A PERCEPTUAL STUDY OF AMERICAN ENGLISH DIPHTHONGS*

THOMAS GAY

Haskins Laboratories, New Haven, Conn.

This study utilized synthetic speech for describing the phonemic boundaries and the effects of formant course on the identification of /ɔi, ai, au/. In the first experiment, synthesized formant transitions appropriate to classes of /ɔi-ai/ and /au-o/ were presented to ten phonetically trained listeners for purposes of phoneme labelling. Results showed that preferred /ɔi/ courses were from [ɔ] or [ʊ] to [i, y, ɪ], /ai/ courses from [a] to [ɪ] and /au/ from [a] to [ɔ].

The purpose of the second experiment was to determine whether the phonetic identity of the targets or the absolute course of the second formant transition serves as the primary identifying cue. These features were separated along the time dimension by synthesizing diphthongs whose second formant frequency course remained fixed but whose durations varied. Results of the listening tests showed shifts in perception from simple vowel to diphthong occurring as a function of duration rather than frequency onset or offset positions. Further, in the case of /ɔi-ai/, second formant rate of change serves as the primary distinguishing cue.

The results suggest /ɔi, ai, au/ are characterized by an invariant second formant rate of change, whose onset and offset values vary across changes in duration.

INTRODUCTION

Experimental data treating the acoustic cues used in speech sound perception have been reported for most sound classes of American English with perhaps only a few exceptions. One such case is the group of falling diphthongs, [ɔi, ai, au, ei, ou, ɪi, ʊu], which have been analysed acoustically (Lehiste and Peterson, 1961; Holbrook and Fairbanks, 1962), but not in terms of isolating those features which provide cues for recognition.

Diphthongs show a gliding movement along a particular path in the vowel space between zones appropriate to two different vowels. This gliding movement, which

* *This paper is based on a Ph.D. dissertation completed at the City University of New York under the direction of Professor Arthur S. Abramson, for whose guidance the author is deeply grateful.*

accounts for the major temporal portion of the diphthong, is evidenced by formants that course between the positions of the adjacent vowel zones. The purpose of this study was to determine the boundaries within which formants of the initial and terminal vowel areas of selected diphthongs contribute to the identification of each diphthong; and to investigate the glide movements, in terms of duration and frequency change, as perceptual cues for diphthong recognition. In this study, the choice of diphthongs was limited to /ɔi, ai, au/ because the diphthongal nature of each is phonemically distinctive in most dialects of American English. [ei, ou, ɪi, ʊu], on the other hand, alternate with non-diphthongal allophones of the phonemes /e, o, i, u/, respectively, thus suggesting that their off-glides carry no phonemic significance. Specifically then, this study was designed to answer the following questions:

(1) At what points along various acoustical continua are /ɔi/, /ai/, and /au/ resolved into three distinct phoneme categories?

(2) What is the phonemic status of the initial and terminal portions of these diphthongs?

(3) What are the differential effects of formant frequency change and duration as perceptual cues for /ɔi/, /ai/ and /au/?

General procedures

The present study is comprised of two experiments concerned with the phonemic boundaries and duration-recognition relationships of the three phonemically distinct diphthongs of American English, /ɔi, ai, au/.

The synthetic speech stimuli used in all experiments were produced by the Haskins Laboratories Pattern-Playback. The Pattern-Playback converts spectrographic patterns, hand-painted on acetate, into corresponding acoustic units by means of an optical transducing system. A light source is modulated by a rotating "tone-wheel" into 50 harmonics whose fundamental frequency is 120 c.p.s. The light is reflected by the spectrographic pattern to a photocell collector which transduces the optical elements into electrical and subsequent acoustic waveforms (Cooper, 1952). In this study various patterns were drawn, synthesized and recorded on to magnetic tape. Stimuli for each experiment were random-ordered into master lists by tape splicing techniques. All stimuli were preceded by the synthesized carrier phrase, "The word is—," set at intervals of 4 or 4.5 seconds (depending on the particular experiment). Stimulus lists were played back to subjects in group sessions through a loudspeaker at intensity levels of approximately 80 db. overall SPL. Testing was done in a quiet but not fully sound-treated room.

Subjects were ten undergraduate speech majors ranging in age from 18 to 20 years. All subjects were second generation born and raised New York City residents whose speech was typical of the dialect area. All subjects had some training in phonetics. Hearing loss was ruled out by routine audiometric screening.

Experiment I

This part of the study is concerned with the differences in formant frequency transitions responsible for separating /ɔi, ai, au/ into distinct phoneme categories. Preliminary investigation with stimuli synthesized on the Pattern-Playback revealed that such distinctions could be made along certain acoustical continua where important cues are provided by variations in the course and extent of the formants, especially the second formant. The /ɔi-ai/ distinction occurs along a continuum where second formant transitions course upward through time; /au/, on the other hand, is separated from /o/ along second formant continua that course downward through time.[1] In addition, further modifications of diphthong identification accompany changes in first formant movements and to a lesser degree, third formant movements. Although these continua produce sounds which are phonemically identifiable as /ɔi-ai/ or /au-o/, they do not specify for example, whether /ɔi/ is characterized by transitions which begin at formant positions appropriate to [ɔ] or extend to positions appropriate to [ɪ] or [i]. To determine the status of these targets, steady state vowels, with formants corresponding to the initial and terminal targets of all diphthong stimuli, were also synthesized.

Stimuli and test batteries

/ɔi-ai/ Continua

Fig. 1 illustrates the acoustical continua used to produce /ɔi, ai/ stimuli. Exploratory work found these ranges appropriate for either /ɔi/ or /ai/ perception without incurring other phonemic impressions. Five different second formant onset values ranging from 840-1320 c.p.s. were extended to terminal values of either 1920 or 2040 c.p.s. Each second formant was combined with two different first formant and two different third formant transitions; first formant transitions began at 600 c.p.s. and terminated at either 480 or 360 c.p.s. The two third formant transitions likewise began at one initial value, 2640 c.p.s., terminating at either 2520 or 2400 c.p.s. All patterns were drawn with duration of 250 msec. and band-widths three harmonics wide (each of the two side harmonics being of lower intensity than the centre frequency). The transitions in each pattern were drawn as straight bands from onset to termination. This enabled greater control of the transition course without incurring any loss in the naturalness of the sample.

In addition to the primary continua, supplementary patterns were drawn for purposes of enhancing both /ɔi/ and /ai/ perceptions. The supplementary patterns appropriate to /ɔi/ were constructed by adding lower first (480-360 c.p.s.) and third (2520-2400 c.p.s.) formants to the existing second formant continua in Fig. 1, except for deletions of all 1200 and 1320 c.p.s. onsets and any replications. The supplementary /ai/

[1] */o/ in this case occurs as the diphthongal variant* [ou], *which unlike* /ɔi, ai, au/ *is characterized by a non-phonemic off-glide.*

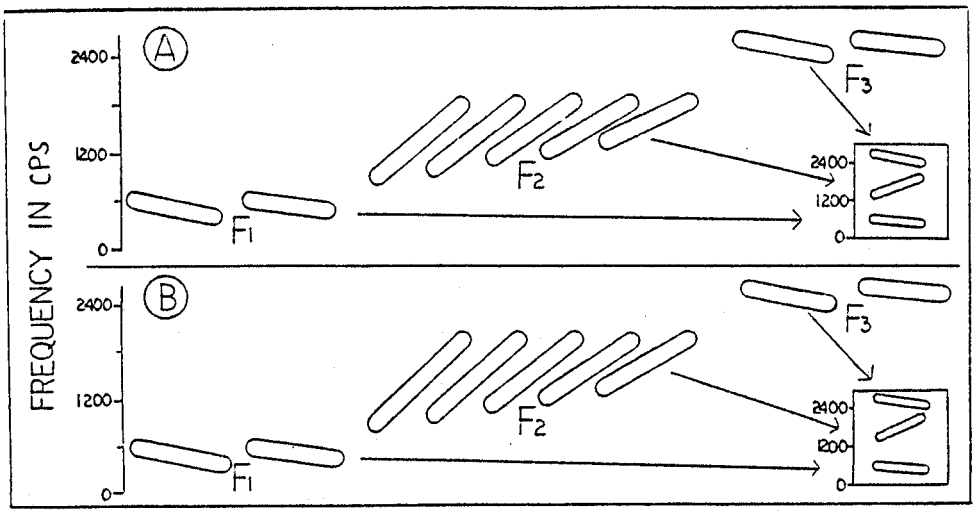

Fig. 1. Schematic illustration of stimuli used to produce primary /ɔi, ai/ continuum. A = second formant offset at 1920 c.p.s., B = 2040 c.p.s. Each first formant was combined with each second and third formant. Insets show complete spectrographic configurations.

patterns were drawn with two higher first formants (720-360, 720-480 c.p.s.) and two higher third formants (2760-2400, 2760-2520 c.p.s.) in combination with the continua in Fig. 1 except for the deletions of all 840-1080 c.p.s. second formant onsets and any replications. These supplementary continua provided a total of 30 additional /ɔi/ and 48 additional /ai/ stimuli.

/au-o/ Continua

For the /au-o/ continua, second formants began at 1080-1320 c.p.s. and terminated at either 960 or 840 c.p.s. First formant transitions of 600-480, 720-480, 720-600 c.p.s. and third formant transitions of 2400-2280, 2520-2280, 2520-2400 c.p.s. were each combined with the six second formant transitions to provide a total of 54 (6 × 3 × 3) stimuli. The actual mechanics of constructing the stimuli were identical to those described above. In addition, the ranges of formant frequencies comprising this continuum were sufficient to provide highly intelligible /au/ stimuli, eliminating the necessity for synthesizing supplementary patterns.[2]

[2] /o/ (= [ou]) *perceptions however, are not limited to these continua but are not explored further since absolute boundaries for* /o/ *are not of primary interest here.*

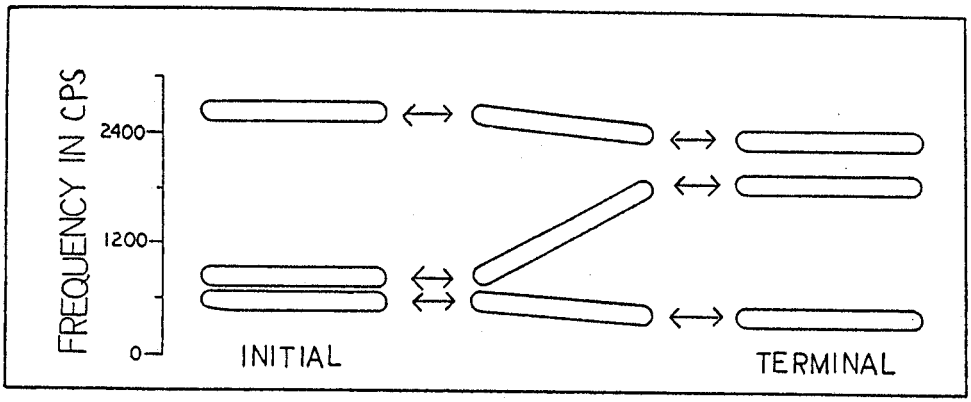

Fig. 2. Illustration of the procedure used for synthesizing initial and terminal target vowels.

Target Vowels

This set of stimuli consisted of steady state vowels whose values correspond to the intial and terminal targets of all diphthong patterns. An example of these patterns is shown in Fig. 2. Here the diphthong is comprised of a first formant of 600-360 c.p.s., a second formant of 840-1920 c.p.s. and a third formant of 2640-2400 c.p.s. The steady state vowel appropriate to the initial target of the diphthong has first, second and third formants of 600, 840, and 2640 c.p.s. respectively. The terminal target vowel had first, second and third formants of 360, 1920, 2400 c.p.s. This procedure was followed for all diphthong patterns in synthesizing a total of 32 initial and 16 terminal target vowels. All vowels were of 250 msec. duration and each formant consisted of a strong centre frequency bounded by two harmonics of lower intensity.

Test Batteries

A total of 172 diphthong stimuli were synthesized on the Pattern-Playback and recorded on to magnetic tape. Four recordings were made of each stimulus for purposes of providing as many replications in the test battery. All diphthongs were randomized into a master list by tape cutting and splicing methods. The synthesized carrier phrase, " The word is," was inserted 0.5 sec. before each diphthong and at successive intervals of approximately 4.5 sec. Listeners' responses were recorded on prepared forms. After hearing each sound, subjects first labelled it as one of the set /ɔi, ai, au, o, a/ and then rated it for quality on a " 1-5 " scale where " 1 " represented highest quality.[3]

[3] *Later analysis of the quality judgments showed them to provide little, if any, significant information, other than a slight positive relationship between higher quality ratings and higher diphthong identification for /ɔi, ai/ only. For this reason quality ratings were eliminated from subsequent test batteries.*

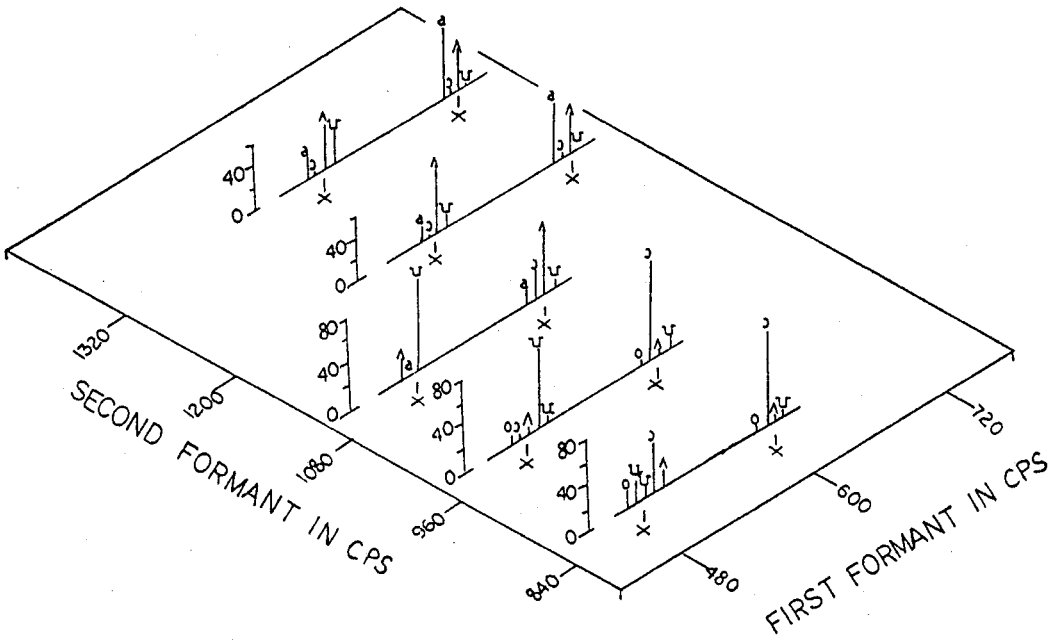

Fig. 3. Target vowel distributions for all initial /ɔi, ai/ stimuli. Data are plotted against appropriate first formant-second formant co-ordinates, in per cent.

Stimuli were arranged in groups of ten with brief rest periods scheduled after every group of fifty stimuli. Testing was carried out over two separate group sessions. Practice items, randomly selected from the master list, were presented before each experimental session.

Initial target and terminal target vowels were arranged in two separate test batteries. The procedures used in preparing the diphthong battery were generally followed with only a few exceptions. Carriers were set at approximately 4 sec. intervals and only labelling responses were obtained. Response choices were /a, ɔ, ʌ, o, ʊ, u/ for the initial target vowels and /i, ɪ, ɛ, u, ʊ, o, ɔ/ for the terminal target vowels. Both tests were presented in sequence during the course of one session.

/ɔi-ai/ Distinction

Figs. 3 and 4 show the initial and terminal target co-ordinates of all /ɔi-ai/ stimuli and the distributions of target vowel responses accompanying them. These co-ordinates and the majority preferences of the accompanying vowel distributions will be used later in plotting the formant movements of the various /ɔi-ai/ stimuli.

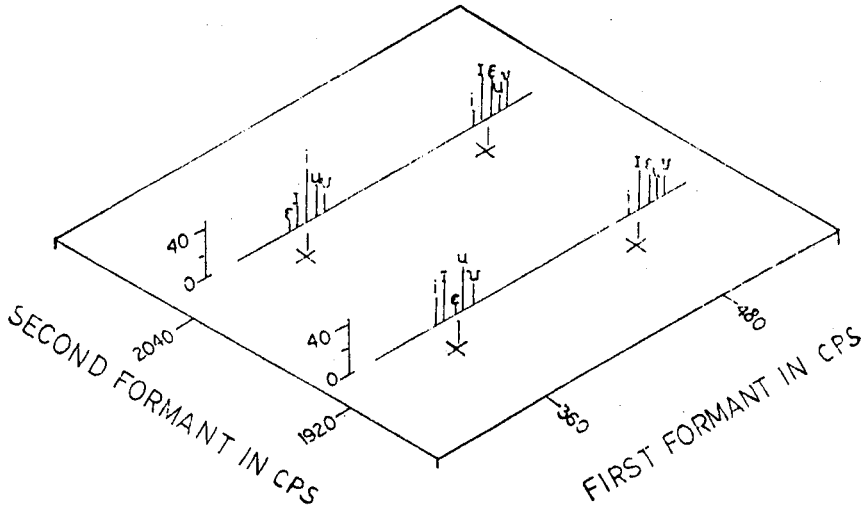

Fig. 4. Target vowel distributions for all terminal /ɔi, ai/ stimuli, in per cent.

The vowel distributions, as would be expected in a grid with closely positioned co-ordinates, show some degree of scattering. This is most evident for the first formant-second formant positions of 480-840, 600-1320, 720-1200 and 720-1320 c.p.s. Even with these scatterings however, clear, if sometimes small, phoneme majorities are evident. Majority preferences are in general alignment with those of earlier studies involving synthetic vowels (Delattre, Liberman, Cooper and Gerstman, 1952; Liberman, Delattre and Cooper, 1952) and with acoustic measurements of real speech (Peterson and Barney, 1952). Terminal target distributions, on the other hand, show widespread scattering across both front and back vowel categories. Of special interest is the majority of [u] responses for the 360-1920 c.p.s. position (Fig. 4). This sound gives an auditory impression of the high-front, rounded vowel, [y]. This impression is supported by Delattre, Liberman, Cooper and Gerstman's (1952) data on synthetic vowels which show first and second formant positions of 250 and 1900 c.p.s. as being most appropriate for [y] perception. Since [y] was not included in the response mode of the present study, it is suspected that subjects randomly assigned this sound to either a front or back vowel category. For these reasons then this co-ordinate might best be described as [y] rather than [u]. Likewise, other [u] and [ʊ] preferences at the three remaining co-ordinates are probably due to some [y] colouring. The front vowel confusions on the other hand, might be explained by their co-ordinates not being aligned with real speech measurements. The 360-2040 c.p.s. co-ordinate, which shows [i] as a majority preference, has a lower second formant than is usually found for real speech [i]. Also, the [ɪ] preferences, which are accompanied by relatively

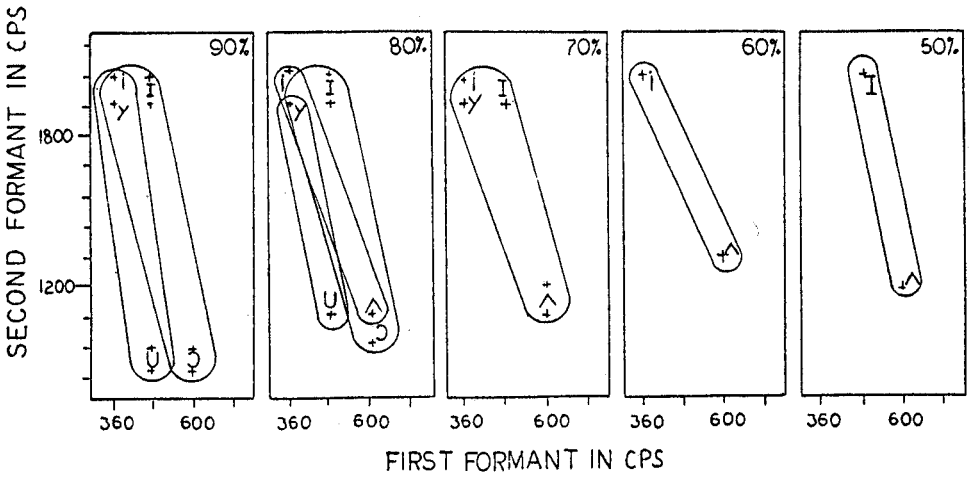

Fig. 5. Formant movements for different percentages of /ɔi/. Contours are based on the continua data and co-ordinate labels are based on the majority preferences of the target vowel distributions.

high [ɛ] preferences, occur at co-ordinates whose real speech values range between those of [ɪ] and [ɛ] (Peterson and Barney, 1952). It should be noted however, that these co-ordinates are aligned with those found for real speech [ɔi, ai].

Based on the majority preferences of the first formant-second formant co-ordinates, frequency tracts between targets appropriate to different percentages of /ɔi/ are plotted as contours in Fig. 5. These contours show that highly preferred /ɔi/ patterns (90-100 per cent) are not necessarily bounded by initial and terminal targets appropriate to [ɔ] and [ɪ]. /ɔi/ can course from either [ʊ] to [y, i] or from [ɔ] to [y, i, ɪ]. Other strong /ɔi/ preferences (80-90 per cent) are characterized by formants coursing from [ʌ] to [i]. In general, as /ɔi/ preference declines, initial targets shift to [ʌ], and terminal targets shift from [y, i] to [ɪ]. The formant movements of the highly preferred /ɔi/ contours generally enclose those routes established by acoustical measurements. Thus, it becomes apparent, from both these data and acoustical measurements, that formant movements appropriate to /ɔi/ recognition course between areas that enclose more than one specific vowel position. For at least the dialect area under study then, a more appropriate description of /ɔi/ might best be made by referring to general rather than specific initial and terminal target areas.

In Fig. 6, where /ai/ contours are plotted, strong preferences for /ai/ (80-90, 90-100 per cent) appear to have more specific boundaries. However, closer inspection of the appropriate initial and terminal targets (cf. Figs 3 and 4) reveals that vowel preferences for these positions are among those showing only small phoneme majorities.

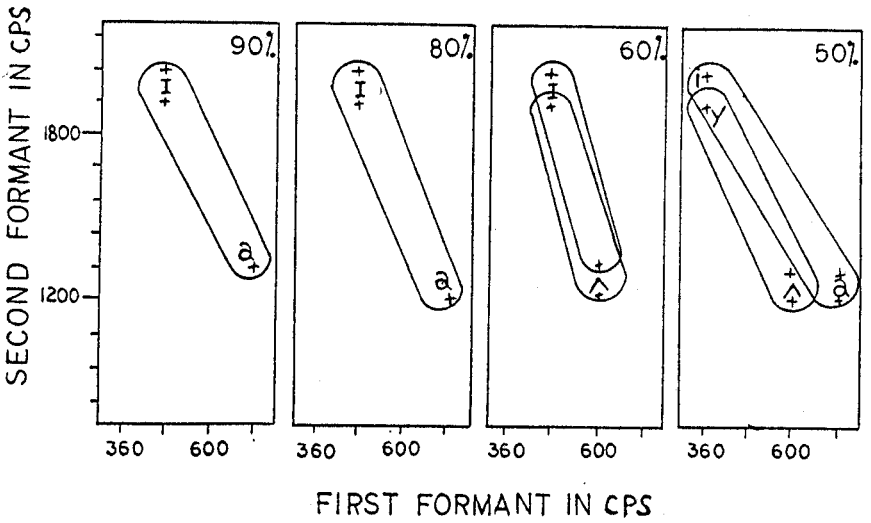

Fig. 6. Formant movements for different percentages of /ai/.

The appropriate [a] positions are accompanied by high [ʌ] responses and the [ɪ] positions are accompanied by high [ɛ] responses. Thus, the explicit status of these targets is perhaps doubtful and formant movements for /ai/, although restricted to a more limited course than those for /ɔi/, might also be described in more general terms. As /ai/ preferences decrease, the initial target moves from [a] to [ʌ]. The effect of a lower terminating first formant incurring lower /ai/ preference is shown here as a shift in terminal target from [ɪ] to [y, i]. Thus, whereas a glide from [a] to [ɪ] provides high /ai/ identification, a glide from [a] to [y, i] provides only marginal /ai/ identification. This effect of faster rate of frequency change has an interesting articulatory correlate. The tongue position for [i] is somewhat higher and slightly more forward that that of [ɪ]. Thus, a glide from [a] to [i] shows greater tongue movement than a glide from [a] to [ɪ] and consequently, if both sounds are of equal duration, the [a] to [i] articulatory movement occurs with greater speed than [a] to [ɪ] articulation. This also applies to the /ɔi/ movement. Since the tongue position for [ɔ] is farther back than that of [a], the articulatory speed of /ɔi/ is that much faster than any /ai/ articulation.

In general then, the shift from /ɔi/ to /ai/ occurs for initial targets that change from [ɔ] to [ʌ] to [a] and for terminal targets that change from [y, i] to [ɪ], a shift which is accompanied by a progressively slower rate of frequency change. Before discussing the phonemic implications of these data, the formant frequency characteristics of /au, o/ will be described, as these sounds bear relationships similar to those of /ɔi, ai/.

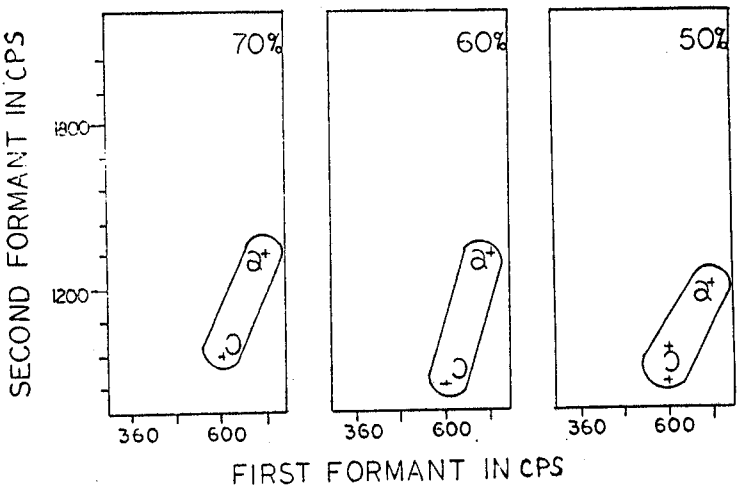

Fig. 7. Formant movements for different percentages of /au/.

/au-o/ Distinction

The initial and terminal targets of /au-o/ are the same as those for initial /ɔi-ai/ (except for the first and second formant co-ordinate of 480-1080 c.p.s.) and thus have distributions as shown in Fig. 3. There is however, an additional initial target used in the /au-o/ continuum. This is the co-ordinate at 720-1080 c.p.s. The responses at this position are distributed similarly to those at the 720-1200 c.p.s. position, showing a majority of [a] responses (62%), followed by smaller [ʌ] (30%) and [ɔ] (6%) judgments.[4]

The contours for /au/, which are plotted in Fig. 7, show that recognition of /au/ requires a glide from [a] to [ɔ]. Moreover, rather large gradations in /au/ identification occur for different formant movements enclosed within the general [a] to [ɔ] areas. Highest /au/ recognition occurs for highest initial and terminal first and second formants with progressively lower /au/ preferences accompanying lower first and second formants. The limitation of the terminal portion of /au/ to [ɔ] is supported by Holbrook and Fairbanks' (1962) acoustical measurements which also found /au/ to terminate squarely at [ɔ]. Potter and Peterson's (1948) data are not so explicit, showing termination beyond [ɔ], and Lehiste and Peterson's (1961) analysis shows /au/ termination closer to but not at [ʊ] (their data also shows /au/ initiation closer to [ʌ] than [a]). Yet, these areas are more restricted than those for /ɔi, ai/.

[4] *This position is probably more appropriate to* [ɑ] *than it is to the more fronted* [a]. *However, since these sounds vary only allophonically, listeners were not required to discriminate between them.*

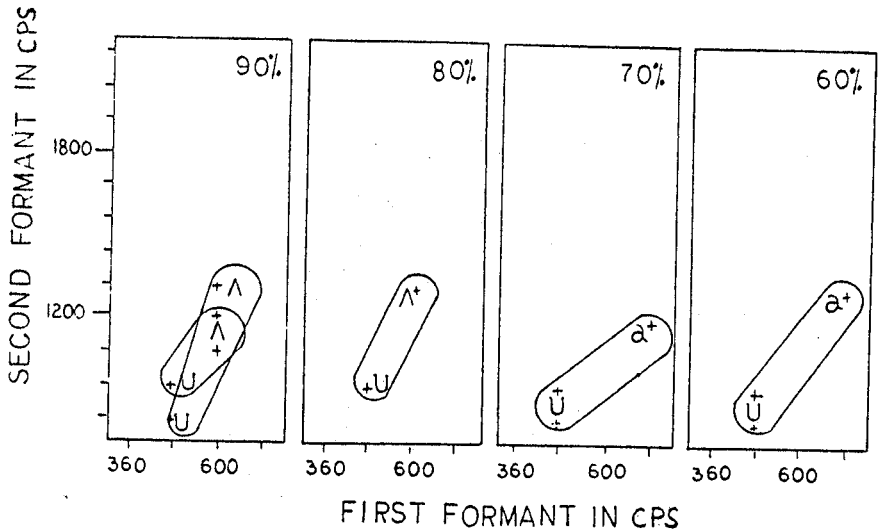

Fig. 8. Formant movements for different percentages of /o/.

The effects of a lower terminating first formant on the /au-o/ distinction are demonstrated clearly in the contours for /o/ (Fig. 8). Here, the various preferences for /o/ are all bounded by terminal targets appropriate to [ʊ]. Initial targets, on the other hand, range from [ʌ] to [a], with as much as 70 per cent /o/ intelligibility occurring for formants that course from [a] to [ʊ]. Thus, whereas /au/ is characterized by formants that course from [a] to [ɔ], /o/ is characterized by formants that course from [ʌ] to [ʊ] or [a] to[ʊ]. Although these data show that rate of formant change is generally greater for /o/ than /au/ (the glide for /o/ travels a greater distance, beyond [ɔ] to [ʊ], during a given period of time), real speech measurements would probably show the reverse is true. In this study, a complete range of /o/ configurations was not constructed. Thus, it is not unexpected to find that the formant movements for /o/ are not in agreement with those of real speech. Interestingly enough however, the highest preferred /o/ patterns at least, seem to be higher frequency extensions of real speech [ou], i.e. the synthesized /o/ formants begin and terminate at higher frequencies but course in the same direction as [ou], with the terminal targets of the /o/ patterns approaching the initial targets of real speech [ou]. Both phonetic and articulatory comparisons between /au-o/ are further complicated by the fact that [ou] occurs as a non-phonemic variant of /o/ and thus has an off-glide of no phonemic significance, as opposed to the off-glide role in /au/.

The results of this experiment have a certain relevance to the major analyses of /ɔi, ai, au/ and, in addition, provide some basis for developing an overall descriptive account of these sounds.

Discussion

The theories that /ɔi, ai, au/ consist of either sequences of vowel plus vowel or vowel plus semivowel require certain acoustic phonetic evidence, much of which is not found in the results of this experiment.

In accordance with the vowel plus vowel theory, /ɔi, ai, au/ would each necessarily contain steady state initial and terminal targets. The stimuli used in this study contained neither, thus showing the non-essential characteristics of such steady states for diphthong identification. However, some form of steady state target is usually found in real speech, but these steady states do not apparently constitute actual vowel sequences. Holbrook and Fairbanks' (1962) data show a steady state present only at onset, not at termination. Lehiste and Peterson (1961) found that /ɔi, ai, au/ "usually" contain both initial and terminal steady states; but the actual status of these steady portions seems questionable in the light of the criteria used for classification. According to their definition, a target of at least 20 msec. is classified as a steady state; however, whether a target of this duration is sufficient for describing a steady state, especially in an utterance totalling as much as 370 msec. duration, seems doubtful. Apparently then, since steady targets are neither required for perception nor found consistently in real speech measurements, a description of /ɔi, ai, au/ as actual sequences of two vowels does not appear justified.

In the vowel plus semivowel theory, a different set of acoustic and phonetic variables is encountered. Here, phonological significance is assigned to a post-vocalic glide, either /j/ or /w/, instead of a terminal vowel or vowel area. Since these glides are posited as allophones of pre-vocalic /j, w/, their gliding movements, both articulatory and acoustic, are similar in course to initial /j, w/, only in reverse order. Thus, for purposes of comparison, the formant characteristics of initial /j, w/ can be aligned with the formant characteristics of terminal /ɔi, ai, au/. One such comparison was made by Lehiste (1964) who found that the target frequencies of initial /j, w/ are not compatible with the terminal target frequencies of /ɔi, ai, au/. For purposes of this study however, a more appropriate comparison might be made with the analysis by O'Connor *et al.* (1957) of initial /j, w/, inasmuch as their study was based on perception and stimuli were produced by the Pattern-Playback. Here too, there are several features of initial /j, w/ not characteristic of /ɔi, ai, au/. First, neither first nor second formant termination values of reversed /jɔ, ja, wa/ correspond to those of /ɔi, ai, au/. First formant offsets for both /j, w/ occur at about 240 c.p.s., while terminal target frequencies range from 360-480 c.p.s. for /ɔi, ai/ and are fixed at 600 c.p.s. for /au/. The second formant terminal values for /j/ are higher than the terminal values for /ɔi, ai/, 2760 c.p.s. v. 2040 and 1920 c.p.s. and the second formant termination of /w/ is lower than that of /au/, 600 c.p.s. for /w/ as opposed to 960 c.p.s. for /au/. Thus, since the terminal frequency positions of hypothetical /ɔi, aj, aw/ are not compatible with those of /ɔi, ai, au/, the suitability of a vowel plus semivowel

description of /ɔi, ai, au/ seems doubtful.[5] In addition, another obvious difference between the sets of patterns occurs along the time dimension. The durations of /j, w/ are approximately 100 msec. as compared to diphthong durations of 250 msec. The effect of this difference is apparently quite relevant. Both O'Connor *et al.* and Liberman *et al.* (1956) found that transition duration acts as a primary cue for separating different sound classes. Specifically, their data show that longer durations of /j, w/ plus a vowel incur an auditory impression of a vowel of changing colour, e.g. a shift from /jɛ/ to /iɛ/ and /wɛ/ to /uɛ/. Thus, the impression for hypothetical /ɔj, aj, aw/ would apparently be distinct from /ɔi, ai, au/ insofar as the duration of /j w/ would not be great enough to produce diphthongal quality.[6] Further, if these glide durations were equal to those shown for /ɔi, ai, au/, the glide itself, without the preceding steady state vowel, would carry the diphthongal quality.

Apparently then, the primary feature of /ɔi, ai, au/ is a gliding movement which in itself is sufficient for providing diphthongal quality. In this experiment, these gliding movements have been described primarily in terms of onset and termination frequencies. However, a glide is also bounded by movement through time, with the rate of this movement bearing a direct relationship to either the duration of the glide (if the target levels are fixed) or the frequency levels of the targets (if duration is fixed). In this experiment, duration was the fixed feature and consequently, /ɔi, ai, au/ each differed in terms of both rate of formant movement and initial and terminal frequency levels. Thus, whether /ɔi, ai, au/ are recognized as such by consequence of their target frequency positions or rate of formant movement cannot be stated without first separating these features along the time dimension.

Experiment II

In the previous experiment, /ɔi, ai, au/ were identified by differences in the course and extent of formant frequency transitions. Accordingly, /ɔi, ai, au/ were each shown to be characterized by a specific course of formant movement and a particular set of phonetically describable targets. The purpose of the next experiment was to determine whether perception of /ɔi, ai, au/ is determined explicitly by the phonemic identity of these targets, with the rate of frequency change of the transition between serving no phonemic role, or by the rate of frequency change of the transition, the phonemic identity of the targets being only consequential.

[5] *It should be noted however, that since post-vocalic and pre-vocalic* [j, w] *are posited as allophones of* /j, w/, *hypothetical* [ɔj, aj, aw] *need not necessarily be acoustic "mirror images" of* [jɔ, ja, wa], *thus suggesting that only tentative conclusions can be based on syllable reversals of this type.*

A similar feature was observed informally by playing the present /ɔi, ai, au/ *tapes backwards. The auditory impressions were* [ɪɔ, ɪa, ɔa] *rather than* [jɔ, ja, wa].

The extension of the formant transitions of /ɔi, ai, au/ toward termination, is controlled by the time dimension which governs frequency change and, at any given point in time, frequency position. Thus, elimination along the time dimension, of either the initial or terminal portions of the transition, concurrently produces a change in the frequency position of the target with no accompanying change in the rate of frequency change of the transition. Exploratory work has shown that progressive reduction of transition duration of /ɔi, ai, au/ causes a shift in perception from diphthong to simple vowel. Whether this shift is a function of transition duration alone or the frequency position at the cut-off (that is, rate of formant change or target position), is the major concern of this experiment.

Stimuli and test batteries

Stimuli

The stimuli used in this experiment consisted of three groups of synthesized patterns each based on the spectrographic configuration most appropriate for /ɔi, ai, au/ identification. In each group, the full duration pattern was, in effect, reduced in duration from 250 msec. to 100 msec. in steps of 10msec., in one case beginning at the terminal target and in the other, beginning at the initial target.[7] Fig. 9 illustrates the procedure used in constructing patterns based on the basic /ai/ configuration. Since the course of the second formant transition shows the greatest rate of frequency change and is primarily responsible for the separation of vowel from diphthong, the control of its time-frequency characteristics constituted the experimental variable. Thus, all first and third formants were drawn as steady states, a procedure which did not affect diphthongal quality. Full duration /ai/ consisted of steady state first and third formants of 720 and 2760 c.p.s. and a second formant which extended from 1320 to 1920 c.p.s. The top row of Fig. 9 shows the patterns for which second formant transitions extend progressively higher through time, until extension is completed at 250 msec. In each case, the rate of change of the second formant transition remains fixed. This series of patterns, in which the onset position of the second formant transition remains preserved throughout changes in duration, produces a shift in perception from /a/ to /ai/. In the second series of patterns the terminal target frequency remains fixed through time, producing a shift from /ɛ/ to /ai/.[8] In each series, duration was varied from 100 msec. to 250 msec. in steps of 10 msec., producing a total of 32 patterns appropriate to shifts from vowel to /ai/.

[7] *The full duration patterns in this experiment are identical in duration to the stimuli of Experiment I.*

[8] *The impression of /ɛ/ rather than /ɪ/ is apparently due to the influence of the higher terminating steady state first formant.*

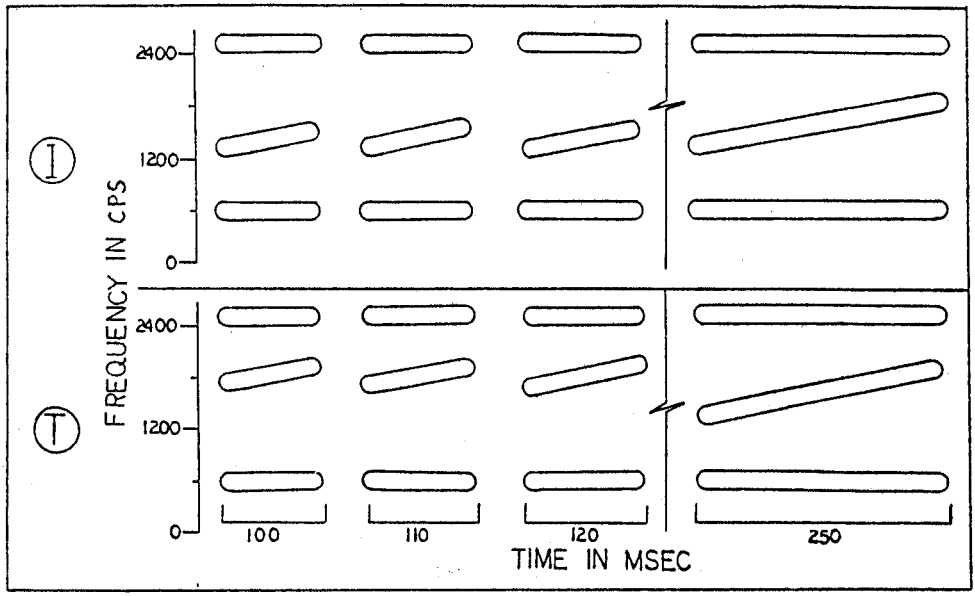

Fig. 9. Schematic illustration of stimuli used to produce /a-ai/ shift. I = patterns whose second formant onsets remain fixed, T = patterns whose second formant offsets remain fixed. In each case, duration is varied from 100 msec. to 250 msec., providing a shift in offset frequency (I) or onset frequency (T) positions with rate of frequency change remaining constant.

The procedures used for varying the duration of /ai/ were followed in varying the duration of /ɔ, au/. Full duration /ɔi/ consisted of steady state first and third formants of 600 c.p.s. and 2520 c.p.s. and a second formant extending from 840 to 2040 c.p.s.; /au/ consisted of 720 and 2400 c.p.s. first and third formants and a 1320-960 c.p.s. second formant.

Test batteries

The procedures used in preparing the /ɔi, ai, au/ stimuli of Experiment I were generally followed here. Fixed-onset and fixed-offset stimuli were arranged in two separate lists. Each list consisted of all appropriate /ɔi, ai, au/ stimuli, replicated 4 times each, for a total of 192 (16 × 3 × 4) test items. The synthesized carrier, " The word is," was inserted 0.5 sec. before each stimulus and at successive intervals of approximately 4 sec. Subjects were required to label the fixed-onset series stimuli as /ɔi, ai, au, ɔ, a/ and the fixed-offset series stimuli as /ɔi, ai, au, ɛ, ʌ/ on prepared

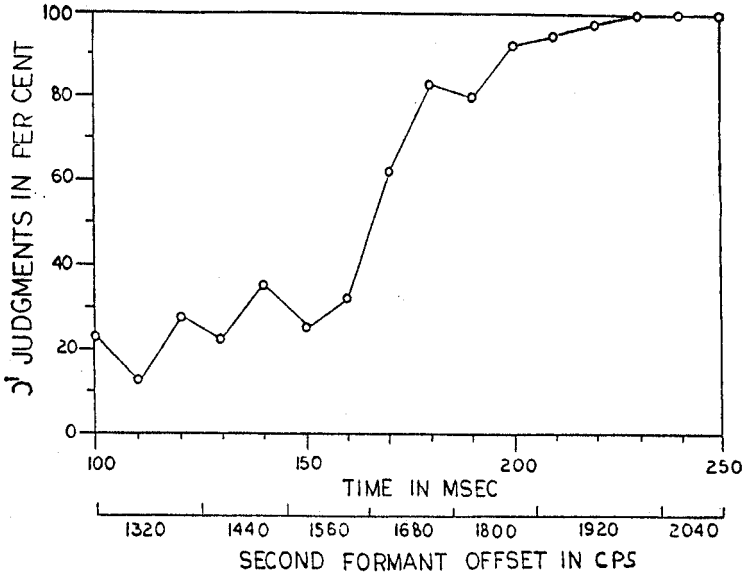

Fig. 10. Effects of duration on /ɔ-ɔi/ shift. Second formant offset varies from 1320-2040 c.p.s. while onset remains fixed at 840 c.p.s. Per cent /ɔ/ = 100 minus per cent /ɔi/.

answer forms.[9] Subjects were given practice items before each test battery. Both tests were presented in sequence during the course of one group session.

Time-frequency effects for /ɔi/

Results of both stimulus series for /ɔi/, shown in Figs. 10 and 11, indicate that duration rather than frequency position provides the primary perceptual cue for the simple vowel-diphthong separation.[10] For the fixed-onset patterns, the /ɔ-ɔi/ shift (more than 50 per cent /ɔi/ recognition) occurs at 170 msec. and a corresponding second formant centre frequency of 1680 c.p.s. Although this shift accompanies a change in frequency

[9] *Nine of the original ten subjects participated in this experiment. The additional tenth however, also met the residence and speech requirements described earlier.*

[10] *For all /ɔi, ai, au/ series, subject responses, except where otherwise noted, were of the two-category type,* e.g. */ɔ/* or */ɔi/. Thus, all graphs are plotted as per cent diphthong recognition (with per cent vowel recognition equalling 100 minus per cent diphthong recognition).*

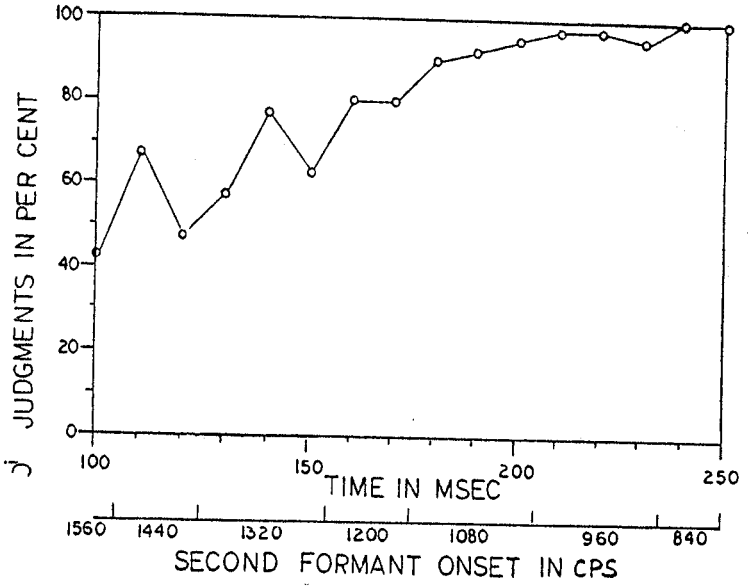

Fig. 11. Effects of duration on /ɛ-ɔi/ shift. Second formant onset varies from 1560-840 c.p.s. while offset remains fixed at 2040 c.p.s.

from 1560-1680 c.p.s., /ɔi/ recognition does not level off at this frequency position but rather increases across it through time. The increase in /ɔi/ recognition across this and other second formant cut-off frequencies would not be expected if the cut-off frequency position were the primary cue. The second formant position at 1680 c.p.s. in combination with the first formant of 600 c.p.s. is appropriate to acoustic positions for /æ/. The glide at this point then, courses from [ɔ] to [æ] but still provides greater /ɔi/ than /ɔ/ intelligibility. As the shift to /ɔi/ continues, second formant termination approaches acoustic positions more appropriate to [ɛ].

The /ɛ-ɔi/ shift, shown in Fig. 11, occurs earlier in time than does the /ɔ-ɔi/ shift, at 130 msec. This curve also shows that the effect of duration is greater than that of onset frequency position. The second formant onset frequency at 130 msec. is 1320 c.p.s., which in combination with the 600 c.p.s. first formant was phonetically described earlier as [ʌ] (cf. Fig. 3). Thus, the targets for these patterns are appropriate to [ʌ] and [ɪ], with the glide between providing approximately 66 per cent /ɔi/ intelligibility across the duration range of 130-150 msec. /ɛ/ identification for this group of patterns is approximately 30 per cent and /ai/ identification, 4 per cent.[11]

[11] *All* /ai/ *judgments, which never exceed 7.5 per cent for this series, were obtained from one subject whose responses in the previous experiment were also out of line with the other subjects' responses.*

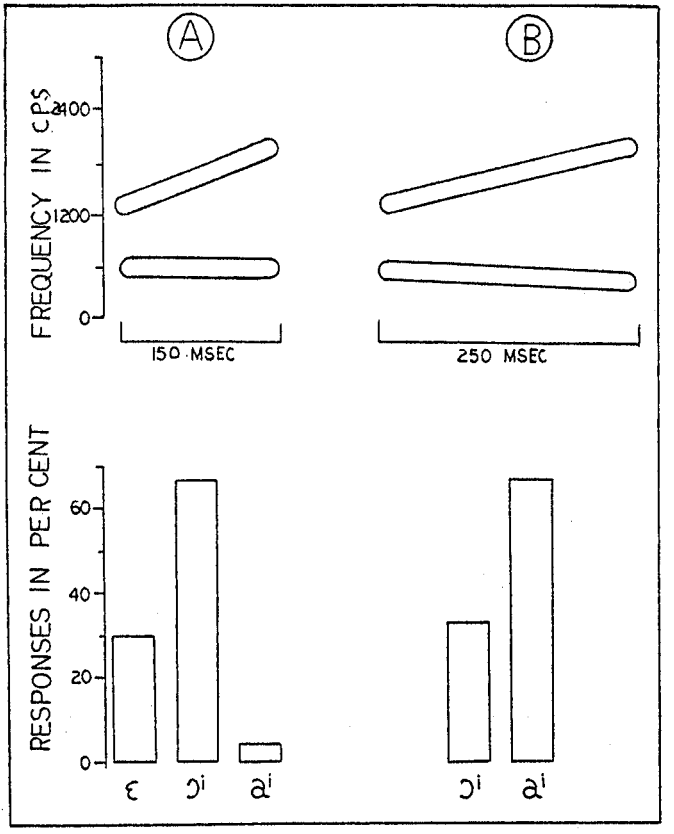

Fig. 12. Comparison of stimuli and responses, A = Experiment I, B = Experiment II.

In the previous experiment however, judgment for a glide between these same two targets favoured /ai/ over /ɔi/, 67 per cent to 33 per cent (cf. Fig. 6).[12] Fig. 12 illustrates this situation in which two patterns originating at similar, if not almost identical, formant frequency positions but differing in absolute duration and more importantly, rate of formant frequency change, are recognized as two separate phonemes. This separation clearly demonstrates that rate of formant frequency change is primarily responsible for separating /ɔi/ from /ai/. The effect is further evident at the 1200 c.p.s. position where /ai/ judgments, although somewhat less expected than at the 1320 c.p.s. position, are nonetheless, virtually absent.

[12] *The spectrographic configurations of these earlier patterns differ from the duration patterns only in that the latter contain a steady state first formant, the effect of which however, should only serve to enhance* /ai/ *perception.*

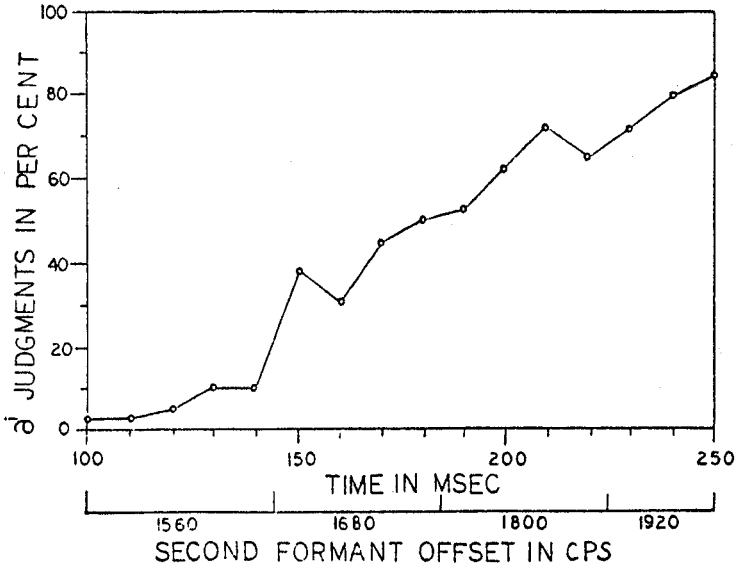

Fig. 13. Effects of duration on /a-ai/ shift. Second formant offset varies from 1560-1920 c.p.s. while onset remains fixed at 1320 c.p.s.

The results of this series of patterns clearly show that the second formant course for /ɔi/ need not necessarily begin or terminate explicitly at the bounding targets for the diphthongal quality of /ɔi/ to be perceived. /ɔi/ is perceived as a diphthong by consequence of duration and further, is separated from other diphthongs, notably /ai/, by its greater rate of formant frequency change.

Time frequency effects for /ai/

The effects of duration for /ai/, shown in Figs. 13 and 14, are similar in nature to those for /ɔi/. The /a-ai/ and /ɛ-ai/ shifts are clearly time based, consistently progressing across the ranges of different cut-off frequencies. The /a-ai/ shift occurs at 180 msec., with the transition at this point extending between acoustic positions appropriate to /a-æ/. Strong /ai/ preferences require the full 250 msec. duration. The /ɛ-ai/ shift is consistently stronger through time than the /a-ai/ shift although the 50 per cent point occurs only 10 msec. earlier, at 170 msec. The onset target configuration at this point has a slightly higher second formant than that appropriate to /a/ but is not high enough to approximate /æ/. These patterns also show that /ai/

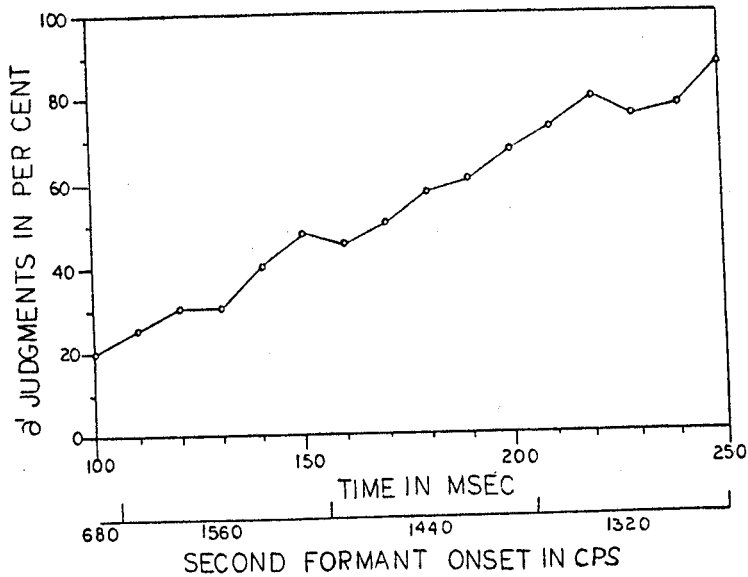

Fig. 14. Effects of duration on /ɛ-ai/ shift. Second formant onset varies from 1680-1320 c.p.s. while offset remains fixed at 1920 c.p.s.

preferences increase to a maximum at full duration. Both the fixed-onset and fixed-offset curves for /ai/ rise at a slower rate than those for /ɔi/, the difference perhaps being attributable to the greater rate of frequency change characteristics of /ɔi/.

Time-freouency effects for /au/

Figs. 15 and 16 show the effects of duration on the perception of /au/. The shift from /a/ to /au/ is rather sharp, occurring at 150 msec, with sharp increases in /au/ continuing until 170 msec. before slowing down. Here, as expected, the shift is clearly time based. Strong /au/ preferences occur at about 3/4 full duration, a duration similar to that for strong /ɔi/. The /ʌ-au/ curve also rises sharply, showing the shift to /au/ occurring at 160 msec. and strong /au/ preferences at 190 msec. or also at close to 3/4 duration. As was shown in the continua data, /au/ is bounded by specific formant positions with reductions in /au/ preference accompanying even small changes in first and second formant onset and termination frequencies. In this stimulus series however, the effect of a higher terminating steady state first formant has no adverse effect on /au/ preference, but rather perhaps enhances /au/. Also, for both fixed-onset

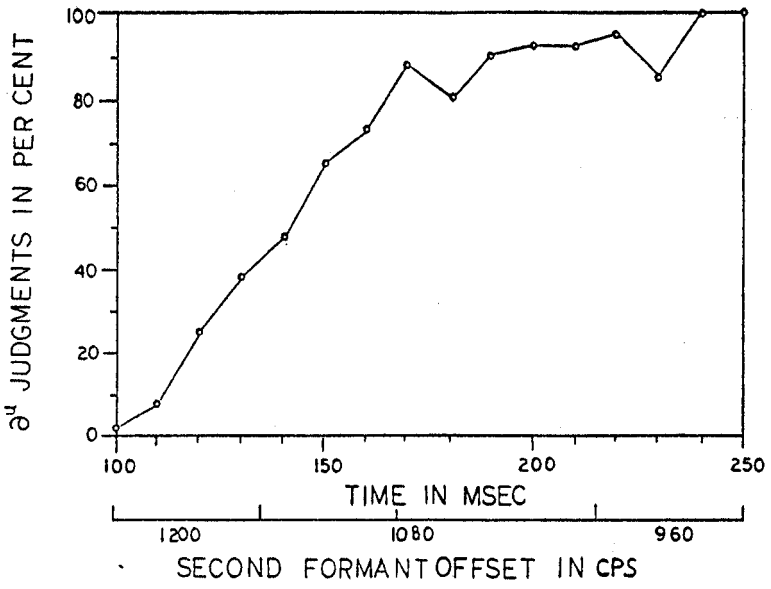

Fig. 15. Effects of duration on /a-au/ shift. Second formant offset varies from 1200-960 c.p.s. while onset remains fixed at 1320 c.p.s.

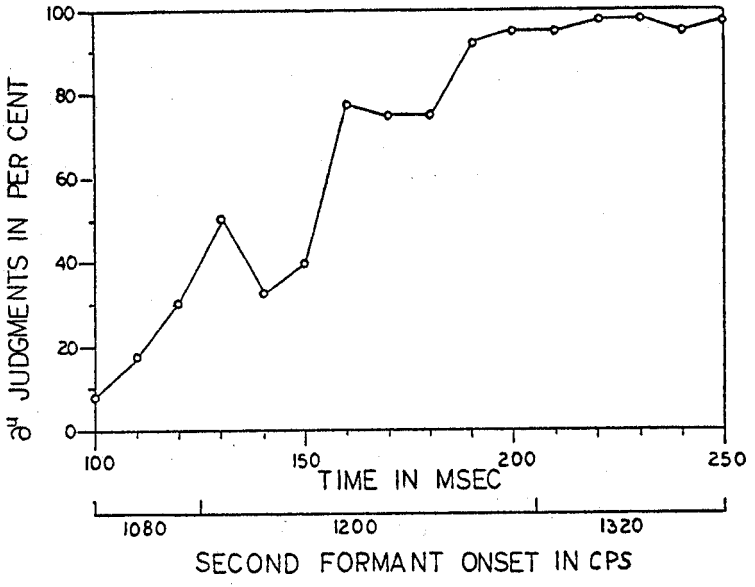

Fig. 16. Effects of duration on /ʌ-au/ shift. Second formant onset varies from 1080-1320 c.p.s. while offset remains fixed at 960 c.p.s.

TABLE 1

Real speech measurements of /ɔi, ai, au/ under three different conditions of rate. Measurements shown are averages for two speakers.

	RATE	DURATION (msec.)	F_2-INITIAL (c.p.s.)	F_2-TERMINAL (c.p.s.)	F_2-CHANGE (c.p.s./msec.)
/ɔi/	Slow	153	655	1520	5.7
	Moderate	113	690	1265	5.3
	Fast	75	720	1200	5.8
/ai/	Slow	163	930	1475	3.4
	Moderate	110	935	1350	3.7
	Fast	88	980	1240	3.8
/au/	Slow	270	1145	800	1.3
	Moderate	165	1110	925	1.1
	Fast	125	1045	905	1.1
/o/	Slow	130	895	835	0.5
	Moderate	123	895	865	0.3
	Fast	105	850	835	0.1

and fixed-offset patterns, a lower second formant cut-off point does not incur lower /au/ responses, except at shorter durations. Since the course of the second formant for /au/ changes rather slowly and mostly within areas bounded generally by /a/, the formant movements of /au/ do not glide through intermediate vowel positions until the glide approaches the positions for /ɔ/, slightly before termination.

DISCUSSION

The results of this experiment show that transition duration rather than change in frequency position provides primary cues for separating vowel *v.* diphthong and also, that the rate of formant frequency change is a fixed feature of the diphthong movement. These data further imply that changes in the rate of production of /ɔi, ai, au/ would be reflected by changes in target frequency positions rather than changes in the speed of articulatory movement. To relate these perceptual implications to real speech articulation, formant frequency and formant rate of change measurements were made for real speech /ɔi, ai, au/ under three different conditions of speech rate. Table 1 shows the results of these measurements as well as those made for /o/. All diphthongs

were produced in a single sentence context, "The *boy* passed *by* the *bow* of the *boat*" with samples obtained from two male informants. For both speakers, the results of all measurements for /ɔi, ai, au/ occurred as would be expected within the frame of the perception results. /ɔi, ai/ behave similarly, each showing rather marked decreases in terminal target positions as a function of increased rate of production. Second formant onset levels increase concurrently but not to the same degree as terminal target levels decrease. In each case however, second formant rate of change remains relatively fixed. The measurements for the different durations of /au/ also show a constant rate of second formant change, with variations occurring in the frequency levels of the initial and terminal targets. These variations, unlike those for /ɔi, ai/, occur uniformly for both the initial and terminal target levels. The second formant rate of change for [ou], on the other hand, tends to slow down with increased rate of production. This is not unexpected in light of the [ou] off-glide incurring no phonemic significance. Thus, these measurements support the earlier findings by showing that second formant rate of change is a fixed feature across changes in duration.

It was shown in the first experiment that the formant movements of /ɔi, ai, au/ are not compatible with those of a vowel plus vowel plus semi-vowel sequence in terms of either target frequency position or glide duration and consequently, that these sounds might best be treated as unit phonemes. The results of this experiment further support this treatment in demonstrating that /ɔi, ai, au/ are characterized primarily by an invariant rate of formant frequency change. Both the vowel plus vowel and vowel plus semi-vowel theories suggest that the distinction between /ɔi/ and /ai/, for example, is attributable to differences in initial target position ([ɔ] versus [a]), with the gliding movement serving only the simple vowel-diphthong separation.[13] The present data however, show that the specific course of the glide, rather than the locations of the targets, serves as the primary distinguishing cue (with glide duration responsible for the simple vowel-diphthongal separation). Apparently then, since the target positions serve no phonemic role, /ɔi, ai, au/ cannot be described as sequences of two phonemes.

The results of the previous experiment specified the formant movements appropriate to identification of /ɔi, ai, au/, while the results of this experiment determined the features of those movements which provide cues for recognition. Taken together, these results permit the following overall phonetic and phonemic description of /ɔ, ai, au/.

The diphthongs /ɔi, ai, au/ are each characterized by gliding movements from one vowel area to another. The onset and termination points of these glides are general for /ɔi, ai/ and more specific for /au/. The locations of these points however, do not in themselves contribute phonemic status to the diphthongs but rather serve as loci from which and toward which articulatory movement is directed within a particular unit of time. The movement through time provides the primary cue for diphthong *vs.* simple vowel perception and the direction of glide separates /ɔi, ai/ from /au/. The

[13] *In the case of* /ai-au/ *however, the direction of the glide, whether fronting or retracting, respectively, is relevant in the distinction.*

/ɔi-ai/ distinction is attributable to rate of frequency change or in articulatory terms, speed of movement. Since /ɔi, ai, au/ glide toward but do not necessarily reach any one specific terminal target, an overall phonetic transcription might best be made by using a superscript form, as in [$ɔ^{i}$, $a^{ɪ}$, $a^{ɔ}$].

References

Cooper, F. S. (1952). Spectrum analysis. *J. acoust. Soc. Amer.*, 22, 761.

Delattre, P., Liberman, A. M., Cooper, F. S. and Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel colour: observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word*, 8, 195.

Holbrook, A. and Fairbanks, G. (1962). Diphthong formants and their movements. *J. Speech Hear. Res.*, 5, 33

Lehiste, I. (1964). Acoustical characteristics of selected English consonants, Bloomington: Indiana Univ. Res. Center in Anthro., Folklore and Ling., P. No. 34.

Lehiste, I. and Peterson, G. E. (1961). Transitions, glides and diphthongs. *J. acoust. Soc. Amer.*, 38, 268.

Liberman, A. M., Delattre, P. and Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *Amer. J. Psychol.*, 65, 497.

Liberman, A. M., Delattre, P., Gerstman, L. J. and Cooper, F. S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *J. exp. Psychol.*, 52, 127.

O'Connor, J., Gerstman, L. J., Liberman, A. M., Delattre, P. and Cooper, F. S. (1957). Acoustic cues for the perception of initial /w, j, r, l/ in English. *Word*, 13, 24.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in study of the vowels. *J. acoust. Soc. Amer.*, 24, 175.

Potter, R. K. and Peterson, G. E. (1948). The representation of vowels and their movements. *J. acoust. Soc. Amer.*, 20, 528.