*Perception & Psychophysics*
*1999, 61 (6), 1200-1210*

# Perception of dynamic information for vowels in syllable onsets and offsets

JAMES J. JENKINS and WINIFRED STRANGE
*University of South Florida, Tampa, Florida*

It has been demonstrated using the "silent-center" (SC) syllable paradigm that there is sufficient information in syllable onsets and offsets, *taken together*, to support accurate identification of vowels spoken in both citation-form syllables and syllables spoken in sentence context. Using edited natural speech stimuli, the present study examined the identification of American English vowels when increasing amounts of syllable onsets *alone* or syllable offsets *alone* were presented in their original sentence context. The stimuli were /d/–vowel–/d/ syllables spoken in a short carrier sentence by a male speaker. Listeners attempted to identify the vowels in experimental conditions that differed in the number of pitch periods presented and whether the pitch periods were from syllable onsets or syllable offsets. In general, syllable onsets were more informative than syllable offsets, although neither onsets nor offsets alone specified vowel identity as well as onsets and offsets together (SC syllables). Vowels differed widely in ease of identification; the diphthongized long vowels /e/, /æ/, /o/ were especially difficult to identify from syllable offsets. Identification of vowels as "front" or "back" was accurate, even from short samples of the syllable; however, vowel "height" was quite difficult to determine, again, especially from syllable offsets. The results emphasize the perceptual importance of time-varying acoustic parameters, which are the direct consequence of the articulatory dynamics involved in producing syllables.

For many years it was the "received view" that the primary information for vowel identification was carried in the quasi-steady-state vocalic portions of syllables and that this information could be characterized in terms of static vowel "targets" (i.e., the relative frequencies of the first two or three formants of the vocal tract). Acoustic descriptions of vowels were traditionally given as a set of formant values measured from a single spectral section taken at the steady-state or durational midpoint of the syllable (Joos, 1948; Ladefoged, 1967; Peterson & Barney, 1952). Such a description now seems to be too limited. There have been hints in the literature for almost 50 years that steady-state vowels did not capture the information used by listeners in normal communicative use. Tiffany (1953) showed that vowels of several durations, gated out of sustained vowel productions (eliminating onsets and offsets) were not as readily identifiable as 200-msec isolated vowels that included onsets and offsets, and that these in turn were not as identifiable as 200-msec stimuli produced in a consonant context (/t/ V /p/). Fairbanks and Grubb (1961) found only 74% identifiability of nine sustained vowels produced by trained speakers and gated to produce 300-msec samples. More recently, Hillenbrand and Gayvert (1993) found that synthetic static vowels matched to the values of the Peterson and Barney stimuli were not well identified, although Peterson and Barney's listeners had high rates of identifiability for the original stimuli.

Over the last 20 years there has been increasing concern with the role of the dynamic (time-varying) acoustic information that is supplied when the speaker produces vowels in coarticulation with consonants, as is normally the case in natural speech (Lehiste & Meltzer, 1973; Lindblom & Studdert-Kennedy, 1967; Shankweiler, Strange, & Verbrugge, 1977; Strange, 1989b; Strange, Edman, & Jenkins, 1979; Strange, Verbrugge, Shankweiler, & Edman, 1976). Indeed, the classic normative work of Peterson and Barney (1952) has been replicated and extended by Hillenbrand, Getty, Clark, and Wheeler (1995) with special attention to the analysis of changes in formants over the course of the syllable in order to provide a more adequate acoustic description of American English vowels.

Earlier studies of the extent of vowel duration that is required for accurate identification of the vowels in American English (e.g., Fairbanks & Grubb, 1961; Gray, 1942; Powell & Tosi, 1970; Robinson & Patterson, 1995; Schwartz, 1963; Stevens, 1959; Suen & Beddoes, 1972) must now be viewed in a new light because they specifically excluded the dynamic sources of information for vowel identification. All of these studies employed gated, sustained vowels by trained speakers or synthe-

This research was supported by NIDCD 00323 and NINCDS 22568. Research assistants for the study were Bruce Goshe, Elizabeth Lewis, Linda Katz, and Salvatore Miranda. Sonja Trent and David Thornton are thanked for assistance with the tables and figures. Correspondence concerning this article should be addressed to either author: J. J. Jenkins, University of South Florida, 4202 Fowler Ave., Tampa, FL, 33620 (e-mail: jenkins@luna.cas.usf.edu) or W. Strange, Speech and Hearing Sciences, The Graduate Center, CUNY, (e-mail: strangepin@aol.com).

—*Accepted by previous editor, Myron L. Braunstein*

 

**Table 1**
**Total Duration (in Milliseconds) of Syllable (Including Final Consonant Cluster) and Number of Pitch Periods in the Vocalic Portion of Each Test Syllable**

| | Duration | | No. Pitch Periods | |
|---|---|---|---|---|
| Vowel | Token 1 | Token 2 | Token 1 | Token 2 |
| Short | | | | |
| ɪ | 231 | 245 | 13 | 13 |
| ɛ | 276 | 247 | 16 | 15 |
| ʌ | 284 | 268 | 15 | 16 |
| ʊ | 208 | 223 | 12 | 14 |
| Midlength | | | | |
| i | 287 | 299 | 17 | 17 |
| u | 234 | 287 | 14 | 19 |
| Long | | | | |
| e | 278 | 280 | 19 | 18 |
| æ | 321 | 321 | 20 | 21 |
| a | 291 | 304 | 19 | 20 |
| o | 315 | 305 | 21 | 20 |

sized steady-state vowels. In some cases the task was not simple identification but, rather, learning to make a particular response to each sound fragment. (One study even reported that untrained listeners did not hear the stimuli as speech sounds!) In no case were the stimuli taken from speech in sentence context and in no case was there any concern with the acoustic dynamics of the signal. It is not surprising that the results varied widely from study to study, from speaker to speaker, and from listener to listener, as well as from vowel to vowel, often in an unpredictable fashion.

Evidence from several sources suggests that the rapidly changing acoustic patterns at syllable onset and offset (as well as changes in vocalic nuclei themselves) play an important role in vowel identification in both natural and artificial contexts. For example, it has been demonstrated that listeners can identify the intended vowel in a consonant–vowel–consonant (CVC) syllable even when the vowel nucleus has been attenuated to silence (Jenkins, Strange, & Edman, 1983; Parker & Diehl, 1984; Strange, 1987, 1989a; Strange, Jenkins, & Johnson, 1983). Nearey and Assmann (1986) similarly showed that isolated vowels produced in citation form could be identified relatively accurately when two 30-msec portions of each utterance were available (one from a point about one quarter of the way through the syllable and the other from about one third of the way from the end of the syllable), but only when they were presented in the appropriate order. If reversed, vowel identification was poor. Thus, the importance of information defined over syllable onsets and offsets together was demonstrated. (These studies have been reviewed in more detail in Strange, 1989a, and Jenkins, Strange, & Miranda, 1994).

Strange (1989a, 1989b), Verbrugge and Rakerd (1986), and Fowler (1987) have argued that the dynamic acoustic information in a coarticulated syllable is informative of the articulatory gestures that produced the syllable. Thus, identification of the vowel is not a matter of detecting specific acoustic (or articulatory) *targets*, as the older view supposed, but, rather, of apprehending acoustic changes that specify the style of articulatory change that produced the specific vowel. As one of the steps in further examining the dynamics of vowel perception, in the present study we sought to evaluate the perceptual information available in different amounts of the acoustic signal at the beginning or end of a syllable. This was accomplished with waveform editing techniques in which successively larger portions of syllable onsets, and successively larger portions of syllable offsets, were chosen for presentation as stimuli.

The purpose of the experiment, then, was to compare the accuracy of the identification of vowels as a function of increasing amounts of the test syllables presented. The first aim was to determine how much information was needed in order to achieve a high level of accuracy for each of 10 vowels of American English. The second aim was to examine the pattern of both correct and incorrect responses to determine what types of information about vowel quality (i.e., the "features" of tongue height and tongue position) were available in the onsets and offsets of the syllables.

## METHOD

### Stimulus Materials

The speaker was a young adult male who was a native of Ohio who had resided in Florida for 15 years at the time of the recording. He spoke with no perceptible regional dialect; his normal rate of speech was quite rapid. (This is the same speaker who was employed in Experiment 3 in Strange, 1989a, and in the study of mixed-speaker syllables in Jenkins et al., 1994. More details on the acoustics of his productions in this and other contexts are available in those sources.) The syllables of interest were /d/ V /d/ syllables embedded in the sentence, "I say the word /d/ V /d/ some more." The consonant /d/ was chosen as the syllable context because it affords considerable coarticulatory variation in the formant patterns of vowels. Ten vowels /i/, /ɪ/, /e/, /ɛ/, /æ/, /ɑ/, /ʌ/, /o/, /ʊ/, /u/ were spoken in this context. (The vowel /ɔ/ was omitted because this speaker, and many of the listeners, did not contrast /ɑ/ and /ɔ/ in their speech.) Each sentence was recorded multiple times at 7½ ips with a Revox (A77) two-track tape recorder and Panasonic low-impedance microphone. Two sentences containing each vowel were chosen from this set as the stimulus corpus.

All 20 test sentences (10 vowels × 2 tokens) were low-pass filtered at 4900 Hz and converted to digital waveform files (10-kHz sampling rate, 12-bit resolution) using a PDP-11/34 computer. Duration measurements were made from waveform displays. The average length of the sentences for the speaker was 1.40 sec, ranging from 1.30 to 1.60 sec. Total duration of the target syllables was measured from the release burst of the initial stop consonant to the beginning of the friction associated with the /s/ in "some more." On average, voice onset time (VOT) was 11.6 msec, ranging from 6.6 to 18.2 msec. Table 1 gives the duration of the test syllables in milliseconds and the number of pitch pulses in the vocalic portion of the syllable. Figure 1 shows the values of the first two formants for each vowel (averaged over the two tokens of the vowel) at three locations in the syllable: (1) the third pitch period from the onset, (2) the durational midpoint of the syllable (not including the final consonant closure), and (3) the fourth pitch period from the end of the vocalic portion of the test stimuli. Significant acoustic change of one or both formants within these syllabic nuclei is readily apparent for 7 of the 10 vowels. Formant movement to-

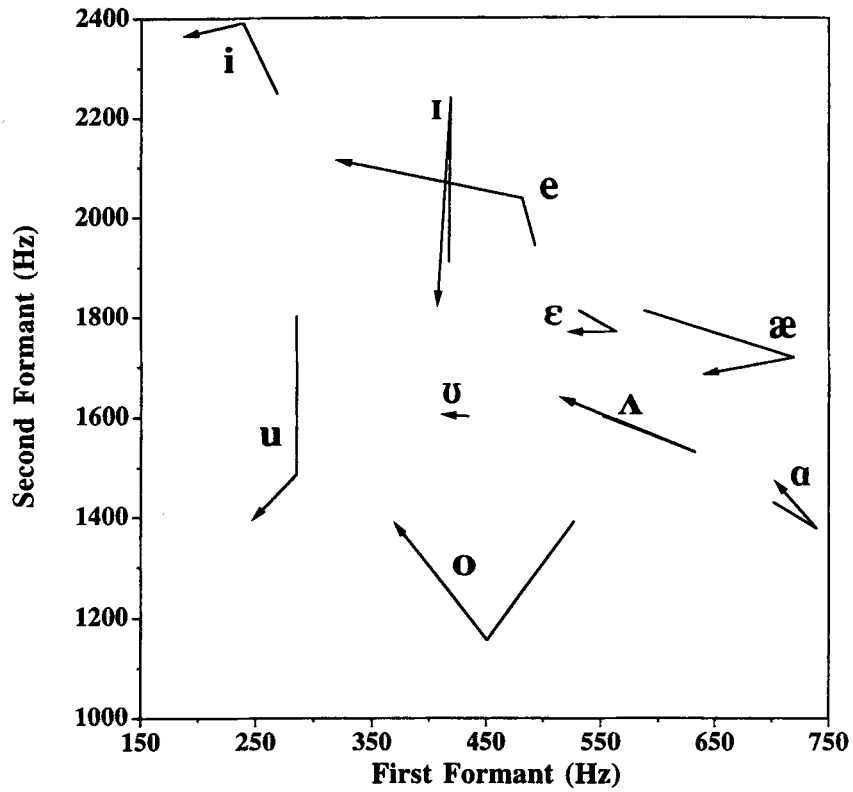


**Figure 1. Formant values (average of two tokens of each vowel) from a point three pitch pulses into the syllable, through the midpoint (50%) of the vocalic interval, to a point four pitch pulses from the end of the vocalic portion (shown by arrowhead).**

ward and away from midpoint formant values includes both (asymmetrical) diphthongal patterns for [eᴵ], [ᵉæ], [ᵘu], [oᵘ], [ⁱi] and (symmetrical) coarticulatory patterns for [ɪ], [ʌ].

Syllables were edited in terms of pitch periods (which averaged 9.5 msec in length). This has the advantage that signals can be cut cleanly at zero crossings without creating noisy transients (clicks and thumps) and without doing injustice to the spectral features of the waveform, which otherwise might be appreciably changed by truncation within a pulse. It must be kept in mind, however, that because natural syllables in English vary in intrinsic duration, the *proportion* of the syllable represented by any particular number of pulses is different, especially for long and short vowels.

Each test syllable was altered electronically to create eight stimulus conditions. Figure 2 illustrates schematically the nature of the editing performed to produce these conditions.

**Silent-center (SC).** Each test syllable was divided into three components: (1) an initial component that included the initial consonant release burst, aspiration (if any), and the first 3 pitch periods of the syllable; (2) a final component that included the last 4 pitch periods prior to final consonant closure plus the closure portion; and (3) a center component that was the portion of the vocalic signal between the initial and final components. The center component ranged in duration from 56 to 134 msec and included from 5 to 12 pitch periods. The SC test utterances were constructed by attenuating to silence the center component of each test syllable, leaving initial and final components (and the carrier sentence) intact. Because duration of the vowel was not of primary concern in this study, the silent interval between onsets and offsets was set equal to the average duration of the center section of all of the vowel tokens. (This condition is a complete replication of one of the earlier *neutral duration silent-center* studies in Strange, 1989a.)

**Initial, one pitch period** (I-1PP). These sentences were constructed by attenuating to silence the remainder of the syllable following the first pitch period. Thus, the test stimuli included the initial burst and voiceless portion plus one pitch period of each syllable. The average length of stimuli (from consonant burst to beginning of silence) was 22 msec (range, 18–28 msec). Here, (as in the case of the SC stimuli), because the focus of the study was on spectral change information and not on intrinsic duration cues for vowel identity, the initial portion of each syllable was followed by a fixed amount of silence between the end of the test stimulus and the words "some more." The silence duration selected was the average duration from the end of the first pitch period of each test syllable to the friction of the words "some more." Because there were slight differences in the durations of the two sets of 10 tokens, the average duration was determined separately for each group of 10 vowels ($M$ = 251 and 259 msec, respectively); the overall average silence duration was 255 msec.

**Initial, three pitch periods** (I-3PP). These sentences were constructed as in the preceding condition except that the remainder of the syllable after the third pitch period was attenuated to silence. The average duration of these initial portions of the syllables was 41 msec (range 36–47 msec). The overall average silence duration following the initial portions was 237 msec.

**Initial, five pitch periods** (I-5PP). For these sentences, the silenced portion began after the fifth pitch period; the average duration of test stimuli was 60 msec (range, 55–66 msec). The average duration of silence was 219 msec.

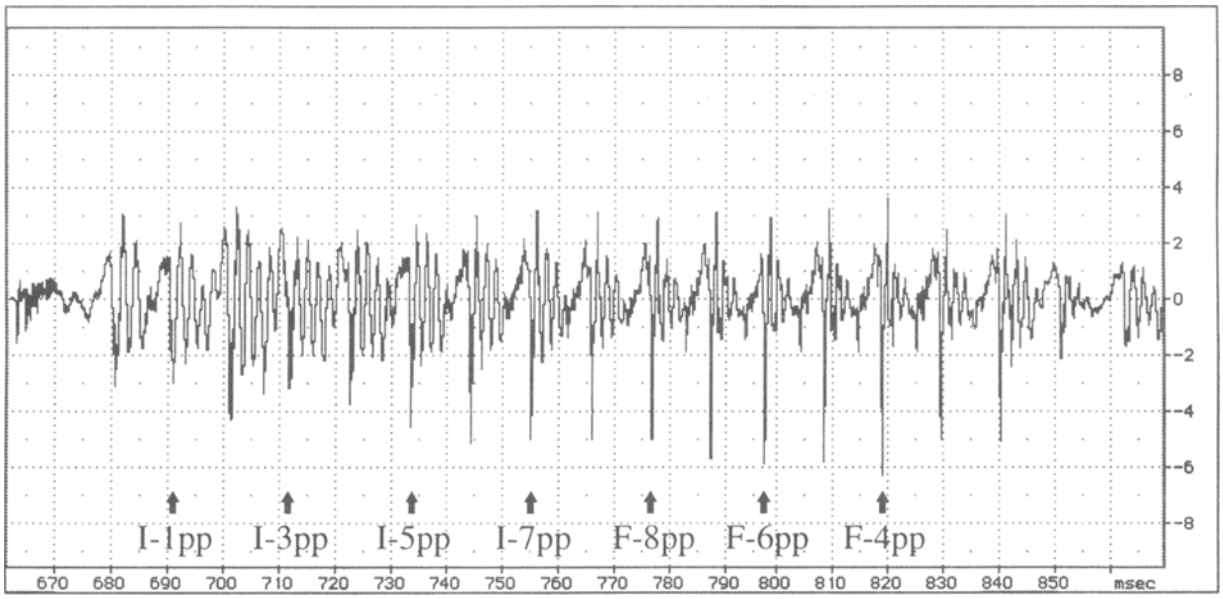


**Figure 2. Sample syllable (one token of /æ/) showing vowel portion and the editing points for each of the syllable conditions used in the experiment. (For some of the short vowels, the initial stimuli and the final stimuli overlapped in the portions of the signal selected.)**

**Initial, seven pitch periods** (I-7PP). For these sentences, the silence began after the seventh pitch period; the average duration of test stimuli was 79 msec (range, 75–85 msec). The average duration of silence was 201 msec.

**Final, four pitch periods** (F-4PP). The poor identification of vowels from syllable offsets alone in previous experiments suggested that a substantial portion of the final segment would be required to achieve accurate vowel identification. Accordingly, more pitch periods were used to evaluate the final conditions. The sentences in the F-4PP condition were constructed by attenuating to silence all of the syllable *preceding* the last four pitch periods. This portion of the syllable was replaced with a fixed duration of silence, as above (144 msec). The average duration of the stimuli was 40 msec (range, 38–44 msec).

**Final, six pitch periods** (F-6PP). These sentences were constructed in the same pattern as above, silencing all but the last six pitch periods. The average duration of stimuli was 59 msec (range, 56–62 msec); preceding silence duration was 126 msec.

**Final, eight pitch periods** (F-8PP). These stimuli were constructed in the same fashion utilizing the last eight pitch periods. Average stimulus duration was 78 msec (range, 75–81 msec); the preceding silence was 108 msec in duration.

Separate listening tests were generated for each of the eight stimulus conditions. The altered waveform files of the sentences were converted back into analogue signals, low-pass filtered at 4900 Hz and recorded in randomly arranged sequences with the restriction that sentences with the same test vowel did not occur in immediate succession. There were 20 different sentences in each condition. Each sentence was repeated six times for a total of 120 stimuli per test. There was a 4-sec interstimulus interval and an interval of 8 sec between blocks of 10 stimulus sentences.

Stimuli for *task familiarization* consisted of 40 unmodified CVC stimuli produced by an adult female in the same carrier sentence. These were recorded in blocks of 10 stimuli each. All 10 vowels occurred four times each.

## Participants

Participants were undergraduate volunteers from courses in introductory speech science and introductory psychology at the University of South Florida. All were native speakers of American English and reported no hearing difficulties. They were naive with respect to formal phonetics training. Sixteen listeners served in each of the experimental conditions for a total of 128 participants. (A total of 195 listeners were tested; 32 did not pass the familiarization criteria, 25 were not native English speakers, and 15 were excluded because of equipment malfunction, history of hearing problems, or administrative errors.)

## Procedures

Listeners in all conditions used response forms on which rows of key words were printed as follows: *ape*, *if*, *eek*, *as*, *heck*, *ah*, *ooze*, *up*, *oh*, *hook*. Listeners identified the vowel in each syllable by marking the key word containing the vowel they heard.

As in our previous studies, an extensive task familiarization procedure was conducted prior to testing. First, the experimenter pronounced each key word, pointing out potential spelling confusions. Then each listener produced the key words, and the experimenter corrected mispronunciations and noted dialectal variations. Following this, the listeners responded to five blocks of 10 trials each. For the first block the experimenter pronounced the syllables in the sentence frame, providing feedback on each trial. Then the 40 recorded familiarization stimuli were presented; feedback was given after each item on the first of these four blocks, at the end of the second block for all items in that block, and at the end of the last two blocks for all items in those two blocks. Performance on the last two blocks was used to establish criteria for inclusion of listeners' test data. Data from listeners who made more than three errors in 20 trials, or more than one error on any given vowel, and data from listeners who were not native English speakers were excluded from analysis in this study.

All participants then listened without responding to the first 20 trials of the relevant test stimuli to become familiar with their particular stimulus condition; no feedback was given. Testing proceeded without further feedback. (It should be noted that there was no training on the experimental stimuli. Training was limited to acquainting the listeners with the general task of vowel identification and teaching the correct use of the response form.)

**Table 2**
**Percent Correct Identification of Vowels for Each Experimental Condition**

| Condition | *M* | *SD* | Range |
|---|---|---|---|
| Silent center | 92.7 | 4.7 | 81.7–100 |
| Syllable initial | | | |
| I-1PP | 38.3 | 10.0 | 9.2–47.5 |
| I-3PP | 59.8 | 6.5 | 47.5–68.3 |
| I-5PP | 70.2 | 9.5 | 54.2–85.8 |
| I-7PP | 74.3 | 8.2 | 59.2–83.3 |
| Syllable final | | | |
| F-4PP | 44.3 | 6.1 | 31.7–53.3 |
| F-6PP | 49.1 | 4.8 | 37.5–56.7 |
| F-8PP | 54.3 | 9.3 | 37.5–68.3 |

Recordings were presented binaurally via tape recorder (Revox A77) and earphones (TDH-39) to listeners at a mean level of about 65 dBA. Participants were tested in groups of 2 to 4 in a sound-attenuated listening room.

## RESULTS AND DISCUSSION

### Overall Identification Accuracy Across Conditions

Any response other than the original vowel intended by the speaker (or an omission) was counted as an error. Because listeners were urged to respond on all trials, there were very few omissions. Table 2 reports the means, standard deviations, and ranges of correct identifications (as a percent of opportunities) computed for the listeners' performance in each condition.

The high proportion of correct responses for the SC condition demonstrates that the familiarization training and screening procedures were successful and that findings concerning the adequacy of information specified over syllable onsets and offsets *together* were replicated. The results in this condition closely duplicate the results obtained in Strange (1989a). There was almost no overlap in scores between the listeners in this condition and the listeners in any other condition.

Performance on the (neutral duration) SC syllables can be compared with performance on I-3PP and F-4PP because these are the syllable fragments that were represented in the SC syllables. If one follows the procedure used by Viemeister and Wakefield (1991) and assumes that the two sources are mutually independent and can be optimally combined, it is possible to estimate what the combination of stimuli should produce. With this procedure, the observed probabilities of hits in the two segments were converted to $d'$ measures, based on the 10-alternative forced-choice procedure (see Swets, 1964, pp. 679–684), the square root of the sum of the squared $d'$s was computed, and the new resulting $d'$ was converted back into probability correct.[1] Following this procedure for each vowel and cumulating the probabilities across the vowels, this presumed "best combination" yielded 76% correct identifications. This is far short of the 93% correct identifications observed in the SC condition. It is apparent that the actual combination of two inadequate sources of information produced "something more"—specifically, a highly adequate source of information for the identification of vowels, a point that we have repeatedly made in our earlier studies.

Table 2 also shows that accuracy of vowel identification increased rapidly (on a negatively accelerated arc) with increasing duration of the initial portions of the syllables. On the other hand, increasing the duration of information available in the final portion of the syllable had only a modest linear influence on accuracy of identification. Remarkably, the initial burst and voiceless portion plus three pitch periods of syllable onsets (about 40-msec stimuli) permitted higher accuracy in identification than the last eight pitch periods from the syllable offsets (about 80-msec stimuli), in spite of the fact that 40% to 60% of the vocalic nuclei of individual vowels were presented in the latter case.

Analyses of variance (ANOVA) on the raw score data of the listeners confirmed that the preceding observations were statistically justified. Significant variation was found between the means of the SC, initial, and final conditions [$F(2,125) = 71.19$, $p < .001$]. Planned comparisons showed that performance in the SC condition was superior to that in the pooled initial and final conditions [$F(1,120) = 326.31$, $p < .001$], and performance in the initial conditions was on average superior to that in the final conditions [$F(1,105) = 59.00$, $p < .001$]. Among the initial conditions, additional pitch periods resulted in significantly improved performance except that the difference between five pitch periods and seven pitch periods failed to reach significance, suggesting that asymptote was being reached. [I-1PP vs. I-3PP, $F(1,30) = 60.92$, $p < .001$; I-3PP vs. I-5PP, $F(1,30) = 14.29$, $p < .001$; I-5PP vs. I-7PP, $F(1,30) = 2.23$, $p = .138$]. Among the final conditions, although there was slight improvement with added pitch periods, the step increments reached only marginal levels of statistical significance [F-4PP vs. F-6PP, $F(1,30) = 2.96$, $p = .088$; F-6PP vs. F = 8PP, $F(1,30) = 3.57$, $p = .061$].

### Identification Accuracy for Individual Vowels

Table 3 presents the percentage of correct identification responses for the 10 vowels in each condition. Vowels are grouped by their intrinsic durations: the four short vowels, /ɪ/, /ɛ/, /ʌ/, /ʊ/, the intermediate vowels, /i/, /u/, and the long vowels, /e/, /æ/, /ɑ/, /o/. The table shows that all vowels but /ʊ/, /o/ were identified with very high accuracy in the (duration-neutralized) SC condition, despite the lack of information about intrinsic vowel length. With respect to the initial and final conditions, however, there was a general advantage for the short, lax vowels, the identification of which was facilitated by increasing amounts of information from either the initial or the final portion of the syllable. These vowels were modestly well identified given five or six pitch periods from either syllable onset or offset. The results for three- and five-initial-pitch-period stimuli were comparable to the results for the

**Table 3**
**Percent Correct Identification for Each Vowel in Each Experimental Condition**

| | Condition | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Vowel | Silent Center | I-1PP | I-3PP | I-5PP | I-7PP | F-4PP | F-6PP | F-8PP |
| Short | | | | | | | | |
| ɪ | 97 | 68 | 94 | 95 | 100 | 98 | 97 | 100 |
| ɛ | 98 | 19 | 62 | 80 | 83 | 71 | 78 | 88 |
| ʌ | 92 | 33 | 45 | 79 | 73 | 43 | 60 | 75 |
| ʊ | 71 | 29 | 46 | 57 | 59 | 76 | 79 | 75 |
| Midlength | | | | | | | | |
| i | 100 | 85 | 98 | 99 | 100 | 73 | 78 | 82 |
| u | 98 | 54 | 75 | 93 | 93 | 72 | 86 | 80 |
| Long | | | | | | | | |
| e | 99 | 13 | 27 | 53 | 50 | 1 | 0 | 7 |
| æ | 94 | 22 | 50 | 50 | 83 | 4 | 2 | 20 |
| a | 100 | 56 | 95 | 94 | 96 | 6 | 11 | 16 |
| o | 76 | 4 | 6 | 1 | 5 | 1 | 0 | 1 |

four- and six-final-pitch-period stimuli (except for /ʊ/, which was better identified in the final sections).

For the midlength vowels, the initial conditions yielded better identification than the final conditions, but vowels in the final conditions were still reasonably well identified. For the long vowels, however, the results were quite different. Although these vowels were generally well identified in the SC syllables, /e/ and /o/ were poorly identified in the syllable-onset conditions and all of the long vowels were at or below chance levels in the syllable-offset conditions. Furthermore, the long vowels were poorly identified in the final segments almost without regard to the length of the segment. Even with eight final pitch periods, /e/ and /o/ were almost never correctly identified. This is undoubtedly because these vowels are ordinarily diphthongized in American English. In a diphthongized vowel (by definition), the initial and final portions alone sound like different vowels. It is only with some information from both portions of the syllable that the intended vowel can be recognized. The vowels /æ/ and /ɑ/ are not typically described as diphthongized in American English. However, Hillenbrand et al. (1995) reported that these vowels were characterized by considerable formant movement in their sample of speakers of Michigan dialect. Acoustic analysis of the stimuli used here indicated that, while /æ/ was quite diphthongized, /ɑ/ was not. Thus, it is surprising that /ɑ/ was not identified better in the final conditions. (See Figure 1 and Jenkins et al., 1994, for further details of formant trajectories for these stimuli.)

If diphthongization were indeed the cause of poor identification of /e/, /ɑ/, /o/, it should be true that the final segments were systematically misidentified as the vowel corresponding to the final vowel of the diphthong. For [eᴵ], one would expect /i/ or /ɪ/; and for [oᵘ], either /u/ or /ʊ/. For [ᵋæ], one might expect /ɛ/ responses in initial conditions, but not in final conditions (although /æ/ varies considerably with dialect). Table 4 gives the modal error response for each of the long vowels and the frequency of that response as a percentage of possible responses. The table clearly supports the supposition that listeners were choosing the appropriate diphthongal ending for /e/ and /o/. The data (surprisingly) suggest that the longer the sample, the more popular the error becomes. For /æ/ and /ɑ/, the modal errors were /ɛ/ and /ʌ/, respectively. These correspond to the spectrally most similar short vowels. However, notice that again, these error responses became more, not less, frequent with increasing stimulus duration.

It can also be argued, of course, that the longer a vowel, the smaller *proportion* of it is available in the altered syllable because the modification of the syllables was in terms of fixed numbers of pitch periods, not in terms of proportions. To explore this case, Figures 3 and 4 present the identification data for each vowel as a function of the proportion of each syllable presented to the listeners.

Figure 3 gives the results for the initial-syllable conditions. The figure shows that *proportion* of syllable presented (as contrasted with the absolute number of pitch periods) adds very little to our understanding of the differences in accuracy of identification. If one examines correct identifications at, say, 30% of the syllable, and draws a vertical line on the graph from that point on the abscissa, the results are virtually the same as when one considers pitch periods. The so-called point vowels (/i/, /ɑ/, /u/) and the high front vowel /ɪ/ form a cluster of vowels that are accurately perceived; the vowels /ɛ/, /æ/, /ʌ/ form the next cluster; /ʊ/, /e/ are at the next level, and /o/ is worst of all. These data suggest that the

**Table 4**
**Modal Identification Error and Percent Error Frequency for Long Vowels in Syllable-Final Conditions**

| | Condition | | | | | |
|---|---|---|---|---|---|---|
| Vowel | F-4PP | | F-6PP | | F-8PP | |
| e | ɪ | 55 | ɪ | 80 | ɪ | 90 |
| o | ʊ | 67 | ʊ | 69 | ʊ | 73 |
| æ | ɛ | 50 | ɛ | 63 | ɛ | 60 |
| a | ʌ | 81 | ʌ | 87 | ʌ | 81 |

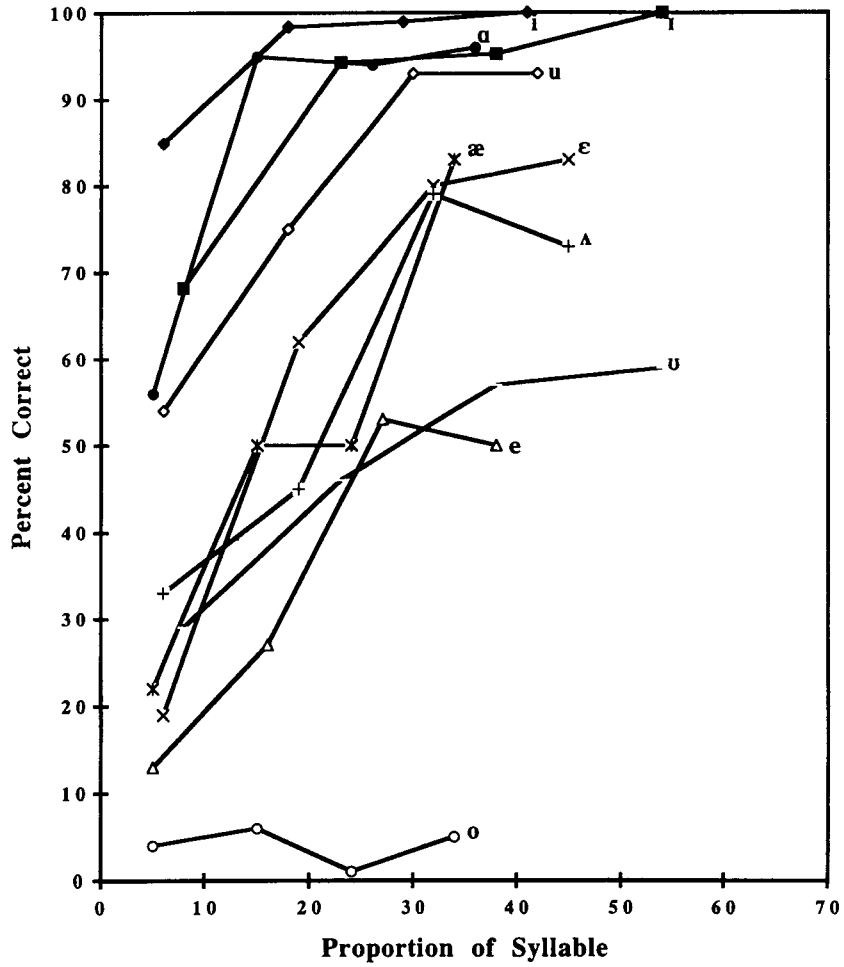


**Figure 3. Accuracy of identification of each vowel as a function of amount of syllable onset presented to the listeners.**

point vowels, which have the most extreme first ($F1$) and second ($F2$) formant values, are readily detected and identified, while the vowels that have intermediate values in the $F1/F2$ acoustic vowel space are much more difficult to identify with reduced dynamic information. Finally, of course, the diphthongized /o/ shows no improvement whatsoever with increasing length of stimulus.

Figure 4 presents a different and rather surprising picture. With the exception of the vowel /ʌ/, the accuracy curves for the vowels show little improvement with increasing duration of the final segments. The vowel /ɪ/ was readily identified at all values. The vowels /i/, /u/, /ɛ/, /ʊ/ formed the next cluster. The vowel /ʌ/ was the only vowel that showed rapidly increasing identification with greater duration, while the vowels /æ/, /ɑ/, /e/, /o/ were poorly perceived even when about 40% of the vocalic portion of those syllables was presented.

The contrast of the two figures strongly suggests that the initial "attack" of the syllable, which characterizes the coarticulated initial consonant release and movement into the vowel, is generally more informative of the identity of the vowel than the release or completion of the syllable. The influence of syllable onsets is also reflected in the work of Kato, Tsuzaki, and Sagisaka (1996), who found that listeners depend on vowel-onset intervals when they try to estimate speaking rates (when the stimuli involve more than two segments). Vowel onset intervals correlated −.91 with estimates of speaking rate, while vowel offset intervals correlated only +.30 with estimates of speaking rate. An analogy with the identification of musical instruments is also suggested; musical acousticians (e.g., Saldanha & Corso, 1964) reported that identification of musical instruments is difficult when the attack portion of a sustained note is deleted, but is unaffected by the presence or absence of the decay transients. The information for identifying the instrument is well represented in the onset but poorly represented in the offset.

### Partial Information

The classic phonetic description of vowels includes two dimensions or features—high/low and front/back—based on the (abstract) characterization of tongue–body

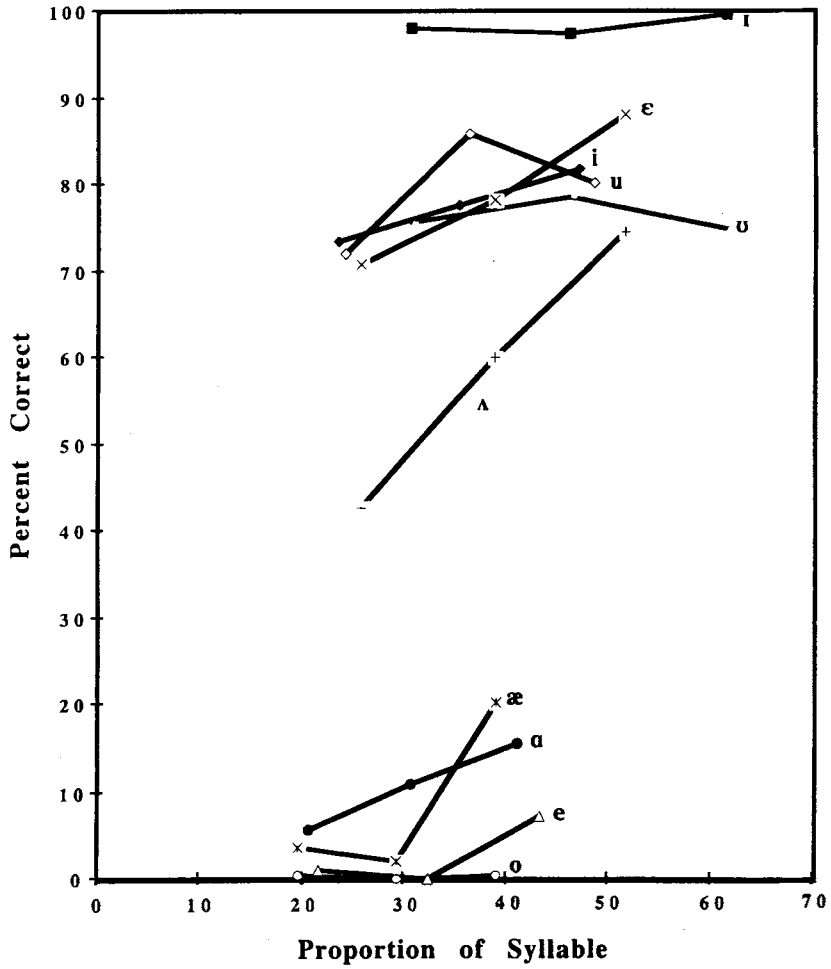


**Figure 4. Accuracy of identification of each vowel as a function of amount of syllable offset presented to the listeners.**

positions in the articulation of sustained, steady-state vowels. When spectrographic analysis of speech became available, it was apparent that these two dimensions paralleled in great part the acoustic values of the first and second formants ($F1$ and $F2$), respectively, giving further substance to the features of height and position. Our next analysis examined both correct responses and errors to determine the extent to which information concerning vowel height and position was conveyed by the syllable fragments used in this study.

For purposes of this analysis, listeners' responses were classified into the three traditional values of height (high, mid, and low) and the two values of position (front and back). The percentages of correct classifications on these features are shown in Table 5 for each experimental group.

The identification responses indicate that the position feature, front versus back, was fairly accurately recognized with as little information as a single initial pitch period. Further, this feature was almost perfectly identified in the I-3PP, I-5PP, and I-7PP conditions. The feature of height, on the other hand, was considerably more difficult to ascertain, and even with stimuli containing seven initial pitch periods, there were still slightly more than 10% errors on this feature.

The data for feature identification when the listeners heard the final segments reflect the same pattern of accuracy with respect to these features. Decisions as to front–back were fairly accurate with four final pitch periods and highly accurate in the F-6PP and F-8PP conditions. Correct decisions as to vowel height were poor in the F-4PP condition and improved surprisingly little even with a doubling of duration of the final segments.

In part, the data in Table 5 reflect the general levels of accuracy of identification because correct vowel identifications are included in the totals of correct feature identification. However, if we examine only the instances when the vowel was incorrectly identified (i.e., the misses), the same pattern of feature identification accuracy is seen. The data in parentheses in Table 5 show the proportion of correct identification of the features in only those cases when the vowel itself was misidentified. It is apparent that

**Table 5**
**Percent Correct Identification (PC) of Vowel Features for Each Experimental Condition: All Responses**

| | Tongue Height | | Tongue Position | |
|---|---|---|---|---|
| | PC | FC | PC | FC |
| Syllable initial | | | | |
| I-1PP | 58 | 32 | 84 | 74 |
| I-3PP | 77 | 45 | 95 | 87 |
| I-5PP | 85 | 51 | 98 | 94 |
| I-7PP | 89 | 58 | 97 | 90 |
| Syllable final | | | | |
| F-4PP | 57 | 23 | 87 | 77 |
| F-6PP | 59 | 20 | 92 | 84 |
| F-8PP | 62 | 18 | 96 | 91 |

Note—FC, feature correct when vowel incorrect.

the pattern is much the same as that in the overall data. The front–back distinction appears to be a salient characteristic that is detected readily even when the vowel itself is not correctly identified. The salience may be enhanced in this study by the fact that the test syllable began and ended in /d/. In general, for the syllable-initial stimuli, the second formant tends to rise from the /d/ "locus" for front vowels and fall from the locus for back vowels. Conversely, for the final /d/, the second formant falls to the locus for front vowels and rises to the locus for back vowels. Whether the front–back feature is as easily detected in other consonant environments awaits further study.

## DISCUSSION

This study contributes additional evidence concerning the dynamic aspects of the information that support the identification of coarticulated American English vowels. Review of the literature revealed that much of the research on vowel identification as a function of the length of the stimulus portions presented (e.g., Fairbanks & Grubb, 1961; Gray, 1942; Powell & Tosi, 1970; Robinson & Patterson, 1995; Schwartz, 1963; Stevens, 1959; Suen & Beddoes, 1972) relied on steady-state conceptions of the relevant information for vowels and used stimuli that must be considered as impoverished by today's standards. Evidence from many sources argues that information for the identification of vowels spoken in sentence context is spread throughout the syllable and is modified by consonantal context in major ways, suggesting that many dynamic sources are available. The present study was aimed at assessing the amount and kind of information for vowel identification that is available in the most rapidly changing parts of the syllable, namely the onsets and offsets.

The present study revealed, as previous work had suggested, that not all sources of dynamic information are equal. A general finding was that even small portions of the syllable onsets were more informative of the identity of the vowels than relatively large portions of the syllable offsets. In particular, the long vowels /e/, /æ/, /ɑ/, /o/ were the most difficult to identify from the offsets of the syllable alone. It is reasonable to suppose that this was due to diphthongization in the case of three of these vowels. Detailed analysis of the identification errors on the final segments of the syllables containing /e/ and /o/ confirmed that listeners responded with the label for the offset portions of [eᴵ], [oᵘ] even when as many as eight final pitch periods were presented. However, the error pattern for /æ/ was not consistent with that expected from formant movement. Furthermore, other vowels with extensive formant movement (Figure 1) did not produce as many errors in final conditions, and the long vowel /ɑ/, which was very poorly identified from final portions, showed very little formant movement. Thus, diphthongization within the vocalic portions of the syllables does not fully account for the discrepancy in identification rates across initial and final conditions for the long vowels.

Short vowels were slightly better identified from final segments of syllables than from initial segments of roughly the same duration, although they were moderately well identified in both conditions. There may, of course, have been some response bias in favor of these particular vowels because all of the stimuli were of short duration, but this study was not designed to separate out those effects. Midlength vowels were generally well recognized, with a small advantage (20% or less) for the initial portions over the final portions.

The most important finding of the study was the superior performance of the listeners in the SC condition over *all* of the initial and final conditions, even those that included roughly the same amount of the original stimuli (79 msec for the I-7PP, 78 msec for the F-8PP, and 81 msec for the initial + final portion for the SC condition) . Although the intrinsic duration differences of the vowels had been removed in the SC condition, the listeners achieved 93% overall correct identification with near-perfect performance on all vowels except /ʊ/ and /o/. Thus, this study demonstrates that the combination of two imperfect sources of dynamic information can produce a complex dynamic signal that is highly informative as to the identification of the vowel. It is in this sense that the present study offers an important alternative to earlier studies, which examined only the duration of steady-state vowels (with the implicit assumption that identification would be some increasing function of duration).

An analysis of responses scored in terms of the two phonetic features of position (front–back) and height (high–mid–low) revealed that the feature front–back was readily extracted from small portions of the signal from either onset or offset, while the feature of height was much more poorly represented. Roughly speaking, this means that the information conveyed largely by the second formant was more salient, more differentiated, or more discriminable than the information from the first formant. These findings must be interpreted with caution for two important reasons. First, it must be remembered that the stimuli used in this study were all /d/ V /d/ syllables. Because initial /d/ consonants produce especially divergent excursions in the second formant across vowels, it

is possible that this finding, although robust in this study, may not be replicated in other consonantal contexts. A second limitation is that all of the stimuli were produced by a single speaker. Given the presence of speaker differences in the older literature (Fairbanks & Grubb, 1961; Schwartz, 1963), it is important that different speakers be studied. Obviously, the present results call for replication both with different consonantal contexts and with other speakers.

The pattern of results found in this study also informs us somewhat about the nature of the dynamic information in coarticulated syllables that is important for vowel identity. In his target + offglide theory, Nearey (Andruski & Nearey, 1992; Nearey, 1989) has suggested that the direction and extent of spectral change within vocalic nuclei (which he termed *vowel-inherent spectral change*) can be specified by two values: a "target" value, which occurs quite early in the syllable, and an "offglide," computed as the difference between target values and values taken at a point somewhere near the end of the vocalic nucleus. He suggested that this information may be sufficient to fully specify the perceptually relevant dynamic information for vowels. The finding that vowels in the SC condition were identified better than in either initial or final conditions of the same duration supports such a characterization. However, when the movement patterns (shown in Figure 1) are inspected, it is difficult to reconcile patterns of identification accuracy on initial and final conditions with the target + offglide characterization of perceptually relevant dynamic information. For example, the formant patterns for the diphthongized vowel [eɪ] indicate that the offglide movement was much more pronounced in the last half of the syllable; however, identification was more accurate in the longest initial condition (about 50% correct) than in the longest final condition (less than 10% correct responses). Likewise, the monophthongal /ɑ/ showed differential performance rates across initial and final conditions, even though both sets of stimuli were characterized by a quasi-steady-state portion defining the "target" and negligible offgliding. As a final example, the vowel /ɪ/ was relatively accurately identified in both initial and final conditions, despite the fact that *F*2 movement was in opposite directions in the first and second halves of the syllable nuclei (i.e., in the initial and final conditions).

In previous works, we have hypothesized that, in addition to vowel-inherent spectral change, temporal trajectories associated with the opening and closing gestures of CVC syllables were a source of information for differentiating so-called tense and lax vowels in American English (Jenkins et al., 1994; Strange & Bohn, 1998; see also Di Benedetto, 1989a). The finding that vowels in neither the initial nor the final conditions were as accurately identified as were vowels in the SC condition supports the conclusion that the *relative* timing of opening and closing gestures is perceptually important. The overall better performance in the initial conditions suggests that the opening gestures are relatively more distinctive. This also accords with the hypothesis offered by Di Benedetto (1989b) that *F*1 onset and (relative) time to *F*1 maximum differentiate vowel height.

Clearly, further research in which dynamic parameters are manipulated in synthetic speech is necessary before we can define with precision just what spectrotemporal parameters are being used by perceivers to differentiate coarticulated vowels. The present study suggests that modeling the (relational) spectrotemporal characteristics associated with opening and closing gestures will be important for a full understanding of how coarticulated vowels are perceived.

## REFERENCES


ANDRUSKI, J. E., & NEAREY, T. M. (1992). On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables. *Journal of the Acoustical Society of America*, **91**, 390-410.

DI BENEDETTO, M.-G. (1989a). Frequency and time variations of the first formant: Properties relevant to the perception of vowel height. *Journal of the Acoustical Society of America*, **86**, 67-77.

DI BENEDETTO, M.-G. (1989b).Vowel representation: Some observations on temporal and spectral properties of first formant frequency. *Journal of the Acoustical Society of America*, **86**, 55-66.

FAIRBANKS, G., & GRUBB, P. (1961). A psychophysical investigation of vowel formants. *Journal of Speech & Hearing Research*, **4**, 203-219.

FOWLER, C. A. (1987). Perceivers as realists, talkers too: Commentary on papers by Strange, Diehl et al., and Rakerd and Verbrugge. *Journal of Memory & Language*, **26**, 574-587.

GRAY, G. W. (1942). Phonemic microtomy: The minimum duration of perceptible speech sounds. *Speech Monographs*, **9**, 75-90.

HILLENBRAND, J., & GAYVERT, R. T. (1993). Identification of steady-state vowels synthesized from the Peterson and Barney measurements. *Journal of the Acoustical Society of America*, **94**, 668-674.

HILLENBRAND, J., GETTY, L. A., CLARK, M. J., & WHEELER, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, **97**, 3099-3111.

JENKINS, J. J., STRANGE, W., & EDMAN, T. R. (1983). Identification of vowels in "vowelless" syllables. *Perception & Psychophysics*, **34**, 441-450.

JENKINS, J. J., STRANGE, W., & MIRANDA, S. (1994). Vowel identification in mixed-speaker, silent-center syllables. *Journal of the Acoustical Society of America*, **95**, 1030-1043.

JOOS, M. (1948). *Acoustic phonetics* (Language Monograph No. 23). Baltimore: Linguistic Society of America.

KATO, H., TSUZAKI, M., & SAGISAKA, Y. (1996, December). *Evidence for the predominance of vowel onsets to offsets in speaking-rate perception*. Paper presented at the joint meeting of the Acoustical Society of America and the Acoustical Society of Japan, Honolulu.

LADEFOGED, P. (1967). *Three areas of experimental phonetics*. New York: Oxford University Press.

LEHISTE, I., & MELTZER, D. (1973). Vowel identification in natural and synthetic speech. *Journal of Language & Speech*, **16**, 356-364.

LINDBLOM, B. E. F., & STUDDERT-KENNEDY, M. (1967). On the role of formant transitions in vowel recognition. *Journal of the Acoustical Society of America*, **42**, 830-843.

NEAREY, T. M. (1989). Static, dynamic and relational factors in vowel perception. *Journal of the Acoustical Society of America*, **85**, 2088-2113.

NEAREY, T. M., & ASSMANN, P. (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, **80**, 1297-1308.

PARKER, E. M., & DIEHL, R. L. (1984). Identifying vowels in CVC syllables: Effects of inserting silence and noise. *Perception & Psychophysics*, **36**, 369-380.

PETERSON, G., & BARNEY, H. (1952). Control methods used in a study of vowels. *Journal of the Acoustical Society of America*, **24**, 175-184.

POWELL, R. L., & TOSI, O. (1970). Vowel recognition threshold as a

function of temporal segments. *Journal of Speech & Hearing Research*, **13**, 715-724.

ROBINSON, K., & PATTERSON, R. D. (1995). The stimulus duration required to identify vowels, their octave, and their pitch chromas. *Journal of the Acoustical Society of America*, **98**, 1858-1865.

SALDANHA, E. L., & CORSO, J. F. (1964). Timbre cues and the identification of musical instruments. *Journal of the Acoustical Society of America*, **36**, 2021-2026.

SCHWARTZ, M. F. (1963). A study of thresholds of identification for vowels as a function of their duration. *Journal of Auditory Research*, **3**, 47-52.

SHANKWEILER, D., STRANGE, W., & VERBRUGGE, R. (1977). Speech and the problem of perceptual constancy. In R. E. Shaw & J. Bransford (Eds.), *Perceiving, acting, and comprehending: Toward an ecological psychology* (pp. 315-345). Hillsdale, NJ: Erlbaum.

STEVENS, K. N. (1959). Effect of duration on vowel identification. *Journal of the Acoustical Society of America*, **31**, 109.

STRANGE, W. (1987). Information for vowels in formant transitions. *Journal of Memory & Language*, **26**, 550-557.

STRANGE, W. (1989a). Dynamic specification of coarticulated vowels spoken in sentence context. *Journal of the Acoustical Society of America*, **85**, 2135-2153.

STRANGE, W. (1989b). Evolving theories of vowel perception. *Journal of the Acoustical Society of America*, **85**, 2081-2087.

STRANGE, W., & BOHN, O.-S. (1998). Dynamic specification of coarticulated German vowels: Perceptual and acoustical studies. *Journal of the Acoustical Society of America*, **104**, 488-504.

STRANGE, W., EDMAN, T. R., & JENKINS, J. J. (1979). Acoustic and phonological factors in vowel identification. *Journal of Experimental Psychology: Human Perception & Performance*, **5**, 643-656.

STRANGE W., JENKINS, J. J., & JOHNSON, T. L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, **74**, 695-705.

STRANGE W., VERBRUGGE, R., SHANKWEILER, D., & EDMAN, T. (1976). Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, **60**, 213-224.

SUEN, C. Y., & BEDDOES, M. P. (1972). Discrimination of vowel sounds of very short duration. *Perception & Psychophysics*, **11**, 417-419.

SWETS, J. A. (Ed.) (1964). *Signal detection and recognition by human observers*. New York: Wiley.

TIFFANY, W. R. (1953). Vowel recognition as a function of duration, frequency modulation and phonetic context. *Journal of Speech & Hearing Disorders*, **18**, 289-301.

VERBRUGGE, R., & RAKERD, B. (1986). Evidence of talker-independent information for vowels. *Journal of Language & Speech*, **29**, 39-57.

VIEMEISTER, N. F., & WAKEFIELD, G. H. (1991). Temporal integration and multiple looks. *Journal of the Acoustical Society of America*, **90**, 858-865.

## NOTE

1. We are grateful to Christopher Darwin for suggesting this method of estimating the effect of combining the two sources of information. Other means of estimating the probabilities of correct identifications with the combination of the probabilities of accuracy of initial and final stimuli resulted in still lower estimates, ranging from 58% to 69% correct, depending on assumptions concerning the correlation between correct identifications of the initial and final portions separately.

(Manuscript received March 17, 1997;
revision accepted for publication June 7, 1998.)