved with the silent-center syllables (and to a lesser degree, the hiss-center syllables). This pattern of results supports the interpretation that the silent-center syllables were perceived as integrated syllables, and that the information available in those syllables is abstract, in that it is specified over the two segments as a unit.

EXPERIMENT 2

The second experiment was designed to acquire more data concerning the modified syllable paradigm while at the same time making the experiment somewhat more analytic. The following changes were made from the conditions and procedures of Experiment 1.

First, the hiss-center syllables condition was dropped for the reasons given in considering the results of Experiment 1. Second, a new condition was added. As noted above, the variable-centers stimuli provided both static spectral information, tradi-

tionally regarded as crucial for vowel identification, and (enhanced) relative duration information for phonetic vowel length. In order to assess the relative contribution of these two kinds of information, a new set of stimuli, one which contained no obvious durational information, was included. This condition, called fixed centers, was constructed by trimming all of the center portions of syllables to a fixed length, namely, the length of the shortest of the variable centers.

Third, the procedure for familiarization of the listeners with the task and stimuli was changed. In experiment 1, there was the possibility that the familiarization stage provided more than training on the task and the response forms. Since there was only one instance of each of the nine vowels, the familiarization procedure with feedback could be regarded as a paired-associate learning situation in which subjects could learn vowel names for the modified stimuli. Although we did not think this had affected the results substantially in the first experiment, it seemed prudent to eliminate this possibility by a procedure in which familiarization with feedback on the task and response forms was accomplished with unmodified control syllables, and then familiarization with the modified stimuli was done without feedback.

Method

Stimulus materials. The stimulus materials for the control, silent-center syllables, variable centers, abutted syllables, initials, and finals conditions were identical to those used in Experiment 1. Stimuli for the fixed centers condition were produced by trimming the stimuli from the variable-centers condition to a uniform length of approximately 60 msec. (Because waveforms were cut at zero crossings, the actual length of the fixed-centers stimuli varied slightly about that value.) Equal portions were trimmed from the beginning and end of each variable center so that the segment corresponding most closely to the target was retained for those vowels that were characterized by movement throughout the vocalic portion.

Procedure. The experimental procedure was the same as that employed in Experiment 1 except that subjects in all experimental conditions were given task and response-form familiarization with full syllables, the control series. Thus, there was no possibility that the subjects were being trained in the interpretation of the stimuli of the experimental series (except, of course, for the control-condition subjects). After this training, the subjects heard 18 tokens (2 of each vowel) of the experimental stimuli they were to be tested on, in random order with no requirement to respond and no informative feedback. Then the 90-item experimental test and the 90-item control test were conducted as before.

Subjects. As in Experiment 1, small groups of subjects were assigned randomly to experimental treatments until 10 to 15 subjects had been tested in each condition. For the fixed-centers condition, however, 21 subjects were tested, because the condition had not been studied before.

Results and Discussion

As in Experiment 1, data from subjects were discarded if the subject made more than 20% errors in identification of the vowels in the (second) control condition. In Experiment 2, records from only three subjects were deleted, all from the fixed-centers condition.

Overall errors in identification of vowels for each of the stimulus conditions are given in Table 2. As before, an error was defined as a vowel response other than the one intended by the speaker (excluding */a/-/ɔ/* confusions) and an omission. Errors were summed over the nine vowels and expressed as a percentage of total response opportunities. It is again apparent that the error rate varied as a function of stimulus condition. A one-way analysis of variance of subjects' average errors yielded a significant difference across conditions [$F(6,96) = 28.89$, $MSe = 69.65$, $p < .001$].

To reveal the clusters of treatments, the critical range was calculated via the least significant difference test ($CR = 8.42$, $p = .01$). Results of this statistical test revealed the following clusters: Controls and silent-center syllables both yielded excellent identification of the vowels. Performance in the variable-centers condition was poorer than performance in the first two conditions and not significantly better than in the abutted syllables. However, the variable-centers condition was significantly better than the fixed centers, finals, and initials conditions. The abutted syllables and fixed centers were significantly better than the initials condition, but not significantly better than the finals condition.

Again, a study of the score distributions in each of the experimental conditions tended to confirm these clusters. The control and silent-center syllables conditions had virtually identical distributions of error scores. No subject in these groups had more errors than the median of the variable-centers condition, and only one subject in the variable-centers condition had fewer errors than the median of the silent-center syllables condition. The variable-centers condition overlapped considerably with the abutted-syllables condition, but the variable-centers condition showed little overlap with the fixed-centers condition (5 of 35 subjects' scores overlapped the median of the other group). The abutted-syllables and fixed-centers scores showed almost complete overlap. Scores in the initials and finals conditions showed lower errors than the median of the abutted-syllables and fixed-centers conditions in only 2 of 21 cases.

These results are very similar to those of the first experiment. The ordering of performance in the experimental conditions was the same. The only inconsistency is that performance in the variable-centers condition clustered with the control and silent-center-syllables conditions in Experiment 1, but with a group of less effective conditions in Experiment 2. Of major importance are two findings: The silent-center syllables were again demonstrated to provide sufficient information to specify the identity of the vowel unambiguously. On the other hand,

Table 2
Identification Errors Over All Vowels, Experiment 2 (Excluding /a/–/ɔ/ Confusions)

Condition	N	Mean	SD	Percent Error	LSD Clusters
Control*	18	6.22	5.87	6.9	
Silent Center Syllables	14	6.00	5.87	6.7	
Variable Centers	14	17.71	8.71	19.7	
Abutted Syllables	15	25.60	11.36	28.4	
Fixed Centers	21	27.19	8.25	30.2	
Finals	11	32.64	8.33	36.3	
Initials	10	35.90	7.33	39.9	

**Data from Experiment 1 for comparison.*

the vowel "target" information by itself, as represented by the fixed-centers condition, was significantly less effective in specifying the vowels. Thus, in two ways the results provide further evidence of the importance of dynamic spectral information and durational information in specifying the vowel for the listener.

The procedure that was adopted in this second experiment in order to prevent paired-associate learning during familiarization trials had little or no effect on the results. For the silent-center syllables and abutted syllables, there was virtually no change from Experiment 1 in the average error rate. For the variable centers, the new procedure appeared to be slightly more difficult, but for the initials and finals, the new procedure yielded somewhat better performance. Since none of these differences was statistically significant, further discussion of the experimental results will be based on pooled data for comparable conditions in Experiments 1 and 2. These data, given as percentages of errors pooled across all nine vowels in each condition, are given in Figure 3.

The overall pattern of results shows that modified-syllable conditions that retained one or more dynamic sources of information yielded the best performance by listeners. The silent-center syllables, which contained both time-varying spectral and durational information, but no vowel targets, were identified as accurately as were the control syllables. Hiss-center syllables and variable centers (the latter of which included enhanced durational information as well as vowel targets) were also well perceived relative to abutted syllables and fixed centers.

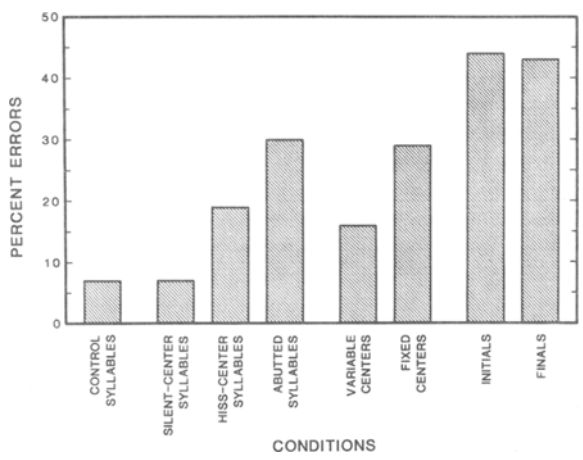

Figure 3. Average identification errors (expressed as percentages of opportunities) for each stimulus condition, averaging over data from Experiments 1 and 2.

That the dynamic spectral and durational sources of information are critical for accurate perception was demonstrated by the poor performance on the fixed-centers condition. These stimuli contained the vowel targets, but neither consonant transitions nor relative duration differences. Subjects made about four times the number of errors on these stimuli as they did on the control syllables. In fact, performance of the fixed centers was no better than on the abutted syllables, which included initial and final transitions, but no targets and no duration differences.

As expected, performance was very poor on the initials and finals, demonstrating that identifiable vowels were not "contained" in either of these components taken by itself. Thus, the extremely good performance in the silent-center syllables must be attributed to the subjects' utilization of relational information that was defined over the two components perceived as an integrated whole.

Before discussing the ramifications of these results, it is informative to inspect the pooled data for each of the nine vowels in each condition, presented in Table 3 as percentages of opportunities to respond. These data afford an opportunity to examine the consistency of the differences between conditions across the individual vowels as well as the opportunity to examine differences in identifiability of the particular vowels across conditions.

As the table shows, particular vowels appeared to be intrinsically less ambiguous than others. The vowels /i/ and /u/, which mark the extremes of the vowel space, produced relatively few errors in all but the finals and initials conditions, respectively. The vowel /I/ was also accurately perceived in all mod-

Table 3
Identification Errors, in Percent, for Each Intended Vowel Combined Data for Experiments 1 and 2 (Excluding /ɑ/–/ɔ/ Confusions)

Intended Vowel	Control Syllable	Silent Center	Hiss-Center	Variable Center	Fixed Center	Abutted Syllable	Initials	Finals
N =	18	30	15	34	21	33	25	30
i	6	3	2	2	<1	5	8	36
ɪ	3	4	6	5	10	5	21	47
ɛ	7	13	44	17	27	49	78	24
æ	<1	4	4	27	68	57	50	69
ɑ	2	<1	10	12	30	29	22	45
ɔ	4	5	21	10	22	45	58	72
ʌ	12	14	35	37	57	35	69	45
ʊ	26	20	40	20	39	33	70	43
u	2	0	6	11	7	8	17	5
Overall Errors	7	7	19	16	29	30	44	43

ified conditions except the initials and finals. In contrast, the vowel /ʊ/ was misidentified relatively often, even in the control condition. The other mid and low vowels yielded differential error rates as a function of the experimental condition. For these vowels, consistently lower error rates were found for conditions in which durational information was available than for conditions in which it was not. This was true both when comparing silent-center syllables with abutted syllables and when comparing variable centers with fixed centers.

Pearson product-moment correlations of errors on each vowel were calculated between all pairs of conditions. While there is little power in this statistic, since it is based on only nine points for each coefficient, the results are of some descriptive interest.

Significant correlations were found for 5 of the 21 coefficients. The only condition that was significantly similar to the control condition was the silent-center syllables condition ($r = 0.89$). These conditions were also highly similar in mean error. The silent-center-syllables condition was also similar in error pattern to the initials condition ($r = 0.82$), although these conditions differed markedly in mean error. The variable centers condition correlated most highly with the fixed centers condition ($r = 0.91$), although they also differed markedly in mean error. Presumably, the positive correlation was a result of the sharing of target formant values, whereas the mean difference was the result of the fact that variable centers contained durational information but fixed centers did not. The remaining significant correlations showed the abutted condition to be related to the initials condition ($r = 0.79$) and to the fixed centers condition ($r = 0.77$). It is possible that these relationships were based on the fact that all three conditions had neutralized durational sources of information for vowel length. The pattern of errors in the finals condition was not significantly related to the error pattern of any other condition.

In summary, the vowel-by-vowel analysis indicated that the overall differences among conditions reflected consistent differences in identifiability of individual vowels, especially the mid and low vowels. Correlations between conditions were suggestive of the sources of information used by listeners in making their identification responses. Durational information appeared to be especially important in determining accurate vowel identification. Because of the way durational information had been neutralized (by shortening the length for the long vowels in fixed centers and the elapsed time for the abutted syllables), errors in these conditions were especially great on intrinsically long vowels. However, there were also increases in errors on short vowels that cannot be accounted for on this basis. (See Strange, Jenkins, & Johnson, 1983, for a further discussion of these issues.)

GENERAL DISCUSSION

In traditional accounts of vowel perception, vowels are characterized acoustically by reference to their canonical targets, that is, those formant frequencies corresponding to the resonances of a static vocal tract held in the position appropriate for the vowel. The problem for understanding the perception of vowels, given this acoustic characterization, is that these canonical targets are not often present in the acoustic signal due to the coarticulation of vowels with consonants in ongoing speech. The lack of invariance in formant frequencies is a function of variations in phonetic context, speaking rate and speaker characteristics. The theorist is thus confronted with a classical constancy problem: how do these variant (and overlapping) acoustic signals give rise to invariant perception of vowels? To cope with this problem, normalization mechanisms have been postulated whereby the inherently ambiguous signals are somehow disambiguated on the basis of other, supposedly independent, information about speaker identity, speaking rate, and phonetic context. (See Summerfield, 1981, for a discussion of speaking rate normalization theories.)

An alternative view, which we have proposed here and elsewhere (see Shankweiler, Strange, & Verbrugge, 1977; Strange et al., 1983) offers a conception of the acoustic (and articulatory) characterization of vowels as intrinsically dynamic in nature. According to this view, coarticulation of consonants and vowels is not to be considered as the introduction of unfortunate "noise" in the acoustic signal. On the contrary, the act of coarticulating phonemes in syllables gives rise to an acoustic array in which the consonants and vowels are cospecified in the time-varying spectral configuration. (See Fowler, 1980, for a presentation of this conception of speech production.) Thus, we would not expect that either the consonants or the vowels were necessarily unambiguously specified in any particular spectral cross-section of the acoustic signal.

While there has been considerable discussion and research dealing with the role of dynamic parameters in the perception of consonants (e.g., Kewley-Port, 1981; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Summerfield, 1981), less effort has been directed toward an understanding of the implications of this approach for theories of vowel perception (but see Verbrugge & Rakerd, 1980). The research reported here is another empirical step toward formulating a model of how coarticulated vowels might be perceived. It demonstrates that perceivers are able to identify vowels in CVC syllables on the basis of dynamic spectral information given by transitions into and out of the "vowel nucleus" in their proper temporal relation. That is, it shows that time-varying sources of information are *sufficient* for accurate vowel identification, even in highly artificial stimuli in which the vowel nuclei are totally absent.

What remains to be accomplished is the specification of the invariant information that supports the identification of coarticulated vowels and an account of how that information is used by perceivers. The present study does not yield definitive answers to those questions, but it does point the way toward a more adequate acoustic characterization of vowels. We can say the following, on the basis of the results reported here. Information for vowel identity is available in the changing acoustic pattern across (at least) an entire syllable. The information is relational, in that it can be specified across two discontinuous segments of energy, neither of which by itself was sufficient to specify the vowel within the experimental paradigm used here. Acoustic parameters that are informative about the timing of articulatory events will be important for an adequate description of how vowels are specified for the perceiver. Relative rates of transitions into and out of the vowel nucleus and relative duration (specified by elapsed time) were implicated in this study as sources of information about timing. More analytical studies using the techniques employed here are underway to determine the exact nature of this information (see Strange et al., 1983). A final test of the adequacy of our descriptions will come from studies using synthetically generated speech in which the dynamic sources of information are manipulated.

REFERENCES

Diehl, R. L., McCusker, S. B., & Chapman, L. S. Perceiving vowels in isolation and in consonantal context. *Journal of the Acoustical Society of America*, 1981, **69**, 239-248.

Fowler, C. A. Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, 1980, **8**, 113-133.

Fujimura, O., & Ochiai, K.Vowel identification and phonetic contexts. *Journal of the Acoustical Society of America*, 1963, **35**, 1889. (Abstract)

Gerstman, L. J. Classification of self-normalized vowels. *IEEE Transactions on Audio- and Electroacoustics*, 1968, **AU-16**, 78-80.

Gottfried, T. L., & Strange, W. Identification of coarticulated vowels. *Journal of the Acoustical Society of America*, 1980, **68**, 1626-1635.

House, A. S., & Fairbanks, G. The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America*, 1953, **25**, 105-113.

Joos, M. A. Acoustic phonetics. *Language Supplement*, 1948, **24**, 1-136.

Kewley-Port, D. *Representations of spectral change as cues to place of articulation in stop consonants.* Unpublished doctoral dissertation, City University of New York, 1981.

Lehiste, I., & Peterson, G. E. Transitions, glides, and diphthongs. *Journal of the Acoustical Society of America*, 1961, **33**, 268-277.

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

Lindblom, B. E. F. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 1963, **35**, 1773-1781.

Lindblom, B. E. F., & Studdert-Kennedy, M. On the role of formant transitions in vowel recognition. *Journal of the Acoustical Society of America*, 1967, **42**, 803-843.

Macchi, M. J. Identification of vowels spoken in isolation versus vowels spoken in consonantal context. *Journal of the Acoustical Society of America*, 1980, **68**, 1636-1642.

Peterson, G. E., & Barney, H. L. Control methods used in a study of vowels. *Journal of the Acoustical Society of America*, 1952, **24**, 175-184.

Shankweiler, D. P., Strange, W., & Verbrugge, R. R. Speech and the problem of perceptual constancy. In R. E. Shaw & J. Bransford (Eds.), *Perceiving, acting and knowing: Toward an ecological psychology*. Hillsdale, N.J: Erlbaum, 1977.

Stevens, K. N., & House, A. S. Perturbation of vowel articulations by consonantal context: An acoustical study. *Journal of Speech and Hearing Research*, 1963, **6**, 111-128.

Strange, W., Edman, T. R., & Jenkins, J. J. Acoustic and phonological factors in vowel identification. *Journal of Experimental Psychology: Human Perception and Performance*, 1979, **5**, 643-656.

Strange, W., Jenkins, J. J., & Johnson, T. L. Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, 1983, **74**, 695-705.

Strange, W., Verbrugge, R. R., Shankweiler, D. P., & Edman, T. R. Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, 1976, **60**, 213-224.

SUMMERFIELD, Q. Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 1981, **7**, 1074-1095.

SUMMERFIELD, A. Q., & HAGGARD, M. P. Vocal tract normalization as demonstrated by reaction times. In G. Fant & M. Tatham (Eds.), *Auditory analysis and perception of speech*. New York: Academic Press, 1975.

VERBRUGGE, R. R., & RAKERD, B. Talker-independent information for vowel identity. *Journal of the Acoustical Society of America*, 1980, **67**, S1, S-28. (Abstract)

VERBRUGGE, R. R., STRANGE, W., SHANKWEILER, D. P., & EDMAN, T. R. What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America*, 1976, **60**, 198-212.

NOTE

1. Diehl, McCuster, and Chapman (1981) and Macchi (1980) have shown that, under some task and stimulus conditions, sustained uncoarticulated vowels may be perceived as well as are vowels coarticulated in CVC syllables. However, in neither study were natural (as opposed to synthesized) isolated vowels found to be perceived better than coarticulated vowels, as would be predicted from target accounts of vowel identity.

APPENDIX
Duration of Syllable Components (in Milliseconds)

Vowel	Total Syllable	Initial (15%)*	Center (50%-65%)	Final (35%-20%)**
		Short		
I	119	18	60	41
ʊ	124	19	62	43
ʌ	132	20	66	46
ɛ	150†	22	75	53†
		Mid		
i	150	22	92	36
u	150	23	90	37
		Long		
æ	175	26	114	35
ɔ	177	27	115	35
a	184	28	120	36
Average	151	23	88	40

**Excludes prevoicing. **Excludes closure. †This exemplar of the vowel /ɛ/ is longer than would be expected on the basis of our prior data or those of Lehiste and Peterson (1961). Thus, the final component is also longer than expected.*

(Manuscript received May 12, 1982;
revision accepted for publication July 22, 1983.)