Dynamic specification of coarticulated vowels

Winifred Strange, James J. Jenkins, and Thomas L. Johnson
Center for Research in Human Learning, University of Minnesota, 75 East River Road, Minneapolis, Minnesota 55455

(Received 21 June 1982; accepted for publication 1 June 1983)

An adequate theory of vowel perception must account for perceptual constancy over variations in the acoustic structure of coarticulated vowels contributed by speakers, speaking rate, and consonantal context. We modified recorded consonant–vowel–consonant syllables electronically to investigate the perceptual efficacy of three types of acoustic information for vowel identification: (1) static spectral "targets," (2) duration of syllabic nuclei, and (3) formant transitions into and out of the vowel nucleus. Vowels in /b/–vowel–/b/ syllables spoken by one adult male (experiment 1) and by two females and two males (experiment 2) served as the corpus, and seven modified syllable conditions were generated in which different parts of the digitized waveforms of the syllables were deleted and the temporal relationships of the remaining parts were manipulated. Results of identification tests by untrained listeners indicated that dynamic spectral information, contained in initial and final transitions taken together, was sufficient for accurate identification of vowels even when vowel nuclei were attenuated to silence. Furthermore, the dynamic spectral information appeared to be efficacious even when durational parameters specifying intrinsic vowel length were eliminated.

PACS numbers: 43.70.Dn, 43.70.Ve

INTRODUCTION

An adequate theory of speech perception must explain how a listener recovers the phonetic segments from the acoustic signal produced by the speaker's articulatory acts. A central goal in understanding this process is to describe the correspondence between parameters of the acoustic signal and phonetic units, that is, to specify the acoustic information that supports the perception of phonetic segments. Several decades of research have shown us that the correspondence is not a simple one-to-one mapping between acoustic features and phonetic features. Many different acoustic patterns give rise to the same phonetic percept; and likewise, the same acoustic pattern may give rise to the perception of different phonetic units, depending upon its relation to the surrounding acoustic context. In other words, speech perception is an instance of the perceptual constancy problem.

The research reported here addresses the problem of the correspondence between the acoustic signal and the phonetic percept for a major class of English phonemes, the vowels. Vowels have traditionally been differentiated in articulatory terms by the static vocal tract shapes attained by positioning the tongue, jaw, and lips in different configurations. These characteristic vocal tract shapes are often referred to as articulatory "targets." The acoustic patterns which are the consequences of these articulatory targets are described in terms of their static spectral characteristics, and are often referred to as acoustic targets. The center frequencies of the first two or three oral speech formants differentiate English vowels when they are spoken as sustained, isolated tokens by a single speaker. Vowels are thus conceived of as points in an acoustic vowel space in which the coordinates are the frequencies of the first and second formants. Multiple tokens of a particular vowel type, spoken by a single speaker as isolated (uncoarticulated) phones, fall into a small region in vowel space, well differentiated from the regions which circumscribe each other vowel type.

The variability in the acoustic patterns for a particular perceived vowel derives from several sources. First, because formant frequencies are a function of the overall size and shape of the supralaryngeal vocal tract, vowels spoken by different speakers vary acoustically in complex ways. The variability is especially great when comparing vowels spoken by men, women, and children, but there is considerable variability even for vowels produced by speakers of the same sex and age (Peterson and Barney, 1952; Strange *et al.*, 1976). Second, when vowels are coarticulated with consonants, as they nearly always are in continuous speech, the spectral characteristics of the acoustic signal vary such that the acoustic targets found in isolated vowels may not be attained in any spectral cross section taken through the changing acoustic pattern (Stevens and House, 1963). This is often referred to as target "undershoot." Third, vowels coarticulated with consonants in ongoing speech may display different amounts of target undershoot, depending upon speaking rate, sentence and word stress, and the individual style of speech (Lindblom, 1963; Gay, 1978).

The influence of any of these factors may result in a set of acoustic patterns in which the static spectral configurations that are characteristic of isolated vowels are not realized. More importantly, the region in vowel space populated by tokens of a particular vowel type produced in all these contexts will often overlap significantly with regions containing tokens of other vowel types. Acoustic vowel targets are thus ambiguous with respect to perceived vowel identity across variations in speakers, phonetic context, speech rate, and stress.

In the face of this perceptual constancy problem, two general types of theories have been offered to account for the

perception of vowels, which we refer to here as (1) target normalization theories and (2) dynamic specification theories. They differ in their characterizations of the acoustic information that supports vowel perception and in their accounts of how that information is detected and used in the process of recovering the phonetic sequence from the speech signal.

Target normalization theories assume that the essential information for vowel identity is contained in the asymptotic spectral cross section within the syllabic nucleus, which most closely corresponds to the canonical (isolated) vowel targets. However, since these static spectral patterns are inherently ambiguous across speakers and contexts, the veridical perception of vowels requires complicated normalization processes through which the variable acoustic "input" is recoded in some way to arrive at the invariant percept (see Joos, 1948; Ladefoged and Broadbent, 1957; Lieberman *et al.*, 1972; Stevens and House, 1963). More recently, researchers have attempted to differentiate vowels acoustically on the basis of transformations of the target formant frequencies (Gerstman, 1968; Skinner, 1977; Nearey, 1977). However, it remains the case that all these models take as their acoustic "raw data" a single spectral cross section through the acoustic signal (but see Assmann *et al.*, 1982).

An alternative approach, taken in our laboratory and elsewhere, seeks a characterization of vowel perception that refocuses attention on the whole complex of acoustic consequences of articulating vowels in ongoing speech. In this view, vowels are conceived of as characteristic *gestures* having intrinsic timing parameters (Fowler, 1980). These dynamic articulatory events give rise to a dynamic acoustic pattern in which the changing spectro-temporal configuration provides sufficient information for the identification of the phonetic units. Perception is conceived of as the pickup of that information as it is specified over time in the acoustic signals (see Shankweiler *et al.*, 1977).

The research reported here examines the nature of the acoustic information used by listeners in identifying American English vowels in consonant–vowel–consonant (CVC) syllables. Previous research has shown that vowels spoken in CVC syllables are identified quite accurately by phonetically naive listeners, despite the presence of considerable ambiguity in the static acoustic configurations (targets) produced by differences in speakers, rate of speech, and phonetic context (Verbrugge *et al.*, 1976; Strange *et al.*, 1976; Macchi, 1980). Further studies explored possible sources of dynamic information by investigating perception of vowels produced in CVC, CV, and VC syllable contexts (Strange *et al.*, 1979; Gottfried and Strange, 1980). Results indicated the importance of two sources of information: (1) formant transitions into and out of the "vowel nucleus," and (2) temporal parameters which specify intrinsic vowel length.[1]

While these studies point to the importance of temporal and dynamic spectral information for vowel perception, a general problem with their interpretation derives from the fact that the vowels presented in the different syllable contexts were actually different productions. Thus differences in vowel identifiability across these syllabic conditions, taken as evidence for the relative perceptual efficacy of the different acoustic parameters, could have resulted from uncontrolled differences in production.

To circumvent this confounding of perception and production, the present experiments used a different technique to explore the sources of acoustic information that specify vowel identity in a CVC syllable. Starting with syllables produced by adult speakers, digitized waveforms of the syllables were electronically modified in order to delete or alter various spectral and temporal parameters of the acoustic signal while holding others constant. The altered waveforms were then converted back to analog signals and presented to listeners who identified the vowels.

We were particularly interested in examining the relative contributions to vowel identification of three sources of information: (1) the quasi-steady-state formant frequencies of the vowel nucleus that correspond most closely to the canonical acoustic vowel targets, (2) temporal information, including the correlated parameters of length of vocalic nuclei and elapsed time between initial consonant release and final consonant closure, and (3) the formant transitions into and out of the vowel nucleus, which provide what we will refer to as dynamic spectral information. The latter two sources of information cannot be characterized by acoustic parameters available in any single spectral cross section of the syllable, but rather must be described with reference to a temporal interval or a change over time in spectral configuration. As such, they are a function of the dynamic articulatory gestures characteristic of coarticulated vowels.

I. SINGLE-SPEAKER EXPERIMENT

In the first study, CVC syllables spoken "briskly" in citation form by one adult male served as the corpus. Ten English vowels were produced twice each in the consonantal context, /b/–vowel–/b/. The second repetition of each syllable was spoken at a somewhat faster rate, in order to introduce some variability in the acoustic patterns associated with the vowels. Several modified syllable conditions were generated by altering the digitized waveforms of these 20 syllables. Before describing in detail how the stimulus materials were generated (see Sec. IA) the basic technique and rationale are described here.

Each syllable was divided into three components, as shown in Fig. 1: (a) an initial component, which included prevoicing (if present) and the initial transitions, (b) a center component, which encompassed the entire quasi-steady-state vowel nucleus, and (c) a final component, which included the transitions out of the vowel nucleus and the final stop release, if present. These components were defined as proportions of the total syllable extent from initial consonant release to final consonant closure. Thus their absolute durations varied across syllable tokens and types.

Seven modified syllable tests were generated by selecting various combinations of components and altering the temporal relationships among them. (1) Silent-center syllables were generated by attenuating to silence the center component, leaving the initial and final components intact and in their original temporal relationships. (2) Variable (duration) centers were the converse—both initial and final compo-

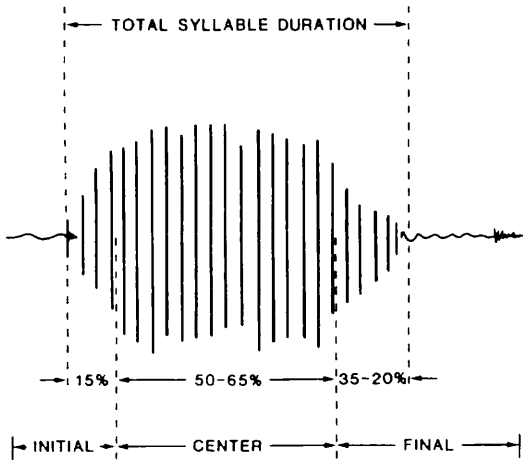

FIG. 1. Schematic representation of the acoustic waveform of a syllable. Each syllable was divided into three components which were proportions of the total syllable length, from initial stop release to final stop closure.

nents were attenuated to silence, leaving the vowel nucleus of each syllable intact. Three additional tests were generated by altering temporal parameters: (3) Fixed (duration) centers were the same as (2) except that all 20 tokens were "trimmed" to be the same length, that of the shortest stimulus of condition (2). (4) Shortened silent-center syllables were the same as (1) except elapsed-time differences were neutralized by substituting the shortest original silent interval between all 20 initial and final components. (5) Lengthened silent-center syllables were generated similarly, by substituting the longest original interval. Two final modified syllable tests consisted of (6) initial components alone, and (7) final components alone. In addition, a control syllables test contained the unaltered syllables. Figure 2(a) and (b) illustrates examples of stimuli in modified conditions 1 through 5.

Let us consider the kinds of information for vowel identity available in each of these sets of stimuli. The original syllables (control condition) included all sources of information: vowel nuclei containing static target information; initial and final transitions, which carry dynamic spectral information about the initiation and completion of the vowel gesture; differences in vocalic duration and elapsed time between initial consonant release and final consonant closure, which are informative of intrinsic vowel length. The silent-center syllables contained two of these sources of information: dynamic spectral information about the entire vowel gesture, available in the consonant transitions (taken together), and temporal information given by elapsed-time differences. The information missing from these stimuli were the static vowel targets. These syllables can be thought of as "vowel-less" from the standpoint of traditional definitions of vowels (see Jenkins *et al.*, in press).

In contrast, the variable center stimuli contained the vowel targets and also information about intrinsic vowel length, specified by vocalic duration differences. Temporal information was minimized in the fixed centers by equalizing vocalic duration, and in the shortened and lengthened

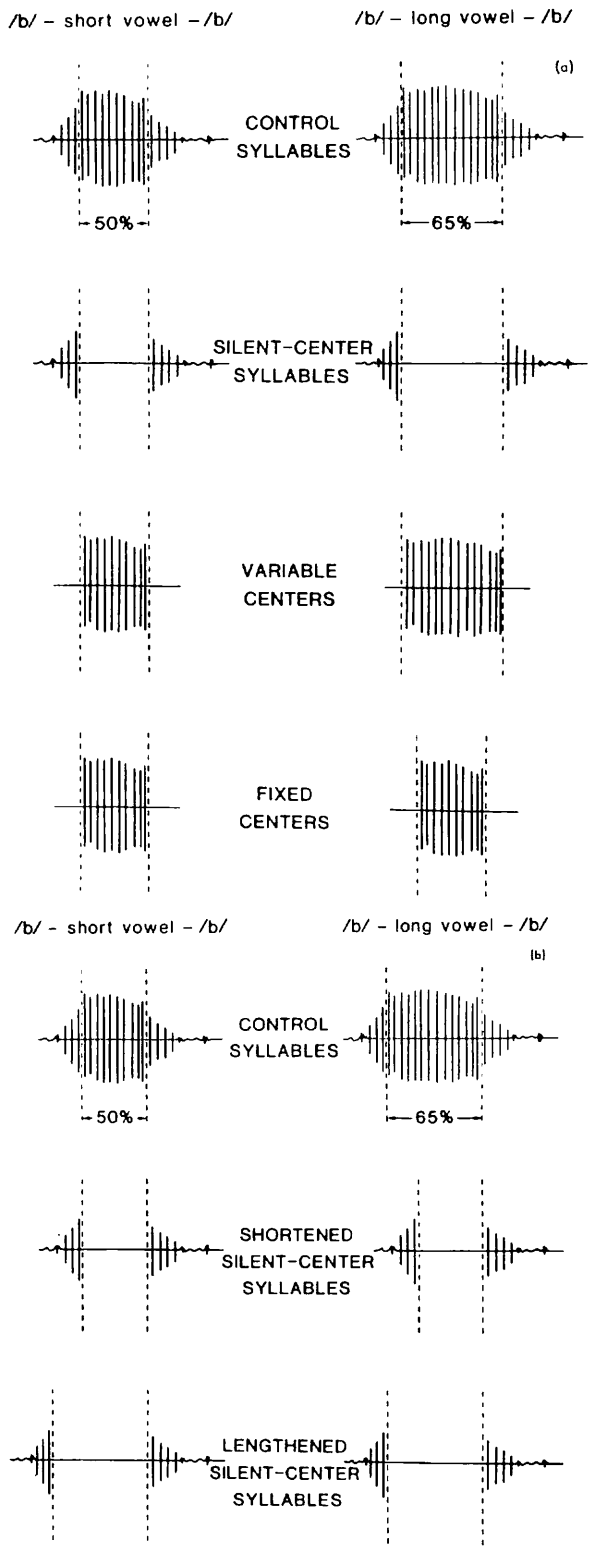

FIG. 2. (a) Schematic representations of the acoustic waveforms of control syllables (top row), silent-center syllables (second row), variable centers stimuli (third row), and fixed centers stimuli (bottom row). (b) Schematic representations of the acoustic waveforms of control syllables (top row), shortened silent-center syllables (middle row), and lengthened silent-center syllables (bottom row).

silent-center syllables by essentially equalizing elapsed time between initial consonant release and final consonant closure. Fixed center stimuli thus contained static target information, but essentially no dynamic spectral or temporal information. Stimuli in the shortened and lengthened silent-center syllables conditions contained dynamic spectral information, but neither vowel targets nor durational information about intrinsic vowel length. Likewise, the initial components and final components each contained dynamic spectral information, but only about the initiation or completion of the vowel gesture, respectively. (Previous research indicated that neither component alone was sufficient to specify vowel identity adequately. Thus these conditions were included as controls for the three silent-center conditions.)

According to dynamic specification theory, the identification of coarticulated vowels is accomplished on the basis of information specified over the temporal course of (at least) the syllable-length utterance. Following this model, we would predict that modified stimuli which retained dynamic spectral and temporal sources of information would yield relatively accurate vowel identification, whereas stimuli in which such information was minimized would yield relatively poor identification. Specifically, we hypothesized that the silent-center syllables, in which both dynamic spectral information (about the entire vowel gesture) and temporal information about intrinsic vowel length were available, would fare best of all the modified syllable conditions. Indeed, if dynamic information is sufficient to specify the vowel gesture, we would expect vowels in silent-center syllables to be identified unambiguously, despite the absence of the vowel targets. Target normalization theories, alternatively, would predict a significant decrement in vowel identification for silent-center syllables, since the vowel targets, thought to be the primary "cue" to vowel identity, are not physically instantiated in these stimuli.

The relative contribution of dynamic spectral and temporal information can be assessed by comparing the three silent-center syllable conditions. Vowels in CVC syllables may be perceived accurately primarily because the consonants provide perceptually salient temporal markers specifying intrinsic vowel length. If this were the case, then we would expect significant decrements in identification accuracy for both the shortened and lengthened silent-center syllables, since neither vocalic duration differences nor elapsed time differences were retained in these stimuli. If, however, the transitional components of the acoustic pattern provide perceptually relevant information about the timing of articulatory gestures, vowel identifiability in these two conditions might remain quite good, despite the absence of (ordinarily) correlated durational parameters. Furthermore, if neither the initials nor finals components alone produced accurate vowel identification relative to these silent-center syllables, we could conclude that the dynamic spectral information specifying the vowel was abstract, i.e., defined as a relation over both initial and final transitional parts of the CVC syllable.

Finally, the relative importance of static targets versus vocalic duration information can be assessed by comparing performance on variable and fixed center stimuli. If, as i often assumed, the targets are the primary cues for vowe identity, we might expect vowel identification in the fixec centers condition to be no worse than in the variable center condition, in which relative duration differences (and some formant movement) were present. (This constitutes a strong target theory prediction which few researchers would advocate today. However, studies in which single cross sections of syllables are taken as the only input for normalization algorithms imply such a strong position.) Alternatively, dynamic specification theory would predict a decrement in performance in the fixed centers condition, because information about the vowel gesture as coarticulated with the consonants is minimized in this condition.

A. Method

1. Stimulus materials

Twenty /b/–vowel–/b/ syllables containing the vowels /i,ɪ,e,ɛ,æ,ɑ,ʌ,o,ʊ,u/ were spoken by an adult male. Two repetitions of the ten syllables, spoken at slightly different speaking rates, were recorded with a Revox A77 tape recorder and a Spherodyne microphone. The 20 syllables were low-pass filtered (3860-Hz cutoff) and digitized at a 10-kHz sampling rate, using the Haskins Laboratories PCM system.

From visual displays of the waveforms, the total duration from the initial consonant release to the final consonant closure (end of high-frequency energy) was determined for each syllable. Durations ranged from 114 to 202 ms with a mean of 167 ms. Each syllable was then divided into three proportional components in such a way that all the quasi-steady-state vowel was encompassed in the center component (Lehiste and Peterson, 1961). The duration of the initial component of each syllable was defined as the first 15% of the total duration (plus any prevoicing, if present). It contained from three to five pitch periods after the release and was from 22- to 30-ms long, not counting prevoicing.

The proportional duration of the center components varied for different vowel types: for intrinsically short vowels, /ɪ,ɛ,ʌ,ʊ/, 50% of the total duration (following the initial component) was designated the center component. For the intermediate vowels, /i,u/, 60% of the total duration was so designated, and for the intrinsically long vowels, /e,æ,ɑ,o/, the center component was 65% of the total duration. The number of pitch periods in the 20 centers varied from six to 15 and durations ranged from 57 to 127 ms.

The final component of each syllable was the remaining 20% (long vowels), 25% (intermediate vowels), or 35% of the total syllable duration, plus the final consonant release, if present. It contained from four to six pitch periods and was from 33- to 42-ms long. In order to minimize transients produced by abrupt onsets and offsets, all cuts in the waveforms were made at the zero crossing closest to the point determined by the above definitions. However, the integrity of pitch periods was not always preserved.

Acoustical analysis performed after stimulus construction confirmed that none of the quasi-steady-state vowel nucleus remained in the initial or final components, with the exception of the final component of one token of /bʊb/. For

all other tokens, one or more of the first three formants were still in transition at the end of the initial and beginning of the final components. That is, formants had not attained their asymptotic frequencies within either component. Many syllables were characterized by formant movement throughout their entire extents. In addition to these spectral criteria, amplitude envelopes showed that all peak amplitude pitch pulses were included within the center components. For initial and final components, amplitude envelopes were rising and falling, respectively.

Silent-center (SC) syllables were produced by attenuating to silence the center component of each syllable, leaving the initial and final components intact and in the appropriate temporal relationship. Thus these syllables each contained a (noticeable) silent interval of from 57- to 127-ms long. Shortened silent-center (ShSC) syllables were constructed by positioning the initial and final components of each of the SC syllables such that the silent interval between them was 57 ms for all 20 syllables. Lengthened silent-center (LoSC) syllables were made by separating the initial and final components of each of the 20 syllables by a 163-ms interval.[2]

Variable (V) centers were constructed by attenuating to silence both initial and final components. Fixed (F) center stimuli were generated by attenuating to silence equal portions from the beginning and end of the V centers such that each stimulus was about 58 ms in duration, and included from five to seven pitch periods.[3] Finally, the Initial (I) stimuli and the final (F) stimuli were generated by attenuating to silence the center and final components and the center and initial components, respectively.

Eight separate test conditions were constructed by randomly arranging four repetitions of each of the 20 appropriate stimuli in an identification test, with a 4-s interstimulus interval, and an 8-s interval between each block of ten stimuli. The control condition consisted of the 20 unmodified syllables each appearing four times in the test order. The SC syllables condition included four repetitions of each of the 20 SC syllables, and so on. Digital waveforms were reconverted to analog signals, low-pass filtered, and recorded on audio tape with a Crown SX tape recorder for playback to subjects.

2. Procedure

Subjects were randomly assigned to the eight stimulus conditions and tested in small groups in a quiet room. Test tapes were presented via a Revox A77 tape recorder, MacIntosh MV49 amplifier, and AR acoustic suspension loudspeaker at a comfortable listening level. Subjects responded by circling key words on a response form. For the V centers and F centers conditions, the response alternatives were key words beginning with the vowel sound: eat, it, ate, Ed, at, odd, up, oat, (h)ook, ooze. For the remaining six conditions, the key words were: beeb, bib, babe, beb, bab, bob, bub, bobe, buub (should), boob.[4] Prior to testing, all subjects were given a familiarization sequence with the task and response forms, using a subset of the control syllables as stimuli. Subjects practiced on the response form on which they would be tested and feedback was given. Following this, each group of subjects was presented 20 stimuli from the test condition in which they were participating, but no feedback was given. The subjects in the seven modified syllables conditions were told that the syllables had been modified electronically and that they were to try to identify the vowel that had been spoken in the original syllable.

After completing the 80-item test of their assigned experimental condition, subjects in all eight conditions completed a second 80-item test in which the control stimuli were presented. (For the subjects assigned to the control condition, this was a retest on the same materials.) No feedback was given for either test. Data from both tests of any subject who had an error rate greater than 20% on the second (control syllables) test were discarded on the grounds that these listeners could not reliably identify the speaker's vowels even when full acoustic information was available.

3. Subjects

A total of 159 subjects were tested. Data from seven subjects were discarded on the basis of their performance on the second test (no more than three from any one group). Nineteen subjects remained in each of the eight stimulus conditions. All were native speakers of American English and reported no hearing loss. Almost all subjects were natives of the upper Midwest area. They were recruited from introductory psychology courses at the University of Minnesota and had received no training in phonetics.

B. Results and discussion

Data from the first test only for the 19 subjects in each group were included in the analyses presented here. Perceptual performance was first analyzed by comparing the mean number of overall errors in vowel identification in each of the eight stimulus conditions. An error was defined as a vowel response other than the one intended by the speaker in the original production or the omission of a response (the latter occurred only rarely). Figure 3 presents the overall error rates for the eight conditions, expressed as a percentage of total opportunities.

It is readily apparent that performance varied markedly across the eight conditions. A one-way analysis of variance showed the overall difference in mean errors between groups to be highly significant, $F(7,144) = 73.40, p < 0.001$. *Post hoc* comparisons were performed, using a Tukey test of honestly significant differences ($p = 0.05$).

Of primary interest is the finding that vowels in the SC syllables were identified relatively accurately (only 6% errors), despite the fact that the vowel nuclei were missing from the signals. Indeed, identification of the vowels in these "vowel-less" syllables was not significantly worse than identification of the unmodified control syllables. This supports our main hypothesis that dynamic sources of information are sufficient for highly accurate identification of coarticulated vowels (see also Jenkins *et al.*, in press).

Performance in the initials and finals conditions was significantly worse than for any of the other conditions. These extremely high error rates corroborated our expectations that sufficient information for vowel identification was not "contained within" either of the components taken by

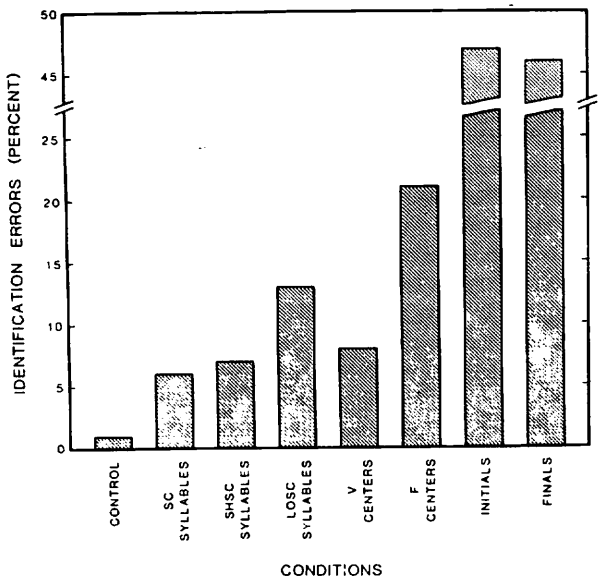

FIG. 3. Average identification errors (expressed as percentages of opportunities) for each stimulus condition in the single-speaker experiment.

itself. We can conclude, then, that the information used by subjects in identifying the vowels in the SC syllables was dynamic (and abstract) in that it was specified as a relational function over the two acoustic components of the stimulus taken as a whole. Indeed, to both subjects and experimenters, the stimuli sounded like single CVC syllables with a "hiccup" or glottal stop in the middle of each vowel.

A comparison of the three SC syllable conditions allows us to separate the relative contribution of temporal information about intrinsic vowel length and dynamic spectral information available in the transitional components of the CVC syllables. If vowel duration, specified in these stimuli by differences in elapsed time from consonantal release to consonant closure, were the primary source of information about the vowel gesture used to differentiate the vowels in the SC syllables, then we would expect error rates on the ShSC and LoSC syllables to be significantly higher, since elapsed time differences among the vowels were essentially neutralized in these conditions. However, Tukey tests showed no significant overall differences among the three conditions, nor did a more lenient test of pairwise comparisons (least significant differences, Keppel, 1973). In fact, errors on ShSC syllables were not significantly greater than errors on control syllables, although vowels in LoSC syllables were misidentified significantly more often than in the control syllables by a Tukey test ($p = 0.05$).

While the overall error rates for the three SC syllable conditions were not different, the pattern of errors shown in a vowel-by-vowel analysis varied somewhat across conditions. Table I gives the percentages of errors for intrinsically short, mid, and long vowels in these three conditions. To the extent that durational information is important for accurate vowel identification, we expected that long vowels in the ShSC syllables condition, and short vowels in the LoSC syllables condition would be misidentified more often than in

TABLE I. Identification errors (in percent) on short, mid, and long vowels in silent-center syllables, shortened silent-center syllables, and lengthened silent-center syllables conditions: single-speaker experiment.

	Vowels			
Condition	Short /ɪ, ɛ, ʌ, ʊ/	Mid /i, u/	Long /e, æ, ɑ, o/	Overall vowels
SC syllables	9%	3%	5%	6%
ShSC syllables	9%	3%	8%	7%
LoSC syllables	21%	6%	9%	13%

the SC syllables condition. As the percentages indicate, these trends occurred in the predicted direction. However, the effect was quite small for long vowels in the ShSC syllable condition; a numerical increase in errors occurred for only two of the four long vowels. The overall increase in errors on short vowels in the LoSC syllables condition was more substantial; three of the four vowels showed a numerical increase in errors over the SC syllables condition. However, two of the long vowels, /e/ and /o/, also showed more errors in the LoSC syllables condition, relative to the SC syllables. Thus it might be that part of the increase in errors in this condition was due to factors other than the neutralization of durational information for vowel length.

One possible reason for the increase in errors on LoSC syllables is that the two components of the LoSC syllables were so spread in time that the integrity of the syllable as a single unit may have been jeopardized. Recall that the silent interval used in all these syllables was even longer than the longest silent interval in the original SC syllables. Thus the average change in temporal extent from the original was more extreme in the LoSC syllables condition than it was in the ShSC syllables condition. This might account, at least in part, for the asymmetry of perceptual results.

The relatively accurate identification of vowels in all three SC syllable conditions is a dramatic result when one considers traditional accounts of vowel perception. Neither the "primary" cue to vowel identity—the targets—nor the "secondary" cue of relative duration (in terms of elapsed time) was present in the ShSC and LoSC syllables, and yet vowel perception remained quite accurate. This finding supports the contention that the rapidly changing acoustic patterns at the beginning and end of a CVC syllable provide important information for vowel identity, independent of their (usually correlated) role of providing temporal information about vocalic duration or elapsed time between consonant gestures.

In order to assess the relative contribution of temporal information for vowel identity when transitional information is not present, or is attenuated, performance on the V centers and F centers conditions was compared. Tukey tests indicated that there were significantly more vowel identification errors in the F centers condition than in the V centers condition. (The latter was not significantly different from the controls.) Relative duration differences between short and long vowels were actually enhanced in the V centers condition because of the way in which these components were defined. (Recall that a larger proportion of the syllable was

taken as the centers for long vowels than for short vowels.) The results suggest that perceivers utilized this enhanced relative duration information to disambiguate spectrally similar vowels.

Acoustical analysis indicated that there was significant movement of formants within the V centers. This was especially noticeable in the case of the /e/ and /o/, and reflects the fact that these vowels were diphthongized in the dialect of the speaker in this study. Thus formant movement provided another dynamic source of information for vowel identity in the V centers stimuli. In the F centers, formant movement was minimal, leaving only information about relatively static spectral targets. In the case of the /e/ and /o/, the F centers encompassed the primary target, but adequate information about diphthongal movement was probably not available within the 57-ms portion.

An inspection of the errors on the short, mid, and long vowels in these two conditions, shown in Table II, indicates that the major source of increased errors on F centers was confusion between spectrally similar short–long vowel pairs. As expected, the long vowels were misidentified as their short counterparts; error rates on /e/ and /o/ were especially high.

In general, the pattern of perceptual results reported here offers strong support for the view that vowels in CVC syllables are specified by dynamic spectral and temporal acoustic parameters. Of the three kinds of information under investigation, we found that both temporal parameters specifying intrinsic vowel length and dynamic spectral information carried in the transitional components of the acoustic signal contributed significantly to the identification of the vowels. When initial and final transitional components were both present, vowel identification was relatively accurate even when elapsed time differences were neutralized. On the other hand, vocalic duration differences appeared to be quite important for accurate perception of vowels when transitional information was not available. Static spectral targets, present in the middle of the syllable, provided relatively impoverished cues to vowel identity, although identification accuracy was still well above chance (see Assmann *et al.*, 1982, for a similar pattern of results with phonetically trained listeners). We can conclude that information for the vowel as a gesture is spread throughout the changing acoustic pattern of the syllable. Portions of the acoustic pattern characterized by relatively rapid spectral change at the beginning and end of the syllable, taken together, appear to provide especially good information about the intended vowels.

TABLE II. Identification errors (in percent) on short, mid, and long vowels in variable centers and fixed centers conditions: single-speaker experiment.

	Vowels			
Condition	Short /ɪ, ɛ, ʌ, ʊ/	Mid /i, u/	Long /e, æ, ɑ, o/	Overall vowels
Variable centers	10%	4%	9%	8%
Fixed centers	12%	3%	39%	21%

II. MULTIPLE-SPEAKER EXPERIMENT

In the above experiment, the original corpus consisted of CVC syllables produced by a single speaker. Although some variability was introduced by a speaking-rate change, the acoustic patterns specifying the vowels did not reflect other sources of variability discussed in the Introduction. In order to provide a more stringent test of the claim that dynamic information specifies vowel identity and to extend the generality of the findings of the first experiment, we replicated the study, using a corpus which included syllables produced by four different speakers, two men and two women.

A. Method

1. Stimulus materials

The 20 tokens used in experiment I were also included in this study. In addition, two repetitions of each of the ten vowels spoken in /b/–/b/ syllables were produced by each of three additional speakers. The speakers were instructed to recite the syllables briskly; the second repetition was spoken more rapidly than the first. The second male speaker was a long-time resident of Minnesota who originally came from St. Louis, Missouri. One of the female speakers was a native of metropolitan Minnesota. The other was originally from Northern California and had resided in Minnesota for seven years at the time of recording. The dialect of all but the last speaker was similar to that spoken by the majority of subjects serving as listeners. The dialect of the second female was somewhat different and has been characterized by trained phoneticians as a variant of Southern Midland.

The 60 new syllables were filtered, digitized, and divided into initial, center, and final components, using the same procedures as described in experiment I. Average total duration of syllables for the four speakers varied from 167 ms for the original male speaker to 196 ms for one of the female speakers; the range in durations of individual tokens was from 114 to 251 ms. Center components ranged from 57 to 163 ms, with an average of 104 ms. Initial components were from 17- to 37-ms long, with an average duration of 27 ms. Final components were from 31 to 62 ms in duration with an average of 46 ms.

Seven modified syllable conditions and a control condition were constructed in the same way as for experiment I. SC syllables included initial and final components in their original temporal relationship. All 80 ShSC syllables had a silent interval of 57 ms; all 80 LoSC syllables contained a 163-ms silent interval. *F* centers were all about 57 ms in length and contained from six to eight pitch periods for the male tokens and from nine to 13 pitch periods for the female tokens.

Eight separate listening tests were constructed by randomly arranging the 80 appropriate syllables, converting digital waveforms to analog signals, filtering, and recording them on audio tape with a 4-s interstimulus interval and an 8-s interval between blocks of ten stimuli.

2. Procedure

Subjects were tested using the same procedures as described above. Task and response form familiarization was accomplished using control stimuli spoken by a single male speaker, after which subjects heard 20 tokens of the modified condition in which they were to be tested, including some tokens produced by each of the four speakers. As in experiment I, all eight groups of subjects were tested on control syllables after completing the experimental condition and performance on this test was used to discard subjects with error rates exceeding 20%.

3. Subjects

A total of 158 subjects were tested; data from six subjects were discarded because of high errors on control syllables on the second test (one in each of six conditions). Thus 19 subjects remained in each of the eight listening conditions. Subjects were native English speaking volunteers from introductory psychology classes, almost all of whom were from the Upper Midwest. They reported no hearing losses, and no expertise in phonetics.

B. Results and discussion

Overall results of perceptual tests for the eight stimulus conditions are presented in Fig. 4. Errors averaged over all 80 tokens in each test are given as percentages of opportunities. As in experiment I, there were marked differences in identification accuracy across the eight conditions. An analysis of variance showed that mean differences between groups were highly significant, $F(7,144) = 172.39, p < 0.001$.

As the figure shows, errors in the control condition were extremely low (5%) despite the presence of considerable acoustic variability contributed by speaker and speaking rate differences. All four speakers' tokens were accurately perceived even though the syllables were randomly arranged and no information about the identity of the speaker was available prior to each test syllable. Identification of the first male speaker's tokens was as accurate in this study as in experiment I. These results corroborate earlier findings (Verbrugge *et al.*, 1976; Strange *et al.*, 1979; Gottfried and Strange, 1980; see also Macchi, 1980; Diehl *et al.*, 1981) and support the claim that vowels can be unambiguously specified within CVC syllables despite speaker-contributed acoustic variations.

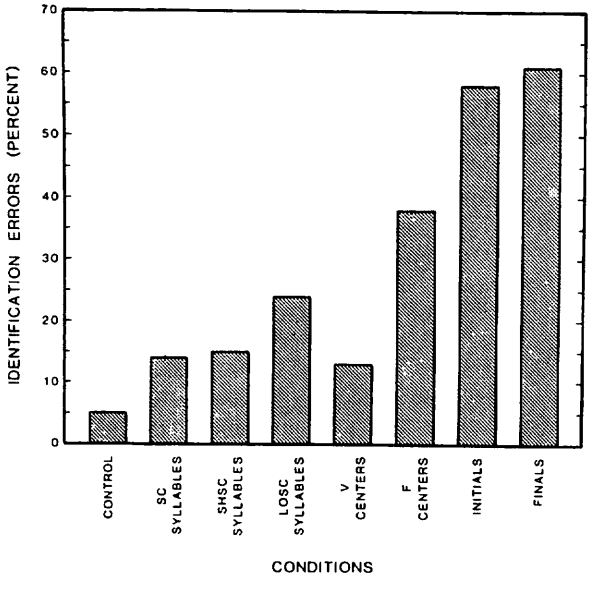

FIG. 4. Average identification errors (expressed as percentages of opportunities) for each stimulus condition in the multiple-speaker experiment.

Tukey tests of honestly significant differences ($p = 0.05$) indicated that error rates for all seven modified conditions were significantly greater than in the control condition. However, some modified syllable conditions yielded fewer errors than others. As in experiment I, vowels in SC syllables were identified relatively well, despite the fact that the vowel nuclei were not present in the signals. Subjects made an average of 11 errors in 80 trials (14%). These errors were not equally distributed across vowel types or speakers. Tokens of the second female speaker, whose dialect varied most from that of the majority of listeners, were misidentified more often (27%) than those of the other three speakers (10%). Vowels which contributed most to the error rate for the former were the vowels, /ɛ,æ,ɑ,ʌ,ʊ/. Thus while listeners were able to identify vowels of a different dialect with equal accuracy when the unmodified CVC syllables were presented (6% errors), they appeared to have more difficulty in identifying these variants when the vowel nuclei were removed.

Error rates for the initial and final conditions were far greater than for all other conditions, and not significantly different from each other. Again, this shows that the intended vowels very often could not be identified on the basis of information contained within either of these components presented alone. The relatively accurate identification of the SC syllables was dependent on information specified over both components as an integrated stimulus.

A comparison of the SC, ShSC, and LoSC syllable conditions showed a pattern of results similar to that found in experiment I. Tukey tests indicated that, while the ShSC syllables were not identified with significantly less accuracy than the SC syllables, the LoSC syllables yielded a significantly higher error rate ($p = 0.05$). Again, this shows an asymmetry in the perceptual consequences of neutralizing elapsed time information for vowel length.

Table III shows the pattern of errors on intrinsically short, mid, and long vowels in these three SC syllables conditions. As expected, errors on long vowels in the ShSC condition and short vowels in the LoSC condition increased relative to the original SC syllables. However, as in experiment I, the effect was minimal for the ShSC syllables. All four short vowels in the LoSC condition were misidentified considerably more often relative to the original SC syllables. But again, as in experiment I, errors also increased for some mid and long vowels, suggesting that factors other than neutralization of elapsed time information were affecting the perceptual results. Again, these findings support the conclusion that dynamic information about vowel gestures, specified

TABLE III. Identification errors (in percent) on short, mid, and long vowels in silent-center syllables, shortened silent-center syllables, and lengthened silent-center syllables conditions: multiple-speaker experiment. Data in parentheses give error rates, excluding tokens by the one female speaker whose dialect was different from the listeners'.

	Vowels			
Condition	Short /ɪ, ɛ, ʌ, ʊ/	Mid /i, u/	Long /e, æ, ɑ, o/	Overall vowels
SC syllables	23 (16)%	3 (4)%	11 (6)%	14%
ShSC syllables	19 (18)%	5 (5)%	16 (10)%	15%
LoSC syllables	42 (39)%	8 (9)%	15 (12)%	24%

over the transitional portions of the syllable (taken as an integral unit), contributes to the accurate identification of vowels even when temporal parameters which usually accompany intrinsic vowel length are neutralized.

The V and F centers conditions were compared as in experiment I to assess the relative efficacy of vocalic duration and target information for vowel identification. The V centers were identified quite accurately overall, whereas vowels in the F centers condition were misidentified significantly more often. The increase in errors occurred for all ten vowel types and ranged from an increase of 6% to 63%. Table IV presents the error rates for intrinsically short, mid, and long vowels. As expected, the errors for the long vowels were especially great in the F centers condition. However, errors on short and mid vowels were also greater than in the V centers condition. This was probably due to the increased variability of vowel targets contributed by speaker and speech rate differences. However, duration information was apparently less affected by differences in speakers and speaking rate. The relatively accurate identification of V centers suggests that this temporal source of information was useful in the disambiguation of vowels despite the variation in absolute duration contributed by differences in speaking rate within and across subjects. (Again, see Assmann *et al.*, 1982, for similar findings with a multiple-speaker corpus.)

It is interesting to note that performance on tokens contributed by the female speaker whose dialect differed was, in general, not different from that on the other speakers' productions in either the V centers or F centers conditions. That is, her vowel nuclei were identified as well (or as poorly) as those of other speakers. It appears then, that the dialect variation present in this corpus changed the nature of the dynamic information present in transitional components rather more than the static target information.

TABLE IV. Identification errors (in percent) on short, mid, and long vowels in variable centers and fixed centers conditions: multiple-speaker experiment. Data in parentheses give error rates, excluding tokens by the one female speaker whose dialect was different from the listeners'.

	Vowels			
Condition	Short /ɪ, ɛ, ʌ, ʊ/	Mid /i, u/	Long /e, æ, ɑ, o/	Overall vowels
Variable centers	13 (13)%	3 (4)%	17 (18)%	13%
Fixed centers	29 (23)%	13 (15)%	60 (56)%	38%

The general pattern of results found in this study replicated that obtained in experiment I with only minor differences. Vowels were identified relatively accurately in conditions which contained one or more sources of dynamic spectral or temporal information. Vowels in "vowel-less" syllables could be identified with relatively few errors, despite variations in the acoustic patterns contributed by speaker differences, speaking rate differences, and even dialect differences. This was true even when elapsed-time differences were neutralized, as in the ShSC syllables condition. Vowel nuclei were also identified quite accurately when initial and final transitions were deleted, but only when (enhanced) relative duration information was available. Identification of fixed duration vowel nuclei was relatively poor, with error rates for individual vowel types ranging from 11% to 68%. The rather large increase in errors from experiment I to experiment II for this condition (21% to 38%) can be attributed to the increase in ambiguity of static target information due to speaker and speaking rate differences.

III. GENERAL DISCUSSION

In these experiments, our goal was to explore three sources of information available in coarticulated CVC syllables which support the perception of vowels. We used a technique by which utterances produced by one or more speakers were modified electronically in order to delete or alter the acoustic pattern presented to perceivers. This paradigm differs from the one employed in our previous studies in which differences in the production of vowels in different syllabic conditions could have confounded the perceptual results. The procedure used here provides better control over the acoustic signals presented, while still using "natural" speech (as opposed to synthetic speech) in which the acoustic consequences of coarticulation are present. This is important if the goal is to discover what acoustic parameters normally carry information for vowel identity.

Three kinds of information for vowel identity were manipulated in these studies: (1) information provided in the vowel nuclei of the syllables, which corresponds most closely to the static spectral targets thought to be the primary differentiating cues for vowels, (2) information provided by durational differences (either vocalic duration or elapsed time from syllable onset to offset), considered a secondary cue for vowel identity in English, and (3) dynamic spectral information defined over the initial and final transitional portions of the syllables, taken together. We found that the presence of the third kind of information was sufficient to maintain accurate identification of the vowels, even when vowel nuclei were attenuated to silence. Further, this source of information appeared to be relatively independent of duration information. When differences in elapsed time between consonant gestures for intrinsically short, mid, and long vowels were neutralized, perceivers were still able to disambiguate the vowels most of the time on the basis of relational information defined over the initial and final transitions taken together.

What is the nature of this information provided in the rapidly changing patterns at the beginning and end of the syllables? We can tentatively rule out the hypothesis that the initial and final transitions specify formant trajectories, the asymptotes of which correspond to the static vowel targets (Lindblom and Studdert-Kennedy, 1967). Changing the time interval between initial and final components, as in the ShSC and LoSC syllables conditions, would also change the asymptotes specified by these two components. Yet vowel identity was not significantly disrupted by the ShSC modification. The LoSC modification did produce a significant increase in errors in experiment II. However, the pattern of errors suggested that the perceptual problem may have been due, at least in part, to the disruption of the integrity of the syllable as a unitary acoustic and articulatory event.[5]

Returning to the perspective presented in the Introduction, we may speculate that the acoustic patterns given in the initial and final parts of a syllable provide especially useful information about the characteristic gestures which differentiate the vowels. That is, if the articulatory movements (as well as the achieved vocal tract state) are essential defining characteristics of a vowel type, then the perceiver must obtain information from the acoustic pattern about those movements. The results reported here suggest that that information, defined relationally, is available in the portions of the acoustic pattern that correspond to the beginning and end of the vowel gesture.

An inspection of the acoustic patterns of syllables containing different vowels (Lehiste and Peterson, 1961) suggests some possible relational dynamic acoustic parameters that may be perceptually relevant. CVC syllables containing long vowels (sometimes referred to as tense vowels) have formant patterns that are nearly temporally symmetrical about the vocalic nucleus. That is, transitions into and out of the quasi-steady-state nucleus tend to be approximately equal in slope and duration (for a particular consonant). In addition, the proportions of the total syllable length taken up by initial and final transitions are approximately equal. In contrast, syllables containing short (lax) vowels are characterized by asymmetrical initial and final transitions. Transitions out of the vowel into the final consonant are more gradual than transitions into the vowel, and take up a relatively greater proportion of the total syllable length.[6] Lehiste and Peterson conclude the following: "Thus it appears that the characteristic difference between the long and short monophthongs may be described as a difference in the articulatory rate of change associated with the movement from target position to the following consonants. The traditional terminology "lax" and "tense" seems appropriate to label this difference. "Lax" vowels, then, are those vowels whose production involves a short target position and a slow relaxation of the hold; for "tense" vowels the target position is maintained for a longer time, and the (articulatory) movement away from the target position is relatively rapid. The relationship of the three stages to the total duration remain approximately constant, regardless of the fluctuation in duration produced by the following consonant" (1961, pp. 274–275).

In summary, we can say that vowels, as gestures, are differentiated by their timing with respect to adjacent segments and syllables, as well as by the positioning of the tongue during the relatively sustained vocalic portion of the syllable. The perceiver must identify the intended vowels on the basis of information in the acoustic pattern about the *timing* of the gesture as well as the vocal tract state attained. Dynamic spectral parameters such as those described above, as well as differences in vocalic duration and elapsed time from closure to closure are all correlated with (or determined by) articulatory timing constraints. As such, these acoustic parameters may serve as information for the perceiver about the identity of the vowels. The results of the present study indicate that perceivers can utilize these abstract acoustic parameters in identifying vowels even when static vowel targets are completely missing from the signal. To the extent that these relational parameters remain invariant over variations in speaker identity, speaking rate, and consonantal context, they may provide especially good information for vowel identity and account for the perceptual constancy evidenced by perceivers.

ACKNOWLEDGMENTS

We wish to express our thanks to Dr. Alvin Liberman and the staff of Haskins Laboratories for their continued support of our research (partially supported by NICHHD Contract 71-2420) and to our research staff at Minnesota (Thomas Edman, Lenief Heimstead, James Nead, Grant Miller, Christopher Jenkins, Elizabeth Balow, David Pollak, Karen Siegel) for their assistance in this project. We also express our appreciation to Robert Verbrugge, Terrance Nearey, and an anonymous reviewer for their critical comments on the manuscript. Portions of this research were reported at the 95th meeting of the Acoustical Society of America. 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 NICHHD (HD-0098) and NSF (BNS-75-03816). Winifred Strange and James J. Jenkins are now at the University of South Florida. Requests for reprints should be sent to Winifred Strange, Department of Communicology, University of South Florida, Tampa, FL 33620.

[1]Vowel length is not considered phonologically distinctive in English. However, phonetically, stressed English vowels vary redundantly in "intrinsic" vowel duration (sometimes referred to as tenseness), which is specified acoustically by systematic differences in the duration of vocalic nuclei (Peterson and Lehiste, 1960). Traditionally, these temporal parameters have been considered "secondary" cues for vowel identity in English. In our view, phonetic vowel length can be considered a control variable in the specification of timing parameters for coarticulated speech, and might thus have acoustic consequences throughout the syllable.

[2]Because of the variation in absolute durations of initial and final components, the total elapsed time of the ShSC syllables ranged from 113 to 126 ms. For LoSC syllables, elapsed times ranged from 219 to 233 ms. While syllables containing intrinsically short vowels were, on the average, slightly shorter than syllables containing long vowels, the ratios of even the shortest to longest syllables in ShSC and LoSC conditions (0.90 and 0.93, respectively) were far greater than in the SC syllables and control syllables (ratio = 0.56) and probably not perceptually relevant.

[3]We were concerned that, despite cutting at zero crossings, the sudden onsets and offsets of these stimuli might spuriously increase identification errors. Thus a second set of F centers stimuli were constructed in which the first and last pitch periods of each stimulus were attenuated by digital multiplication, thus shaping the amplitude contours of these stimuli. However, perceptual tests indicated a significant increase in errors on these shaped F centers, relative to the unshaped ones, especially for /e/ and /o/. Therefore the data reported below are those obtained on the unshaped F centers stimuli.

[4]Several studies have shown that the type of response forms used, and the compatibility of stimuli and responses, can have significant effects on vowel identification performance (Macchi, 1980; Diehl *et al.*, 1981; Assmann *et al.*, 1982). Pilot studies using both score sheets described here on the V centers and F centers conditions showed a small (4%–8%) but significant advantage in performance with the "eat, it,..." score sheets over the "beeb, bib,..." score sheets. No differences between score sheets were found for control stimuli; we did not redo the other modified syllables conditions. While the differences in performance attributable to score sheets in the V centers and F centers did not significantly affect the pattern of results described below, we report performance on V centers and F centers conditions using the better score sheets. Note that the major comparisons of interest are between groups who used the same score sheets.

[5]More research, including both acoustical analyses and perceptual studies, is needed before ruling out this hypothesis. However, the fact that stimuli which actually contained the asymptotic vowel nuclei were not well perceived argues quite strongly against the trajectory hypothesis in its simple form. In addition, Verbrugge and Rakerd (1980) presented SC syllables in which the initial and final portions were contributed by different speakers (a man and woman, respectively). Vowel identification for these "hybrid" syllables was not significantly worse than identification of vowels of SC syllables in which both initial and final portions were contributed by the same speaker. This provides strong evidence against the formant trajectory hypothesis.

[6]Formant transition rate and duration varies as a function of the place of articulation of the consonant, especially for stops. The data reported by Lehiste and Peterson (1961) included many different consonants and their conclusions about proportionality are independent of variations due to consonant identity.

Assmann, P. F., Nearey, T. M., and Hogan, J. T. (**1982**). "Vowel identification: Orthographic, perceptual, and acoustic aspects," J. Acoust. Soc. Am. **71**, 975–989.

Diehl, R. L., McCusker, S. B., and Chapman, L. S. (**1981**). "Perceiving vowels in isolation and in consonantal context," J. Acoust. Soc. Am. **68**, 239–248.

Fowler, C. A. (**1980**). "Coarticulation and theories of extrinsic timing," J. Phonet. **8**, 113–133.

Gay, T. (**1978**). "Effect of speaking rate on vowel formant movements," J. Acoust. Soc. Am. **63**, 223–230.

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

Gottfried, T. L., and Strange, W. (**1980**). "Identification of coarticulated vowels," J. Acoust. Soc. Am. **68**, 1626–1635.

Jenkins, J. J., Strange, W., and Edman, T. R. (in press). "Identification of vowels in vowel-less syllables," Percept. Psychophys.

Joos, M. A. (**1948**). "Acoustic phonetics," Lang. Suppl. **24**, 1–136.

Keppel, G. (**1973**). *Design and Analysis: A Researcher's Handbook* (Prentice-Hall, Englewood Cliffs, NJ).

Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.

Lehiste, I., and Peterson, G. E. (**1961**). "Transitions, glides, and diphthongs," J. Acoust. Soc. Am. **33**, 268–277.

Lieberman, P., Crelin, E., and Klatt, D. (**1972**). "Phonetic ability and the related anatomy of the newborn, adult human, Neanderthal man and the chimpanzee," Am. Anthropol. **74**, 287–307.

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.

Macchi, M. J. (**1980**). "Identification of vowels spoken in isolation versus vowels spoken in consonantal context," J. Acoust. Soc. Am. **68**, 1636–1642.

Nearey, T. M. (**1977**). "Phonetic feature systems for vowels," PhD. thesis, University of Connecticut, 1977 (reproduced by Indiana University Linguistics Club, 1978).

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of 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. **30**, 693–703.

Shankweiler, D. P., Strange, W., and Verbrugge, R. R. (**1977**). "Speech and the problem of perceptual constancy," in *Perceiving, Acting and Knowing: Toward an Ecological Psychology*, edited by R. E. Shaw and J. Bransford (Erlbaum, Hillsdale, NJ), pp. 315–345.

Skinner, T. E. (**1977**). "Speaker invariant characterizations of vowels, liquids, and glides using relative formant frequencies," J. Acoust. Soc. Am. Suppl. 1 **62**, S5.

Stevens, K. N., and House, A. S. (**1963**). "Perturbation of vowel articulations by consonantal context: An acoustical study," J. Speech Hear. Res. **6**, 111–128.

Strange, W., Edman, T. R., and Jenkins, J. J. (**1979**). "Acoustic and phonological factors in vowel identification," J. Exp. Psychol.: Human Percept. Perform. **5**, 643–656.

Strange, W., Verbrugge, R. R., Shankweiler, D. P., and Edman, T. R. (**1976**). "Consonant environment specifies vowel identity," J. Acoust. Soc. Am. **60**, 213–224.

Verbrugge, R. R., and Rakerd, B. (**1980**). "Talker-independent information for vowel identity," J. Acoust. Soc. Am. Suppl. 1 **67**, S28.

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.