67-12,560

GAY, Thomas John, 1940-

A PERCEPTUAL STUDY OF AMERICAN ENGLISH DIPHTHONGS.

The City University of New York, Ph.D., 1967
Speech

University Microfilms, Inc., Ann Arbor, Michigan

A PERCEPTUAL STUDY OF AMERICAN ENGLISH DIPHTHONGS

by

JOHN
THOMAS GAY

A dissertation submitted to the Graduate Faculty in Speech in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1967

This manuscript has been read and accepted for the University Committee in Speech in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

May 2, 1967
date

Arthur S. Abramson
Chairman of Examining Committee

May 2, 1967
date

Executive Officer

Arthur S. Abramson
Arthur S. Abramson

Katherine S. Harris
Katherine S. Harris

Arthur J. Bronstein
Supervisory Committee

The City University of New York

ACKNOWLEDGEMENTS

The efforts of the Supervisory Committee, Professors Arthur S. Abramson, Katherine S. Harris and Arthur J. Bronstein, were largely responsible for the shaping and completion of this project.

Thanks are also due Dr. Franklin S. Cooper, President of Haskins Laboratories, New York City, for use of the laboratory facilities.

CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

I. INTRODUCTION

Adequate description of speech sounds requires physiological, acoustical and perceptual analyses within a linguistic framework. Such data have been reported for most sound classes of American English with perhaps only a few exceptions. One such case is the group of diphthongs, [$ɔ^i$,a^i,a^u,e^i,o^u,$ɪ^i$,$ʊ^u$],[1] which have been analyzed acoustically (Lehiste and Peterson, 1961; Holbrook and Fairbanks, 1962), but not treated in terms of those features which provide cues for recognition. These diphthongs are used in widespread varieties of English in such words as <u>boy</u>, <u>buy</u>, <u>bough</u>, <u>bay</u>, <u>boat</u>, <u>beet</u> and <u>boot</u>.

Purpose of the Study

Diphthongs show a gliding movement along a particular path in the vowel space between zones appropriate to two different vowels. This gliding movement, which accounts for the major temporal portion of the diphthong,

---

[1]The transcription followed here implies a glide from a general vowel area toward, but not necessarily reaching, a higher vowel area. This notation will be used except when references are made to specific phonemic theories. In addition, the use of square brackets shows, as is conventional, that in this instance the transcription reflects a phonetic rather than a phonemic description.

is evidenced by formants that either rise or fall sharply, depending on the values 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 /ɔ^{i},a^{i},a^{u}/ because the diphthongal nature of each is phonemically distinctive in most dialects of American English. [e^{i}, o^{u},ɪ^{i},ʊ^{u}], on the other hand, alternate with nondiphthongal allophones of the phonemes /e,o,i,u/, respectively, thus suggesting that their offglides carry no phonemic significance. Specifically then, this study was designed to answer the following questions:

1. At what points along various acoustical continua are /ɔ^{i}/, /a^{i}/ and /a^{u}/ 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 /ɔ^{i}/, /a^{i}/ and /a^{u}/?

Phonemic Status

According to traditional phonetic theory, diph-

thongs are produced by a continuous change in articulation from one vowel to another. In the case of /ɔ^{i}/ for example, the glide is said to be from a vowel position of [ɔ] to one of [ɪ]. Articulation begins with an open mouth, low-back tongue position and some lip rounding and changes gradually to a more closed mouth, high-front tongue position with the lips becoming more spread. The position for /a^{i}/ begins lower and more central, at [a], and glides higher and toward the front to [ɪ]. /a^{u}/ likewise begins at [a] but changes to a high-back, mouth-closed position with some liprounding ([ʊ]). The usual transcription, both phonetic and phonemic, is a sequence of two vowels, /ɔɪ,aɪ,aʊ/. The treatment of a diphthong as a sequence of two vowels is held more commonly for /ɔ^{i},a^{i},a^{u}/ than for [e^{i},o^{u}], the distinction being based largely on the phonemic stability of /ɔ^{i}, a^{i},a^{u}/ (Pike, 1947) and acoustic data (Lehiste and Peterson, 1961) which show only /ɔ^{i},a^{i},a^{u}/ as having explicit initial and terminal vowel areas.

Perhaps a more widely followed system is the Trager and Smith (1951) analysis which treats diphthongs as a sequence of a vowel plus a semivowel. Trager and Smith describe the vowel system of American English as consisting of nine simple vowels and twenty-seven complex nuclei. These nuclei are formed by the combination of any of the nine simple vowels with one of the three

off-glides, /y/, /w/ or /h/.[2] (These glides are considered allophones of pre-vocalic /y,w,h/.) The semivowel /y/ glides higher and toward the front; /w/ glides higher and toward the back and is more rounded; and /h/ is a more central and unrounded glide. The transcription of /ɔ^{i},a^{i},a^{u}/ according to this system would be /ɔy, ay, aw/. Although all twenty-seven complex nuclei are not found in each regional dialect, they are represented when the different dialects are viewed collectively. Francis (1958, p. 143) agrees in general with the Trager-Smith system but adds a fourth glide, /r/. In addition, he uses the terms, "fronting," "retracting" and "centering" in describing diphthongs ending in /y/, /w/ and /h/ or /r/, respectively. Gleason (1961) is also in close agreement with Trager and Smith but considers /h/ as only a pre-vocalic fricative. He posits a separate phoneme, /H/, to represent a centering off-glide as well as length.

The Trager-Smith system however, is not without criticism. Sledd (1954), for example, suggests that a nine vowel-three glide system is inadequate for describing all regional dialects and that "pure long vowels" (as in "beat" and "boot") can occur as such, and not only as complex forms as Trager and Smith suggest. Kurath (1964, pp. 17-19) also disagrees with the Trager-

---

[2]The Trager-Smith [y] corresponds to the IPA [j].

Smith viewpoint in suggesting a system of six "checked" and eight or nine (depending on regional dialect) "free" vowels. This dichotomy is based on the traditional tense (free) - lax (checked) vowel distinction. Checked vowels are either monophthongal or centering while free vowels tend to glide to or toward a higher position. Also, checked vowels do not occur in word-final positions whereas free vowels are found in all positions of words. The diphthongs /$ɔ^i$,a^i,a^u/ are classified among the upgliding free vowels and thus, are considered as unitary phonemes rather than vowel plus vowel or vowel plus semivowel sequences.

It is apparent that the alignment of diphthongs within the phonological structure of American English is unclear in terms of both phonetic and phonemic descriptions. This absence of widely adopted descriptions has, in part, led to the analysis of these sounds in terms of their acoustical characteristics.

Acoustical Characteristics

The excitation source of all vowel sounds is normally at the glottis where vocal fold vibrations generate a quasi-periodic tone whose spectrum reveals a rich harmonic structure. This glottal tone is applied to the vocal tract which acts as a resonator modifying the tone into a distinct spectral envelope. This envelope is revealed by several amplitude peaks located at different

frequencies. The frequency locations at which these amplitude peaks occur depend on the overall shape of the vocal tract. Thus, changes in articulation produce changes in vocal tract configuration and subsequent changes in resonant frequency location. These resonances are called formants and the center frequencies at which they occur are known as formant frequencies. Since vowel articulation is relatively steady through time, formant frequencies are likewise relatively unchanged. Diphthongs, on the other hand, are produced by continuous articulatory change and thus, are characterized by continuous formant frequency change (the most marked of which occurs for the second formant). Results of research by Potter, Kopp and Green (1947), Joos (1948) and Delattre, Liberman, Cooper and Gerstman (1952), among others, have shown that the formant frequencies of the different vowel sounds are responsible for individual vowel quality. The lowest three formants are the primary identifying cues although even one- or two-formant synthetic vowels can be recognized.

The published acoustical research on diphthongs, based largely on the premise that these sounds are sequences of two vowels each, is concerned primarily with specifying the acoustic positions of initial and terminal target areas by means of spectrographic analysis. A sound spectrogram is a graphic representation of the

time, frequency and intensity characteristics of a short-time speech sample. Formants are seen as dark bands (with degree of darkness corresponding to intensity) which course through time. Sample spectrograms of /ɔ^{i}, a^{i},a^{u}/ are shown in Figure 1.1.

Figure 1.2 summarizes the results of three different spectrographic analyses (Potter and Peterson, 1948; Lehiste and Peterson, 1961; Holbrook and Fairbanks, 1962) in a graph where first formant frequencies are plotted against second formant frequencies relative to vowel positions reported by Peterson and Barney (1952).[3] It can be seen from this graph that, except for onset values of /ɔ^{i}/, the initial and terminal targets of each diphthong are not necessarily bounded by specific vowel positions.[4] The course for /ɔ^{i}/ begins at formant positions appropriate to [ɔ] and terminates at positions bounded by [ɛ] and [ɪ]. /a^{i}/ begins at or higher than [ɑ], perhaps at [a], and terminates in areas close to [ɛ] and [ɪ]. /a^{u}/ shows a general movement from formant positions between [ɑ] and [æ] to areas between [ɔ] and [ʊ].

---

[3]Although these vowel positions are used primarily for summary purposes, their values vary only slightly from those of each individual study and thus, can be regarded as general reference boundaries for initial and terminal diphthong positions.

[4]It is interesting to note that the above three studies, along with one carried out by Peterson and Coxe (1953), found similar "phonetic ambiguities" of initial and terminal vowel areas for the diphthongs, [e^{i}] and [o^{u}].

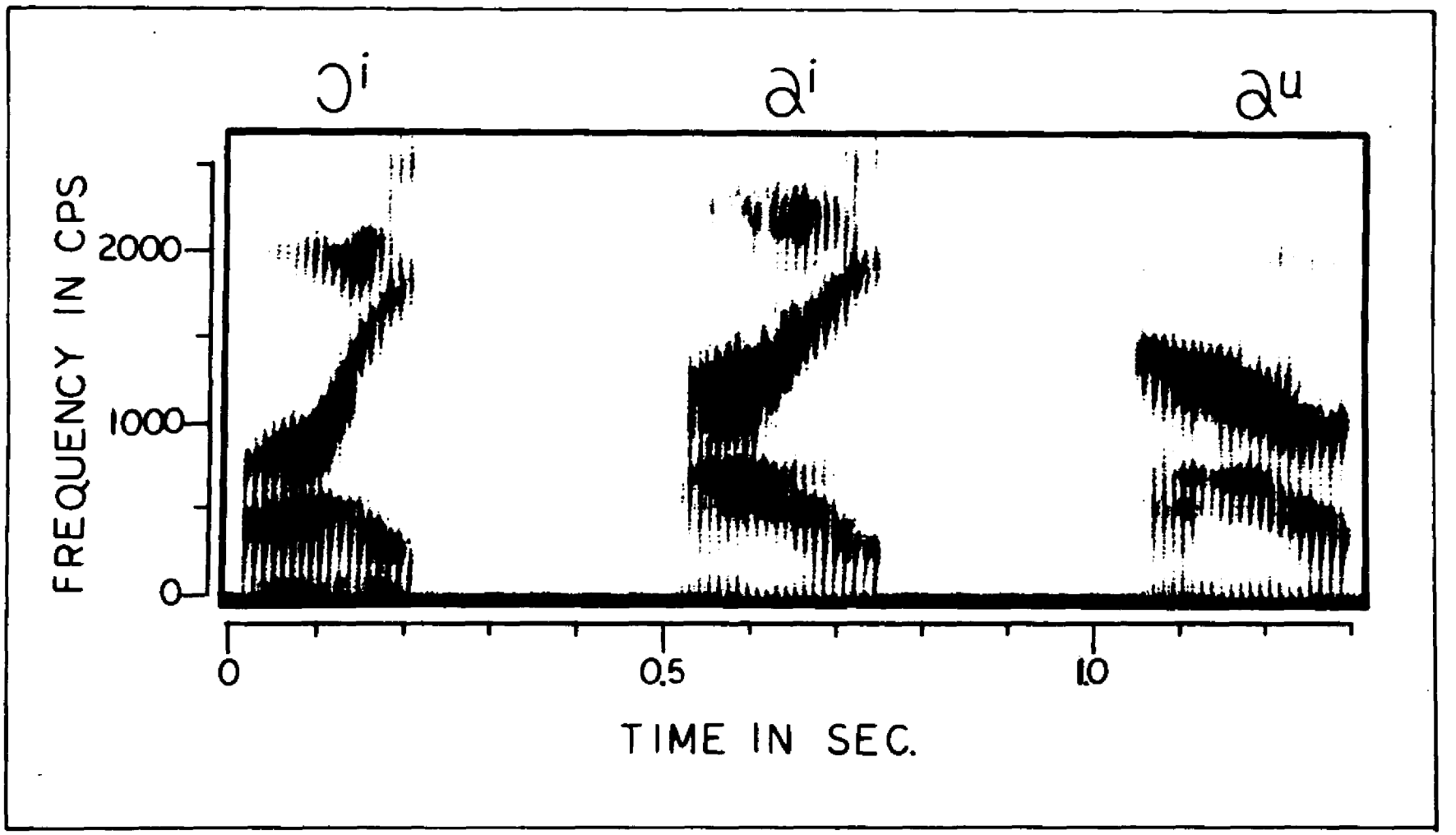

Figure 1.1.--Sample spectrograms of /ɔ^{i},a^{i},a^{u}/. Frequency is plotted along the vertical axis, time along the horizontal axis and relative intensity is revealed by the darkness of the sound pattern. Diphthong formants are characterized by a continuous change in frequency through time.

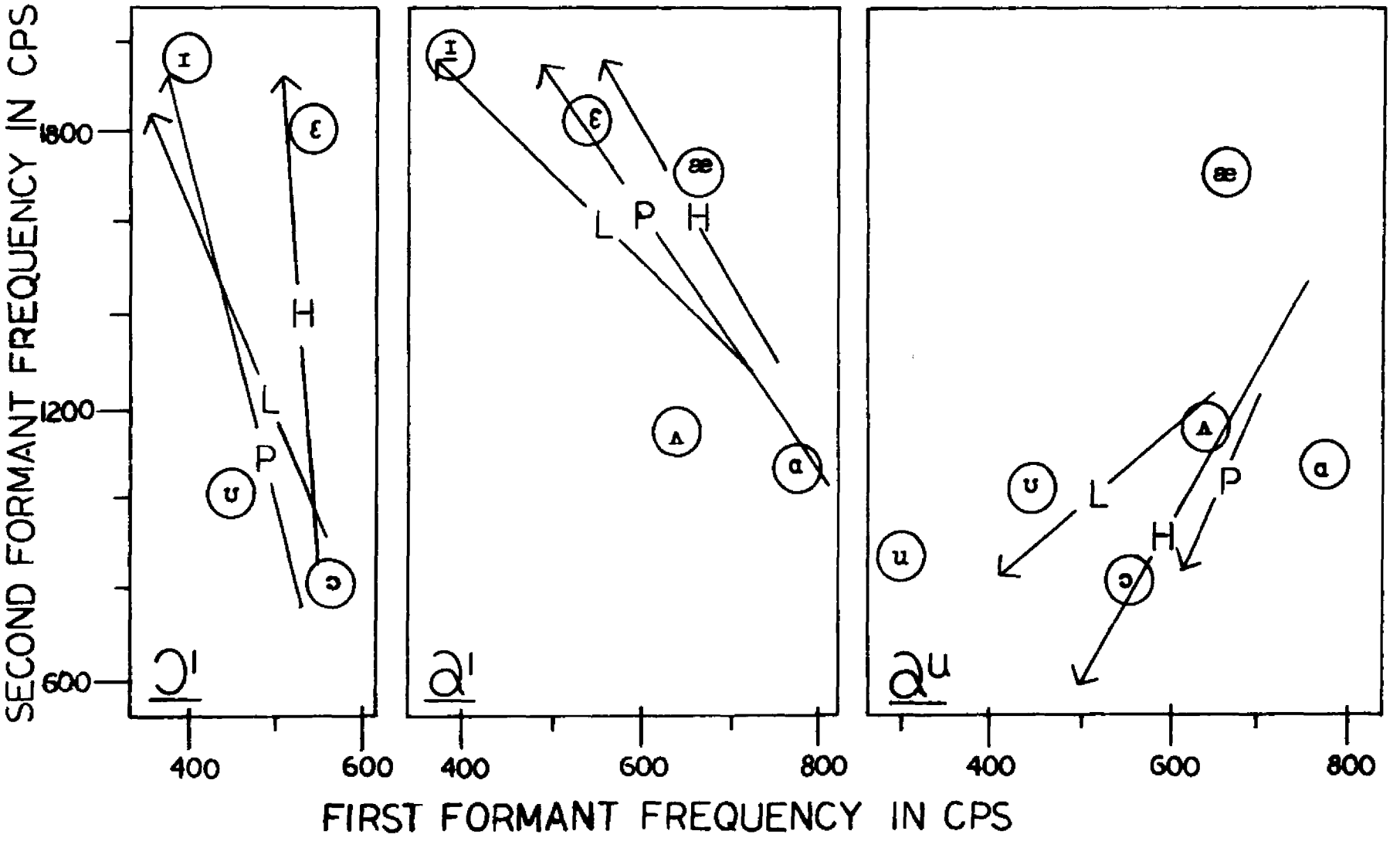

Figure 1.2.--Graph of direction and extent of diphthong movements according to values of formants one and two. Arrows indicate direction of movement. H=Holbrook and Fairbanks (1962), L=Lehiste and Peterson (1961), P=Potter and Peterson (1948). Vowel positions are taken from Peterson and Barney (1952).

All three diphthongs course through fairly wide areas following more or less straight routes from initial to terminal targets. The status of these targets is apparently variable. The data of Holbrook and Fairbanks show that the major formant movements occur during the final half of the syllable with a prominent steady state present only at initiation. Lehiste and Peterson, on the other hand, found that /ɔ^{i},a^{i},a^{u}/ are each characterized by two explicit target areas (defined arbitrarily as the frequency position where the formant is parallel to the time axis for at least 20 msec), the first longer than the second but both shorter than the transition. In summary, the /ɔ^{i}-a^{i}/ distinction can be attributed to lower initial first and second formant values for /ɔ^{i}/. First and second formants for both /ɔ^{i}/ and /a^{i}/ diverge toward termination, with greater degrees of divergence evident for /ɔ^{i}/. /a^{u}/, on the other hand, glides in a different direction, with first and second formant values converging toward termination.

The effects of certain portions of the transitional components as perceptual cues for diphthongs were observed by Wise (1965). He states that the elimination of 30 msec of the transition immediately before the terminal target is "scarcely perceptible," incurring "no phonemic change." With the entire transition (115 msec) eliminated and the two targets brought together (by means

of a gating circuit and a dual-loop tape recorder), a diphthong rather than two separate vowels is still perceived (this is apparently due to the proximity in time of the two segments). Perception of the terminal target areas for /a^i/ and /a^u/ favor [i] or [ɪ] and [u] or [ʊ], respectively. These judgments correspond to acoustical measurements.

Although acoustical analyses have contributed important data regarding diphthong movements, experimental results are varied and inadequate for purposes of identifying the primary perceptual features of these sounds. In this respect, psychoacoustic experiments involving perception of synthetic speech have been especially applicable in studying glide-like sounds (Liberman et al, 1956; O'Connor et al, 1957; Lisker, 1957).

General Procedures

The present study is comprised of three experiments concerned with the phonemic boundaries, target area perceptions and duration-recognition relationships of the three phonemically distinct diphthongs of American English, /$ɔ^i$,a^i,a^u/.

Synthetic speech stimuli were used in all experiments. These stimuli were produced by a speech synthesizer, the Pattern-Playback of Haskins Laboratories, New York City. The Pattern-Playback converts spectrographic patterns, hand-painted on acetate, into corresponding

acoustic units by means of an optic transducing system. A light source is modulated by a rotating "tone-wheel" into 50 harmonics whose fundamental frequency is 120 cps. 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 onto 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.

II. EFFECTS OF FORMANT FREQUENCY MOVEMENTS ON THE PERCEPTION OF /ɔ^{i},a^{i},a^{u}/

This part of the study is concerned with the differences in formant frequency transitions responsible for separating /ɔ^{i},a^{i},a^{u}/ 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 /ɔ^{i}-a^{i}/ distinction occurs along a continuum where second formant transitions course upward through time; /a^{u}/, 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 /ɔ^{i}-a^{i}/ or /a^{u}-o/, they do not specify for example, whether /ɔ^{i}/ is characterized by transitions which begin at

[1] /o/ in this case occurs as the diphthongal variant, [o^{u}], which unlike /ɔ^{i},a^{i},a^{u}/ is characterized by a non-phonemic offglide.

formant positions appropriate to [ɔ] or extend to positions appropriate to [ɪ]. 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

/ɔ^{i}-a^{i}/ Continua.--Figure 2.1 illustrates the acoustical continua used to produce /ɔ^{i},a^{i}/ stimuli. Exploratory work found these ranges appropriate for either /ɔ^{i}/ or /a^{i}/ perceptions without incurring other phonemic impressions. Five different second formant onset values ranging from 840-1320 cps were extended to terminal values of either 1920 or 2040 cps. All patterns with second formant transitions terminating at 1920 cps will subsequently be referred to as "A" patterns and those terminating at 2040 cps, "B" patterns. Each "A" and "B" set was combined with two different first formant and two different third formant transitions. The two first formant transitions each began at 600 cps and terminated at either 480 or 360 cps. The two third formant transitions likewise began at one initial value, 2640 cps, terminating at either 2520 or 2400 cps. All patterns were drawn with durations of 250 msec and bandwidths three harmonics wide (each of the two side harmonics being of lower intensity than the center frequency). The transitions in each pattern were drawn as straight bands from onset to termin-

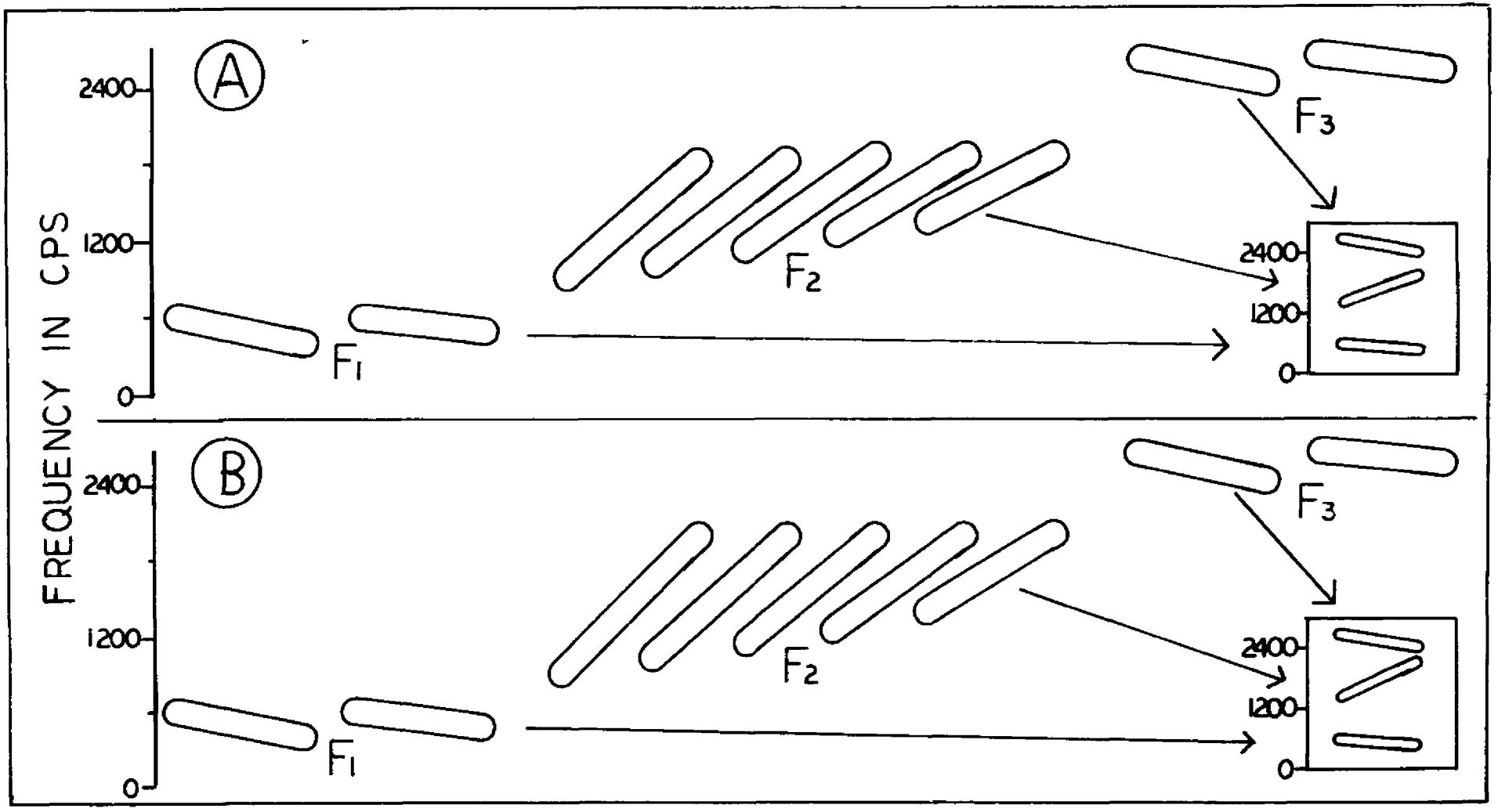

Figure 2.1.--Schematic illustration of stimuli used to produce primary /ɔ^{i},a^{i}/ continuum. A=second formant termination at 1920 cps, B=2040 cps. Each first formant was combined with each second and third formant. Insets show complete spectrographic configurations.

ation. 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 /ɔ^{i}/ and /a^{i}/ perceptions. The supplementary patterns appropriate to /ɔ^{i}/ were constructed by adding lower first (480-360 cps) and third (2520-2400 cps) formants to the existing "A" and "B" continua in Figure 2.1, except for deletions of all 1200 and 1320 cps second formant onsets and any replications. The supplementary /a^{i}/ patterns were drawn with two higher first formants (720-360, 720-480 cps) and two higher third formants (2760-2400, 2760-2520 cps) in combination with the continua in Figure 2.1 except for deletions of all 840-1080 cps second formant onsets and any replications. These supplementary continua provided a total of 15 additional /ɔ^{i}/ and 24 additional /a^{i}/ stimuli for each "A" and "B" set.

/a^{u}-o/ Continua.--The acoustical continua appropriate to /a^{u}-o/ perceptions are shown schematically in Figure 2.2. Second formants begin at 1080-1320 cps and terminate at either 960 cps ("A") or 840 cps ("B"). First formant transitions of 600-480, 720-480, 720-600 cps and third formant transitions of 2400-2280, 2520-2280, 2520-2400 cps were each combined with the three second formant transitions in each set to provide a total of 27 (3x3x3)

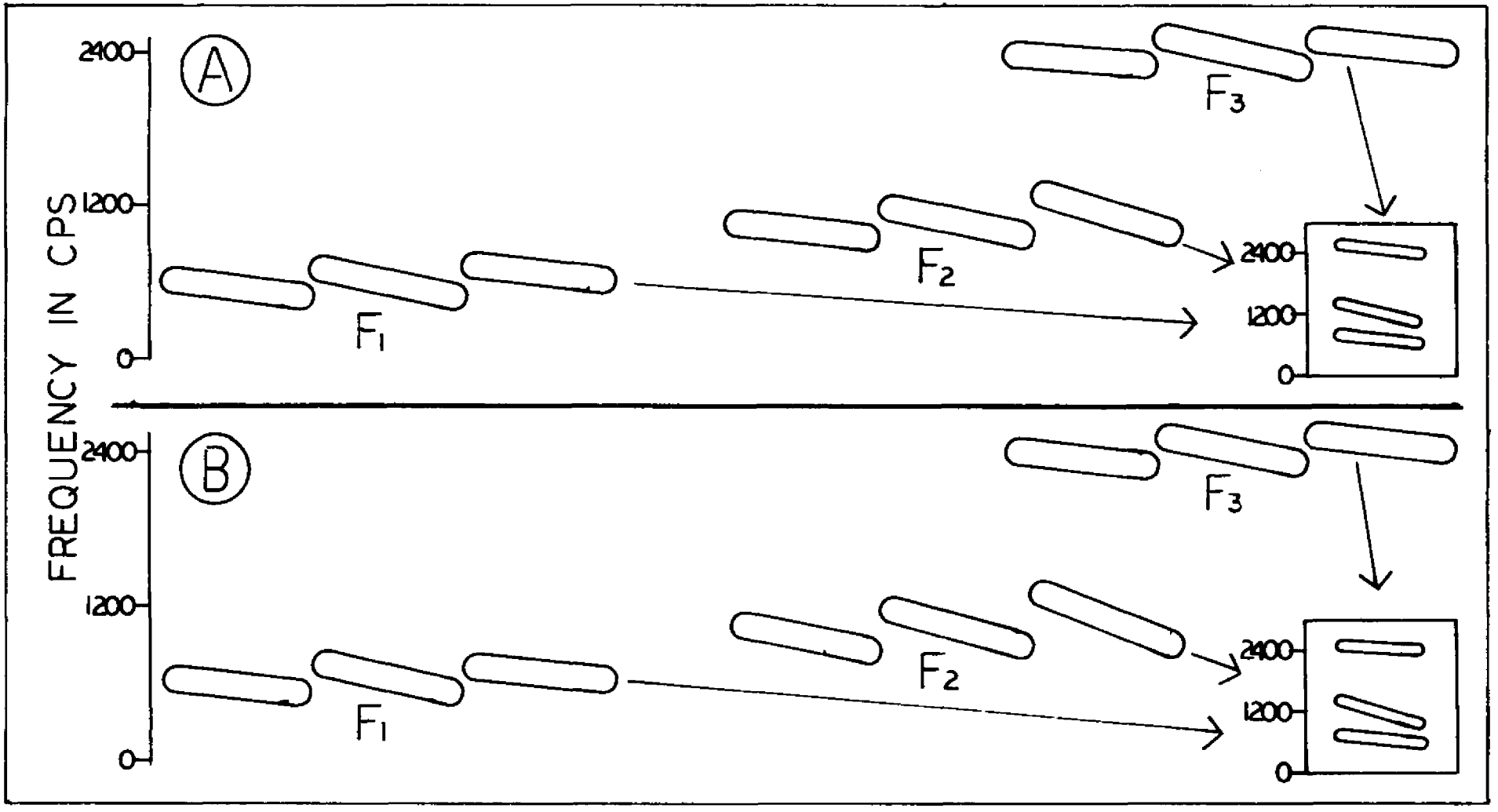

Figure 2.2.--Schematic illustration of stimuli used to produce the /a^{u}-o/ continuum. A=second formant termination at 960 cps, B=840 cps. Each first formant was combined with each second and third formant. Insets show complete spectrographic configurations.

"A" and "B" 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 /a^{u}/ stimuli, thus, eliminating the necessity for synthesizing supplementary patterns.[2]

Target vowels.--This set of stimuli consisted of steady state vowels whose values correspond to the initial and terminal targets of all diphthong patterns. An example of these patterns is shown in Figure 2.3. Here the diphthong is comprised of a first formant of 600-360 cps, a second formant of 840-1920 cps and a third formant of 2640-2400 cps. The steady state vowel appropriate to the initial target of the diphthong has first, second and third formants of 600, 840 and 2640 cps, respectively. The terminal target vowel has first, second and third formants of 360, 1920, 2400 cps. 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 center frequency bounded by two harmonics of lower intensity.

Test batteries.--A total of 172 diphthong stimuli were

---

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

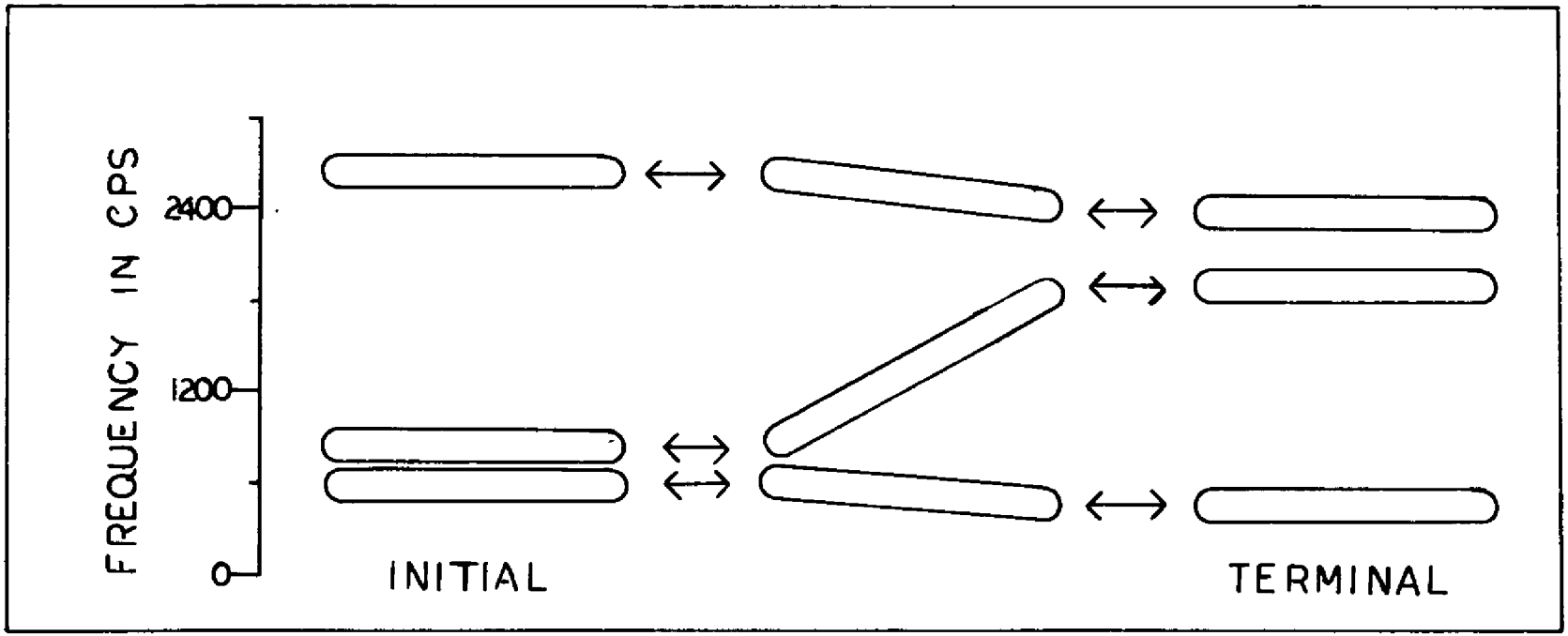

Figure 2.3.--Illustration of the procedure used for synthesizing initial and terminal target vowels. Initial target vowels are based on the initial target values of the diphthong and terminal target vowels are based on the terminal target values.

synthesized on the Pattern-Playback and recorded onto 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 seconds before each diphthong and at successive intervals of approximately 4.5 seconds. Listener responses were recorded on prepared forms. After hearing each sound, subjects first labelled it as one of the set /ɔ^{i}, a^{i},a^{u},o,a/ and then rated it for quality on a "1-5" scale where "1" represented highest quality. Later analysis of these quality judgments showed them to provide little, if any, significant information. Figure 2.4 shows the relationship between diphthong identification and quality ratings. The scattergrams for /ɔ^{i}/ and /a^{i}/ show only a slight positive relationship between higher quality ratings and higher diphthong identification. This does not hold for /a^{u}/ and /o/ where quality ratings are generally higher than for /ɔ^{i},a^{i}/ but are scattered more or less uniformly throughout the range of diphthong intelligibility. Except for these results, the quality ratings provided no further information and for this reason were eliminated from subsequent test batteries.

Stimuli were arranged in groups of ten with brief rest periods occurring between groups. Longer rest

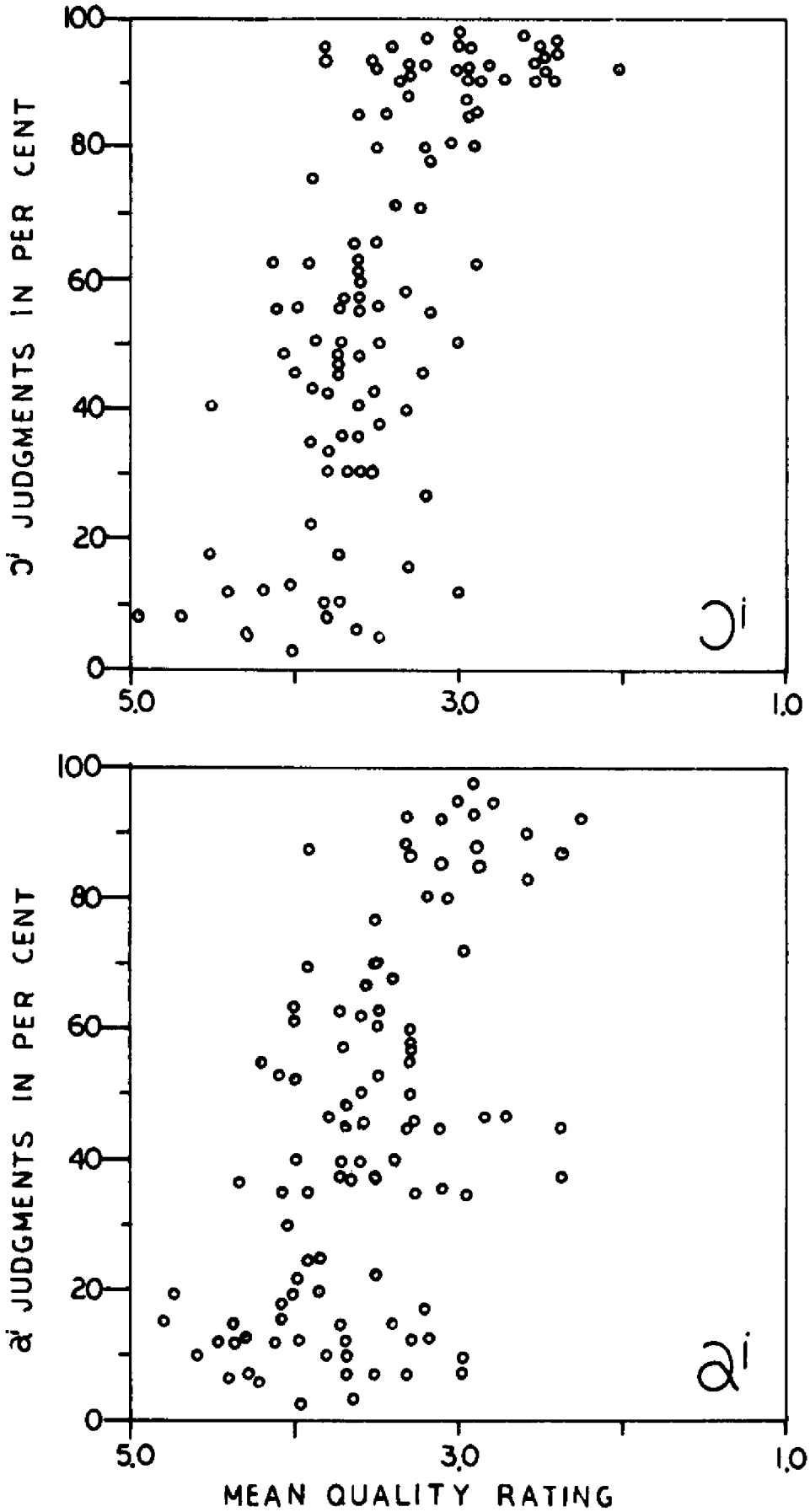

Figure 2.4.--Relationship between identification and quality ratings for /ɔ^{i},a^{i},a^{u},o/. Total number of observations per response=40.

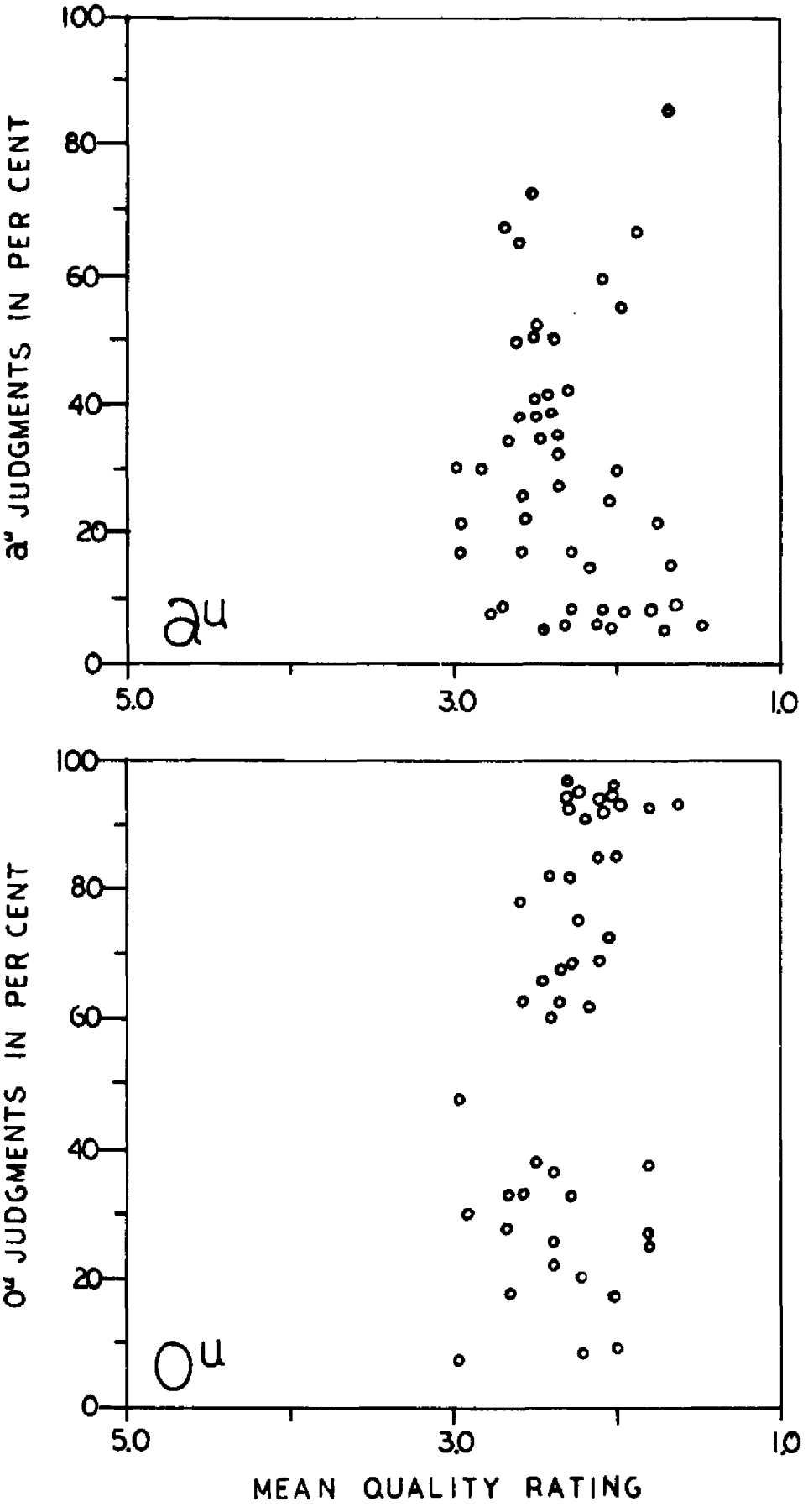

Figure 2.4 (cont'd)

periods were scheduled after every group of 50 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 batteries. The procedures used in preparing the diphthong battery were generally followed with only a few exceptions. Carriers were set at approximately four second intervals and only labelling responses were obtained. Response choices were /a,ɔ, o,ʊ,ʌ/ 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.

/ɔ^{i}-a^{i}/ Distinction

Results of the continua.--For all /ɔ^{i}-a^{i}/ continua, the effects of third formant course are slight. Figure 2.5 shows these effects for all patterns appropriate to both /ɔ^{i}/ and /a^{i}/ judgments.[3] Results for /ɔ^{i}/ are least variable with curves indicating only negligible changes in /ɔ^{i}/ perception attributable to changes in third formant values. For those patterns appropriate to /a^{i}/, some changes in /a^{i}/ judgments are evident for changes in third formant values but these changes are slight and show no real consistency in terms of either

---

[3]For the entire /ɔ^{i}-a^{i}/ continua, all stimuli not heard as /ɔ^{i}/ were heard as /a^{i}/ and vice versa.

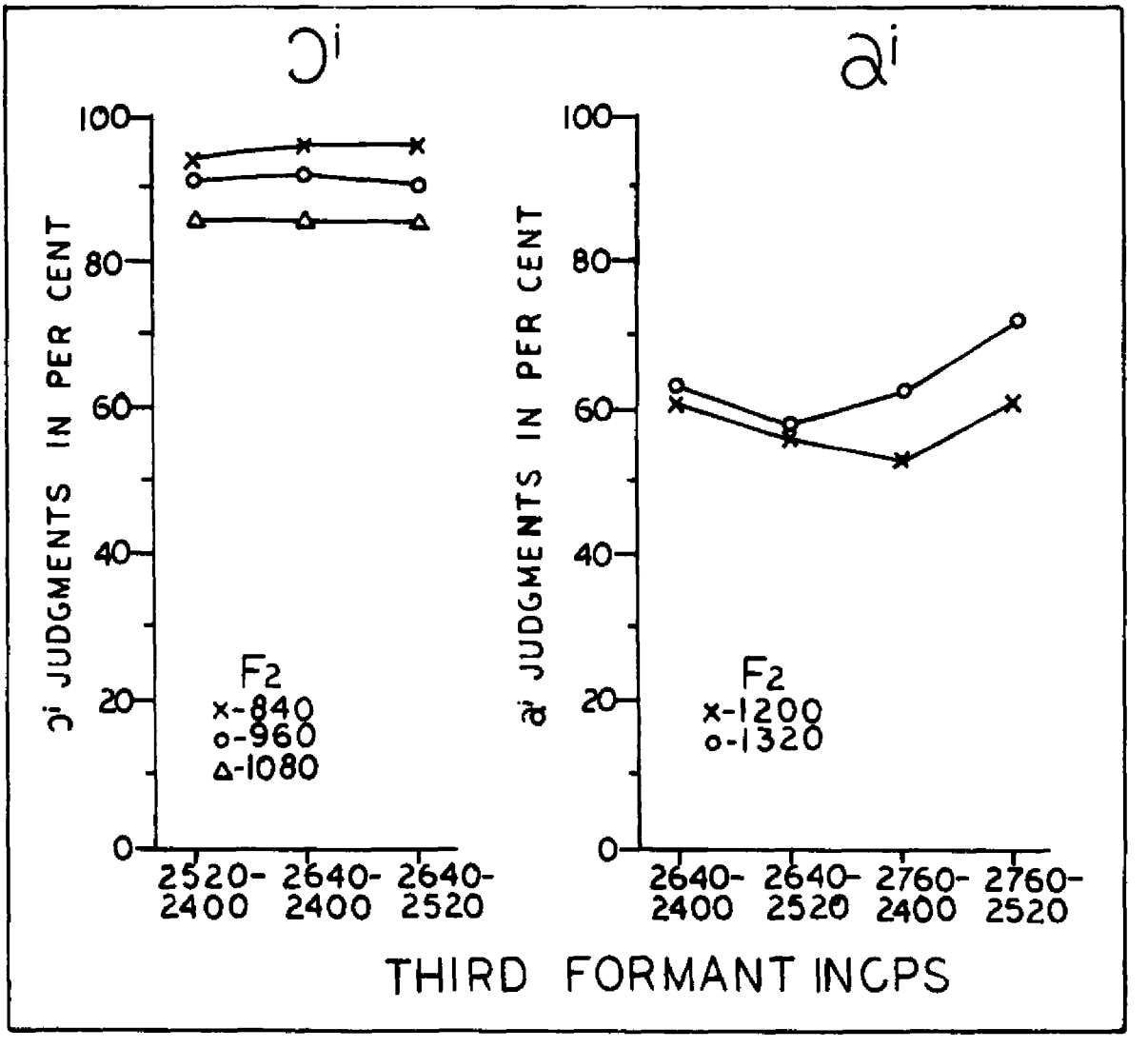

Figure 2.5.--Effects of third formant values on /ɔ^{i},a^{i}/ identification. Each graph is based on both the primary and appropriate supplementary continua. All data are pooled across first formant values and "A" and "B" patterns.

onset or termination values. Since these movements appear to have only a stabilizing effect on /ɔ^{i}-a^{i}/ identification, subsequent data for the remaining continua will be pooled across all third formant results.

The results of the primary /ɔ^{i}-a^{i}/ continua, shown in Figure 2.6, indicate that in general, /ɔ^{i}/ and /a^{i}/ are distinguishable by onset values of second formant transitions. High /ɔ^{i}/ recognition is maintained for second formant transitions that originate between 840-1080 cps. Up to this point, differences in first formant transitions have little effect. For second formant onsets of 1200 and 1320 cps, /ɔ^{i}/ identification is sharply decreased (/a^{i}/ identification increased) for three of the four first formant-second formant combinations. For these two onset frequencies, /ɔ^{i}/ identification is generally lower for a lower terminating second formant ("A" patterns) and consistently lower for a higher terminating first formant (480 cps versus 360 cps). This last effect is most obvious for the "B" patterns where at a second formant onset of 1320 cps, for example, a higher first formant offset (480 cps) is is accompanied by a decrease in /ɔ^{i}/ intelligibility of approximately 35 per cent. Apparently then, the shift from /ɔ^{i}/ to /a^{i}/ not only occurs for a higher second formant onset but also for a higher first formant offset in combination with a lower second formant offset

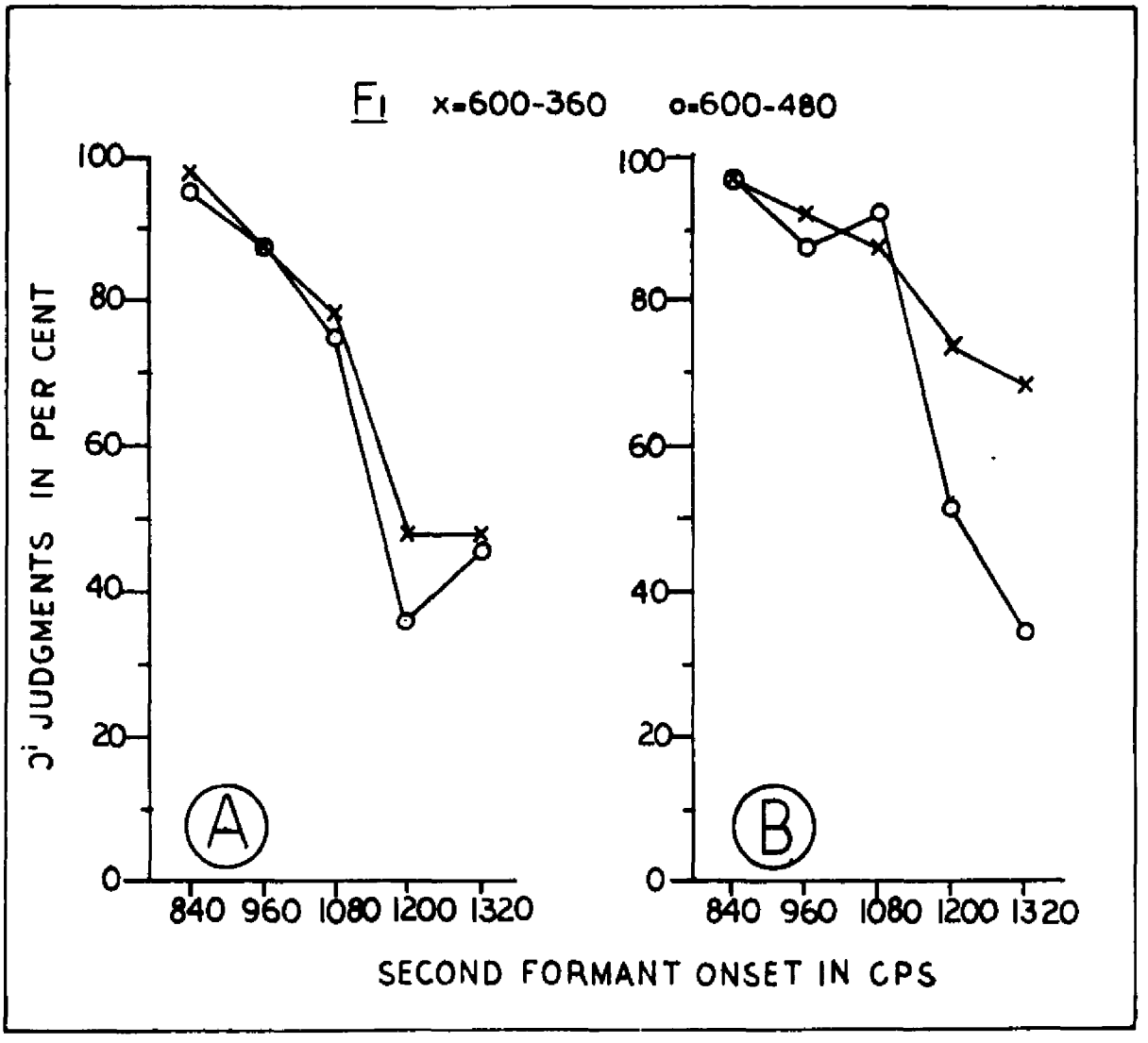

Figure 2.6.--Identification of the primary /ɔ^{i}-a^{i}/ continuum. A=second formant termination at 1920 cps, B= 2040 cps. Total N per stimulus=80. Per cent /a^{1}/=100-per cent /ɔ^{1}/.

(the combined effect at offset being less divergence of first and second formants and consequently a slower rate of formant frequency change).

Similar effects are also evident in the supplementary stimuli (Figures 2.7, 2.8). For these patterns, the only effect of a lower first formant is to maintain high /ɔ^{i}/ identification across the range of all second formant transitions. /a^{i}/ judgments, while remaining relatively stable across all 1200 and 1320 cps second formants, show as much as a 38 per cent increase for the 720-480 cps first formant as opposed to the lower terminating, faster changing 720-360 cps first formant.

This effect, while not deducible from the usual phonemic analysis of English, is nonetheless supported, to varying degrees, by each of the three acoustical analyses summarized earlier (Figure 1.2). These measurements each show a higher first formant offset value for /a^{i}/ as opposed to /ɔ^{i}/. Thus, two factors apparently operate in separating /ɔ^{i}/ from /a^{i}/. High /ɔ^{i}/ recognition requires low initial first and second formants while high /a^{i}/ recognition requires both high initial first and second formants and high terminal first formants. These characteristics of /ɔ^{i},a^{i}/ present certain relevant phonetic and articulatory implications both of which will be demonstrated more clearly in the following section.

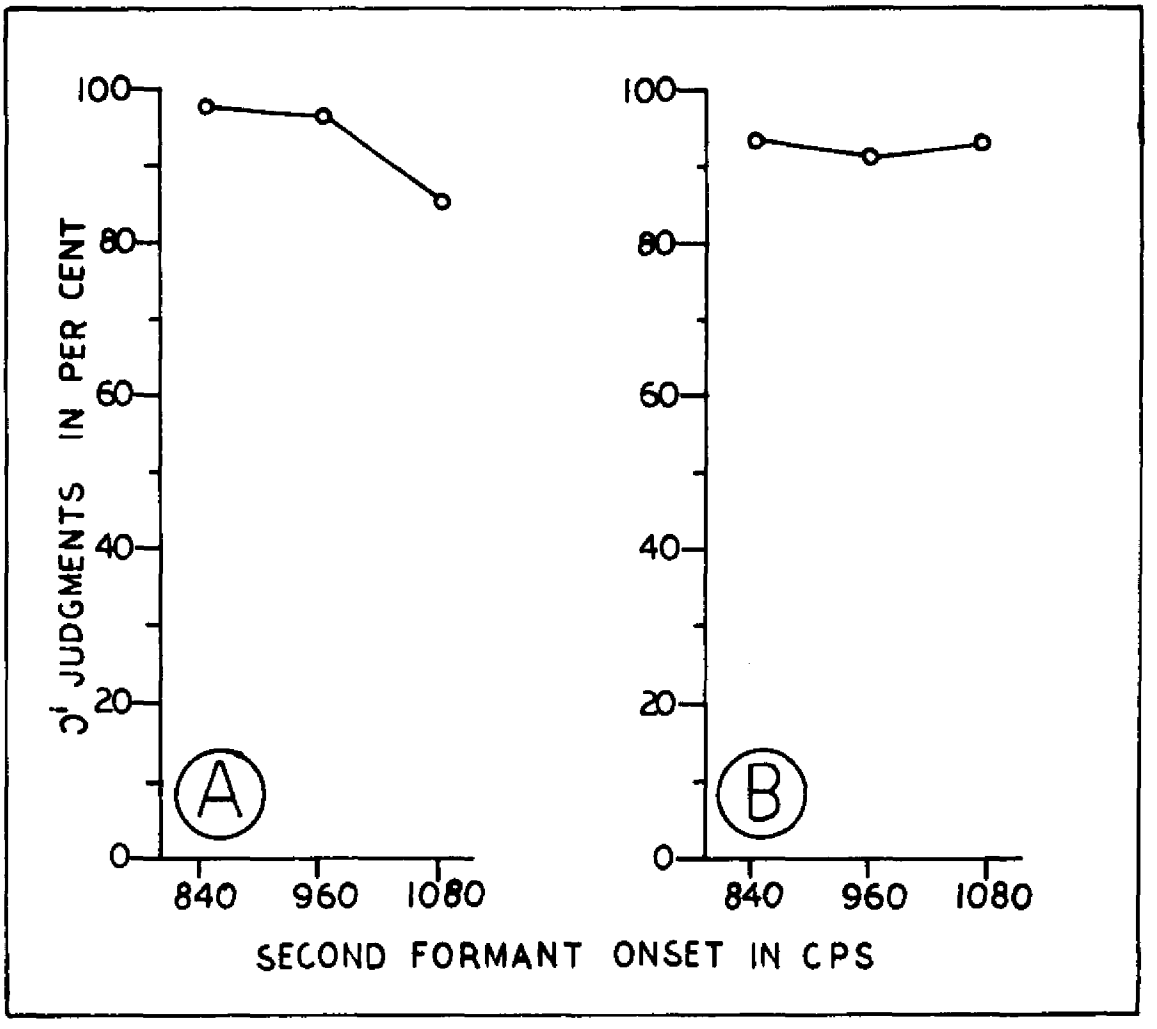

Figure 2.7.--Identification of the supplementary /ɔ^{i}/ continua. A=second formant termination at 1920 cps, B=2040 cps. First formant=480-360 cps. Total N per stimulus=120.

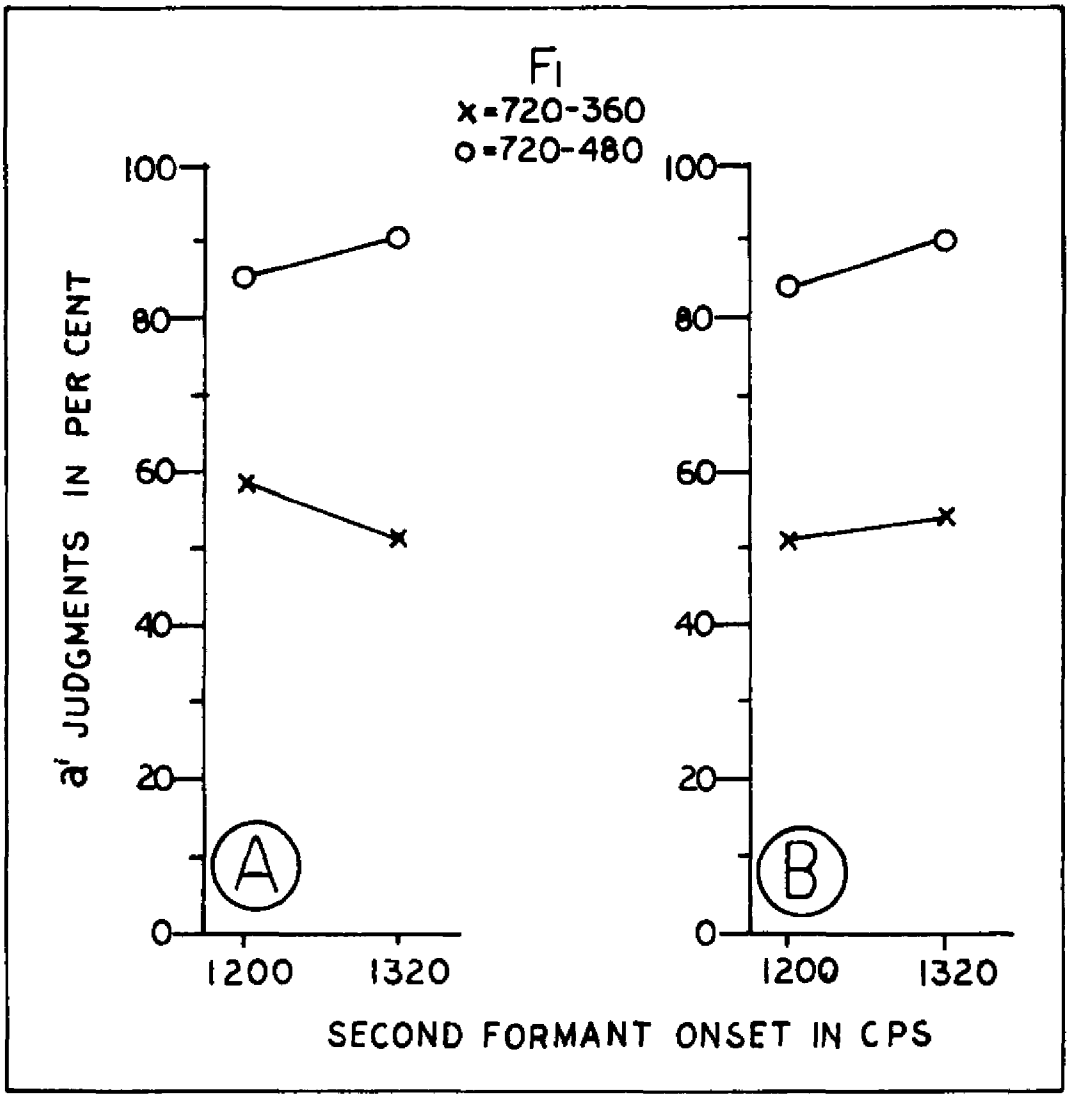

Figure 2.8.--Identification of the supplementary /a^{i}/ continua. A=second formant termination at 1920 cps, B=2040 cps. Total N per stimulus=160.

Phonetic description.--In this section, a broader, phonetic description of the /ɔ^{i},a^{i}/ continua will be presented. Figures 2.9, 2.10 show the initial and terminal target coordinates of all /ɔ^{i},a^{i}/ stimuli and the distributions of target vowel responses accompanying them. These coordinates and the majority preferences of the accompanying vowel distributions will be used later in plotting the formant movements of the various /ɔ^{i}-a^{i}/ stimuli.

Results of these vowel distributions, as would be expected in a grid with closely positioned coordinates, 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 cps. 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 cps position. This sound gives an auditory impression of the high-front, rounded vowel, [y]. This impression

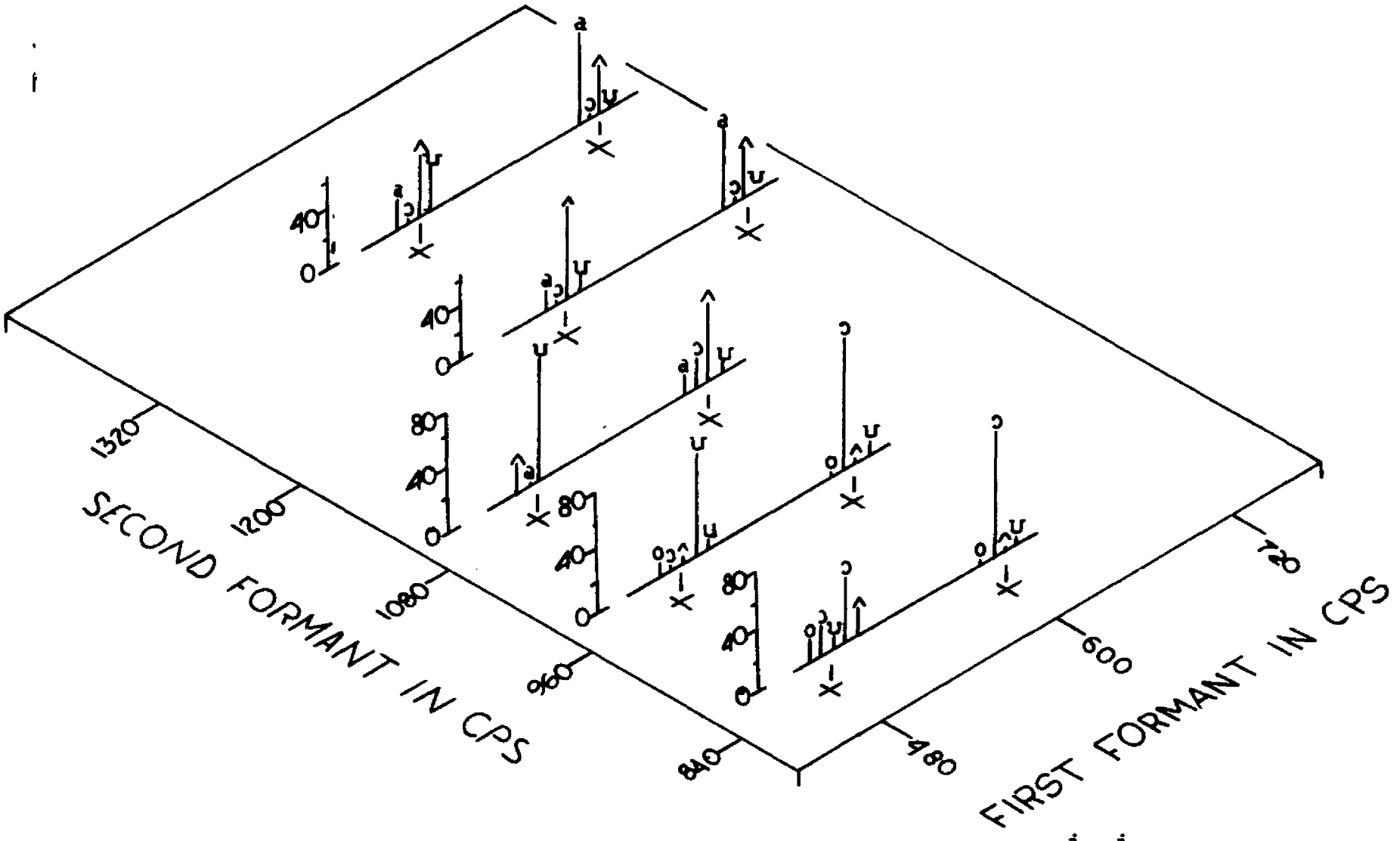

Figure 2.9.--Target vowel distributions for all initial /ɔ^{i},a^{i}/ stimuli. Data are plotted against appropriate first formant-second formant coordinates, in per cent.

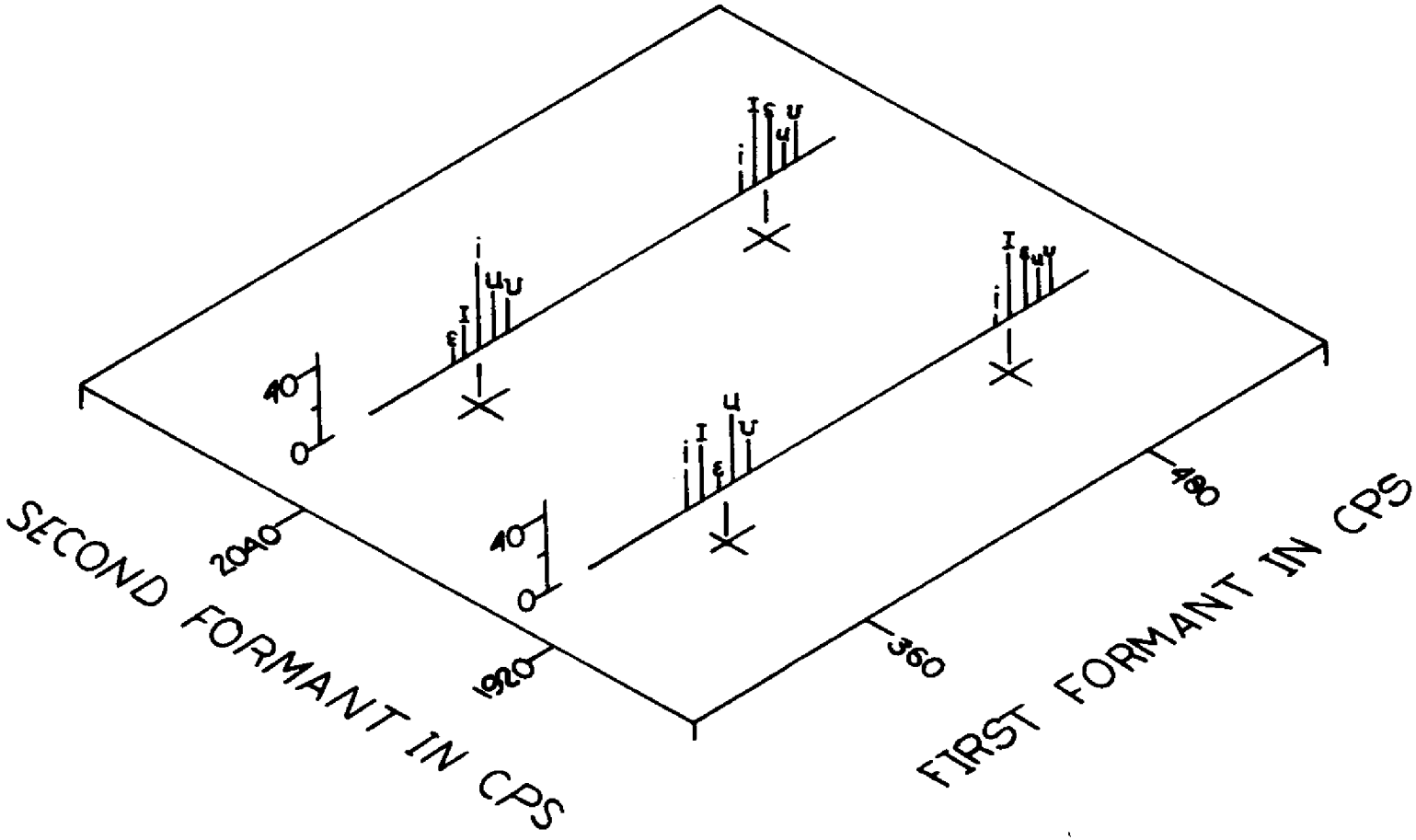

Figure 2.10.--Target vowel distributions for all terminal stimuli, in per cent.

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 cps 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 coordinate might best be described as [y] rather than [u]. Likewise, other [u] and [ʊ] preferences at the three remaining coordinates are probably due to some [y] coloring. The front vowel confusions on the other hand, might be explained by their coordinates not being aligned with real speech measurements. The 360-2040 cps coordinate, 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 high [ɛ] preferences, occur at coordinates whose real speech values range between those of [ɪ] and [ɛ] (Peterson and Barney, 1952). It should be noted however, that these coordinates are aligned with those found for real speech [ɔ^{i},a^{i}].

Based on the majority preferences of the first formant-second formant coordinates, frequency tracts between targets appropriate to different percentages of /ɔ^{i}/, are plotted as contours in Figure 2.11. These

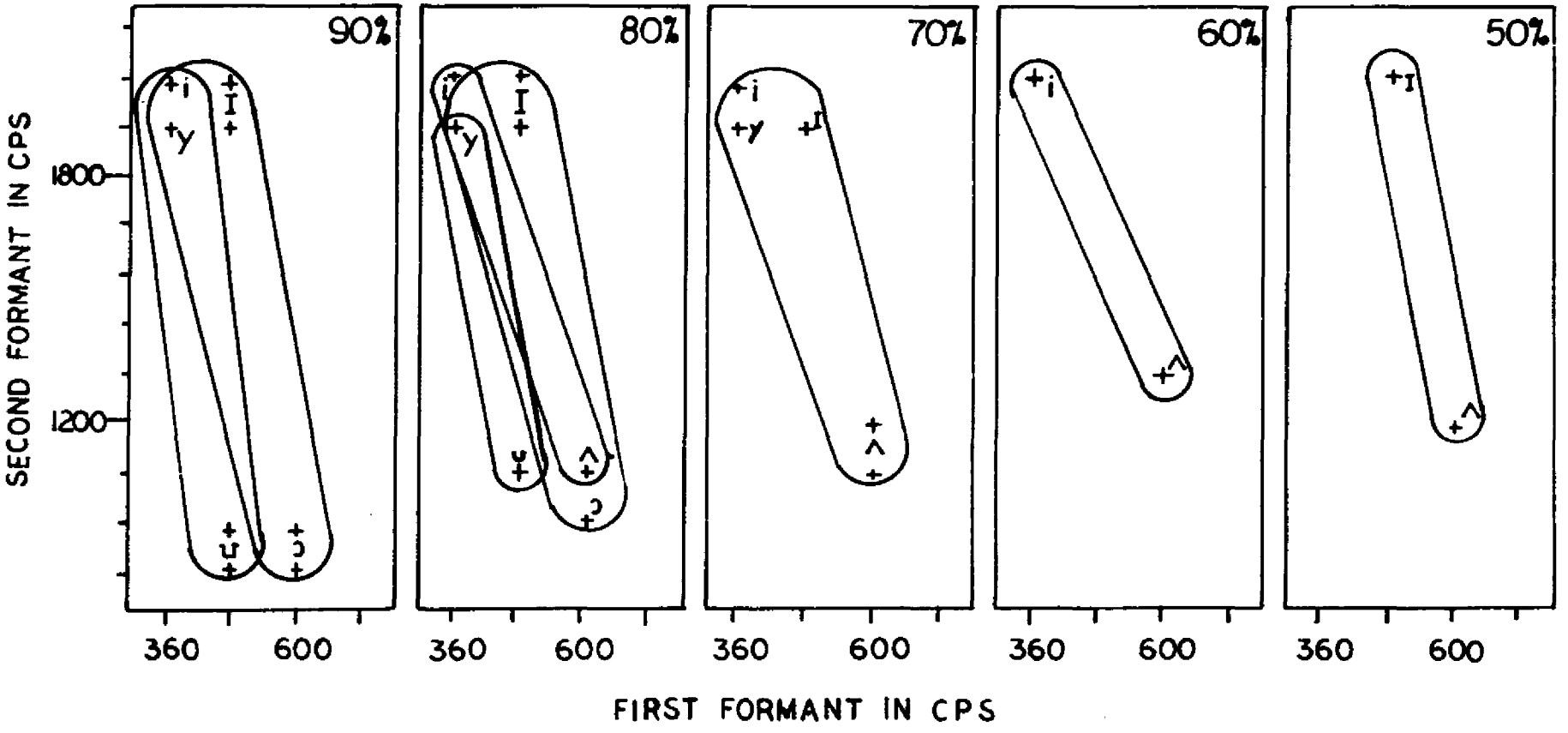

Figure 2.11.--Formant movements for different percentages of /ɔ^{i}/. Contours are based on the continua data and coordinate labels are based on the majority preferences of the target vowel distributions.

contours show that highly preferred /ɔ^{i}/ patterns (90-100 per cent) are not necessarily bounded by initial and terminal targets appropriate to [ɔ] and [ɪ]. /ɔ^{i}/ can course from either [ʊ] to [y,i] or from [ɔ] to [y, i,ɪ]. Other strong /ɔ^{i}/ preferences (80-90 per cent) are characterized by formants coursing from [ʌ] to [i]. In general, as /ɔ^{i}/ preference declines, initial targets shift to [ʌ], and terminal targets shift from [y,i] to [ɪ]. The formant movements of the highly preferred /ɔ^{i}/ 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 /ɔ^{i}/ recognition course between areas that enclose more than one specific vowel position. For this dialect area then, a more appropriate description of /ɔ^{i}/ might best be made by referring to general rather than specific initial and terminal target areas.

In Figure 2.12, where /a^{i}/ contours are plotted, strong preferences for /a^{i}/ (80-90, 90-100 per cent) appear to have more specific boundaries. However, closer inspection of the appropriate initial and terminal targets (Figures 2.9, 2.10) reveals that vowel preferences for these positions are among those showing only small phoneme majorities. The appropriate [a] positions are accompanied by high [ʌ] responses and the [ɪ] positions are accompanied by high [ɛ] responses.

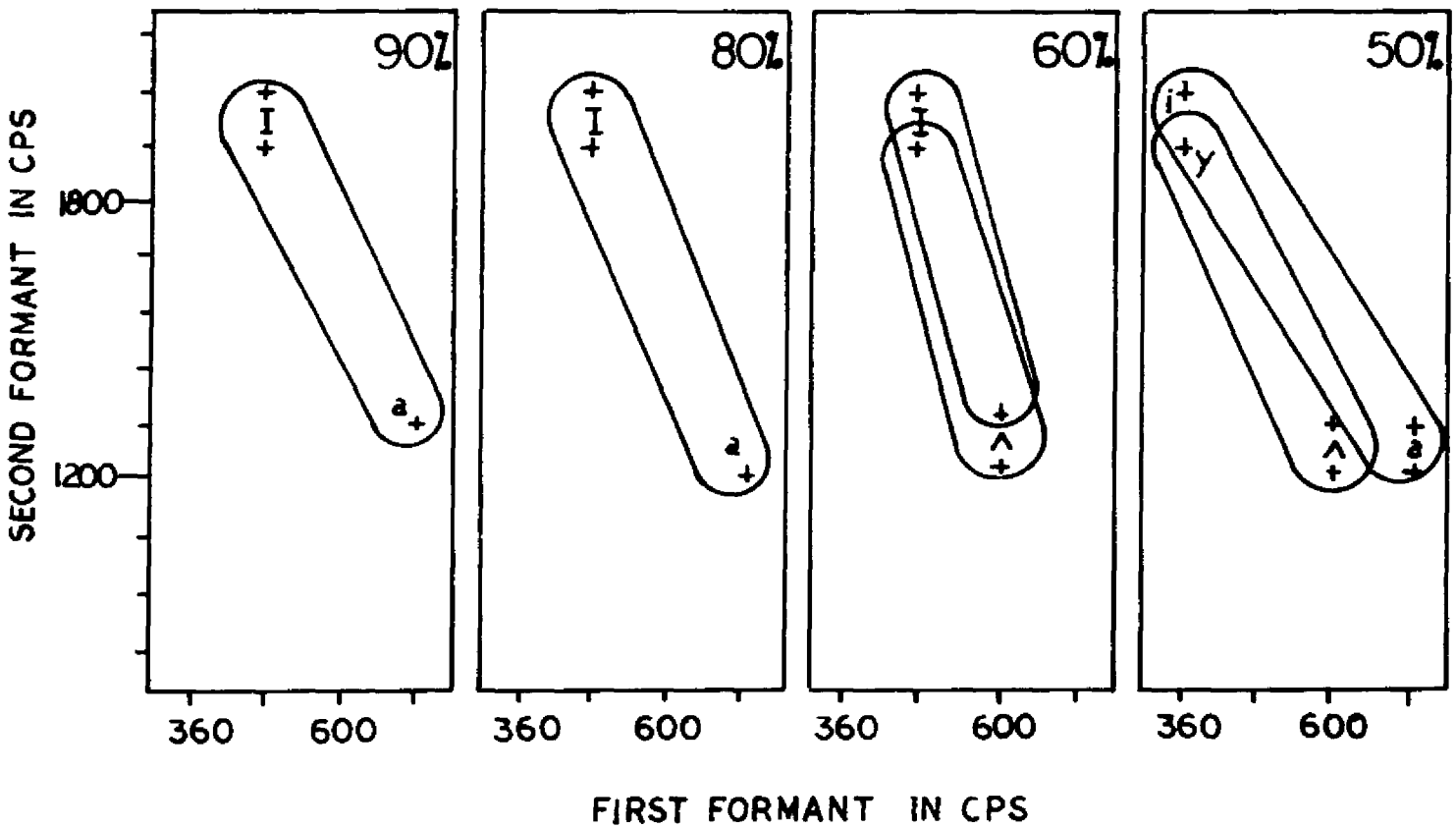

Figure 2.12.--Formant movements for different percentages of /a^i/.

Thus, the explicit status of these targets is perhaps doubtful and formant movements for /a^i/, although restricted to a more limited course than those for /$ɔ^i$/, might also be described in more general terms. As /a^i/ preferences decrease, the initial target moves from [a] to [ʌ]. The effects of a lower terminating first formant incurring lower /a^i/ 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 /a^i/ identification, a glide from [a] to [y,i] provides only marginal /a^i/ 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 than 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 the [a] to [ɪ] articulation. This also applies to the /$ɔ^i$/ movement. Since the tongue position for [ɔ] is farther back than that of [a], the articulatory speed of /$ɔ^i$/ is that much faster than any /a^i/ articulation.

In general then, the shift from /$ɔ^i$/ to /a^i/ 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 speed of articulation. Before discussing the phonemic implications of these data, the formant frequency characteristics of /a^u,o/ will be described first, as these sounds bear relationships similar to those of /$ɔ^i$,a^i/.

/a^u-o/ Distinction

Results of the continua.--As was apparent for the /$ɔ^i$-a^i/ continua, different third formants contribute little to the separation of /a^u-o/. This is shown in Figure 2.13 where second formant curves vary only slightly across the range of third formants. The maximum variation for any second formant is about 8 per cent but this variation is not consistent in terms of either third formant onset or extent. Thus, the earlier procedure of pooling responses across third formant data will be followed here.

The results of the /a^u-o/ continua are shown in Figure 2.14. Here, /a^u/ identification depends on covariation of first and second formant course. For both "A" and "B" patterns, primary contributions are made by both higher initial and higher terminal first formants, with highest /a^u/ preference accompanying the 720-600 cps first formant. The data for this formant are somewhat complicated by the number of /a/ responses which occurred along with /a^u,o/ responses. The /a/ curves for both "A" and "B" patterns are shown in Figure 2.15. The relatively strong /a/ preferences at the second formant

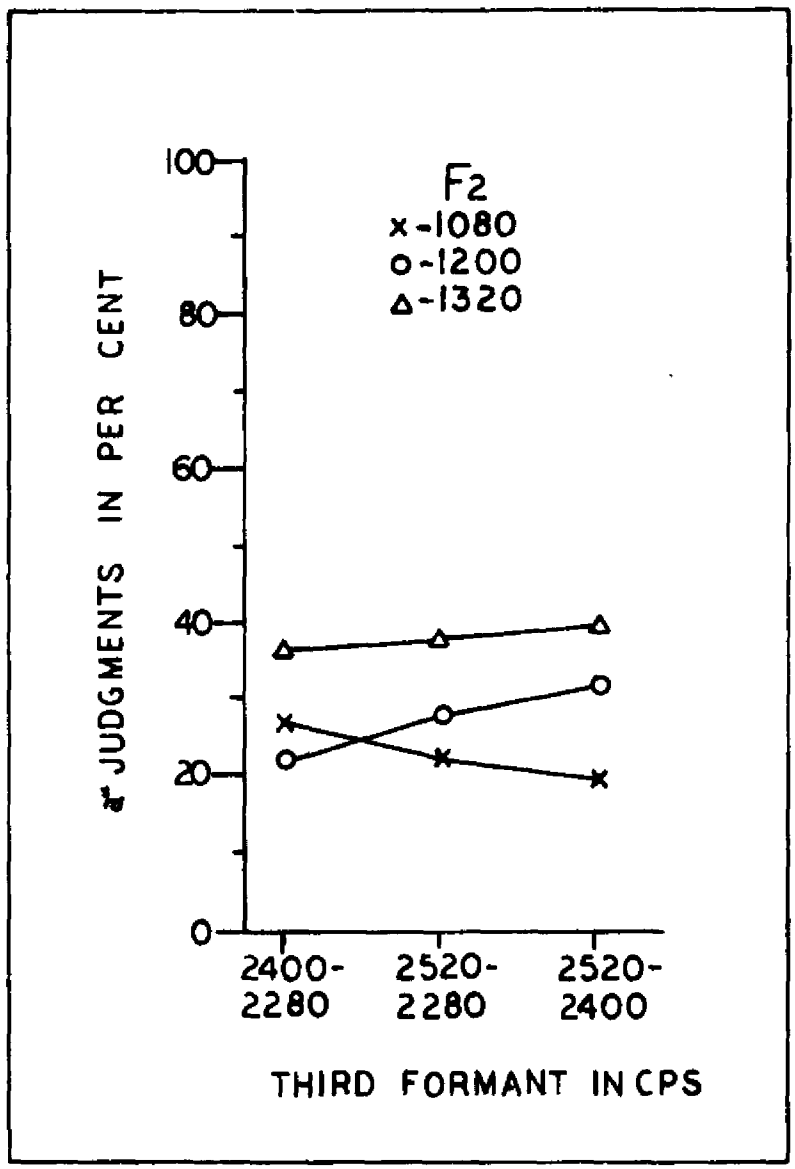

Figure 2.13.--Effects of third formant values on /a^{u}/ identification for the /a^{u}-o/ continuum. All data are pooled across first formant values and "A" and "B" patterns. Total N per stimulus=240.

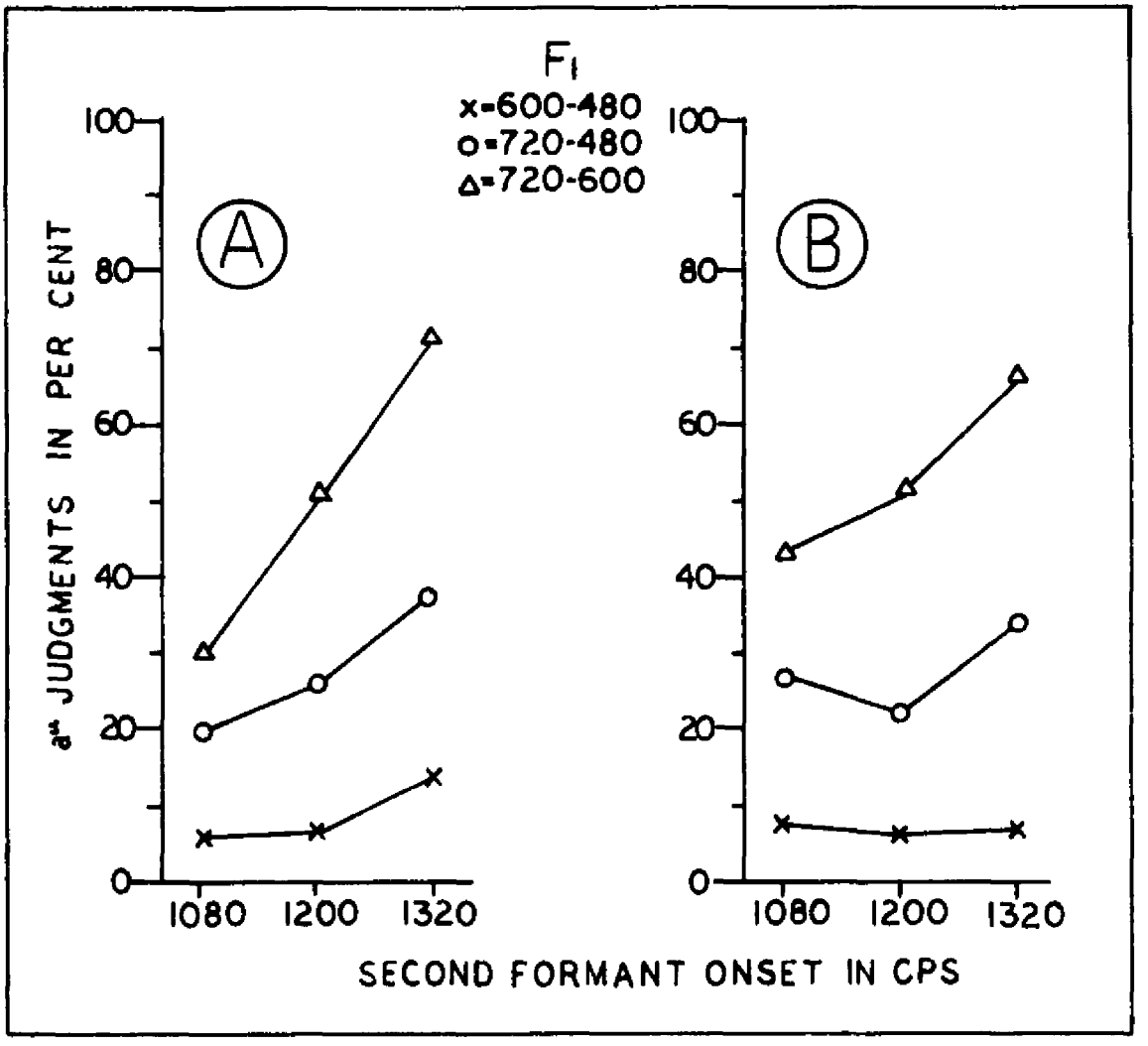

Figure 2.14.--Identification of the /a^{u}-o/ continuum. A=second formant termination at 960 cps, B=840 cps. Total N per stimulus=120.

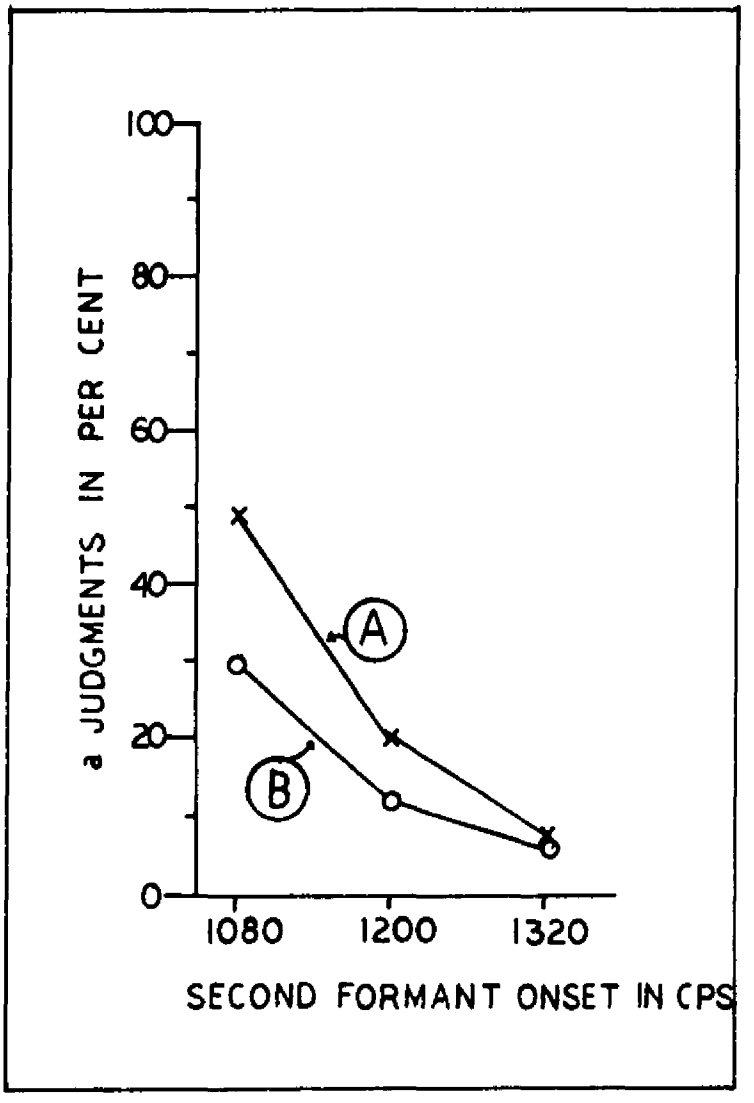

Figure 2.15.--/a/ responses for the /aᵘ-o/ continuum. A=second formant termination at 960 cps, B= 840 cps. First formant=720-600 cps. Total N per stimulus=120.

onset values of "A"-1080 cps can be expected inasmuch as the course of this formant extends only one harmonic lower at termination. After this point, progressively greater differences in second formant onset and termination values are accompanied by consistently fewer /a/ responses. It would be difficult to speculate on what effect two category judgments (/a^u-o/) would have had on listener responses, i.e. to which category, /a^u/ or /o/, listener's would have assigned these stimuli.

The effect of a higher terminating 720-600 cps first formant in the /a^u-o/ continua operates in a somewhat reverse fashion to a higher terminating first formant in the /$ɔ^i$-a^i/ continua. Second formants in this continua, unlike those for /$ɔ^i$-a^i/, extend down toward termination. Thus, a higher terminating first formant here, serves to increase the degree of convergence of first and second formants at termination. The perception of /a^u/ then, behaves along lines similar to the perception of /$ɔ^i$,a^i/, insofar as cues being assignable to extent of the first formant. As will be shown below however, this effect plays a more important role in separating /a^u-o/ than it does for /$ɔ^i$-a^i/.

Phonetic description.--The initial and terminal targets of /a^u-o/ are the same as those for initial /$ɔ^i$-a^i/ (except for the first and second formant coordinate of 480-1080 cps) and thus have distributions as shown in

Figure 2.9. There is however, an additional initial target used in the /a^{u}-o/ continuum. This is the coordinate at 720-1080 cps. The responses at this position are distributed similarly to those at the 720-1200 cps position, showing a majority of [a] responses (62 per cent), followed by smaller [ʌ] (30 per cent) and [ɔ] (6 per cent) judgments.[4]

The contours for /a^{u}/, which are plotted in Figure 2.16, show that recognition of /a^{u}/ requires a glide from [a] to [ɔ]. Moreover, rather large gradations in /a^{u}/ identification occur for different formant movements enclosed within the general [a] to [ɔ] areas. Highest /a^{u}/ recognition occurs for highest initial and terminal first and second formants with progressively lower /a^{u}/ preferences accompanying lower first and second formants. The limitation of the terminal portion of /a^{u}/ to [ɔ] is supported by Holbrook and Fairbanks' (1962) acoustical measurements which also found /a^{u}/ 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 /a^{u}/ termination closer to but not at [ʊ] (their data also show /a^{u}/ initiation closer to [ʌ] than [a]). Yet, these areas are

---

[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.

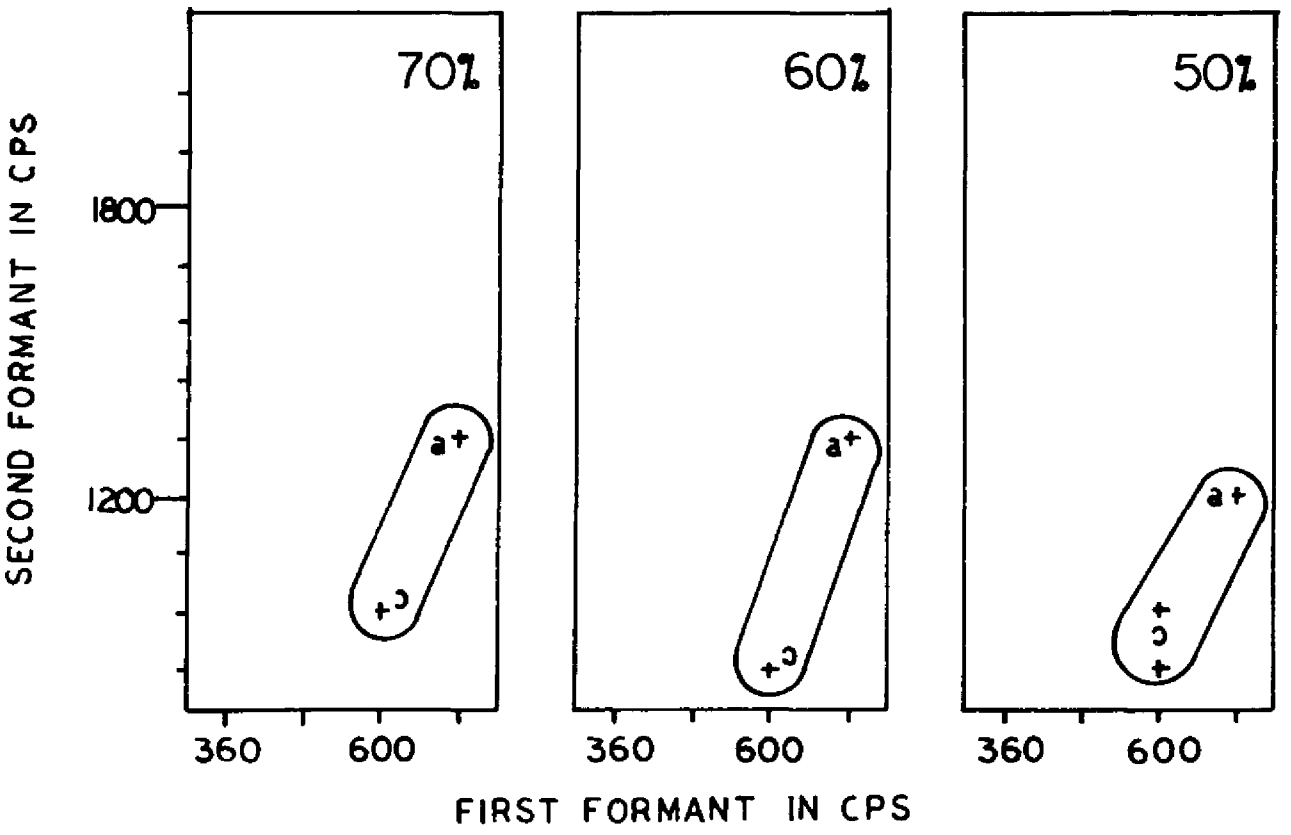

Figure 2.16.--Formant movements for different percentages of /a^{u}/.

more restricted than those for /ɔ^{i},a^{i}/.

The effects of a lower terminating first formant on the /a^{u}-o/ distinction are demonstrated clearly in the contours for /o/ (Figure 2.17). 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 /a^{u}/ 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 articulatory speed is generally greater for /o/ than /a^{u}/ (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 [o^{u}], i.e. the synthesized /o/ formants begin and terminate at higher frequencies but course in the same direction as [o^{u}], with the terminal targets of the /o/ patterns approaching the initial targets of real speech [o^{u}]. Both phonetic and articula-

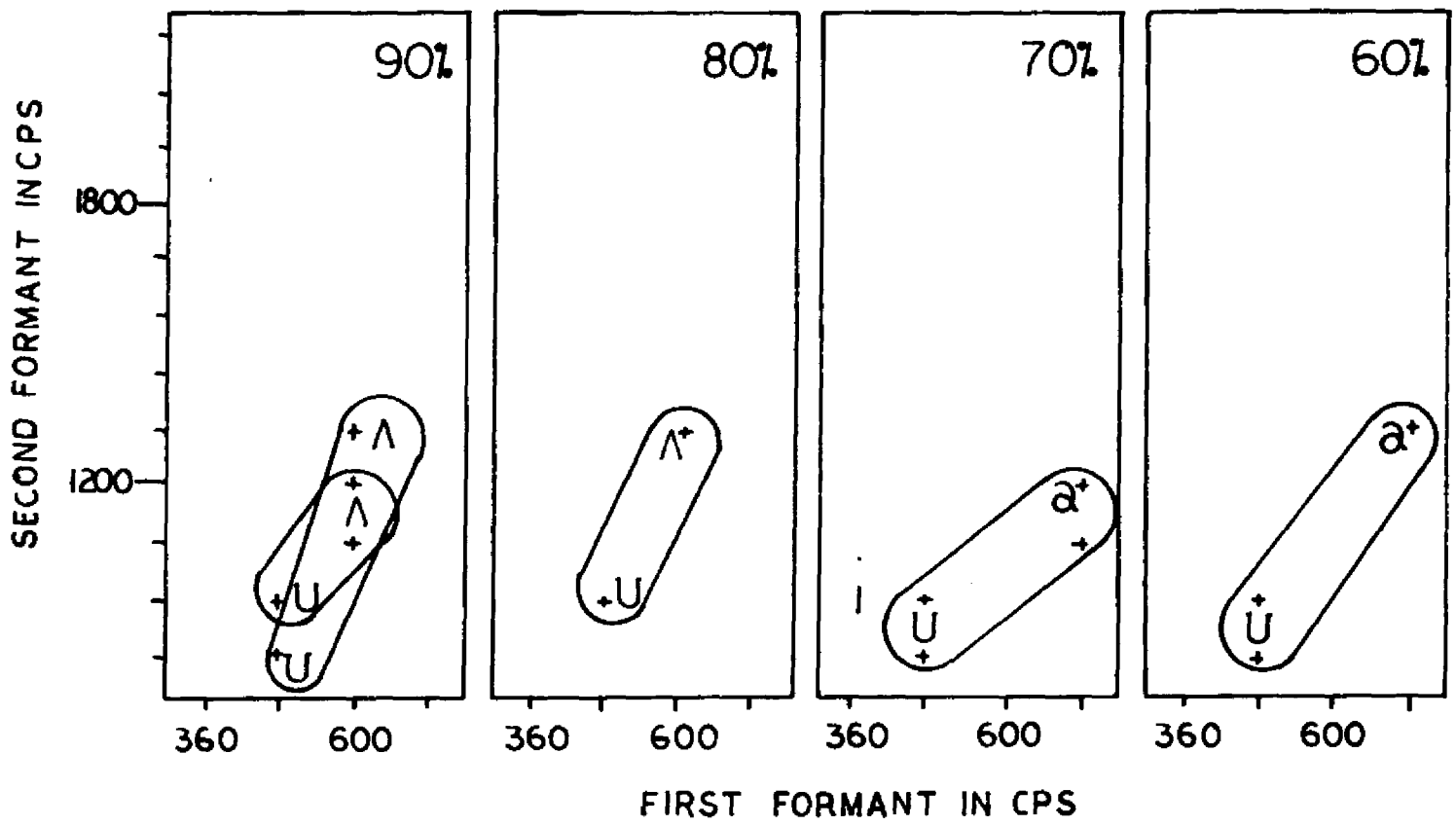

Figure 2.17.--Formant movements for different percentages of /o/.

tory comparisons between /a^u-o/ are further complicated by the fact that [o^u] occurs as a nonphonemic variant of /o/ and thus, has an offglide of no phonemic significance, as opposed to the offglide role in /a^u/.

Figure 2.18 summarizes the areas within which /$ɔ^i$,a^i,a^u,o/ are identified at least 50 per cent of the time. Both /$ɔ^i$,a^i/ course through relatively wide areas, showing some overlap at both onset and termination. Initial targets are generally distinct with /$ɔ^i$/ beginning at lower first and second formant areas. Covarying with /$ɔ^i$-a^i/ target shifts are shifts in articulatory speed with the rate of glide movement progressively slowing down for /a^i/. /a^u,o/ on the other hand, show areas which overlap at onset but separate toward termination. Also, /o/ courses through wide areas while the route for /a^u/ is limited. Thus, in strict phonetic terms, a description of /$ɔ^i$,a^i/ as bounded by glides from [ɔ] to [ɪ] and [a] to [ɪ], respectively, might be too limiting. On the other hand, an [a] to [ɔ] description of /a^u/ might be more justified.

The results of this experiment have a certain relevance to the major phonemic analyses of /$ɔ^i$,a^i,a^u/ and, in addition, provide some basis for developing an overall descriptive account of these sounds.

Phonemic Interpretation

The theories that /$ɔ^i$,a^i,a^u/ consist of either

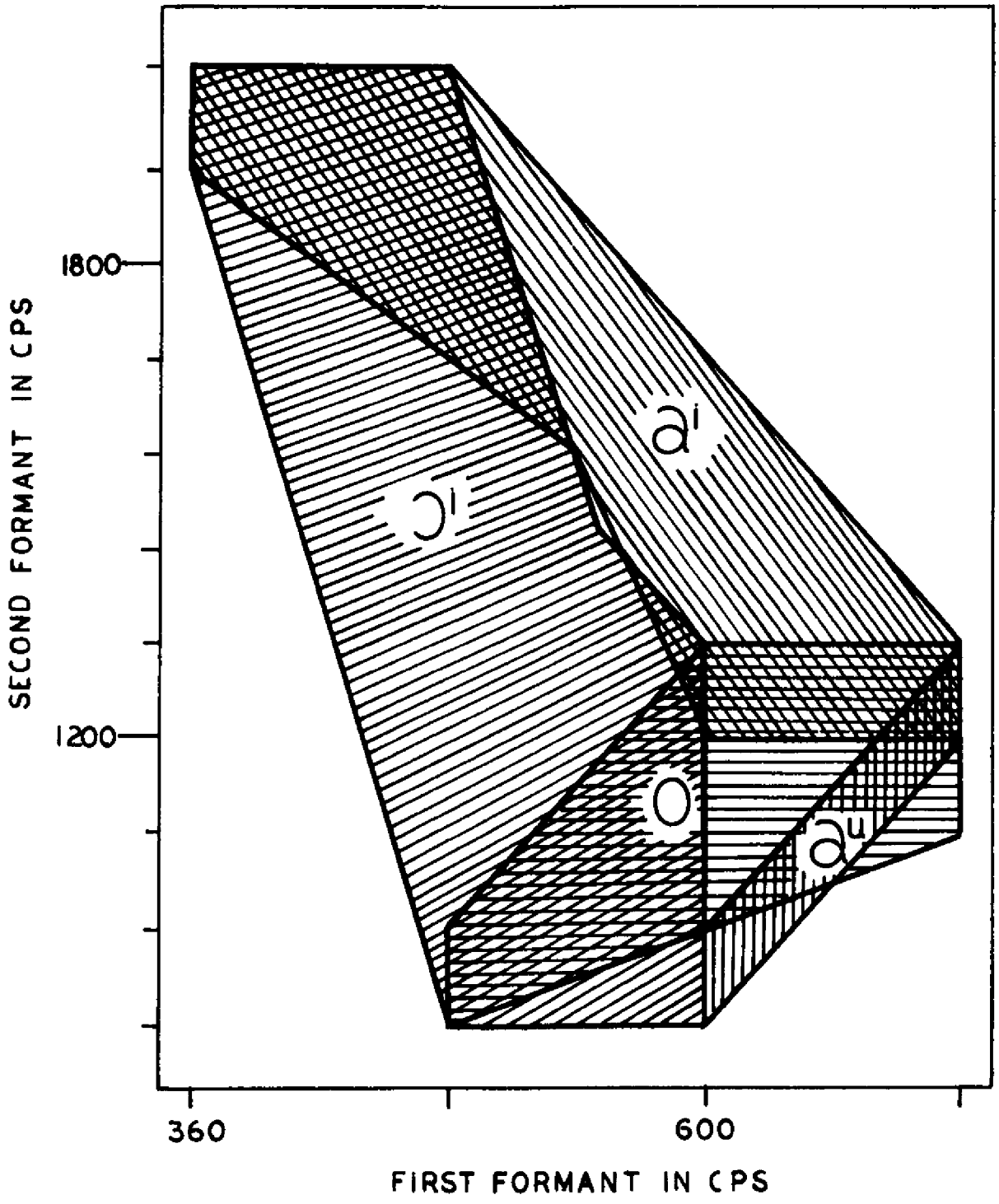

Figure 2.18.--Summary of formant areas enclosing /ɔ^{i},a^{i},a^{u},o/. Boundaries enclose formant movements identified at least 50 per cent of the time.

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, /ɔ^{i},a^{i},a^{u}/ would each necessarily contain steady state initial and terminal targets. The stimuli used in this study contained neither, thus showing the nonessential 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 /ɔ^{i},a^{i},a^{u}/ "usually" contain both initial and terminal steady states; but the actual status of these steady portions seems questionable in light of the criteria used for classification. According to their definition, a target of at least 20 msec is classified as 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. Thus, since steady targets are neither required for perception nor found consistently in real speech measurements, a description of /ɔ^{i},a^{i},a^{u}/ as actual sequences of two vowels does not seem justified.

In the vowel plus semivowel theory, a different set of acoustic and phonetic variables are 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 /ɔ^{i},a^{i},a^{u}/. 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 /ɔ^{i},a^{i},a^{u}/. 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. Figure 2.19 shows how O'Connor's preferred /jɔ,ja,wa/ in reverse order compare with /ɔ^{i}, a^{i},a^{u}/. This illustration shows several features of initial /j,w/ not characteristic of /ɔ^{i},a^{i},a^{u}/. First, neither first nor second formant termination values of reversed /jɔ,ja,wa/ correspond to those of /ɔ^{i},a^{i},a^{u}/. First formant offsets for both /j,w/ occur at about 240 cps, while terminal target frequencies range from

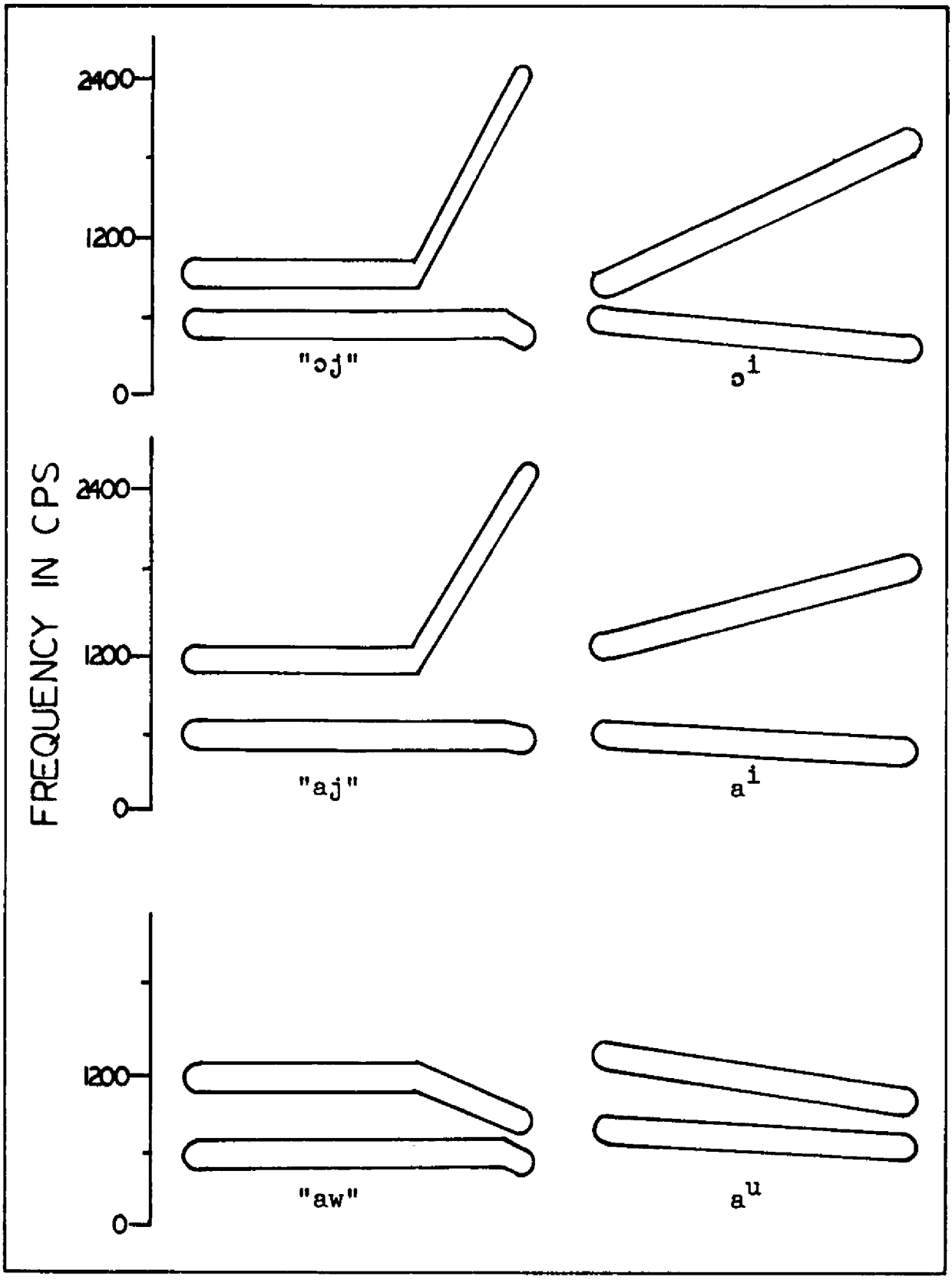

Figure 2.19.--Spectrographic comparisons of reverse order /jɔ,ja,wa/ (as adapted from O'Connor et al, 1957) and /ɔ^{i},a^{i},a^{u}/. Total duration of "/ɔj,aj,aw/"=300 msec, /ɔi,a^{i},a^{u}/=250 msec.

360-480 cps for /ɔ^{i},a^{i}/ and are fixed at 600 cps for /a^{u}/. The second formant terminal values for /j/ are higher than the terminal values for /ɔ^{i},a^{i}/, 2760 cps versus 2040 and 1920 cps, respectively, and the second formant termination of /w/ is lower than that of /a^{u}/, 600 cps for /w/ as opposed to 960 cps for /a^{u}/. Thus, since the terminal frequency positions of hypothetical /ɔj,aj,aw/ are not compatible with those of /ɔ^{i},a^{i},a^{u}/, the suitability of a vowel plus semivowel description of /ɔ^{i},a^{i},a^{u}/ seems doubtful.[5] In addition, another obvious difference between the two 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 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 color, e.g. a shift from /jɛ-iɛ/ and /wɛ-iɛ/. Thus, the impression for hypothetical /ɔj, aj,aw/ would apparently be distinct from /ɔ^{i},a^{i},a^{u}/ insofar as the duration of /j,w/ would not be great

---

[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.

enough to produce diphthongal quality.[6] Further, if these glide durations were equal to those shown for /ɔ^{i}, a^{i},a^{u}/, the glide itself, without the preceding steady state vowel, would carry the diphthongal quality.

Apparently then, the primary feature of /ɔ^{i},a^{i}, a^{u}/ 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 its 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, /ɔ^{i},a^{i},a^{u}/ each differed in terms of both rate of formant movement and initial and terminal frequency levels. Thus, whether /ɔ^{i},a^{i},a^{u}/ 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.

---

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

III. EFFECTS OF DURATION ON THE PERCEPTION OF /ɔ^{i},a^{i},a^{u}/

In the previous experiment, /ɔ^{i},a^{i},a^{u}/ were identified by differences in the course and extent of formant frequency transitions. Accordingly, /ɔ^{i},a^{i}, a^{u}/ were each shown to be characterized by a specific course of formant movement an a particular set of phonetically describable targets. The purpose of the next experiment was to determine whether perception of /ɔ^{i},a^{i},a^{u}/ is cued 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.

The extending of the formant transitions of /ɔ^{i},a^{i},a^{u}/ toward termination, is controlled by the time dimension which governs frequency change and, at any given point, 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 /ɔ^{i},a^{i},a^{u}/ 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 cut-off (consequently, 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 /ɔ^{i},a^{i},a^{u}/ identification. In each group, the full duration pattern was, in effect, reduced in duration from 250 msec to 100 msec, in steps of 10 msec, in one case beginning at the terminal target and in the other, beginning at the initial target.[1] Figure 3.1 illustrates the procedure used in constructing patterns based on the basic /a^{i}/ 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

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

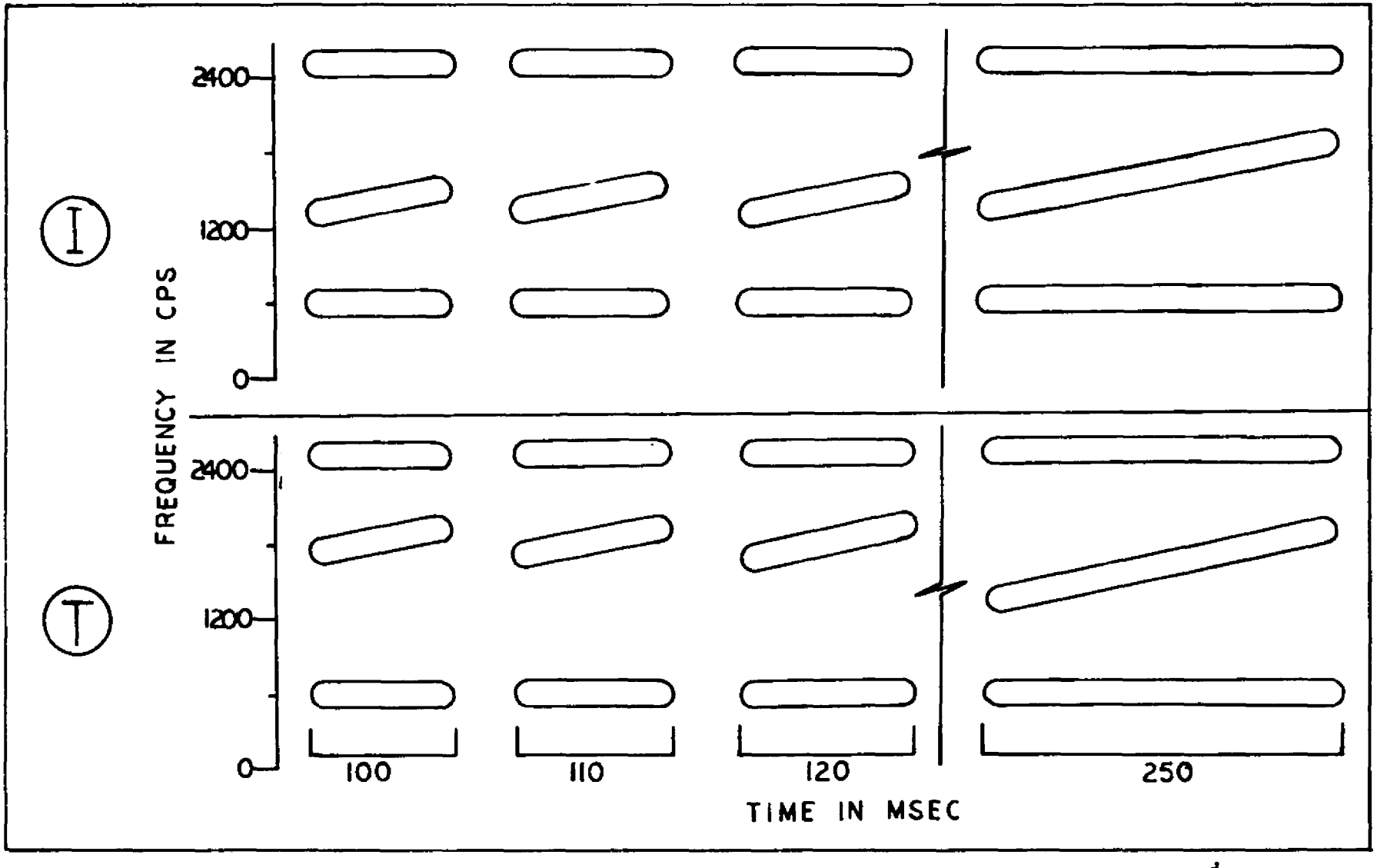

Figure 3.1.--Schematic illustration of stimuli used to produce /a-a^{i}/ shift. I=patterns whose second formant initial targets remain preserved, T=patterns whose second formant terminal targets remain preserved. In each case, duration is varied from 100 msec to 250 msec, providing a shift in terminal target frequency (I) or initial target frequency (T) positions with rate of frequency change remaining fixed.

quality. Full duration /a^i/ consisted of steady state first and third formants of 720 and 2760 cps and a second formant which extended from 1320 cps to 1920 cps. The top row of Figure 3.1 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, will subsequently be labelled "I". These patterns produce an /a-a^i/ shift. In the second series of patterns, labelled "T", the terminal target frequency remains fixed through time, thus producing an /ɛ-a^i/ shift.[2] In each series, duration was varied from 100 msec to 250 msec in steps of 10 msec, producing a total of 16 "I" and 16 "T" patterns appropriate to shifts from vowel to /a^i/.

The procedures used for varying the duration of /a^i/ were followed in varying the durations of /$ɔ^i$,a^u/. Full duration /$ɔ^i$/ consisted of steady state first and third formants of 600 cps and 2520 cps and a second formant extending from 840 cps to 2040 cps; /a^u/ consisted of 720 cps and 2400 cps first and third formants and a

---

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

1320-960 cps second formant.

Test batteries.--The procedures used in preparing the /ɔ^{i},a^{i},a^{u}/ stimuli of Experiment I were generally followed here. "I" and "T" series stimuli were arranged in two separate master lists. Each list consisted of all appropriate /ɔ^{i},a^{i},a^{u}/ stimuli, replicated 4 times each, thus containing a total of 192 (16x3x4) test items. The synthesized carrier, "The word is," was inserted 0.5 second before each stimulus and at successive intervals of approximately four seconds. Subjects were required to label the "I" series stimuli as /ɔ^{i},a^{i},a^{u},ɔ,a/ and the "T" series stimuli as /ɔ^{i},a^{i},a^{u},ɛ,ʌ/ on prepared answer forms.[3] The choice of simple vowel responses was based on exploratory work. 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 /ɔ^{i}/

Results of the "I" and "T" series patterns for /ɔ^{i}/ are shown in Figures 3.2 and 3.3.[4] Both curves

---

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

[4]For all /ɔ^{i},a^{i},a^{u}/ series, subject responses, except where otherwise noted, were of the two-category type, e.g. /ɔ/ or /ɔ^{i}/. Thus, all graphs are plotted as per cent diphthong recognition.

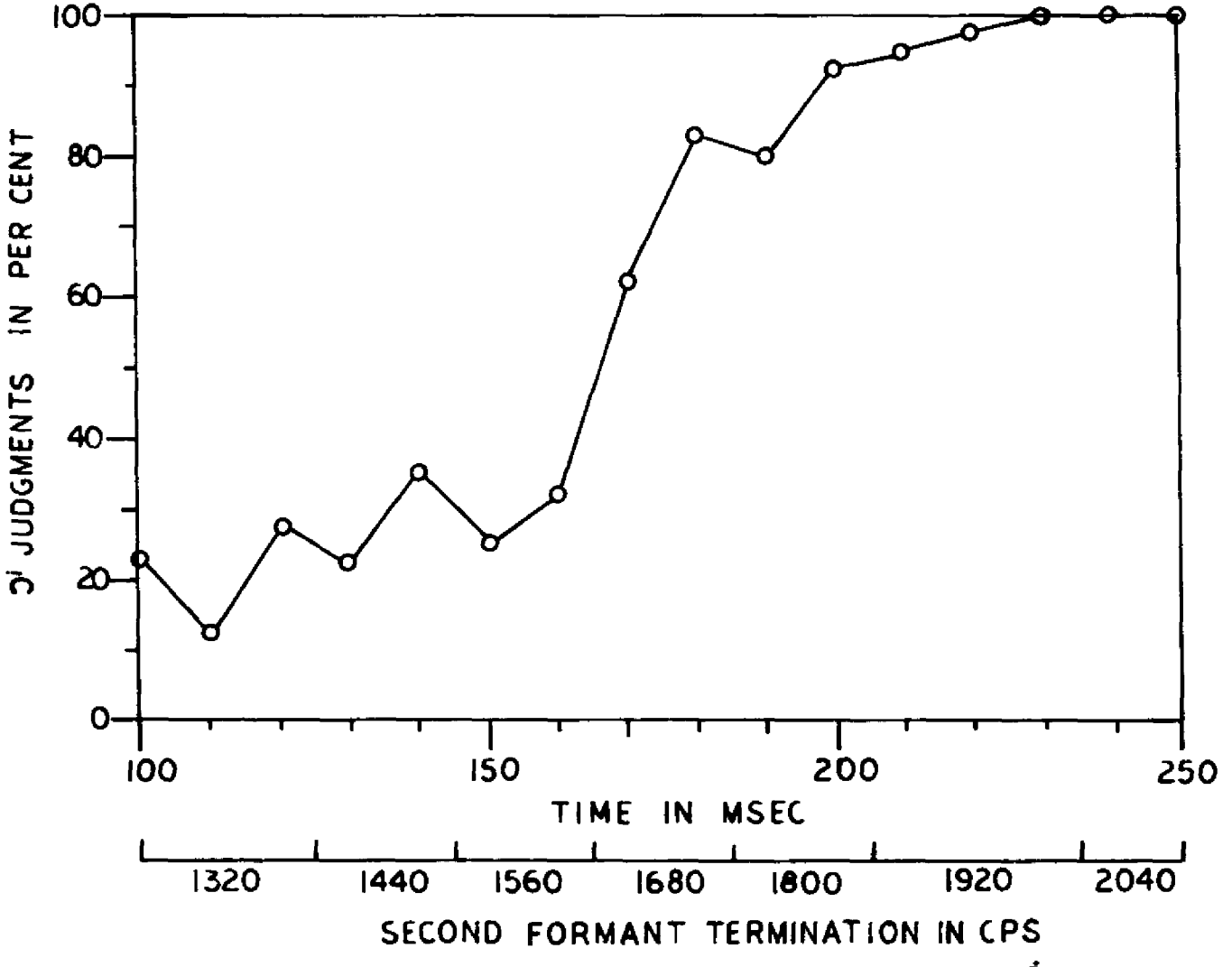

Figure 3.2.--Effects of duration on /ɔ-ɔ^{i}/ shift. Second formant termination varies from 1320-2040 cps while onset remains fixed at 840 cps. Per cent /ɔ/= 100 minus per cent /ɔ^{i}/. Total number of observations for this and all other I and T series patterns=40 per stimulus.

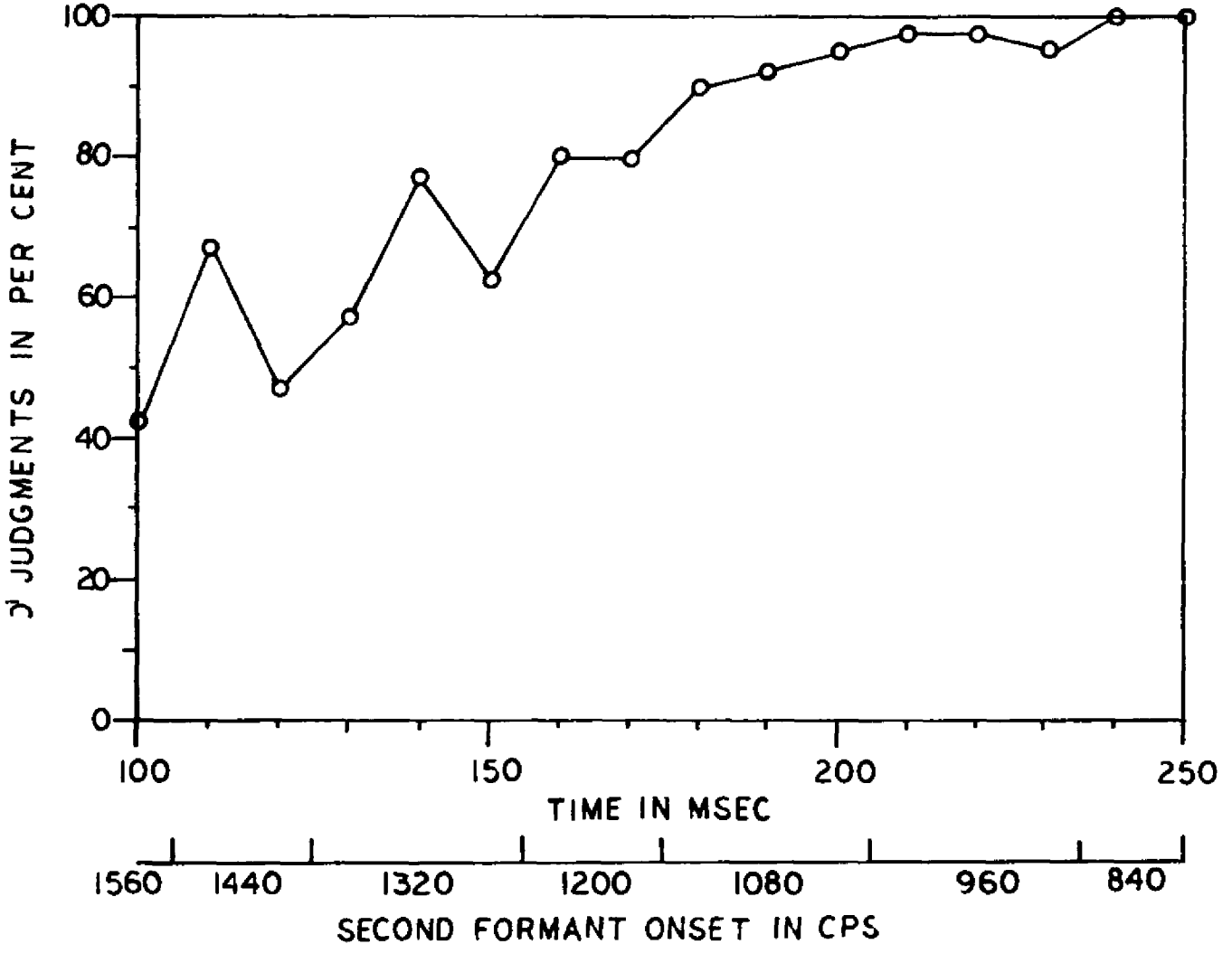

Figure 3.3.--Effects of duration on /ɛ-ɔ^{i}/ shift. Second formant onset varies from 1560-840 cps while the terminal target remains fixed at 2040 cps.

indicate that duration rather than frequency position provides the primary perceptual cue for separating /ɔ/ from /ɔ^{i}/. For the "I" patterns, the /ɔ-ɔ^{i}/ shift (more than 50 per cent /ɔ^{i}/ recognition) occurs at 170 msec and a corresponding second formant center frequency of 1680 cps. Although this shift accompanies a change in frequency from 1560-1680 cps, /ɔ^{i}/ recognition does not level off at this frequency position but rather increases across it through time. The increase in /ɔ^{i}/ recognition across this and other second formant cut-off frequencies would not be expected if the cut-off frequency position were the primary cue. This 1680 cps position in combination with the first formant of 600 cps is appropriate to acoustic positions for [æ]. The glide at this point then, courses from [ɔ] to [æ]. but still provides greater /ɔ^{i}/ than /ɔ/ intelligibility. As the shift to /ɔ^{i}/ continues, second formant termination approaches acoustic positions more appropriate to [ɛ].

The /ɛ-ɔ^{i}/ shift, shown in Figure 3.3, occurs earlier in time than does the /ɔ-ɔ^{i}/ shift, at 130 msec. This curve also clearly demonstrates that the effect of duration is greater than that of onset frequency position. The second formant onset at 130 msec is 1320 cps, which in combination with the 600 cps first formant was phonetically described earlier as [ʌ] (Figure 2.9). Thus, the targets for these patterns are appro-

priate to [ʌ] and [ɪ], with the glide between providing approximately 66 per cent /ɔ^{i}/ intelligibility across the duration range of 130-150 msec. /ɛ/ identification for this group of patterns is approximately 30 per cent and /a^{i}/ identification, 4 per cent.[5] In the previous experiment however, judgments for a glide between these same two targets favored /a^{i}/ over /ɔ^{i}/, 67 per cent to 33 per cent (Figure 2.6).[6] Figure 3.4 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 /ɔ^{i}/ from /a^{i}/. This effect is further evident at the 1200 cps position where /a^{i}/ judgments, although somewhat less expected than at the 1320 cps position, are nonetheless, virtually absent.

The results of this series of patterns clearly demonstrate that the second formant course for /ɔ^{i}/

---

[5]All /a^{i}/ judgments, which never exceeded 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.

[6]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 /a^{i}/ perception.

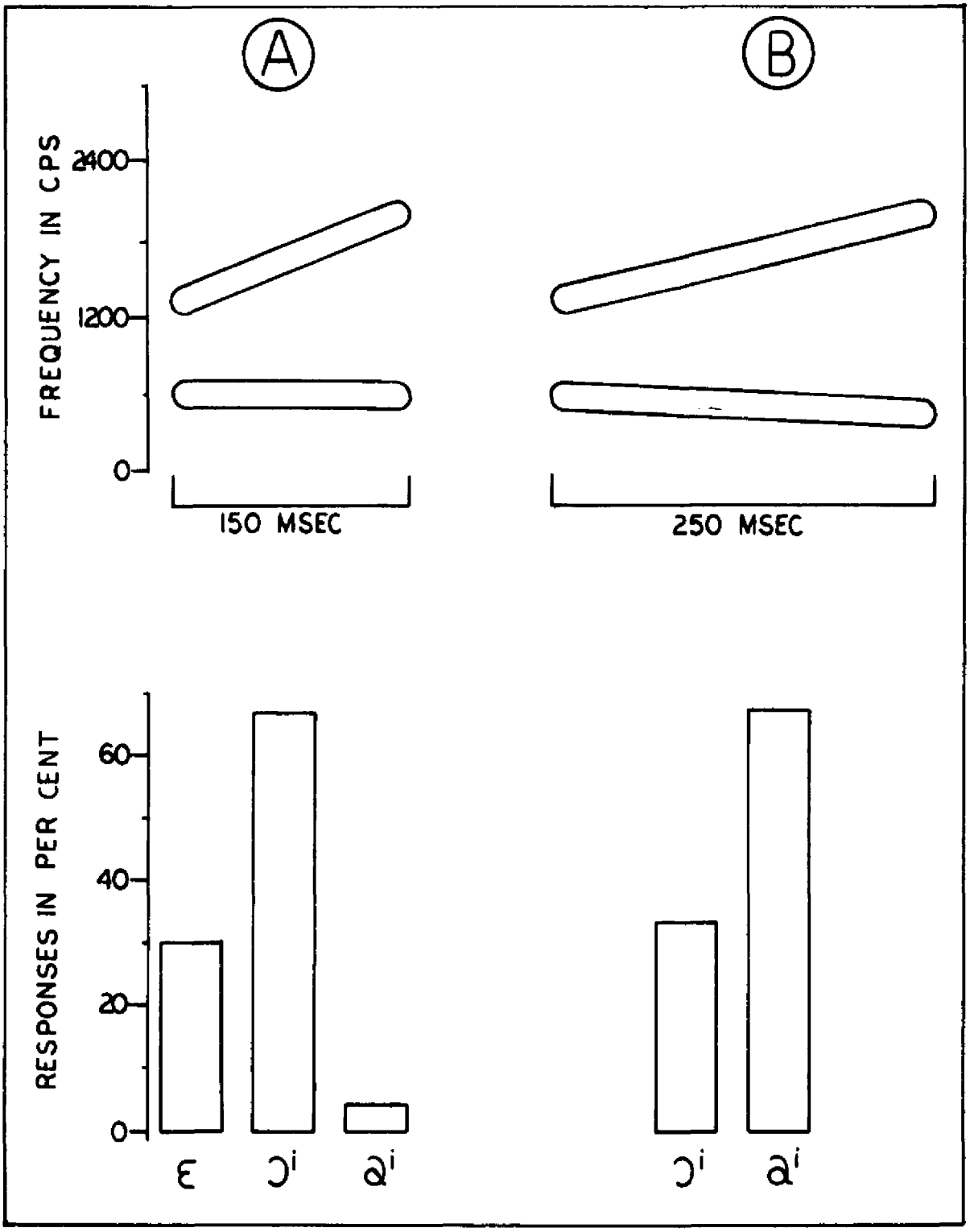

Figure 3.4.--Comparison of stimuli and responses, each showing almost identical formant values but different rates of second formant change. A=Experiment I (Figure 2.6), B=Experiment II (Figure 3.3).

need not necessarily begin or terminate at the bounding targets for the diphthongal quality of /ɔ^{i}/ to be perceived. /ɔ^{i}/ is perceived as a diphthong by consequence of duration and further, is separated from other diphthongs, notably /a^{i}/, by its greater rate of formant frequency change or correspondingly, faster speed of articulation.

Time-Frequency Effects for /a^{i}/

The effects of duration for both "I" and "T" patterns appropriate to /a^{i}/, shown in Figures 3.5 and 3.6, are similar in nature to those for /ɔ^{i}/. The /a-a^{i}/ and /ɛ-a^{i}/ shifts are clearly time based, consistently progressing across the ranges of different cut-off frequencies. The /a-a^{i}/ shift occurs at 180 msec, with the transition at this point extending between acoustic positions appropriate to [a] and [æ]. Strong /a^{i}/ preferences require the full 250 msec duration. The /ɛ-a^{i}/ shift is consistently stronger through time than the /a-a^{i}/ 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 /a^{i}/ preferences increase to a maximum at full duration. Both the "I" and "T" curves for /a^{i}/ rise at a slower rate than those for /ɔ^{i}/, the difference perhaps being

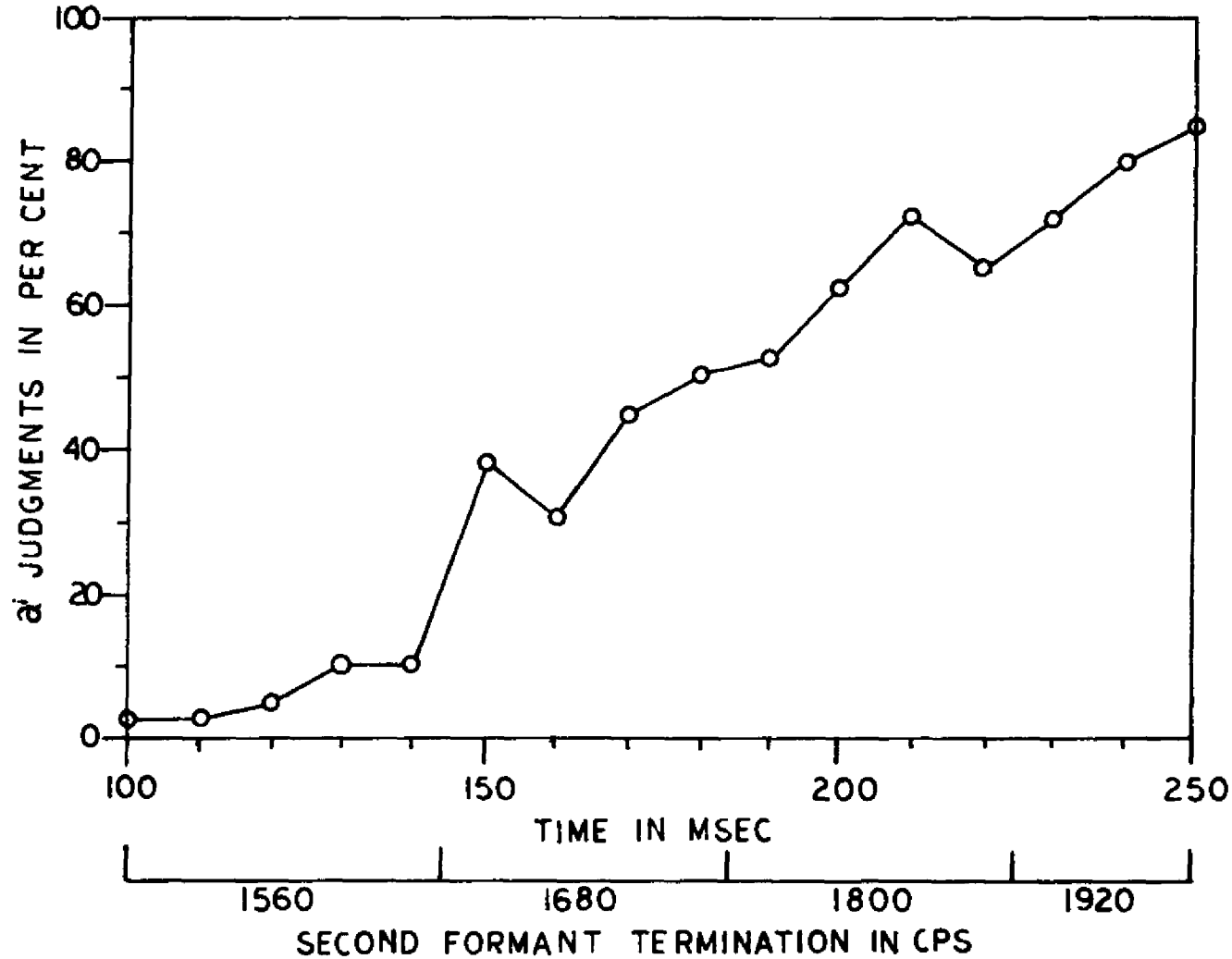

Figure 3.5.--Effects of duration on /a-a^{i}/ shift. Second formant terminal target varies from 1560-1920 cps while onset remains fixed at 1320 cps.

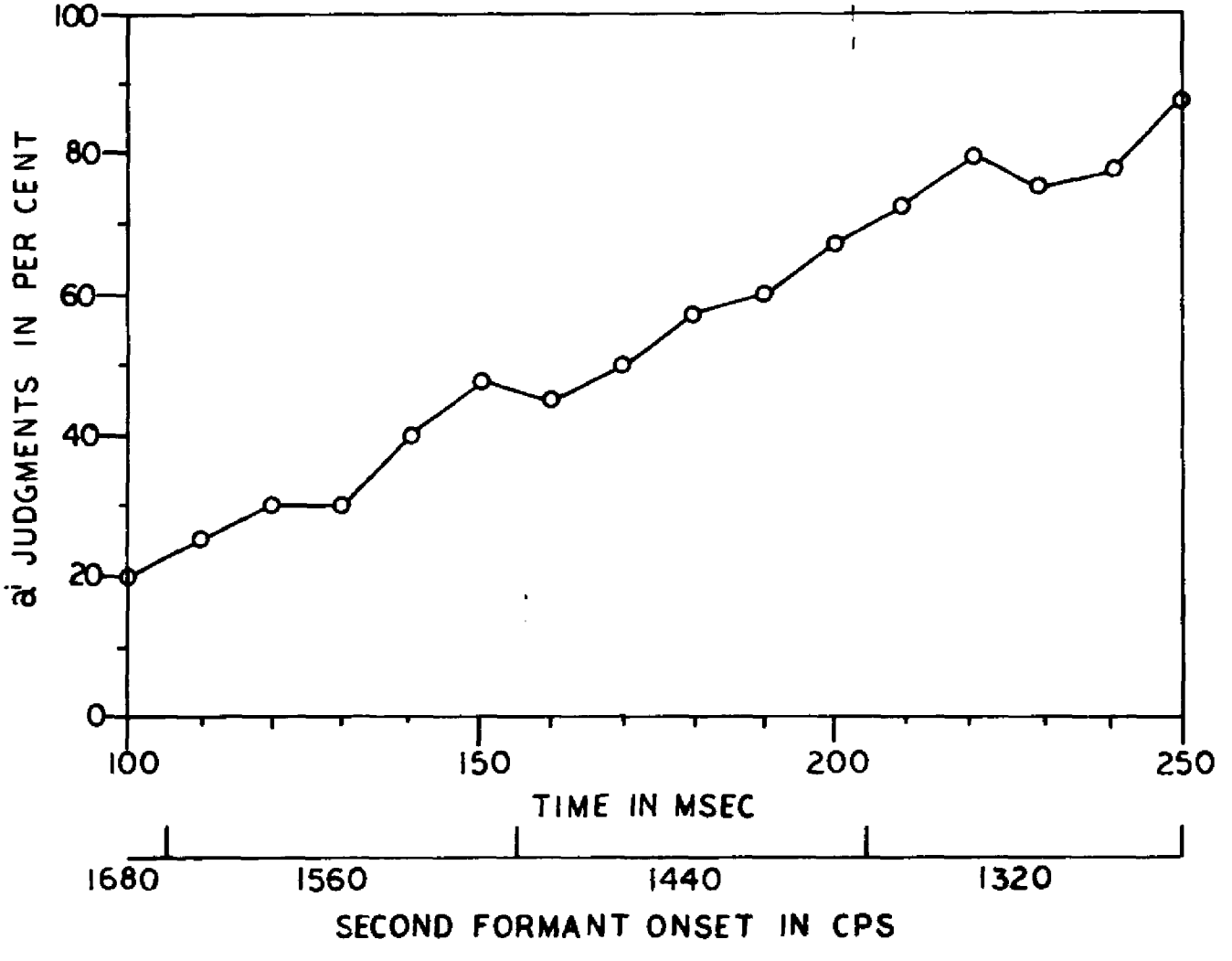

Figure 3.6.--Effects of duration on /ɛ-a^{i}/ shift. Second formant onset varies from 1680-1320 cps while the terminal target remains fixed at 1920 cps.

attributable to the greater rate of frequency change characteristic of /ɔ^{i}/.

Time-Frequency Effects for /a^{u}/

Figures 3.7 and 3.8 show the effects of duration on the perception of /a^{u}/. The shift from /a/ to /a^{u}/ is rather sharp, occurring at 150 msec with sharp increases in /a^{u}/ continuing until 170 msec before slowing down. Here, as expected, the shift is clearly time based. Strong /a^{u}/ preferences occur at about 3/4 full duration, a duration similar to that for strong /ɔ^{i}/. The /ʌ-a^{u}/ curve also rises sharply, showing the shift to /a^{u}/ occurring at 160 msec and strong /a^{u}/ preferences at 190 msec or also at close to 3/4 full duration. As was shown in the continua data, /a^{u}/ is bounded by specific formant positions with reductions in /a^{u}/ preference accompanying even small changes in first and second formant onset and termination frequencies. In this stimulus series however, the presence of a higher terminating, steady state first formant has no adverse effect on /a^{u}/ preference, but rather perhaps enhances /a^{u}/. Also, for both "I" and "T" patterns, a lower second formant cut-off point does not incur lower /a^{u}/ responses, except at shorter durations. Since the course of the second formant for /a^{u}/ changes rather slowly and mostly within the general area of [a], the formant movements of /a^{u}/ do not glide through intermediate vowel posi-

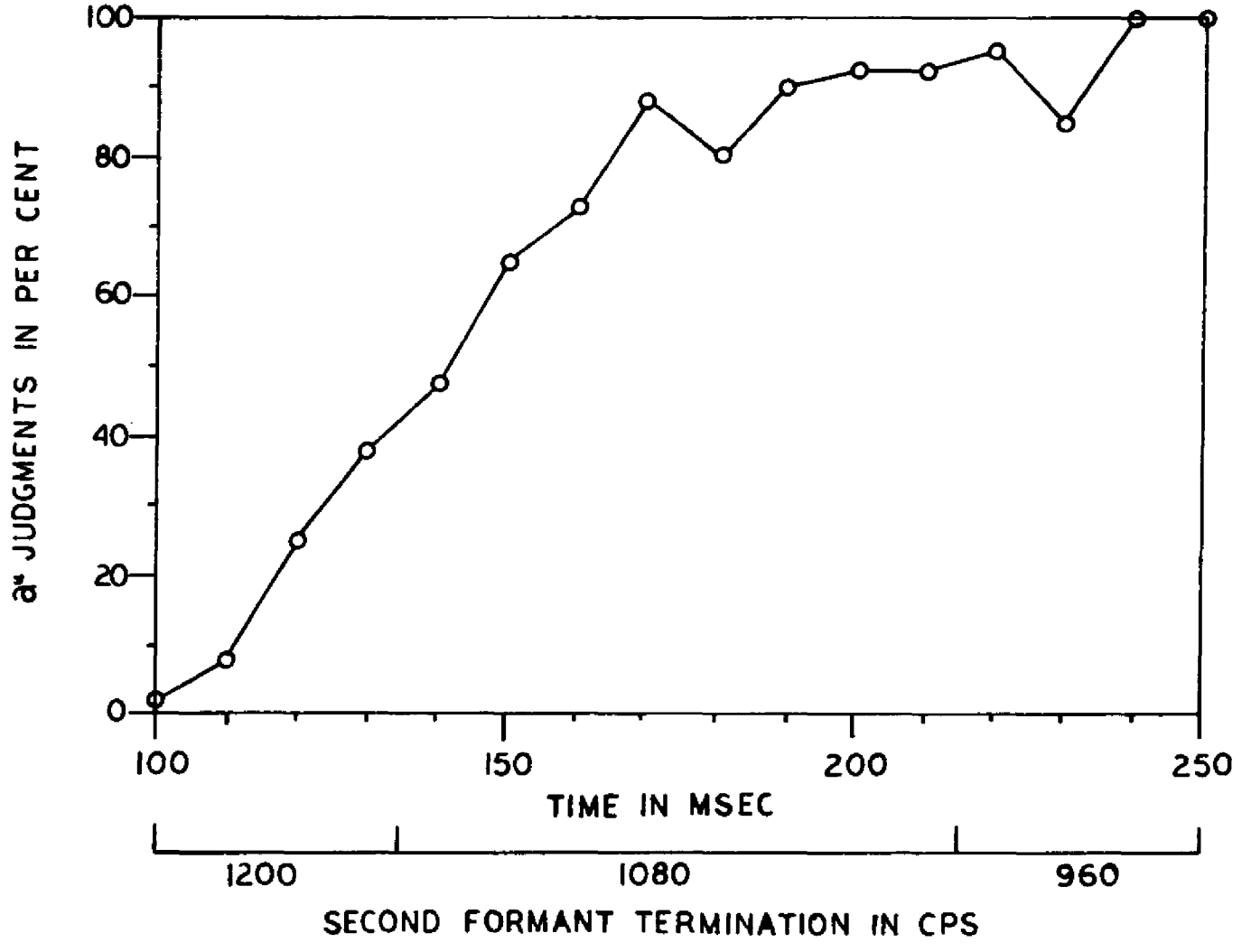

Figure 3.7.--Effects of duration on /a-a^{u}/ shift. Second formant termination varies from 1200-960 cps while onset remains fixed at 1320 cps.

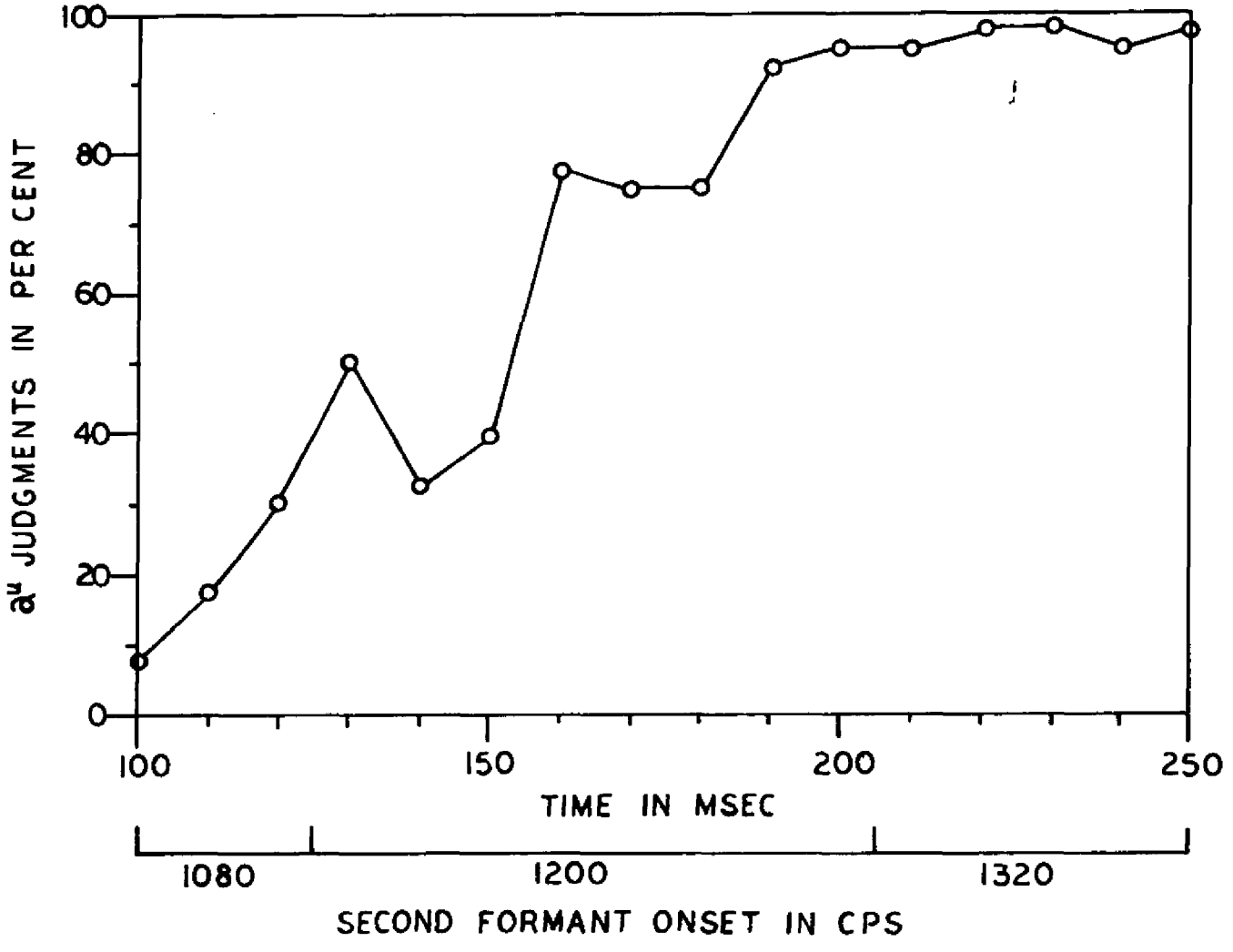

Figure 3.8.--Effects of duration on /ʌ-a^{u}/ shift. Second formant onset varies from 1080-1320 cps while the terminal target remains fixed at 960 cps.

tions until the glide approaches the positions of [ɔ], slightly before termination.

Discussion

The results of this experiment clearly demonstrate that transition duration rather than change in frequency position provides primary cues for separating vowel versus diphthong and consequently, 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 /ɔ^{i},a^{i},a^{u}/ 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 /ɔ^{i},a^{i},a^{u}/ under three different conditions of speech rate. Table 3.1 shows the results of these measurements as well as those made for /o/. All diphthongs were produced in a single sentence context, "The <u>boy</u> passed <u>by</u> the <u>bow</u> of the <u>boat</u>," with samples obtained from two male informants. For both speakers, the results of all measurements for /ɔ^{i},a^{i}, a^{u}/ occurred as would be expected within the frame of the perception results. /ɔ^{i},a^{i}/ 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

TABLE 3.1.--Real speech measurements of /ɔ^{i}, a^{i}, a^{u}, o/ under three different conditions of rate. Measurements shown are averages for two speakers.

	Rate	Duration (msec)	F2-Initial (cps)	F2-Terminal (cps)	F2 Change (cps/msec)
/ɔ^{i}/	Slow	153	655	1520	5.7
	Moderate	113	690	1265	5.3
	Fast	75	720	1200	5.8
/a^{i}/	Slow	163	930	1475	3.4
	Moderate	110	935	1350	3.7
	Fast	88	980	1240	3.8
/a^{u}/	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

not to the same degree as terminal target levels decrease. In each case however, second formant rate of change remains fixed. The measurements for the different durations of /a^{u}/ 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 /ɔ^{i},a^{i}/, occur uniformly for both the initial and terminal target levels. The second formant rate of change for [o^{u}], on the other hand, tends to slow down with increased rate of production. This is not unexpected in light of the [o^{u}] 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 suggested in the first experiment that the formant movements of /ɔ^{i},a^{i},a^{u}/ are not compatible with those of a vowel plus vowel or vowel plus semivowel 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 /ɔ^{i},a^{i},a^{u}/ are characterized primarily by an invariant rate of formant frequency change. Both the vowel plus vowel and vowel plus semivowel theories suggest that the distinction between /ɔ^{i}/ and /a^{i}/, for

example, is attributable to differences in initial target position ([ɔ] versus[a]), with the gliding movement serving only the simple vowel-diphthong separation.[7] 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-diphthong separation). Apparently then, since the target positions serve no phonemic role, /ɔ^{i},a^{i},a^{u}/ cannot be described as sequences of two phonemes.

The results of the previous experiment specified the formant movements appropriate to identification of /ɔ^{i},a^{i},a^{u}/, 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^{i},a^{i},a^{u}/.

The diphthongs /ɔ^{i},a^{i},a^{u}/ are each characterized by gliding movements from one vowel area to another. The onset and termination points of these glides are general for /ɔ^{i},a^{i}/ and more specific for /a^{u}/. The location 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

---

[7]In the case of /a^{i}-a^{u}/ however, the direction of the glide, whether fronting or retracting, respectively, is relevant in the distinction.

movement is directed within a particular unit of time. The movement through time provides the primary cue for diphthong versus simple vowel perception and the direction of glide separates /ɔ^{i},a^{i}/ from /a^{u}/. The /ɔ^{i}-a^{i}/ distinction is attributable to rate of frequency change or in articulatory terms, speed of movement. Since /ɔ^{i}, a^{i},a^{u}/ 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^{I},a^{ʊ}].

IV. SUMMARY AND CONCLUSIONS

The diphthongs /ɔ^{i},a^{i},a^{u}/ are each characterized by a pronounced gliding movement through a particular path in the vowel space. The gliding movements are evidenced by formants that either rise or fall sharply depending on the values of the adjacent targets. This study was concerned with the effects of both the course and duration of these movements on the recognition of /ɔ^{i},a^{i},a^{u}/. In the first part of the study, various formant transitions appropriate to classes of /ɔ^{i}-a^{i}/ and /a^{u}-o/ were synthesized on the Haskins Laboratories Pattern-Playback, converted to sound and presented to ten phonetically trained listeners for purposes of phoneme labelling. To determine the phonemic status of the diphthong targets, steady state vowels, with formants corresponding to the initial and terminal targets of all diphthong stimuli, were also produced. The results of these two listening tests provided an acoustic-perceptual description of each phoneme. Both /ɔ^{i},a^{i}/ course through relatively wide formant areas showing some overlap at onset and termination. The glide for preferred /ɔ^{i}/ begins at [ɔ] or [ʊ] and courses to [i,y,ɪ]. The course for /a^{i}/ is somewhat more specific, beginning at

[a] and terminating near [ɪ]. The formant course for /a^u/ is limited to a glide from [a] to [o]. The results of this experiment support a treatment of /$ɔ^i$,a^i,a^u/ as unit phonemes in that the gliding movements of each are not compatible with those of a vowel plus vowel or vowel plus semivowel sequence in terms of either frequency course or duration. In addition, a gliding movement alone, exclusive of steady state targets, is sufficient for providing diphthongal quality.

The purpose of the second experiment was to determine whether the phonemic identity of the initial and terminal targets or the rate of frequency change of the second formant transition cues the perception of /$ɔ^i$, a^i,a^u/. These features were separated along the time dimension by synthesizing diphthongs whose absolute second formant frequency course remained fixed but whose durations were extended from 100 msec to 250 msec in steps of 10 msec, in one case beginning at the terminal target and in the other, at the initial target. Each series was presented to ten subjects who provided labels of either vowel or diphthong. For each series of /$ɔ^i$, a^i,a^u/, the shift in perception from simple vowel to diphthong occurred as a function of duration rather than second formant terminal frequency position, indicating that transition duration rather than target position is the primary perceptual feature. Further, in the case of /$ɔ^i$-a^i/, the rate of frequency change of the second

formant transition and not the onset frequency position serves as the primary distinguishing cue. This invariability of second formant rate of frequency change (or in articulatory terms, speed of movement) is also evident in real speech spectrographic measurements.

The results of this study suggest a treatment of /ɔ^{i},a^{i},a^{u}/ as unit phonemes each characterized primarily by an invariant speed of articulatory movement. Although each begins and terminates at phonetically identifiable zones in the vowel space, the actual positions of these targets are neither perceptually relevant nor fixed across changes in duration. Thus, since target positions are only implicit, an overall phonetic transcription might best be made by using a superscript form, as in [ɔ^{i},a^{ɪ},a^{ɔ}].

In suggesting that the formant rate of change of the gliding movements of /ɔ^{i},a^{i},a^{u}/ is a fixed feature, the results of this study provide a framework for examining the effects of such restraints as stress and phonetic environment on the diphthong movement. Perhaps the most significant questions might be whether a change in intensity or fundamental frequency or the presence of a steady state target (especially initially) serves suprasegmentally as the stress bearing element of the syllable, whether different consonant environments affect formant frequency extension and if these effects are evident in the same way for both /ɔ^{i},a^{i},a^{u}/

and the group of diphthongal variants, [e^i,o^u,$ɪ^i$,$ʊ^u$]. Finally, it might be speculated that the parameter of articulatory speed need not be unique to diphthong perception but in addition, might be relevant in describing other speech events characterized by articulatory movement, as in the case of semivowels and certain consonant-vowel sequences.[1]

As was stated at the outset of this study, an adequate description of speech sounds is dependent on physiological, acoustical and perceptual analyses. All of these features however, may not be viewed by the linguist as necessary for the phonological description of a language. He might, for example, prefer a strategy that makes considerations of economy and symmetry paramount. The phonemic role of the diphthongs that distinguish words such as boy, buy and bough is not in question. What is in question is the use of an experimental (or some other) basis for fitting these diphthongs into the English vowel system. The view implicit in this study is that if a phonological model is to have any explanatory power in the production and perception of speech, the parameters of the model should be testable.[2] It is hoped that the experimental data presented here contribute to such an undertaking.

---

[1]For an example of recent work on the dynamics of consonant-vowel coarticulation, see Öhman (1966).

[2]For a fuller discussion of the relationship between experimentation and language description, see Lisker, Cooper and Liberman (1962).

REFERENCES

Cooper, F. S. 1952. Spectrum analysis. J. Acous. Soc. Amer., 22:761-762.

Delattre, P., Liberman, A. M., Cooper, F. S. and Gerstman, L. J. 1952. An experimental study of the acoustic determinants of vowel color: observations on one- and two-formant vowels synthesized from spectrographic patterns. Word, 8:195-210.

Francis, W. N. 1958. The structure of American English. New York: Ronald.

Gleason, H. A. 1961. An introduction to descriptive linguistics. New York: Holt, Rhinehart and Winston.

Holbrook, A. and Fairbanks, G. 1962. Diphthong formants and their movements. J. Speech Hear. Res., 5:33-58.

Joos, M. 1948. Acoustic phonetics. Lang., Monograph 23.

Kurath, H. 1964. A phonology and prosody of Modern English. Ann Arbor: Univ. Mich. Press.

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

_____ and Peterson, G. E. 1961. Transitions, glides and diphthongs. J. Acous. Soc. Amer., 38:268-277.

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., 65:497-516.

_____, 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., 52:127-137.

Lisker, L. 1957. Minimal cues for separating /w,r, l,y/ in intervocalic position. Word, 13:256-257.

_____, Cooper, F. S. and Liberman, A. 1962. The uses of experiment in language description. Word, 18:82-106.

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, 13:24-43.

Öhman, S. 1966. Coarticulation in VCV utterances: spectrographic measurements. J. Acous. Soc. Amer., 39:151-168.

Peterson, G. E. and Barney, H. 1952. Control methods used in a study of the vowels. J. Acous. Soc. Amer., 39:175-184.

_____ and Coxe, M. 1953. The vowels /e/ and /o/ in American speech. Quart. J. Speech, 39:33-41.

Pike, K. L. 1947. On the phonemic status of English diphthongs. Lang., 23:151-159.

Potter, R. K., Kopp, G. A. and Green, H. C. 1947. Visible speech. New York: Van Nostrand.

_____ and Peterson, G. E. 1948. The representation of vowels and their movements. J. Acous. Soc. Amer., 20:528-535.

Sledd, J. 1954. Review of Trager, G. L. and Smith, H. L. An outline of English structure. Lang., 31: 312-335.

Trager, G. L. and Smith, H. L. 1951. An outline of English structure. Norman: Battenburg Press.

Wise, C. M. 1965. Acoustic structure of English diphthongs and semi-vowels vis-a-vis their phonemic symbolization, in Zwirner, E. and Bethge, W. (eds.). Proceedings of the fifth international congress of phonetic sciences, 5th, Munster. Basel: Karger.

AUTOBIOGRAPHICAL STATEMENT

The writer received his bachelor's degree from The City College, The City University of New York in 1962 and his master's from Adelphi University in 1964. He spent one year at the University of Illinois from 1963 to 1964. From 1964 to 1965 he was Assistant Director, Speech Pathology-Audiology Clinic, City Hospital at Elmhurst, New York and from 1964 to 1967 he was Instructor of Speech Pathology and Audiology at Teachers College, Columbia University.

He is presently Assistant Professor of Speech, Hunter College, The City University of New York and a member of the staff at Haskins Laboratories, New York.