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

What information enables a listener to map a talker's vowel space?*

Robert R. Verbrugge† and Winifred Strange

Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455

Donald P. Shankweiler

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

Thomas R. Edman

Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455
(Received 23 December 1975; revised 30 March 1976)

Prior experience with a talker's speech contributes little to success in vowel identification. Adult listeners averaged only 12.9% errors on 15 vowels in /h–d/ syllables spoken in mixed order by 30 talkers (men, women, and children), and 17.0% errors on 9 vowels spoken in /p–p/ syllables by 15 talkers. When the /p–p/ test series was spoken by single talkers, errors decreased by less than half to 9.5%. Experience with known subsets of a talker's vowels did not significantly reduce errors on subsequent test tokens: following the point vowels (/i/, /a/, /u/), errors averaged 12.2% on vowels in /h–d/ context and 15.2% in /p–p/ context; following three central vowels (/ɪ/, /a/, /ʌ/), errors averaged 14.9% in /p−p context. Precursors mainly influenced listeners' response biases, rather than facilitating true improvements in vowel identifiability. These results did not support the hypothesis that point vowels provide listeners with unique information for normalizing a talker's "vowel space." Errors on vowels in rapid, destressed /p–p/ syllables (excised from sentence context) averaged 23.8%. Errors jumped to 28.6% when point-vowel precursors were introduced, while presentation of syllables in the original sentences reduced errors to 17.3%. Sentence context aids vowel identification by allowing adjustment primarily to a talker's tempo, rather than to the talker's vocal tract.

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

INTRODUCTION

The acoustic structure of speech varies markedly from one talker to another. Peterson and Barney's (1952) spectrographic measurements showed that center frequencies of vowel formants vary widely across men, women, and children, and that considerable variation also exists among talkers of the same sex and age group. Similar results were found by Peterson (1961). This acoustic variation is attributed to differences in the sizes and shapes of talkers' vocal cavities. Since each talker's vowels are idiosyncratic in their acoustic composition, it has been thought that a listener needs an extended sample of a talker's speech in order to identify vowel tokens accurately. In general terms, such experience would enable listeners to adjust to each voice they encounter.

Instead of supplying typical frequency values for each vowel, experience with a voice is thought to result in a more general adjustment to the talker's "vowel space." This assumes that a listener identifies a particular vowel of a given talker in terms of the relation between its acoustic structure and the acoustic structure of other vowels produced by the same person (Joos, 1948; Ladefoged and Broadbent, 1957; Ladefoged, 1967). The first sample of a talker's speech will calibrate (or "normalize") the framework to which the listener refers later vowel tokens for identification. Ladefoged and Broadbent (1957) tested this idea with synthetically produced stimuli and found that the perception of an acoustically fixed test word varied predictably as the formant frequencies of a carrier sentence were shifted up or down. They interpreted this result within the framework of adaptation level theory (Helson, 1948), which assumes that perceivers regularly gauge the range of a stimulus continuum in the process of formulating psychophysical judgments.

There have been few explicit hypotheses about how much precursory speech from a talker is required for accurate calibration and what phonetic information is most effective. The most common suggestion, dating back to Joos (1948), is that the point vowels /i, ɑ, u/ are the primary calibrators of vowel space. The most recent proponents of this view are Lieberman and his colleagues (Lieberman, Crelin, and Klatt, 1972; Lieberman, 1973). They argue that experience with the point vowels (or the related glides, /j, w/) is a necessary condition for accurate identification of syllables produced by a novel talker. They note that the point vowels are exceptional in several ways: (a) they represent the extreme positions in a talker's articulatory vowel space, (b) they represent the extremes of formant frequency values in a talker's acoustic vowel space, (c) they are acoustically stable for small changes in articulation (Stevens, 1972), and (d) they are the only vowels in which an acoustic pattern can be related to a unique vocal tract area function (Lindblom and Sundberg, 1969; Stevens, 1972). Other vowels are ambiguous unless calibration to a vocal tract has taken place.

There is little evidence to support the claim of a special role for the point vowels. Suggestive evidence is provided by Gerstman (1968), who developed a com-

Copyright © 1976 by the Acoustical Society of America

puter algorithm for recognition of vowels. Gerstman's algorithm used the extreme values of a talker's formant frequencies (usually those of /i, ɑ, u/) to scale all of the talker's vowels. The algorithm operated on these normalized values and classified the vowels produced by the Peterson and Barney (1952) panel with a high level of accuracy. However, it must be recognized that such an algorithm is not a perceptual strategy, but only a logically possible strategy. There is no evidence that human listeners perform the computations found in Gerstman's algorithm (such as scaling formants or computing their sums and differences). Ladefoged and Broadbent's (1957) results provide no assistance on the question of point vowels, since their study did not systematically vary the phonetic content of the precursory speech.

More generally, there is reason to doubt whether a preliminary normalization step plays the major role in vowel perception that is commonly attributed to it. Remarkably low error rates have been found when human listeners identify single syllables produced by human talkers. Peterson and Barney (1952) and Abramson and Cooper (1959) found average error rates of 4% to 6% when listeners identified the vowels in h-vowel-d words spoken in random order by a group of talkers. The test words were spoken as isolated syllables and in most conditions the listeners had little or no prior experience with the talker's voice. On the face of it, these low observed error rates seem inconsistent with any theory that stresses the need for extended prior experience with a talker's vowel space. However, it is difficult to assess the full significance of these findings, since several vowels were substantially more ambiguous than the mean error rates would suggest, and the possible role of point vowels in reducing those ambiguities was not explored.

For these reasons, it is worth investigating what information listeners actually rely upon in natural speech for identifying the vowels produced by a variety of talkers. There is currently no consensus about the perceptual problem posed by vowels in the context of a single syllable, nor about the information gained during experience with a voice. In particular, there is no perceptual evidence that the point vowels play a special role as calibrators of a talker's vowel space. The experiments reported here represent a systematic investigation of these questions.

I. EXPERIMENT I: PERCEPTION OF VOWELS IN /h–d/ ENVIRONMENT

Identifying a vowel in a naturally-spoken syllable should be most difficult when a listener has had no prior experience with the talker's voice. Thus, the need for normalization over several syllables can best be assessed by presenting listeners with a series of single syllables, each spoken by a different talker. The presence of many natural sources of talker-related acoustic variation (e.g., differences in age, sex, vocal tract size, and characteristic pitch level) should maximize the difficulty of such a test. These test conditions were approximated in the perceptual experiments of Peterson and Barney (1952), who presented 20 tokens from each of 10 talkers (men, women, and children) in each block of trials, and Abramson and Cooper (1959), who used 15 tokens spoken by each of 8 adult talkers. Both experiments studied vowels in a fixed /h–d/ consonantal frame.

Our first experiment also used /h–d/ syllables and addressed two major issues: (a) the need for extended familiarization with a talker's vowel space, and (b) the possible role of the point vowels as calibrators of that space. Compared to earlier studies, a greater effort was made in this study to eliminate any potential contribution of familiarity with individual talkers' voices. Thirty talkers each spoke only three syllables distributed throughout the test. In addition, five diphthongs were added to the ten vowels studied by Peterson and Barney in order to make all perceptual alternatives available to the listeners: /i, I, ɛ, æ, ɑ, ɔ, ʌ, ʊ, u, ɝ, eɪ, oʊ, aɪ, aʊ, ɔɪ/.

There were two test conditions in the experiment. The no-precursor test contained a long series of /h–d/ syllables; vowel identity and talker identity were unpredictable from one syllable to the next. In the point-vowel-precursor test, each /h–d/ test syllable was preceded by a string of three syllables containing the point vowels /i, ɑ, u/ spoken by the same talker. The three vowels were spoken in a /k–p/ consonantal environment; thus, the precursor string contained real words that were different from the test words. The listeners' task in each condition was to identify the vowel in the test syllables. A comparison of the errors made in the two conditions provides a direct measure of the information supplied by exposure to a talker's point vowels. If the point vowels serve as primary calibrators of vowel space, one would expect significantly better vowel identification in the point-vowel-precursor condition than in the no-precursor condition.

A. Method

1. Stimulus materials

Thirty talkers of varying ages, physical sizes, and characteristic pitch ranges were selected. The group included 13 men, 12 women, and 5 children. The children ranged from 4 to 10 years of age. All talkers spoke English as their native language, but they were heterogeneous in dialect.

The talkers were recorded individually in a sound-attenuated experimental room with a ReVox A77 stereo tape recorder and Spher-o-dyne microphone. Each talker recorded the full list of 15 test syllables twice, plus two repetitions of the precursor string. The syllables in each precursor string were read at a rate of 1 sec^{-1}. The first utterance of each syllable or precursor string was used in the listening tests, unless the talker had clearly mispronounced it.

The test series for each condition contained 90 test syllables, presented in three blocks of 30 syllables each. Each talker contributed only three syllables containing different vowels to the test, one syllable to each block. Each vowel appeared a total of six times, twice within

each block. Vowels were assigned to talkers randomly. The order of presentation of syllables within blocks was random with the constraints that (a) no less than 10 trials intervened between tokens produced by the same talker in one block and the next and (b) no vowel appeared more than twice in succession.

The point-vowel-precursor test was constructed first. Test trials were assembled in the order just described. For each trial, a precursor string was rerecorded, followed by the appropriate test syllable for the same talker. A 1-sec pause was inserted between the last precursor syllable and the test syllable. The same precursor string preceded all three of a talker's test syllables. Peak intensity for each precursor string and test syllable was equalized within $\frac{1}{2}$ dB as monitored on the VU meter of the tape recorder. A 4-sec intertrial interval was inserted between each test syllable and the following set of precursors, and a 10-sec interval was inserted between blocks of 30 syllables.

The no-precursor test was constructed by rerecording the test syllables and deleting the precursors. Thus, the two tests contained identical test syllables; the order of presentation, the intervals between successive test syllables, and the intensity of the syllables were all the same.

2. Procedure

Tests were presented to small groups of subjects in a quiet experimental room via a Crown CX 822 tape recorder, MacIntosh MC40 amplifier, and AR acoustic suspension loudspeaker. The output level was the same for both tests as monitored by a Heathkit ac VTVM placed just ahead of the output of the loudspeaker. The level was clearly audible in all parts of the room. Subjects responded on score sheets which contained 15 response alternatives, all written out in full and arrayed in rows as follows: "hood, head, hoed, heard, who'd, hide, heed, how'd, hud, hayed, hod, hoyed, had, hid, howed." They were told that they would hear "several different talkers." Subjects in the point-vowel-precursor condition were informed that each test word would be preceded by three other words spoken by the same person, and that listening to those three words might help them identify the fourth. Subjects listened to the full test series twice, for a total of 180 judgments per subject, 12 on each intended vowel.

3. Subjects

The listeners were 37 paid volunteers from undergraduate psychology classes at the University of Minnesota. All were native speakers of English and most were native to the upper midwest region of the United States. Seventeen were subjects in the no-precursor condition, while 20 were subjects in the point-vowel-precursor condition.

B. Results and discussion

Errors in vowel identification were tabulated for each condition. An error was defined as a failure to select the vowel intended by the talker; the error category included omissions, i.e., failures to select any alternative. In the no-precursor condition, subjects made an average of 12.9% errors, and in the point-vowel-precursor condition, subjects averaged 12.2% errors on the test syllables. Contrary to the prediction that point-vowel precursors would substantially reduce errors, the error rates for the two conditions were not significantly different, $t(35) = 0.57$.

The error rate in the no-precursor condition was somewhat higher than the error rates found in the two earlier studies using /h–d/ syllables. Peterson and Barney (1952) reported an overall error rate of 5.6%. Their lower observed rate may be due to the smaller number of response alternatives in their study (10 instead of 15), the smaller number of talkers appearing in a particular block of trials (10 instead of 30), and the larger total number of tokens from each talker (20 instead of 6). Abramson and Cooper (1959) reported an error rate of 4.0% in a study involving 15 vowel alternatives and eight adult talkers. In contrast to the present study, talkers carefully selected tokens they considered typical, and the listeners were familiar with the talkers (in fact, the group of listeners included the talkers). In addition, the number of talkers in the Abramson and Cooper study was smaller (8 instead of 30) and the total number of tokens from each talker was larger (15 instead of 6). Thus there are several possible sources for the higher error rate observed in the no-precursor condition of this study. But whatever the source, it must not be overlooked that 12.9% is a remarkably low error rate for a 15-alternative response set, especially if one believes that a single syllable from a novel talker is a highly ambiguous entity.

Though experience with talkers' point vowels did not reduce overall errors, it is important to determine whether the precursors influenced the perception of individual vowels. The percentage of errors made on each intended vowel is presented in Table I for each test

TABLE I. Mean percent error in identification of /h–d/ syllables. Parenthesized figures present the mean percent error when confusions between /ɑ/ and /ɔ/ are excluded.

	Condition	
Intended vowel	No precursor	Point-vowel precursor
i	1.0	0.0
ɪ	20.1	29.6
ɛ	19.1	9.2
æ	12.3	9.6
ɑ	48.5 (9.3)	43.3 (4.6)
ɔ	18.1 (9.3)	42.9 (19.2)
ʌ	14.7	3.8
ʊ	14.7	18.3
u	8.3	1.7
ɝ	0.0	0.0
eɪ	2.4	2.1
oʊ	12.7	4.6
aɪ	2.0	0.0
aʊ	16.2	17.9
ɔɪ	3.9	0.0
Overall	12.9 (9.7)	12.2 (8.0)

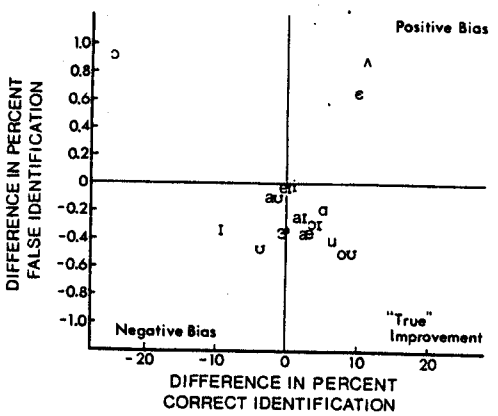

FIG. 1. Changes in correct and false identification attributable to /kip, kɑp, kup/ precursors (/h—d/ syllables). Each axis plots the difference between the point-vowel-precursor condition and the no-precursor condition.

condition. (Confusion matrices for these conditions are presented in Tables A-I and A-II in the Appendix.) Several results are worth noting. First, errors tended to be very high on the intended vowels /ɑ/ and /ɔ/. Most of these errors involved confusions between the two vowels. In fact, confusions between /ɑ/ and /ɔ/ account for 39% of all errors made by listeners in the no-precursor condition, compared to 28% of all errors in Peterson and Barney's experiment (1952). Thus, the phonetic confusion between /ɑ/ and /ɔ/ may have contributed to the higher overall error rate observed in this study. The degree of confusability is not surprising since little distinction is made between /ɑ/ and /ɔ/ in upper midwestern dialects; most of the listeners (and many of the talkers) were native to that region. The error rates for identifying these two vowels, *excluding* /ɑ/—/ɔ/ confusions, are included in parentheses in Table I.

Second, several vowels were identified very accurately, even in the no-precursor condition. This is true for two of the three point vowels (/i/ and /u/), for /ɝ/, and for three of the diphthongs (/eɪ/, /aɪ/, and /ɔɪ/). Low error rates for /i/, /u/, and /ɝ/ were also observed by Peterson and Barney. The presence of two point vowels in this group verifies predictions that they should be relatively unambiguous (cf. Lieberman *et al.*, 1972), although their role as calibrators remains in question. The low error rates for diphthongs suggests that their addition to the response set did not contribute much to the higher overall error rate in this study. The error rate for the five diphthongs averaged only 6% across the two conditions.

Third, and most importantly, there was no consistent pattern of change when test syllables were preceded by point-vowel precursors. This was true even for the relatively ambiguous vowels. Of the seven vowels showing a greater-than-average number of errors in the no-precursor condition, three showed an apparent improvement following precursors (/ɛ/, /ɑ/, /ʌ/), while four showed an increase in errors (/ɪ/, /ɔ/, /ʊ/, /aʊ/). Thus, in terms of overall errors on individual vowels, there was no consistent support for the hypothesis that experience with a talker's point vowels allows a listener to disambiguate troublesome vowels.

The differences in error rate for individual vowels need to be interpreted with caution. Differences in response biases in the two conditions could have been responsible for some of the apparent changes in identifiability. That is, a vowel could have been correctly identified more often simply because it was more popular as a response. One indication of such a response bias is how often a vowel is used as an incorrect response to other vowels; when the vowel becomes more popular, the frequency of these false identifications increases. Figure 1 depicts the results of a preliminary analysis for response biases. The horizontal axis indicates the change in correct identification (in percent) between the point-vowel-precursor and no-precursor conditions. Placement to the right of the central vertical line represents superior performance in the point-vowel-precursor condition compared to that in the no-precursor condition. The vertical axis indicates the change in false identification. (This is defined as the percentage of vowel tokens incorrectly identified as a particular vowel.) Placement above the central horizontal line represents a greater frequency of false identifications in the point-vowel-precursor condition relative to the no-precursor condition.

In this preliminary analysis, "true" improvements attributable to precursors may be defined by an increase in correct responses, coupled with a decrease in false identifications.[1] Of the vowels which were most ambiguous in the no-precursor condition, only /ɑ/ showed genuine improvement by this measure. Several less ambiguous vowels also showed genuine improvement: /æ, u, oʊ, aɪ, ɔɪ/. On the other hand, a change in correct identification that corresponds in sign with a change in false identification may be referred to descriptively as a "positive" or "negative bias." Two vowels, /ɝ/ and /ʌ/, showed a clear positive bias, while /ɪ/, /ʊ/, and /aʊ/ showed a negative bias. The remaining ambiguous vowel, /ɔ/, showed no sign of improvement: a large increase in false responses was associated with a large *decrease* in correct responses.

The analysis displayed in Fig. 1 cannot indicate which changes are significant departures from chance variability, nor can it fully disentangle changes in stimulus identifiability from changes in response biases. The number of false identifications of a vowel x might increase, not because of an increased response bias toward x, but because the perceptual similarity (confusability) of x with another vowel y may have increased. Correct and false identification scores for x will reflect the combined impact of changes in the similarity of x to several other vowels (some similarities may increase, while others decrease) and changes in response biases of all vowels concerned. Luce's choice axiom (Luce, 1959, 1963) provides one means of modeling these interactions in a confusion matrix. The model assigns a similarity parameter $\eta_{x,y}$ to each pairwise combination of stimuli and a response bias parameter

β_y to each response alternative. The combined action of these parameters determines a predicted distribution of responses in the confusion matrix.

The Luce model is useful because it allows one to assess the significance of changes in a similarity parameter from one condition to another.[2] In the present experiment, any beneficial effect of hearing point vowel precursors should manifest itself in a decrease in pairwise similarity measures (i.e., pairwise confusions should decrease). Of the 105 possible pairwise combinations of 15 stimuli, 12 pairs accounted for 81% of the errors in the no-precursor condition and 88% of the errors in the point-vowel-precursor condition. Similarity measures were determined for each of these pairs, and a t statistic was computed to assess the significance of the difference between the measures for the two conditions. Only two of the pairs showed a significant change in similarity following point-vowel precursors: /ɑ–ɔ/ and /ɔ–ɑu/; both were cases of *increased* confusability and both involved the vowel /ɔ/. This was a genuine decrement in performance on /ɔ/ which cannot be attributed to an overall change in response biases (as might be expected from Fig. 1). None of the other confusable pairs showed significant changes in similarity.

These results have direct implications for the six vowels in Fig. 1 which showed change in the direction of "true" improvement: /æ, ɑ, u, ou, aɪ, ɔɪ/. The confusion pairs for which similarity measures were obtained include the major sources of error for each of these vowels. With one exception, none of these sources of error showed a significant effect of point-vowel precursors. The exception was the confusability of /ɑ/ and /ɔ/, which showed a large increase. (The increase appeared mainly in incorrect /ɔ/ responses to /ɑ/, possibly due to a contrast between tokens of /ɑ/ in the precursor strings and the test syllables.) In general, then, even the "true" improvements cannot be interpreted as anything more than expressions of chance variability.

Thus, the patterns of error with and without point-vowel precursors were similar, showing major differences only in the identification of /ɔ/. The presence of these differences indicates that the precursors did have an impact on subjects' judgments; the nonsignificant difference in overall errors between the two conditions cannot be due to inattention to the precursor strings. Even so, there is no support in these results for the point-vowel hypothesis; the major differences involved increases in ambiguity and shifts in response biases.

Perhaps the most striking result is that subjects generally had little difficulty identifying the test syllables, even when there was no prior information about talkers' vocal tracts. It is possible that the level of identification was so high in the no-precursor condition that there was little room for improvement: 87% may represent a ceiling on identifiability of these test syllables under any conditions. Thus the failure to find a precursor effect in this experiment might indicate (a) that point vowels do not bear the kind of information hypothesized or (b) that there may be no *need* for such information, if there are no errors that are a function of uncertainties in normalization. It is necessary to know what component (if any) of the 12.9% error rate is due to subjects' uncertainty about the vocal tracts to which they are listening. This would define the maximum improvement in identification that could be contributed by the presence of precursors. The next experiment was designed to measure the error component attributable to vocal tract uncertainty and to reassess the potential value of sample vowels in reducing that uncertainty.

II. EXPERIMENT II: THE PERCEPTION OF VOWELS IN /p–p/ ENVIRONMENT

Two conditions in this experiment were designed to measure the error component in vowel perception that is attributable to talker variation. In the mixed-talker condition a large number of talkers spoke a series of syllables; on each test syllable the listener encountered a voice that was unfamiliar and unpredictable. (This condition is comparable to the no-precursor condition of Experiment I.) In the segregated-talker condition subjects heard the same series of syllables spoken by one person, so there was ample opportunity to become familiar with the voice and the talker was fully predictable from one syllable to the next. The difference between the error rates in these two conditions provides a measure of the increment in perceptual error introduced by talker variation.

Two additional mixed-talker conditions were included to reassess the role of precursory information in reducing perceptual errors. In each condition, the test syllables of the mixed-talker test were preceded by a precursor string from the appropriate talker. In the point-vowel-precursor condition, the precursor string was /hi, hɑ, hu/ (/h-/ syllables were chosen to facilitate articulation, while minimizing nonvocalic sources of information). In the central-vowel-precursor condition, each syllable was preceded by /hɪ, hæ, hʌ/.[3] As was argued in Experiment I, point-vowel precursors should substantially reduce errors if they are privileged carriers of information for normalization. A comparable set of non-point vowels should produce little or no improvement in identification, by the same hypothesis. Finally, if the information available in point vowels is essentially that gained during extended familiarization with a vocal tract, then performance in the point-vowel-precursor condition should resemble that in the segregated-talker condition.

Several changes made in the design of this experiment were intended to increase the average level of errors beyond that found in Experiment I. First, the consonantal context for the vowels was changed from /h–d/ to /p–p/. The /p–p/ environment was chosen because vowel duration tends to be shorter in voiceless stop contexts than in voiced contexts (Stevens and House, 1963). Second, an effort was made to reduce syllable duration and increase coarticulation effects by encouraging talkers to speak rapidly when recording the syllables. Third, the five diphthongs and /ɝ/ were eliminated from the vowel set, since they tended to produce few errors and would be relatively uninformative in the

present design.

A. Method

1. Stimulus materials

A panel of 15 talkers (five men, five women, and five children) were chosen to produce the test syllables for the mixed-talker conditions. They were selected to represent a wide variety of vocal tract sizes and characteristic fundamental frequencies. None were phonetically trained speakers. In the judgment of the experimenters, the talkers represented a fairly homogeneous dialect group, that of the upper midwest region from which the listeners were also drawn.

The mixed-talker tests consisted of 45 tokens, 5 tokens of each of the nine syllables: /pip/, /pɪp/, /pɛp/, /pæp/, /pɑp/, /pɔp/, /pʌp/, /pʊp/, and /pup/. Each talker contributed three test syllables. Vowels were randomly assigned to talkers with the constraint that each talker contributed three *different* vowels, only one of which was a point vowel (/i/, /ɑ/, or /u/). Thus, the five tokens of each syllable type were spoken by different talkers. In addition to three test syllables, each talker produced two sets of precursors: /hi, hɑ, hu/ and /hɪ, hæ, hʌ/. The syllables in each triplet were read at a rate of one per second. No attempt was made to control the intonation pattern of the three-syllable utterance.

The 45 recorded syllables for the mixed-talker test were arranged in a random presentation order with the constraints that (a) the same intended vowel did not appear more than twice consecutively, and (b) tokens produced by the same talker were separated by not less than 8 tokens. A 4-sec interval was inserted between tokens, and a 10-sec interval was inserted after each block of 15 tokens.

The point-vowel-precursor test was constructed by inserting copies of each talker's point-vowel triplet in front of the appropriate three test syllables in a copy of the mixed-talker test. In each case a 1-sec interval was inserted between the offset of the final precursor syllable and the test syllable.

The central-vowel-precursor test was constructed using each talker's central-vowel triplet, according to the same procedures. Thus all three mixed-talker tests contained identical test syllables; the order of presentation, the intensity levels, and the intertrial intervals were all the same.

For the segregated-talker test, one representative man, one woman, and one child were selected from the full panel of talkers.[4] For each component test (man, woman, child) the talker produced the full series of 45 test syllables, five different tokens of each of the nine syllable types. The 45 tokens were arranged in the same order as in the mixed-talker test.[5]

2. Procedure

Tests were presented to small groups of subjects under the same listening conditions as in Experiment I. Subjects responded on score sheets which contained nine response alternatives in each row: "pip, pup, pap, peep, pop, pep, poop, pawp, puup." The experimenter pronounced each word, drawing special attention to the last word, "puup," which stood for the syllable /pʊp/. The three mixed-talker tests were presented to independent groups of subjects. Subjects completed two repetitions of the 45 test trials, for a total of 90 judgments per subject, 10 on each intended vowel. Three additional groups of subjects listened to the segregated-talker tests; each group completed all three tests: man (M), woman (W), and child (C). The order of presentation of the tests was counterbalanced across groups in the orders: MWC, WCM, and CMW. For each group of subjects, data from only the first *two* tests were analyzed. Thus the total number of judgments for the segregated-talker condition was equal to that for each mixed-talker condition (90 judgments per subject) and any effects of fatigue or task familiarity were equally distributed across the three talkers in the segregated-talker tests.

3. Subjects

The listeners were 79 paid volunteers from undergraduate psychology classes at the University of Minnesota. All were native speakers of English and most were native to the upper midwest region. In mixed-talker conditions, 19 subjects heard the mixed-talker test, 15 heard the point-vowel-precursor test, and 12 heard the central-vowel-precursor test. The remaining 33 subjects served in the segregated-talker condition; 11 subjects heard each of the counterbalanced orders.

B. Results and discussion

In the mixed-talker condition (without precursors), subjects made an average of 17.0% errors in identifying vowels produced by the panel of randomly ordered talkers, while in the segregated-talker condition, listeners averaged 9.5% errors for the vowels of the three single talkers. [The mean error rates for the individual tests were 9.8% (M), 6.8% (W), and 11.8% (C).] Familiarity with a talker's voice significantly improved the accuracy of identification, $t(50) = 5.14$, $p < 0.01$. Even so, this factor accounts for less than half of the errors in the mixed-talker condition.

There are two ways to look at the error percentages for /p–p/ syllables. First, on the segregated-talker test, 9.5% is a relatively high error rate, considering the complete predictability from trial to trial of both the talker's voice and the consonantal frame. There are sources of vowel ambiguity not attributable to uncertainties in calibration. Second, on the mixed-talker test, 17% is a relatively low error rate, given that each judgment is made with no familiarity with the voice and without the benefit of sentence context. This error rate is not substantially greater than the overall 12.9% rate found for /h-d/ syllables in a similar mixed-talker test (no-precursor condition, Experiment I), though several changes were made which were intended to increase errors.[6] There is clearly a great deal of information within a single syllable which specifies the identity of

TABLE II. Mean percent error in identification of citation-form /p–p/ syllables.

Intended vowel	Condition: Mixed talker	Segregated talker	Point-vowel precursor	Central-vowel precursor
i	1.1	0.3	3.3	3.3
ɪ	1.6	3.6	2.7	1.7
ɛ	26.8	12.1	4.7	10.8
æ	18.9	1.8	20.7	18.3
ɑ	20.0 (10.0)	22.7 (3.9)	43.3 (26.7)	29.2 (12.5)
ɔ	27.4 (3.2)	18.5 (1.8)	18.7 (12.7)	13.3 (2.5)
ʌ	15.3	7.6	9.3	22.5
ʊ	38.9	17.6	26.7	29.2
u	2.6	0.9	7.3	5.8
Overall	17.0 (13.2)	9.5 (5.5)	15.2 (12.7)	14.9 (11.9)

its vowel nucleus.

The data for the mixed- and segregated-talker conditions challenge the assumption that extended familiarization with a vowel space is the primary factor controlling vowel identification. Even so, *some* information must be available in a series of utterances from a single talker, since listeners correctly identified more vowels in the segregated-talker test than in the mixed-talker test. A vowel-by-vowel analysis of subjects' errors indicates that this improvement was not distributed evenly among the nine vowels. The first two columns in Table II present the error rate for each intended vowel in the mixed- and segregated-talker conditions. Three of the vowels, /i, ɪ, u/, showed little change, since almost all tokens were correctly identified in both conditions. Of the six relatively ambiguous vowels, only /ɑ/ failed to show improvement, while familiarization aided perception of /ɛ, æ, ɔ, ʌ, ʊ/. (Confusion matrices for these two conditions are presented in Tables A-III and A-IV.)

As in Experiment I, it is important to isolate the contribution of response biases and to discover whether any of the changes in vowel similarity reflect factors other than chance variation. Again, both a graphic analysis and the Luce choice model were applied to the data from the segregated-talker and mixed-talker conditions. The first analysis (presented in Fig. 2) showed "true" improvement in the identification of /ɛ/, /æ/, /ʌ/, /ʊ/, and /u/ in the segregated-talker condition. The apparent improvement for /ɔ/ was associated with a large positive bias, while /ɑ/ showed a negative bias. The Luce similarity analysis showed significantly reduced confusions between the following pairs: /ɛ–æ/, /ɑ–ʌ/, /ʌ–ʊ/, and /ʊ–u/. These four confusable pairs were major sources of error for the five vowels showing true improvement. Thus, the increases in correct identification for these vowels reflect more than chance variation. They represent genuine compensation for confusions due to talker variation.

The failure to find true improvement for either /ɑ/ or /ɔ/ or a significant decrease in their pairwise confusion reflects their somewhat ambiguous status in upper midwestern dialects. On the average, errors for /ɑ/ and /ɔ/ were almost as frequent for a single talker as they were for a mixed group of talkers. Thus, the similarity of /ɑ/ and /ɔ/ is apparently a function of the dialect, not of unfamiliarity with talkers' voices.

The kind of improvement resulting from familiarization with a talker's vowel space may be summarized as follows: overall errors drop somewhat (7.5% in this experiment), genuine overall improvement is found for several ambiguous vowels, and there is a significant decrease in similarity for several vowel pairs. If the point vowels specify efficiently the kind of information gained during extended familiarization, we would expect a similar pattern of improvement in the point-vowel-precursor condition.

The results did not support this hypothesis. Exposure to a talker's point vowels aided listeners only slightly, reducing overall errors from 17.0% to 15.2%; the difference was not statistically significant, $t(32) = 0.97$. In the central-vowel-precursor condition, overall errors also dropped slightly, to 14.9%, though again the change was not significant, $t(29) = 1.21$. In other words, not only was there no evidence for a gain attributable to point vowels, but there was no difference between the point vowels and a set of nonpoint vowels. In general, experience with specific sets of vowels seems to make little contribution to the total reduction of errors attributable to prior experience with a person's voice.

It is important to determine whether these conclusions are affected by the results for individual vowels. The right-hand columns in Table II present the errors on each intended vowel following point-vowel and central-vowel precursors. (Confusion matrices for these conditions are presented in Tables A-V and A-VI.) A comparison of errors in the point-vowel-precursor condition and the mixed-talker condition (without precursors) is presented in Fig. 3. In general, the point vowels did not produce a "true" improvement in the perception of ambiguous vowels like that found in the segregated-talker condition. Where similar *apparent* improve-

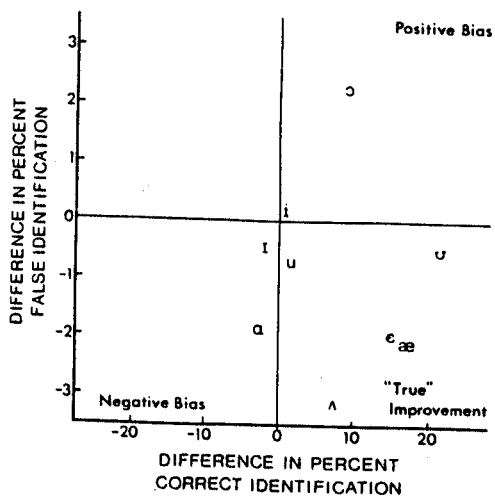

FIG. 2. Changes in correct and false identification attributable to keeping the talker constant throughout a test (citation-form /p–p/ syllables). Each axis plots the difference between the segregated-talker condition and the mixed-talker condition.

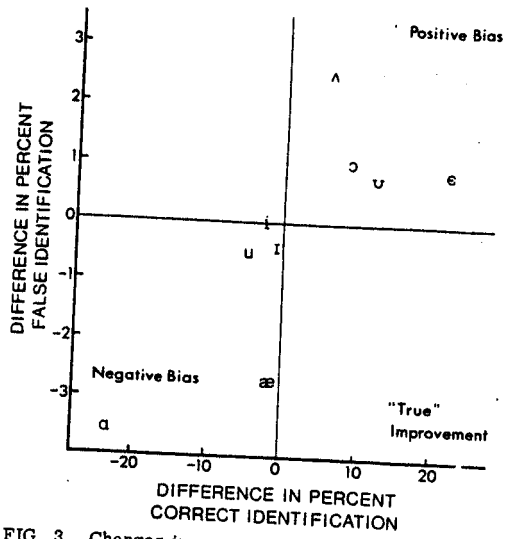

FIG. 3. Changes in correct and false identification attributable to /hi, hɑ, hu/ precursors (citation-form /p–p/ syllables). Each axis plots the difference between the point-vowel-precursor condition and the mixed-talker condition.

ments were found, they tended to be associated with much higher relative levels of false identification in the point-vowel-precursor condition (compare Figs. 2 and 3). In other cases, apparent improvements found for the segregated-talker condition were not found with the point-vowel precursors. A Luce analysis indicated that the only comparable change in pairwise similarities was a substantial reduction in /ɛ–æ/ confusions in both conditions. None of the other reductions found with segregated talkers were found with point-vowel precursors. In addition, the /ɔ–ʌ/ confusion, which showed no change with segregated talkers, showed a sharp *increase* in the point-vowel-precursor condition.

When the central-vowel-precursor condition was compared to the mixed-talker condition on a vowel-by-vowel basis, virtually the same results were obtained. No vowel showed more than a marginal change in the direction of true improvement, and a significant decrease in pairwise similarity was observed for /ɛ–æ/. However, the increase in the /ɔ–ʌ/ confusion that was observed with point-vowel precursors was not observed here. Thus, to the limited extent that improvements are found at all with precursors, there is no evidence that the three point vowels are unique as sources of information about a talker's vowel space.

In general, however, neither set of vowel precursors were efficient carriers of the kind of information available in more extended experience with a talker's voice. Sets of vowels of known identity did not produce reductions in overall errors, errors on specific vowels, or pairwise similarities comparable to those produced by extended experience.

An extension of the Luce model allows one to make comparisons between the overall error patterns for two experimental conditions. Specifically, one may ask whether the same set of stimulus similarity and response bias parameters is sufficient to describe both patterns, or whether different sets provide a closer fit. In the latter case, one may test models in which only the similarity parameters for each condition differ, in which only the bias parameters are different, or in which both parameter sets differ.

Joint models for the mixed- and segregated-talker conditions suggest that the dominant impact of extended familiarization is on perceptual similarity. The different-parameters model ($\chi^2/df = 3.54$), in which both sets differ, provides a closer fit than the same-parameters model ($\chi^2/df = 5.43$).[7] This improvement is contributed largely by different similarity parameters: the different-similarities model ($\chi^2/df = 3.83$) fits both conditions more successfully than the different-biases model ($\chi^2/df = 5.27$). This means that the main effect of hearing a single talker is on a listener's ability to discriminate the vowels themselves, not on the listener's response biases.

A different result is found when the mixed-talker condition is compared with each precursor condition. In each case, estimating different similarity parameters fails to improve the overall goodness-of-fit; different bias parameters, on the other hand, do improve the model. When errors in the mixed-talker and point-vowel-precursor conditions are jointly modeled, the same-parameters model ($\chi^2/df = 3.68$) fits substantially better than the different-similarities model ($\chi^2/df = 5.78$), but not as well as the different-biases model ($\chi^2/df = 2.46$). Similarly, when errors in the mixed-talker and central-vowel-precursor conditions are jointly modeled, the same-parameters model ($\chi^2/df = 2.66$) is not improved by the addition of different similarity parameters ($\chi^2/df = 4.55$), but is improved by different bias parameters ($\chi^2/df = 2.35$). Thus, the precursors not only produced a pattern of similarity changes different from that hypothesized, but produced change of a different kind altogether. Precursors predominantly affected listeners' preferences for various response alternatives, rather than their ability to distinguish among intended vowels.

A possible shortcoming of the design of this experiment is that the test syllables were not sufficiently "natural": since they were spoken in citation form, the formant frequencies of their vocalic centers would not show the degree of variability found for destressed vowels in rapidly articulated sentences. It is possible that the task of perceiving rapidly spoken syllables places a higher premium on information about the vocal tract. Experiment III was designed to determine whether point vowels would benefit listeners on a mixed talker task involving rapidly articulated vowels.

III. EXPERIMENT III: PERCEPTION OF VOWELS IN DESTRESSED /p–p/ SYLLABLES

In the rapidly articulated syllables of connected speech, vowel durations tend to be short and vowel formants are not likely to reach steady-state values. Formant values at the center of syllables in connected speech are different from those found in single syllables spoken in citation form, the degree of deviation depending systematically on the rate of articulation and

the amount of destressing (Tiffany, 1959; Shearme and Holmes, 1962; Lindblom, 1963; Gay, 1974). If vowel perception involves relating vowels to a "space" (defined by some transformation on formant frequencies), then the frequency variation contributed by speaking rate should considerably enhance a listener's difficulty in calibrating to a talker's space. This experiment explores the perceptual problem posed when both talker-dependent and rate-dependent variation are present. The error rate for single, rapidly articulated syllables excised from carrier sentences should be substantially greater than that found for syllables spoken in isolation. Given the (presumably) more difficult task of identifying a rapid, destressed syllable, information about a talker's point vowels may play a larger role than was found in preceding experiments.

The experiment consisted of three test conditions. In the no-precursor condition, listeners heard a mixed-talker test containing /p-p/ syllables spoken by the same panel of talkers used in Experiment II. The syllables were spoken in destressed position in the context of a full carrier sentence and were excised for use in the test. In the point-vowel-precursor condition, each test syllable was preceded by a point-vowel precursor string spoken by the appropriate talker. In the sentence-context condition, each test syllable was heard in the context of the carrier sentence in which it was originally produced. One would expect the error rate in this condition to be lower than that in the no-precursor (and no context) condition, since more information is available about the talkers prior to the test syllables. If so, the degree of improvement provides a measure of the information supplied by sentence context, when no semantic factors are involved. The pattern of improvement following point-vowel precursors should be similar, if the predominant effect of both types of context (precursor and sentence) is to allow calibration to a talker's vowel space.

A. Method

1. Stimulus materials

Each of the 15 talkers contributed the same three syllables they had produced for the mixed-talker tests in Experiment II. In all three conditions of this experiment, the order of talkers and test syllables was the same as in the earlier experiment. The tests contained five tokens of each of nine /p-p/ words; each of the five tokens was produced by a different talker and each talker contributed only one point vowel. The test syllables were spoken in the following carrier sentence: "The little p–p's *chair* is red." Talkers were instructed to read each sentence rapidly, stressing the word "chair."

The test syllables were excised from copies of the carrier sentences for use in the no-precursor and point-vowel-precursor tests. Each recording was monitored and the audio tape was cut within the silent interval just preceding the release burst of the initial /p/ and during the silent closure interval of the final /p/. Thus, the final /p/ of the test syllables did not include a release from closure. To produce the no-precursor test, the 45 excised syllables were assembled in the presentation order and then rerecorded as in Experiment II. The point-vowel-precursor test was constructed by inserting copies of each talker's point-vowel triplet in front of the appropriate three test syllables in a copy of the no-precursor test, using the same precursor strings and recording procedure as in Experiment II. Thus, the no-precursor and point-vowel-precursor tests contained identical test syllables, with the same order of presentation, intensity levels, and intertrial intervals, and each was comparable in these respects to the mixed-talker conditions of Experiment II. The sentence-context test was constructed using copies of the original carrier sentences. The order of talkers and component test syllables was the same as that in the other two tests. A 4-sec interval was inserted between each sentence.

2. Procedure

Tests were presented to small groups of subjects under the same conditions as in previous experiments. Subjects in the sentence-context condition were told that each test word would be spoken in the middle of the same sentence: "The little (something)'s chair is red." The three tests were presented to independent groups of subjects. Subjects completed two repetitions of the 45 test trials, for a total of 90 judgments per subject, 10 on each intended vowel.

3. Subjects

The listeners were 52 paid volunteers from undergraduate psychology classes at the University of Minnesota. All were native speakers of English and most were native to the upper midwest region. Twenty were subjects in the no-precursor condition, 17 in the point-vowel-precursor condition, and 15 in the sentence-context condition.

B. Results and discussion

Listeners averaged 23.8% errors in identifying the vowels in the excised syllables without precursors. As expected, this error rate is higher than the 17.0% rate found for citation-form syllables in the comparable mixed-talker test in Experiment II; the difference between these two conditions is significant, $t(37) = 3.88$, $p < 0.01$.

Given the increased ambiguity when both talker- and rate-dependent variation are present, it might be expected that listeners would make greater use of a talker's point vowels to reduce that ambiguity. Contrary to this expectation, the average error rate in the point-vowel-precursor condition was 28.6%, which is significantly *higher* than the 23.8% rate found when no precursors are present, $t(35) = 2.85$, $p < 0.01$. This is a startling result: it does not fulfill the expectation that greater improvement would be found where more was needed, nor does it even replicate the minor improvements found with point-vowel precursors in Experiments I and II.

In contrast to these results for point-vowel precur-

TABLE III. Mean percent error in identification of destressed /p–p/ syllables.

Intended vowel	Condition: No precursor	Point-vowel precursor	Sentence context
i	11.5	11.2	6.7
ɪ	0.5	1.8	0.7
ɛ	7.9	3.5	20.0
æ	24.5	44.1	2.0
ɑ	62.5 (43.0)	95.9 (92.4)	36.7 (12.7)
ɔ	49.5 (25.5)	50.6 (45.9)	31.3 (4.0)
ʌ	33.0	27.6	33.3
ʊ	19.0	18.2	23.3
u	4.5	4.7	1.3
Overall	23.8 (18.9)	28.6 (27.7)	17.3 (11.6)

sors, a substantial *decrease* in errors was found when the test syllables were heard in their original sentence context. Listeners made an average of 17.3% errors in the sentence-context condition; this is significantly lower than the 23.8% error rate found for the test syllables in excised form, $t(33) = 3.31$, $p < 0.01$. Thus, a carrier sentence contains information which makes vowels in component syllables less ambiguous.

Error rates for individual vowels are presented in Table III for each of the three test conditions. A comparison of the results for excised syllables (first column, Table III) and for citation-form syllables (first column, Table II) suggests that listeners in the no-precursor condition may not have accommodated completely to the rapid pace at which the excised syllables were spoken. In general, errors on these syllables were in the direction of hearing vowels in the periphery of two-formant space as more "centralized" or "reduced" (cf. confusion matrix, Table A-VII). (a) Two point vowels, /i/ and /u/, which produced very few errors in citation-form syllables, were somewhat ambiguous in the destressed syllables. The errors on /i/ generally involved misperceiving it as /ɪ/. The vowel /u/ tended to be misperceived as /ʊ/. (b) Errors more than doubled on /ɑ/ and /ɔ/. By far the most common error on both /ɑ/and /ɔ/ was to perceive them as /ʌ/. As a consequence, /ʌ/ showed a large increase in false identification. (c) The vowels /æ/ and /ʌ/ were also more ambiguous in destressed syllables. They were most frequently misperceived as /ɛ/ and /ʊ/, respectively. (d) In exception to this general pattern of increased error rates, the vowels /ɛ/ and /ʊ/ showed substantially *fewer* errors in destressed syllables. However, both vowels were popular false responses, and the apparent improvement was associated with a positive bias in each case. It is relevant that /ɛ/ and /ʊ/ are the most "central" vowels in two-formant space, in that they are intermediate in first formant frequency and therefore reduction toward schwa does not tend to produce formant combinations typical of other vowels. The tendency for listeners to select more "central" vowel responses suggests that they underestimated the tempo at which the excised syllables were spoken.

Rather than enabling listeners to compensate for errors introduced by tempo uncertainty, the point-vowel precursors served only to increase the errors (see Table III and the confusion matrix in Table A-VIII). Listeners tended to hear vowels more centralized than those intended, and did so with even greater frequency than in the no-precursor condition. The trend was so strong for /ɑ/ and /ɔ/ that confusions between them accounted for only 6% of errors on the two vowels themselves and only 3% of all errors on the point-vowel-precursor test. Relatively low error rates occurred on the two most "central" vowels, /ɛ/ and /ʊ/, as was found on the no-precursor test.

It seems likely that the precursor syllables (spoken in citation form) established an expected tempo inappropriate for perception of the subsequent test syllables. Instead of calibrating listeners to the formant ranges of a talker's vowel space, the precursors calibrated listeners to the tempo of the talker's speech. If the test syllable had truly been spoken in isolation with a stress equal to that of the precursors, the prior adjustment to talker tempo would have been appropriate. This condition was met in the point-vowel-precursor test of Experiment II, where errors averaged only 15%. However, the comparable test in Experiment III juxtaposed syllables spoken with radically different rates and stresses, and the contrast produced a large increase in erroneous judgments. As in the no-precursor condition, the pattern of errors reflected the contraction of acoustic vowel space found for rapid, destressed speech (cf. Lindblom, 1963).

In contrast to the results following precursors, error rates for individual vowels dropped when the destressed test syllables were heard in sentence context (see Table III and the confusion matrix in Table A-IX). Error rates for /i/, /æ/, /ɑ/, /ɔ/, and /u/ were all lower in the sentence-context condition than in the no-precursor condition, where the syllables were heard in isolation. While errors on /ɛ/ and /ʊ/ were relatively infrequent in the excised syllables, they *increased* when heard in sentence context. In general, the pattern of changes was complementary to that observed for the excised syllables. The marked "centralization" of vowel responses disappeared when syllables were heard in sentence context.

These results suggest that a carrier sentence aids identification of vowel targets by allowing listeners to adjust to talker tempo, rather than by allowing them to compensate for talker variation. The observed changes in identification have little in common with those found after extended familiarization with a talker's speech (cf. Fig. 2). When errors in the sentence-context and no-precursor conditions were compared, there were no vowels which showed "true" improvement in identification. The main effect of sentence context was to reverse a pattern of positive biases toward /ɛ/ and /ʊ/ (and to a lesser extent /ɪ/ and /ʌ/), a pattern which has more to do with tempo uncertainty than with talker variation.

Luce analyses for the three experimental conditions corroborate the conclusions drawn from the less formal error analyses. Most pairwise confusions were greater

for destressed syllables (no-precursor condition) than for citation-form syllables (mixed-talker condition, Experiment II). In two cases, /ɑ–ɔ/ and /ɔ–ʌ/, the increases were large and significant. Thus, tempo uncertainty produced some genuine increases in vowel confusability. However, one significant decrease was also observed: the /ɛ–æ/ confusion, largest source of errors on citation-form syllables, was substantially smaller for rapid, destressed syllables. It is possible that rapid articulation produces tokens of /ɛ/ which would also have been produced with high probability in citation form—i.e., rapid articulation may affect /ɛ/ more by reducing its acoustic variance than by shifting its typical formant composition. If this effect were large enough, the overall discriminability of /ɛ/ and /æ/ would increase, as observed.

Pairwise confusions for the point-vowel-precursor condition showed little systematic change relative to the no-precursor condition. The only significant change was an increase in the confusability of /ɑ/ and /ʌ/. The /ɛ–æ/ confusion was more asymmetric than in the no-precursor condition (/ɛ/ was never perceived as /æ/ following precursors), and the similarity showed a further, though nonsignificant decrease.

Pairwise confusions in the sentence-context condition tended to be lower than in the no-precursor condition, though only one of the decreases (/ɔ–ʌ/) was significant. Thus, sentence context reversed one of the two significant increases in confusability found for the excised syllables. The other vowel pair, /ɑ–ɔ/, also showed a reversal, but the decrease was not significant.

While the observed changes in pairwise similarities were usually in the expected direction, they were also few in number. The *predominant* effect of misperceiving tempo was not a change in vowel similarities, but an error-producing shift in response biases. Joint Luce models for the citation-form syllables (mixed-talker condition, Experiment II) and destressed syllables (no-precursor condition) verify that the main impact of tempo uncertainty was on response biases. A same-parameters model ($\chi^2/df = 6.14$) was not improved by different similarity parameters ($\chi^2/df = 7.36$), but was substantially improved by different biases ($\chi^2/df = 3.86$). Joint Luce models comparing the destressed syllables in isolation (no-precursor condition) with those in sentence context yield similar results: a same-parameters model ($\chi^2/df = 4.18$) was not improved by different similarities ($\chi^2/df = 6.58$), but was improved by different biases ($\chi^2/df = 2.27$). Again, these results for the sentence-context condition contrast sharply with those for the segregated-talker test (Experiment II), where the predominant effect was on pairwise similarities, not biases.

It is interesting to note that the error rate for syllable-medial vowels in sentence context (17.3%) was very close to that for medial vowels in citation-form syllables (17.0%); the difference was not significant, $t(32) = 0.16$. This suggests that there is a very stable level of error for vowels in /p–p/ words when heard in a unit of articulation sufficient to specify tempo. The only additional assumption required is that a syllable spoken in isolation specifies its own tempo.

These results provide strong evidence that the perceptual system adjusts to the ongoing tempo of a talker's utterance. However, it remains an open question whether this adjustment involves transforming or calibrating a relational vowel space for individual talkers. No evidence for a talker-specific space of this kind was found in earlier experiments, nor was any found in the precursor condition of this experiment. In addition, the effect of sentence context on identification was very different from the effect of extended familiarization with individual vocal tracts. Thus, this experiment provides no evidence that sentence context aids vowel identification by allowing compensation for talker differences.

Little is currently known about how formant contours are transformed by variations in speaking rate and stress, or how listeners adjust to these changes. Lindblom (1963) has attempted to characterize the variation in vowel center formant frequencies as a function of speaking rate. Lindblom and Studdert-Kennedy (1967), in turn, have demonstrated that listeners are sensitive to these variations when identifying vowels in isolated, synthetic syllables. If two syllables reach the same formant frequency values at the syllable centers, but simulate different rates of articulation, listeners adopt different criteria for identification of the two medial vowels. These preliminary efforts suggest that the formant transitions, which are generally understood to carry consonantal information, must also aid in specifying the *vowel*. They apparently do so, at least in part, by limiting the range of possible talker tempos. The sentence context condition of this experiment suggests that factors beyond the syllable also shape the acoustic specification of vowels and are therefore important to accurate identification. A major function of a carrier sentence is to specify the tempo and stress of component syllables.[8]

IV. SUMMARY AND CONCLUSIONS

These experiments lead to the following conclusions about the perception of vowels in natural speech:

(1) Talker-dependent acoustic variation does not pose a major perceptual problem within a common dialect group. Listeners can identify a high proportion of vowels spoken in citation-form syllables by talkers with whom they have little or no previous experience. In Experiment I, listeners identified 87% of /h–d/ syllables spoken in random order by 30 talkers representing the full natural range of acoustic variation. In Experiment II, they identified 83% of /p–p/ syllables spoken by 15 talkers. Of the errors made in this mixed-talker condition, no more than half can be attributed to talker-dependent sources of ambiguity. Correct identification in segregated-talker tests averaged 90.5% for vowels in /p–p/ syllables (Experiment II). There *was* genuine improvement in the identification of specific vowels, but only a small portion of correct identification could be attributed to familiarization (the difference between 83% and 90.5%). Thus, experience with a voice plays a secondary role in specifying vowel iden-

tity. A single syllable contains substantial information about its medial vowel, whether a talker's voice is familiar or not.

(2) Contrary to the speculations of Joos (1948), Lieberman (1973), and Lieberman *et al.* (1972), the point vowels do not play a major and privileged role as calibrators of a talker-specific vowel space. Experience with a talker's point vowels does not significantly reduce the overall ambiguity of vowels in a subsequent syllable. This result was found for all three types of test syllables studied: /h–d/, citation-form /p–p/, and destressed /p–p/. The pattern of changes following point-vowel precursors did not resemble the pattern resulting from extended experience with a talker's voice (Experiment II). Extended experience produced consistent reductions in pairwise similarities, while experience with a talker's point vowels mainly affected the pattern of response biases, with no consistent effects on vowel identifiability. Point vowels did produce a significant decrease in the confusability of /pɛp/ and /pæp/, but they were not unique in this respect: a significant reduction was also found when test syllables were preceded by central vowels (Experiment II) and when tempo uncertainty was introduced (Experiment III). In general, there was little evidence that sample subsets of a talker's vowels enable listeners to adjust to the talker's idiosyncratic "space" (defined by ranges of acoustic values or by sizes of vocal tract cavities). This conclusion, like the first, does not support the proposal of Ladefoged and Broadbent (1957) and Ladefoged (1967) that vowel perception can be regarded as a problem in establishing an adaptation level (cf. Shankweiler, Strange, and Verbrugge, in press).

(3) Listeners adjust their perceptual criteria for syllable-medial vowels according to the perceived rate of articulation. When destressed /p–p/ syllables were excised from sentence context and presented in isolation (Experiment III), there was a tendency to perceive them as if they had been spoken in citation form: the pattern of errors showed insufficient compensation for the acoustic effects of rapid articulation. When citation-form precursor strings preceded the excised syllables, the contrast of expected and actual tempos enhanced the original pattern of errors and increased the overall error rate. When the excised syllables were heard in their original temporal environments (the carrier sentences), the pattern of errors reversed and the overall error rate decreased. Carrier sentences apparently enabled listeners to adjust continuously to a talker's tempo and to compensate for the acoustic effects of vowel reduction. Information about a talker's ongoing tempo produced a qualitatively different pattern of improvement from that produced by long-term familiarization with citation-form syllables. This confirmed the results of Experiment II (where citation-form test words were heard in the context of prior citation-form syllables) in the more natural situation of words in sentence context. In neither case was there evidence that listeners acquired a scaling function for adjusting a talker's speech to a normative dialectal space. In contrast to the conclusions of Ladefoged and Broadbent (1957), a naturally produced carrier sentence may aid vowel identification more by establishing the tempo of speech than by delimiting an individual vowel space.

How *do* listeners cope with talker-related acoustic variation? One possibility is that a single syllable (with consonants of known identity) carries sufficient information for normalization to take place. Fourcin (1968) and Rand (1971) both have demonstrated that listeners adjust their perceptual criteria for stop consonants to compensate for talker-dependent variation in the consonants' acoustic structure. If the consonants in a test syllable are known in advance, a single syllable could provide relatively unambiguous information about the talker's vocal tract. This information, in turn, could be used in disambiguating the vowel.

A second possibility is that a talker-normalization procedure is not necessary for human perception of vowels. Vowel identity may be specified by properties of the acoustic signal that are relatively invariant across talkers and that do not require a prior calibration process to be accurately detected. The results for destressed syllables suggest that the dynamic properties of speech are especially critical: vowel identification seems to be at least as sensitive to tempo variation as it is to variation in talkers' center formant frequencies. Adjustment to talkers may have more to do with tracking the dynamics of ongoing articulation than with normalization as traditionally defined.

ACKNOWLEDGMENTS

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

APPENDIX A: CONFUSION MATRICES

Tables report the frequency with which each intended vowel x was identified as response alternative y. In addition, summary statistics for each condition are provided: the percent error for each intended vowel, the overall percent error for each repetition (rep.) of the test series, the overall percent error pooling both repetitions, the total number of trials for the two repetitions, the mean number of trials on which listeners made an error ($\bar{x}$), the standard deviation of this mean (s), and the number of listeners (N).

TABLE A-I. /h–d/ syllables: No-precursor condition. Overall percent error: 12.94 (pooled), 14.97 (rep. 1), 10.92 (rep. 2); 180 trials, $\bar{x}=23.29$, $s=8.56$, $N=17$.

Intended vowel	Response																Percent error
	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	ɝ	eɪ	oʊ	aɪ	aʊ	ɔɪ	None	
i	202		1												1		1.0
ɪ	1	163	39										1				20.1
ɛ		12	165	24			1			2							19.1
æ			4	179	12	1	2			1	1		1	1	1	1	12.3
ɑ				5	105	80	4				1		4	2	2	1	48.5
ɔ					18	167	10	1				1		4	3		18.1
ʌ			1	1	24	2	174						2				14.7
ʊ					7		14	174	1		1	4			3		14.7
u					1			13	187			3					8.3
ɝ										204							0.0
eɪ			5								199						2.4
oʊ					1			1	16	1	3	178		1	1	2	12.7
aɪ		2			1						1		200				2.0
aʊ				1	1	3						26		171	1	1	16.2
ɔɪ					1					7					196		3.9

TABLE A-II. /h-d/ syllables: Point-vowel-precursor condition. Overall percent error: 12.19 (pooled), 13.17 (rep. 1), 11.22 (rep. 2); 180 trials, $\bar{x}=21.95$, $s=5.79$, $N=20$.

Intended vowel	Response																Percent error
	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	ɝ	eɪ	oʊ	aɪ	aʊ	ɔɪ	None	
i	240																0.0
ɪ		169	68								1	1	1				29.6
ɛ		3	218	17						1		1					9.2
æ			11	217	5	5	2										9.6
ɑ				4	136	93	6				1						43.3
ɔ		1		1	57	137	22				3	4		5	1	9	42.9
ʌ					7	2	231										3.8
ʊ					1	1	36	196	2		1	1		1	1		18.3
u								1	236			3					1.7
ɝ										240							0.0
eɪ		1	2			1					235					1	2.1
oʊ						4		1	4			229			2		4.6
aɪ													240				0.0
aʊ				2	1	26					1	13		197			17.9
ɔɪ															240		0.0

TABLE A-III. Citation-form /p–p/ syllables: Mixed-talker condition. Overall percent error: 16.96 (pooled), 18.48 (rep. 1), 15.44 (rep. 2); 90 trials, $\bar{x}=15.26$, $s=4.53$, $N=19$.

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

TABLE A-IV. Citation-form /p–p/ syllables: Segregated-talker condition. Overall percent error: 9.46 (pooled), 10.57 (rep. 1), 8.35 (rep. 2); 90 trials, $\bar{x}=8.52$, $s=4.77$, $N=33$.

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

TABLE A-V. Citation-form /p–p/ syllables: Point-vowel-precursor condition. Overall percent error: 15.19 (pooled), 17.48 (rep. 1), 12.89 (rep. 2); 90 trials, $\bar{x}=13.67$, $s=5.26$, $N=15$.

Intended vowel	Response										Percent error
	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	
i	145		5								3.3
ɪ		146	3							1	2.7
ɛ		1	143	4		1	1				4.7
æ			30	119			1				20.7
ɑ				1	85	25	36	3			43.3
ɔ				1	9	122	14	4			18.7
ʌ					3	7	136	4			9.3
ʊ						2	31	110	7		26.7
u								11	139		7.3

TABLE A-VI. Citation-form /p–p/ syllables: Central-vowel-precursor condition. Overall percent error: 1/.91 (pooled), 15.00 (rep. 1), 14.81 (rep. 2); 90 trials, $\bar{x}=13.42$, $s=3.78$, $N=12$.

Intended vowel	Response										Percent error
	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	
i	116	3								1	3.3
ɪ	1	118								1	1.7
ɛ			107	12						1	10.8
æ			22	98							18.3
ɑ					85	20	12	3			29.2
ɔ					13	104	1	1		1	13.3
ʌ					10	8	93	9			22.5
ʊ						6	24	85	5		29.2
u								7	113		5.8

TABLE A-VII. Destressed /p–p/ syllables: No-precursor condition. Overall percent error: 23.84 (pooled), 25.19 (rep. 1), 22.49 (rep. 2); 90 trials, $\bar{x}=22.00$, $N=9$; 88 trials, $\bar{x}=20.55$, $N=11$; pooled scores: $\bar{x}=21.20$, $s=4.98$, $N=20$.

Intended vowel	Response										Percent error
	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	
i	177	16	6						1		11.5
ɪ		199							1		0.5
ɛ[a]		2	164	7		1	2		2		7.9
æ			48	151		1					24.5
ɑ					75	39	76	10			62.5
ɔ				2	48	101	43	6			49.5
ʌ		8	5		1	15	134	35	1	1	33.0
ʊ			1		1	2	22	162	12		19.0
u							2	7	191		4.5

[a]Two trials lost for 11 subjects.

TABLE A-VIII. Destressed /p–p/ syllables: Point-vowel-precursor condition. Overall percent error: 28.63 (pooled), 28.89 (rep. 1), 28.37 (rep. 2); 90 trials, $\bar{x}=25.76$, $s=4.70$, $N=17$.

Intended vowel	Response										Percent error
	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	
i	151	6	1					2	10		11.2
ɪ		167						3			1.8
ɛ		2	164				3	1			3.5
æ			74	95		1					44.1
ɑ		1	2	1	7	6	151	2			95.9
ɔ					8	84	69	9			50.6
ʌ		1	3	1	1	13	123	25	3		27.6
ʊ					4	1	11	139	15		18.2
u					1		1	6	162		4.7

TABLE A-IX. Destressed /p–p/ syllables: Sentence-context condition. Overall percent error: 17.26 (pooled), 18.22 (rep. 1), 16.30 (rep. 2); 90 trials, $\bar{x}=15.53$, $s=5.08$, $N=15$.

Intended vowel	Response										Percent error
	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	
i	140	10									6.7
ɪ		149								1	0.7
ɛ			120	29		1					20.0
æ			2	147							2.0
ɑ				2	95	36	15	1		1	36.7
ɔ					41	103	3	3			31.3
ʌ				1	8	17	100	24			33.3
ʊ					1	4	20	115	10		23.3
u					1			1	148		1.3

*Requests for reprints should be addressed to Robert Verbrugge, Haskins Laboratories, 270 Crown Street, New Haven, CT 06510. A partial summary of these results was presented at the 87th meeting of the Acoustical Society of America, New York, 25 April 1974, and published in R. Verbrugge, W. Strange, and D. Shankweiler, "What information enables a listener to map a talker's vowel space?" Haskins Lab. Status Report on Speech Res. SR-37/38, 199-208 (1974). A more complete exposition of the problem of perceptual constancy in speech perception may be found in Shankweiler, Strange, and Verbrugge (in press).

†Present address: Department of Psychology, University of Connecticut, Storrs, CT 06268.

[1]It is important to note that the relationship between the scales on the horizontal and vertical axes is arbitrary. For example, if a vowel appears in the upper right-hand quadrant on a 45° line passing through the origin, this *cannot* be interpreted as an increase in correct responding which is "perfectly correlated" with the increase in false responding. In Figs. 1, 2, and 3, the aspect ratios have been chosen so that the *ranges* of values on each dimension are given roughly equal weight. It is also important to note that the differences plotted are *linear* functions of error scores. On either axis, the differences indicate the relative contribution of each vowel to the overall change in percent identification. However, the values plotted give no indication of the *proportionate* change in identification on each vowel. For example, if vowel x increased in correct identification from 50 to 55%, and vowel y increased from 94% to 99%, each would appear along the horizontal axis at +5%, though the proportionate improvement is larger for y. The primary goal of these figures is their heuristic value in visualizing relative directions of change in two variables. Choice of the linear transform should not be interpreted as a claim about what differences represent "equivalent" changes in the recognition system.

[2]The predicted frequency of identifying an intended vowel x as the response alternative y, e_{xy}, is defined by the formula:

$$e_{xy}=\beta_y \eta_{xy} n_x / \sum_{y=1}^{N} \beta_y \eta_{xy} \ ,$$

where N is the number of vowel categories (15 in Experiment I) and n_x is the total number of intended vowels which were presented (12 per subject in Experiment I). These "expected values" were estimated for each cell of the confusion matrices, using an algorithm developed by J. E. Keith Smith at the University of Michigan. At theoretical limit, the procedure outputs the set of maximum likelihood estimators for the observed pattern of errors. The xy similarity parameters were estimated as follows: $\eta_{xy}=(e_{xy}e_{yx}/e_{xx}e_{yy})^{1/2}$. Since $-\ln\eta_{xy}$ closely approximates a normal distribution, similarity parameters for two conditions may be compared using $t=2(\ln\eta_2-\ln\eta_1)/(V_1+V_2)^{1/2}$, where V is the estimated variance. A full development of this general procedure may be found in

Goodman (1969, 1970).

[3]The term "central vowel" is used only in contrast to "point vowel," not in the more restricted sense found in traditional phonetic taxonomies. Of the six central vowels so defined, a set of three with fairly wide dispersion in two-formant space were chosen for this condition.

[4]The man, woman, and child chosen as "representative" were individuals in each group of talkers whose test syllables produced a close-to-average number of errors on the mixed-talker test, and who were available for further recording sessions.

[5]Acoustic measurements of vowels in the mixed- and segregated talker tests are reported in a companion study (Strange, Verbrugge, Shankweiler, and Edman, 1976). Average formant frequency and relative duration values were comparable to those reported by Peterson and Barney (1952), Stevens and House (1963), and Peterson and Lehiste (1960).

[6]The shift to a /p–p/ consonantal frame apparently had little effect on the error rate for the nine vowels studied here. Errors on those nine vowels averaged 17.4% in /h–d/ syllables (with 15 response alternatives), compared to 17.0% in /p–p/ syllables.

[7]For ease of comparison, the goodness-of-fit for each model has been characterized by the ratio of the maximum-likelihood chi-square value to the number of degrees of freedom. Most of the chi-square values are significant, and the Luce models appear to be rejected. However, these significance tests assume that the observed frequencies manifest stable population probabilities. Analysis of the variability among subjects revealed significant heterogeneity in their responses to several vowel categories. Thus, the reported chi-square values reflect substantial heterogeneity among subjects, as well as deviations of the expected values from underlying population values. When adjustments are made for the observed heterogeneity, the fit of the Luce models is much improved. The unadjusted chi-square ratios provide a useful measure for present purposes, since the degree of heterogeneity was roughly constant across the experimental conditions being compared.

[8]Gay's (1974) acoustic measurements suggest that the critical feature of destressed syllables in natural sentences is that they are destressed, not that they are rapidly spoken. Point vowels in rapidly spoken syllables did not show the reduction toward schwa which is found in destressed speech (Lindblom, 1963). It is not clear what implications this has for the perceptual studies of Lindblom and Studdert-Kennedy (1967) or the studies presented here. In both cases, tempo variation has provided a plausible basis for explanation. Further research is needed to determine whether perceived pace and syllable duration are secondary to perceived stress in determining the pattern of listeners' identifications.

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

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

Gay, T. (1974). "A cinefluorographic study of vowel production," J. Phonetics 2, 255–266.

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

Goodman, L. A. (1969). "How to ransack social mobility tables and other kinds of cross-classification tables," Am. J. Sociol. 75, 1–40.

Goodman, L. A. (1970). "The multivariate analysis of qualitative data: Interactions among multiple classifications," J. Am. Stat. Assoc. 65, 226–256.

Helson, H. (1948). "Adaptation level as a basis for a quantitative theory of frames of reference," Psychol. Rev. 55, 297–313.

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

Ladefoged, P. (1967). *Three Areas of Experimental Phonetics* (Oxford University, New York).

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

Lieberman, P. (1973). "On the evolution of language: A unified view," Cognition 2, 59–94.

Lieberman, P., Crelin, E. S., and Klatt, D. H. (1972). "Phonetic ability and 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.

Lindblom, B. E. F., and Sundberg, J. (1969). "A quantitative model of vowel production and the distinctive features of Swedish vowels," Q. Prog. Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) STL-QPSR 1, 14–32.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis* (Wiley, New York).

Luce, R. D. (1963). "Detection and recognition," in *Handbook of Mathematical Psychology*, edited by R. D. Luce, R. R. Bush, and E. Galanter (Wiley, New York), Vol. I, pp. 103–189.

Peterson, G. E. (1961). "Parameters of vowel quality," J. Speech Hearing Res. 4, 10–29.

Peterson, G. E., and Barney, H. L. (1952). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. 24, 175–184.

Peterson, G. E., and Lehiste, I. (1960). "Duration of syllable nuclei in English," J. Acoust. Soc. Am. 32, 693–703.

Rand, T. C. (1971). "Vocal tract size normalization in the perception of stop consonants," Haskins Lab. Status Report Speech Res. SR-25/26, 141–146.

Shankweiler, D., Strange, W., and Verbrugge, R. R. (in press). "Speech and the problem of perceptual constancy," in *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, edited by R. Shaw and J. Bransford (Lawrence Erlbaum Associates, Hillsdale, NJ).

Shearme, J. N., and Holmes, J. N. (1962). "An experimental study of the classification of sounds in continuous speech according to their distribution in the formant 1-formant 2 plane," in *Proceedings of the Fourth International Congress of Phonetic Sciences* (Mouton, Hague), pp. 234–240.

Stevens, K. N. (1972). "The quantal nature of speech: Evidence from articulatory-acoustic data," in *Human Communication: A Unified View*, edited by E. E. David, Jr., and P. B. Denes (McGraw-Hill, New York), pp. 51–66.

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

Strange, W., Verbrugge, R. R., Shankweiler, D. P., and Edman, T. R. (1976). "Consonant environment specifies vowel identity," J. Acoust. Soc. Am. 60, 213–221.

Tiffany, W. R. (1959). "Nonrandom sources of variation in vowel quality," J. Speech Hearing Res. 2, 305–317.