Reprinted from WORD
Vol. 13, No. 2, August, 1957
Printed in France

MINIMAL CUES FOR SEPARATING /w, r, l, y/ IN INTERVOCALIC POSITION*

LEIGH LISKER

The experiments[1] to be described attempt to specify the major acoustic differences among the members of the set of American English phones /w, r, l, y/ in intervocalic position. They represent a first systematic effort to synthesize these consonants intervocalically by the use of the Pattern Playback.[2] The choice of consonants and acoustic variables was determined in the main by the results of a large number of preliminary experiments designed to answer a general question concerning the relation between a certain kind of acoustic shift and its phonetic evaluation. The question arose from the well-known observation that while some portions of the speech signal are relatively steady-state in respect to important acoustic features, nowhere is the signal steady-state for intervals of more than a few milliseconds in "ordinary speech." Moreover this holds true even for portions that are heard as vowels whose perceived quality is steady over substantially greater durations.

Since phonetically steady-state but acoustically shifting vowels are found in ordinary speech contexts, the question was raised

* This research was supported in part by the Carnegie Corporation of New York and in part by the Department of Defense in connection with Contract DA49-170-sc-1642.

[1] Concurrently a group of my colleagues at the Haskins Laboratories were working on the /w, r, l, y/ set in initial position. Their findings, reported in O'Connor, Gerstman, Liberman, Delattre and Cooper, "Acoustic cues for the perception of initial /w, r, l, y/ in English," *Word* 13 (in press), and those here presented, were reached in large measure independently of one another. Under the circumstance the almost complete agreement between the two sets of data may be taken as evidence of their reliability.

[2] The Pattern Playback has been described in several papers; see, for example, F. S. Cooper, "Spectrum analysis," *Journal of the Acoustical Society of America*, 22.761-762 (1950).

as to whether such acoustic shifts would be perceived as changes in vowel quality, or perhaps as consonants of one or another type,[3] if their contexts were independently varied. For test purposes this question was replaced by the much more manageable one: To what extent can an isolated synthetic vowel depart from the acoustic steady-state condition without concomitant phonetic shift? To answer this question both two- and three-formant vowel patterns, having durations of about 500 msec. each, were varied systematically in their middle portions with respect to the frequencies, intensities and durations of steady-state intervals and transitional movements of each of the formants. It turned out that any frequency departure from steady state had a perceptible effect, though not always one that would unquestionably be considered a phonetic one. Changes in duration and relative intensity were also perceptible, but only when they were of sizeable magnitude. One interesting finding was that when the acoustic shifts introduced were restricted in respect to certain features listeners limited their phonetic interpretations of these shifts to the /w, r, l, y/ set.

The fact that /w, r, l, y/ were heard when small pattern changes were introduced within steady-state vocalic stretches suggested that these speech sounds might be distinguishable by differences in a limited number of acoustic features or dimensions. Experiments were therefore carried out to determine the features needed to synthesize the four phones and to establish their boundaries in terms of these features. Each of the patterns tested consisted of five segments (Fig. 1), and in the preliminary experimentation each segment was studied in respect to the following properties: duration, formant frequency structure, and relative intensity of the formants.

The preliminary experiments showed that if segments 1 and 5 (Fig. 1) are each 150 msec. in duration, and if each of the medial segments is about 50 msec., then satisfactory vowel-/w, r, l, y/-vowel sequences can be synthesized. To be sure, these durations

[3] It has been shown that in initial position such shifts are interpreted as consonants of various types, depending on their duration, extent and starting frequencies. See Delattre, Liberman and Cooper, "Acoustic loci and transitional cues for consonants," *Journal of the Acoustical Society of America* 27.765-773 (1955); Liberman, Delattre, Cooper and Gerstman, "The role of consonant-vowel transitions in the perception of the stop and nasal consonants," *Psychological Monographs* 1954, 68, No. 8; Liberman, Delattre, Gerstman and Cooper, "Tempo of frequency change as a cue for distinguishing classes of speech sounds," *Journal of Experimental Psychology*, Vol. 52, no. 2, 1956.

represent a compromise in that they are best values for the set of sequences as a whole rather than for any particular one of them.[4] Durations of much less than the values selected for segments 2, 3 and 4 result in flap-like phones falling outside the /w, r, l, y/ set, while durations of much more than 150 msec. for these three segments combined are heard as geminates of /w, r, l, y/

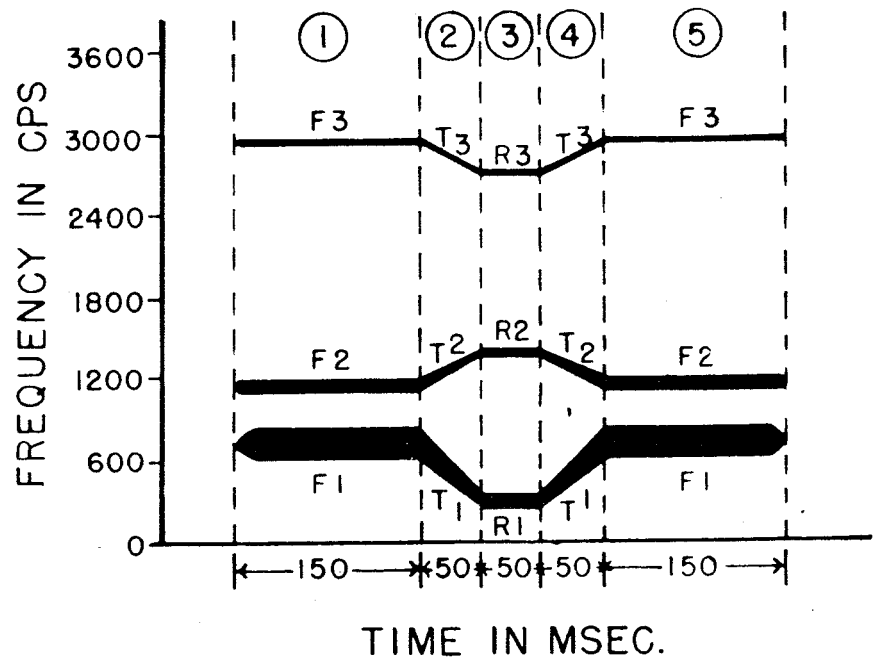


Fig. 1. Sample stimulus showing the five segments and the acoustic features which were manipulated in the experiments.

Formant frequencies for segments 1 and 5 (F1, F2, F3) were selected to yield good American English vowels (but not diphthongized!). Their values were generally set according to the findings of earlier studies.[5] In the case of the vowel /u/ however, the second formant frequency (F2) had to be raised to 840 cps in order to achieve a convincing /uwu/. Formant

[4] For example, segments 2 and 4 of somewhat briefer duration slightly improve the quality of /l/, but at the expense of the other phones of the set.

[5] These were: for /i/, 240-2520-3000 cps; for /a/, 780-1200-2520; for /u/, 240-720-2520 cps.

1 frequencies in segment 3 (R1) that yielded best /w, r, l, y/ center at 360 cps for both /i/ and /a/, and at 180 cps for /u/. Values of R1 below 180 cps sometimes yielded /m/ and /b/ judgments; values much above 360 cps resulted in phones of the /ə-a/ range. Formant frequencies of segments 2 and 4 (T1, T2, T3) were of constantly changing values (i.e. constant slope), with initial and terminal values (and rate of change, in view of the fixed duration) equal to the formant frequencies of the adjacent segments.

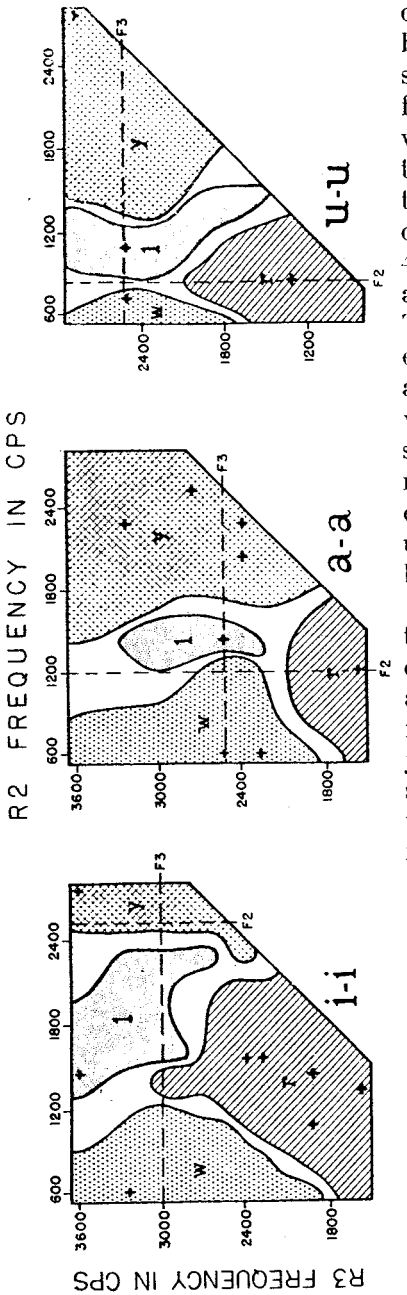
The relative intensities of the formants, both within segments and from one segment to another, were fixed in general accord with spectrographic evidence, in the absence of any more pressing criterion.

The two remaining features of the test patterns, the frequencies of R2 and R3, appeared to be both necessary and sufficient for distinguishing among the members of the /w, r, l, y/ set, given of course those features which mark off the set from other phones of the language.

It should be emphasized that the labelling convention of Fig. 1 is merely one possible way of describing the test patterns; it does not imply necessarily that segments 2 and 4 are of less importance than segment 3 in the perception of /w, r, l, y/. There are in fact reasons for rather describing R1, R2 and R3 as having frequencies determined by the "interior" terminal frequencies of the formants of segments 2 and 4—T1, T2 and T3 respectively; for example, a discontinuity introduced into segment 3 may produce phones belonging to the class of stops, but no features wholely contained within that segment serve in discriminating among the members of the /w, r, l, y/ group. Nevertheless I shall refer to R1, R2 and R3 as features of segment 3, in keeping with the convention of Fig. 1.

The extended experimentation that led to the selection of the last-named features involved typically the phonetic observations of a single listener (myself), with only occasional assistance from other listeners. Since the purpose of the enterprise was to establish phonemic boundaries,[6] the judgments were of significance

[6] Strictly, phoneme boundaries separate utterances which occur in a language and which differ both phonetically and semantically from each other; but they are not defined with regard to differences between nonsense sequences, which are by definition not part of the language. However, when subjects are asked to assign phones in nonsense sequences to categories already established as phonemes, it is not unreasonable to assume that the phonetic boundaries observed are the same as those which separate meaningful utterances in their language.



only insofar as they would be corroborated by a larger number of native speakers of American English. Therefore three sets of test patterns were drawn, one for each of the contexts /i-i, a-a, u-u/, and were submitted, as acoustic stimuli in randomized order, to a jury consisting of about 40 students of elementary psychology at the University of Connecticut. The jury was instructed to judge each stimulus as containing one or another of /w, r, l, y/ in the intervocalic position. Then for each stimulus the distribution of judgments among the four categories was examined in relation to the particular values assumed by the variables R2 and R3.

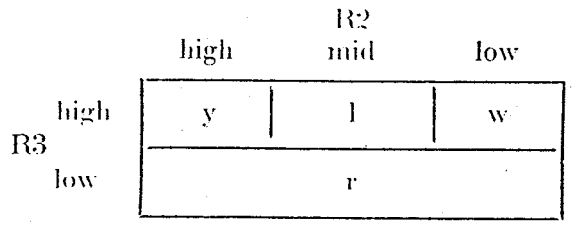
The test results are presented in two ways. In Fig. 2 zone maps for each vowel environment mark out areas of our two-dimensional space for which fifty percent or more of the jury were in agreement, while the small crosses within these areas mark values of R2 and R3 for which agreement as to the classification of the perceived phone was maximum. In Fig. 3 we have drawn the twelve patterns which best represent, on the

Fig. 2. Shaded areas include R2-R3 frequency pairs for which 50 percent or more of the jury was in agreement. For each context about 110 R2-R3 frequency pairs were chosen so as to cover the two dimensions uniformly. The number of listeners participating in the tests were: /i-i/, 33; /a-a/, 44; /u-u/, 45. Crosse indicate stimuli for which listener consensus was maximum.

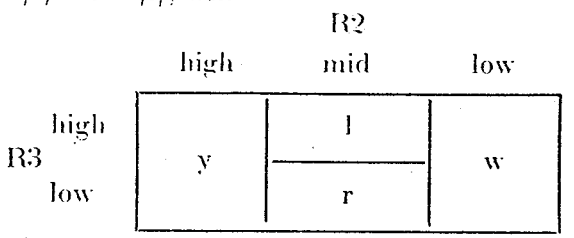
basis of percentage agreement, each of the four consonants with each of the vowel contexts.

From the percentages achieved by the best patterns it appears that they were not equally good approximations to speech, ranging as they do from one-hundred percent agreement for /iwi, awa, ara, aya, uyu/ to only seventy percent for /ili/.[7] Some such differences were perhaps to be expected in view of the very limited number of acoustic dimensions explored in the experiments. The fact remains, however, that I have at this point no real explanation for the relatively poor showing of the /l/ phones.

The maps of Fig. 2 indicate that the relations among /w, r, l, y/ as functions of R2 and R3 may be stated in more than one way. We may say that in each of the contexts /r/ differs from the other three phones in having a low R3, while /w, l, y/ are separable on the basis of their different R2 values, i.e.:



We may, on the other hand, state that R2 serves to divide /w, r, l, y/ into the three subsets /w/, /r, l/ and /y/, and that R3 distinguishes /r/ from /l/, i.e.:



Both statements are oversimplifications in that they pretend that the phonetic effects of R2 and R3 are more independent of each other than in fact they are. For example, the first chart would seem to imply that, given any particular frequency for R3, the

[7] Similarly /l/ is less successfully synthesized in initial position than are the other members of the set (see reference cited in fn. 1).

jury either heard only /r/ for all values of R2, or shifted from /w/ to /l/ to /y/ with rising R2. This is of course not true; from Fig. 2 (i-i) we see that for an R3 of 3000 cps all four phones were heard, being distinguished solely by their R2 frequencies. Moreover there exist values of R2 for which /l/ and /y/ were distinguished on the basis of their R3 frequencies. Both appear to be tenable, however, as statements of general tendency.

To decide which chart more correctly reflects the relative importance of R2 and R3, I tested patterns identical with those of the Fig. 1 type, but from which all third formants had been deleted. The jury heard the two-formant patterns as the following vowel-consonant-vowel sequences:

iyi	aya	uyu
iri	ala	ulu
iwi	awa	uwu

These data indicate that R3 information is not indispensable for distinguishing between /y/ and /w/ in any of the contexts examined, but that it is of crucial importance for the /r/-/l/ distinction. (The distribution of /r/ and /l/ relative to the vowel contexts suggests that the slope of T2, of one or both of segments 2 and 4, has some cue value for these phones.) R2 suffices to distinguish among /w/, /r/ or /l/, and /y/. The second chart is therefore preferred.

If we compare the three zone maps of Fig. 2 it is clear that the phone areas shift in position, shape and size, depending on the contexts. To some extent their positions maintain a fixed relation to the F2 and F3 values of segments 1 and 5, i.e. we may describe, for each phone, its range of appropriate R2 and R3 values as these relate to F2 and F3 respectively. Thus the /r/-region occupies an R3 range lying below F3, while in the R2 dimension it straddles F2; in the /i-i/ context, however, where F2 is at a very high frequency, this /r/-region has R2 values which center instead at about 1300 cps. R3 values for each of the /w/, /l/ and /y/-regions lie in a range which includes F3; in the R2 dimension, the /w/-region lies below F2, /l/ lies just above it, and /y/ is considerably higher yet. Again there is an exception in the case of /i-i/ with its high F2, where the lower boundary of the /y/-region is a bit lower than F2. If, instead of zones, we consider only the positions of the best phones (Fig. 2) relative to the vowel formants, then the relation between /w, r, l, y/ and F2 and F3 is even more striking.

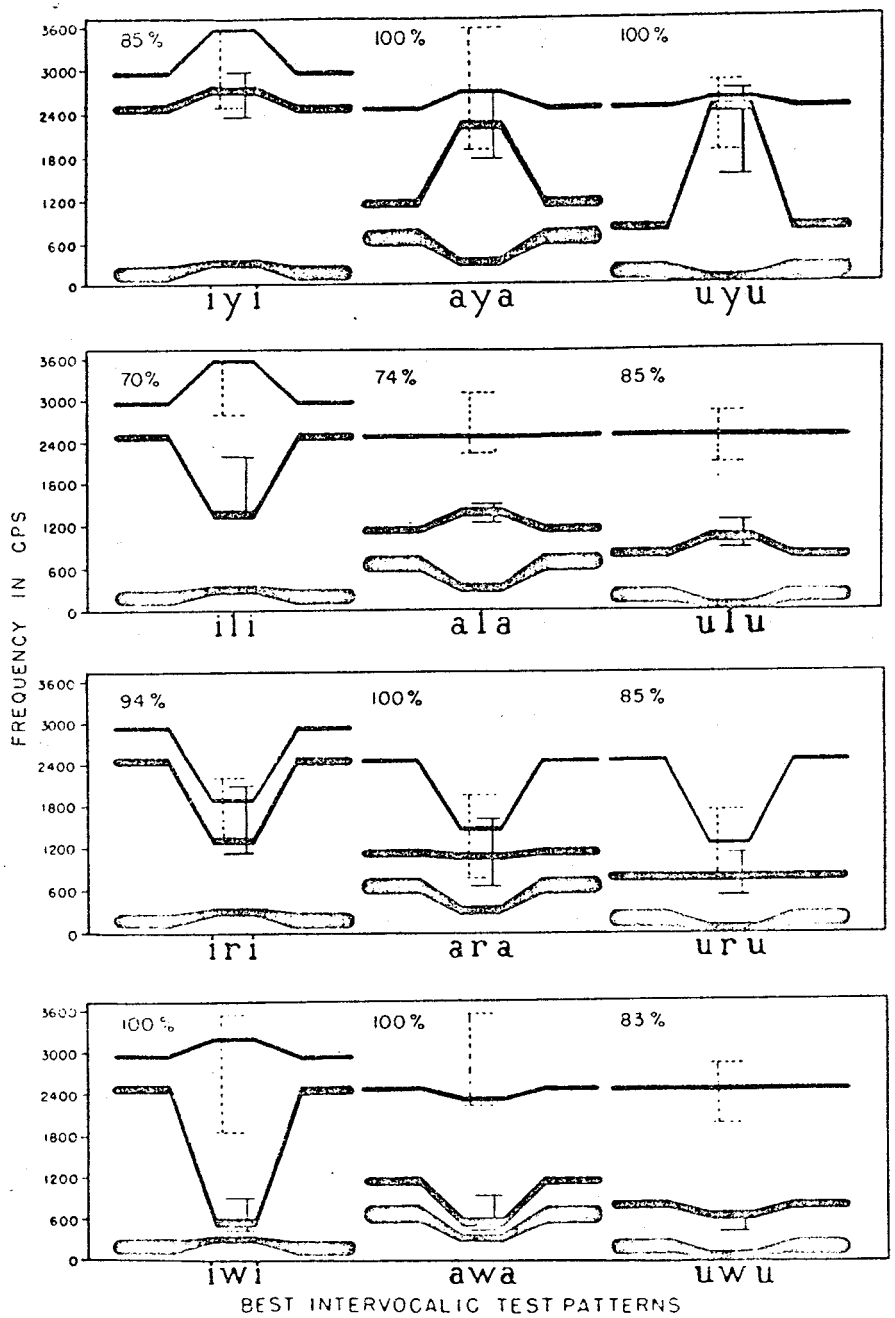


Fig. 3. Patterns for which listener consensus was maximum. (Where more than a single maximum was found the pattern shown has R2 and R3 values obtained by averaging the R2 and R3 frequencies of the maxima.) Also indicated are the frequency ranges over which R2 and R3 varied independently without reducing listener consensus below 50 percent. The percentages given are the actual values of the consensus maxima.

We have seen that the /w, r, l, y/ zones occupy R2-R3 positions which can in part be described as bearing a constant relation to the formant frequencies of neighboring vocalic segments. To determine the extent to which the positions of each of our phones are independent of context we next look for frequency areas which are common to a phone in all its contexts. In Fig. 4 we find the areas "www", "rrr" and "yyy", representing the intersections of all zones of each phone. For these regions 50 percent or more of the jury agreed in identifying segment 3 as the same phone in all three contexts. The approximate centers of these regions can be located at R2-R3 values of 2500-3100 cps for /y/, 850-1300 cps for /r/, and 600-2500 cps for /w/. Somewhat perversely, these values, which might be thought of as the "hard core" characteristics of /w, r, y/, do not lie near the best values marked in Fig. 2. Of more interest, however, is the fact that for /l/ there is no area common to its three zones, a fact which may be related to the difficulty already referred to in connection with this phone.

In Fig. 4 are included also areas that delimit frequency ranges identified with one of our phones by 50 percent or more of the jury, but where the phone identified was not the same one for all contexts. The existence of these areas demonstrates most forcefully the fact that R2 and R3 are not always sufficient of themselves to determine how a test pattern is perceived.

Inasmuch as the identification of the intervocalic phone depends in part upon the composition of segments 1 and 5 (and hence also of 2 and 4), it seemed interesting to discover how identification is affected by a context consisting of dissimilar vowels. Patterns of the kind already used in the /u-u/ test were altered by shifting F2 and F3 of segment 1 to values appropriate to /i/. These new patterns were played to a small group of listeners who made judgements of the consonant in the context /i-u/, and then in the context /u-i/ (by reversing the direction of movement of the patterns on the Playback). These judgments were then compared with the judgments obtained for the /i-i/ and /u-u/ situations. In all, 105 different patterns were evaluated in each of the asymmetric contexts. For 65 of the patterns the same judgments were recorded for both the /i-u/ and /u-i/ orders. In the remaining 40 cases the identifications varied with the order in which the vowels were heard. There were four kinds of such "double entendre":

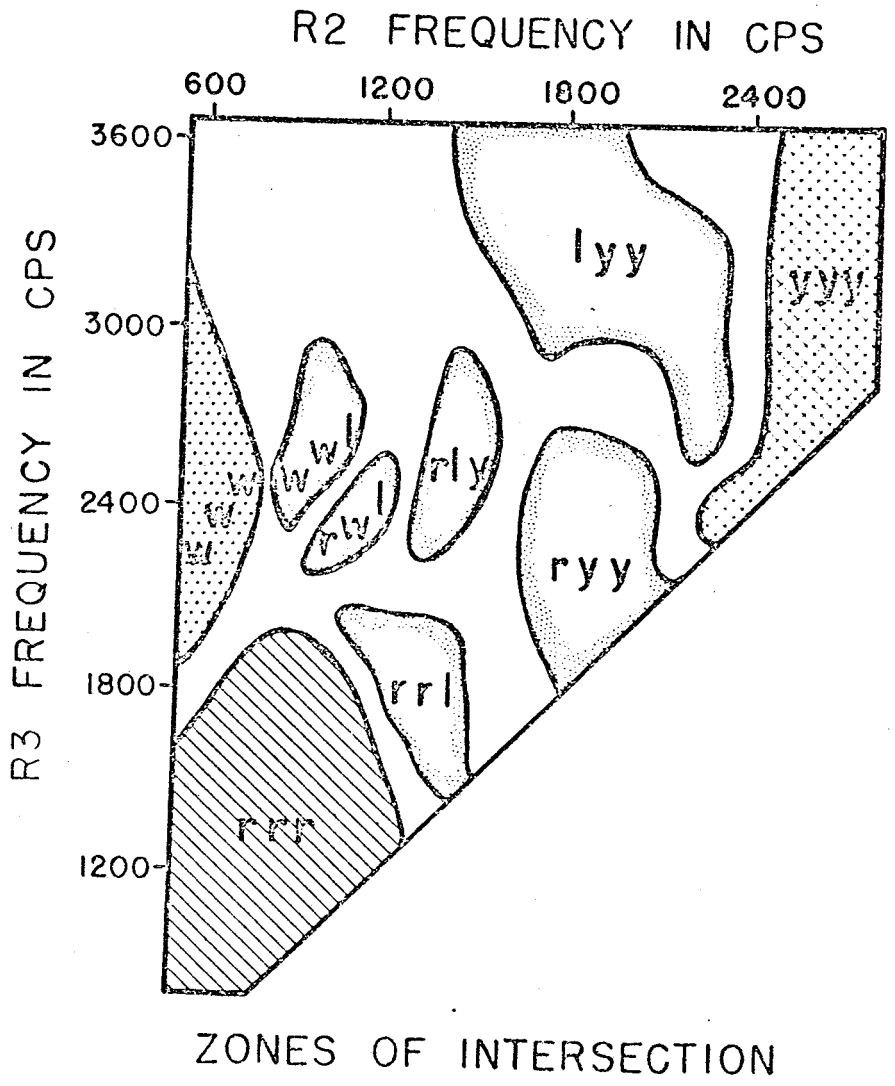


Fig. 4. Each enclosed area includes R2-R3 frequency pairs for which listener consensus was 50 percent or more in all contexts studied. The three letters labelling each area represent, in left-to-right order, identifications for the /i-i/, /a-a /and /u-u/ contexts. (E.g., the area marked "rly" includes R2-R3 pairs judged as /r/ in /i-i/, as /l/ in /a-a/, and as /y/ in /u-u/.)

uwi — iru
uwi — ilu
uri — ilu
uli — iyu

In 38 of these 40 cases the responses to /u-i/ were the same as the responses to the /i-i/ patterns having the same R2 and R3 values; in 30 of the cases responses to /i-u/ were the same as those to the corresponding /u-u/ patterns. In other words, for the majority of cases where identification was dependent in some measure upon the vowel order, the identifications can be accounted for on the basis of the relation between segments 3 and 5 (i.e. T2 and T3 of segment 4). This interpretation is in agreement with the findings of other researchers, which show that of the two transitional segments flanking an intervocalic consonant[8] it is the transiton following which contributes more to its identification.

Conclusion

All the test data show that intervocalic /w, r, l, y/ may be synthesized from acoustic patterns consisting of five segments each. For each segment a limited number of acoustic features require specification: duration; initial and terminal frequencies of three resonance bands. Certain of these features, in some of the segments, may be assigned values that are fixed for the entire set of phones; these are: all features of segments 1 and 5; durations of segments 2, 3 and 4; the frequency of the first resonance band of segment 3.[9] In addition, segments 2 and 4 may be assigned frequencies which are "automatic" in relation to the values assigned segments 1, 3 and 5. Alternatively we may call the following features fixed: segments 1 and 5 in toto; the durations of segments 2, 3 and 4. Frequencies of segment 3 may be described as determined by the terminal frequencies of segment 2 and/or the initial frequencies of segment 4. The second formulation has the advantage that it assigns features common to the set and features marking each member of the set to different segments; i.e. the features of segment 3 mark membership in

[8] More accurately, they are identifying, not segment 3, but the entire stimulus pattern.

[9] The frequency of R1 was constant for the entire /w, r, l, y/ set in each of the environments studied, but was not fixed for all environments.

the /w, r, l, y/ set, and the differences among /w, r, l, y/ are statable as differences within segments 2 and 4. In whichever way the patterns are described, all five segments must be specified in order to describe adequately each of the sequences synthesized.

The fact that five acoustic segments are needed to synthesize sequences of three phones reflects only what we already know—namely, that the phonetic evaluation of a segment is not made independently of its neighbors. Our data might suggest that segments 1 and 5 were judged without regard to their neighbors (i.e. "vowels are relatively stable"), but the structure of the experiments does not allow us to make any statements about those segments. So far as the /w, r, l, y/ judgments are concerned, it is quite clear that each of the segments makes some contribution. If we say that segment 3 (or segments 2 and 4, or segments 2 and 3 and 4) may be considered the acoustic counterpart of /w, r, l, y/, this can mean only that the particular segment or segments contain more information about the intervocalic articulation than does any other segment, not that all the information on this score is to be found in the named segments alone. The information on which the /w, r, l, y/ judgments were based is distributed, unevenly no doubt, throughout the five segments of the test patterns. It is accordingly impossible to segment the patterns into partials that are mutually independent in their phonetic consequences. To the linguist this kind of situation is familiar, and he handles it by considering segments which are explicitly not independent (allophones and allomorphs), and establishing, by distributional analysis, classes of these segments which show a much higher degree of mutual independence. Granting of course that the "phonetic meaning" of an acoustic segment must be ascertained before segment classes can be established, the same kind of treatment is obviously applicable to the /w, r, l, y/ patterns.

Haskins Laboratories and University of Pennsylvania

Reprinted from WORD
Vol. 13, No. 2, August, 1957.
Printed in France.