

TITLE: Complex relation between F1 and F0 in determining vowel height: Acoustic and perceptual evidence.

AUTHOR: Maria-Gabriella Di Benedetto

AFFILIATION:

Department of Information and Communication (INFOCOM)

Faculty of Engineering- University of Rome 'La Sapienza'

Via Eudossiana, 18

00184 Rome- ITALY. a)

a) this work was carried out while the author was with the Speech Communication Group of the Massachusetts Institute of Technology, Cambridge, MA, USA.

Abstract

Results of the acoustic analysis of five vowels of American English [I,ɛ,æ,a,ʌ], spoken by three speakers, in the (F1-F0) vs F2 space (F1 and F0 are expressed in Bark) show that, in the dimension representing vowel height, individual differences for low vowels are reduced when the vowels are represented by the (F1-F0) difference rather than by F1. On the contrary, individual differences for high vowels are increased by the use of the same distance. Results of perceptual experiments which have been carried out using synthetic CVC and one-formant stimuli are in agreement with the observations based on the acoustic analysis. They suggest that the relation between F1 and F0 depends on both the range of F0 and F1 values. A possible interpretation consists in giving to F0 the role of an anchor point when both F0 and F1 are sufficiently high, while the extreme end of the scale would serve as the reference point when F0 and/or F1 assume values in the low frequencies range.

1. Introduction

Characterizing vowel segments in terms of acoustic parameters is a long-standing problem. Formant frequencies have been widely used as acoustic correlates of distinctive vowel features. In particular, the first formant frequency (F1) has been related to vowel height and the second formant frequency (F2) to vowel backness; Vowels which are coded as being [+high] are characterized by lower F1 values than vowels which are coded as [-high], and vowels [+back] are characterized by lower F2 values than vowels [-back].

However, it is well-known that different F1 and F2 values can correspond to the same vowel, when this vowel is pronounced by different speakers. Results of several analyses on vowels (Stevens and House, 1963; Lisker, 1985; Di Benedetto 1989a) have shown that, comparing the vowel areas in the F1 vs F2 space of two different speakers (for example one male and one female), the vowel α -dispersion area for one of the two speakers may be characterized by F1 values similar to those of the vowel β -dispersion area for the other speaker, $[\alpha]$ and $[\beta]$ being both [-back] or both [+back]; A normalization mechanism must take place somewhere in the vowel process, either at the production or at the perception level. For example, considering vowel produced by different speakers, Wakita (1977) proposed to normalize the formant values according to a measure of the vocal tract length. In this way, similar normalized formant values would correspond to the same vowel for all speakers.

The interest of the present study was, according to different experimental approaches, to analyze the hypothesis that vowels are normalized according to F0. The first approach was to carry out acoustical analysis of vowels. The second approach, which was motivated by the first, was to carry out perceptual experiments in which synthetic stimuli, characterized by different acoustic properties, were presented to listeners for identification or for comparison of vowel pairs. The possible influence of F0 time variations on the perception of vowel height was not investigated. Although F0 and the formant frequencies are associated with different elements of the speech production

mechanism, temporal properties of F0 might be important in determining vowel height, as already shown for F1 (Di Benedetto, 1989a).

Several other studies showed that F0 can act as a normalizing factor of formant displacements by influencing the perception of vowel quality (Potter and Steinberg, 1950; Miller, 1953; Fant, Carlson and Granström, 1974; Traünmüller, 1981). Perceptual experiments carried out by Potter and Steinberg (1950) showed that a synthesized [æ] characterized by F1 and F2 values corresponding to a male voice but with a child's fundamental frequency is perceived to be somewhere in between a synthesized child's [æ] and [ɛ]. This result supported the hypothesis of an association of fundamental frequency and formant position, although the association of adult formants and child's F0 gave rise to unnatural sounds.

The influence of F0 in the perception of vowel quality was also systematically analyzed by Miller (1953). In his experiments, two sets of synthetic two-formant vowels stimuli differing only in the value of F0 (144 Hz in the first set and 288 Hz in the second set) were considered. Miller observed that to a considerable extent the vowel areas remained fixed for stimuli characterized by different F0 values, but that there was a shift toward higher frequencies of the boundaries in F1 between different vowel areas when F0 was higher.

Fant, Carlson and Granström (1974) analyzed the relation between parameters that might affect a shift in the phonetic boundary between [e] and [ø] in Swedish. In Swedish, the male [e] and the female [ø] have approximately the same F1 and F2 values and similar F3 values. In their experiment, F0 was switched from the "male" F0=110 Hz to the "female" F0=220 Hz, and the extent to which one parameter or a group of parameters must vary in order to keep the same boundary between [e] and [ø] was examined. It was found that the increase of the values of F3 and higher formants shifted the perceptual quality of the stimuli in the [ø] direction while the opposite effect was expected as this change would increase F2' (this parameter is a combination of F2, F3 and F4, which was shown to be, together with F1, sufficient to synthesize Swedish vowels (Carlson, Granström and Fant, 1970)) and consequently favor [e] responses.

Fant et al. showed how this shift provokes a loss of spectral energy above F2 and suggested that F2' for [ø] might be very close to F2 for female speakers. Other experiments (Carlson, Granström and Fant, 1970), had shown that F2' for the male [e] is located halfway between F2 and F3. The conclusions were that the ambiguity between a female [ø] and a male [e] can be solved by considering the perceived timbre "flattening" effect due to the higher F0 for the female speaker, and, to a smaller extent, to F3 and/or F4 and higher formants.

More recently, Traunmüller (1981) examined the role of intrinsic factors, such as F1 and F0, in the determination of perceptual degree of openness. In one of his experiments, one-formant stimuli, in which F1 and F0 covered the ranges of variation observed in natural speech, were given phonetic judgements on their openness by listeners having a Bavarian dialect in which there appear to be five degrees of openness. In some other experiments, Traunmüller also investigated the perceptual importance of (F1-F0) using synthetic versions of natural vowels in order to analyze the influence of the higher formants on the perception of the (F1-F0) cue. The general conclusion was that the distance between F1 and F0, expressed in Bark, is the prevailing criterion for the perception of height. The higher formants play a marginal role in this regard. In a general model of the auditory representation of American English vowels in hVd and hVC syllables (Syrdal, 1985, Syrdal and Gopal, 1986), the distance between F1 and F0 was incorporated together with the distance between F2 and F1, F3 and F2, F4 and F3, and F4 and F2, all expressed in Bark. These distances are an application of the categorical perceptual effect (Spectral Center of Gravity (SCG)) found by Chistovich and her colleagues (Chistovich, Sheikin and Lublinskaya, 1979). Syrdal's findings were that high vowels are separated from mid and low vowels by the critical distance of the Bark-transformed (F1-F0) of about 3 Bark and that the (F1-F0) distance increased with increasing vowel openness.

The aim of the present study was first to verify whether the use of a parameter such as the difference between F1 and F0, expressed in Bark, is more appropriate than the traditional measurement F1 to represent vowels, pronounced by different speakers

and in several consonantal contexts, according to their height. To this aim, the effectiveness of the use of the (F1-F0) distance was examined on the basis of an acoustic analysis of five vowels of American-English [I, E, æ, a, ʌ] (unrounded and non-diphthongized vowels of American-English), uttered by three speakers. In the results obtained, presented in section 2, the vowels were represented in the (F1-F0) vs F2 space (F1 and F0 are expressed in Bark, F2 is expressed in Hertz). The values of F0 and F1, reported in the present paper, were converted into a critical band tonality scale, according to Zwicker and Terhardt's (1980) mathematical approximation as adopted by Syrdal (1985), and Syrdal and Gopal (1986)¹.

Secondly, the way F0 influences the perception of vowel height was investigated. For this purpose, perceptual experiments were carried out, using synthetic CVC and one-formant stimuli. These experiments will be described in section 3. The agreement of the results obtained with the findings of the acoustic analysis and interpretation of the results will be discussed in the last section.

¹According to Zwicker and Terhardt (1980), the critical band rate B (in Bark) is expressed by:

$$B = 13 \arctan(0.76F) + 3.5 \arctan(F/7.5)^2$$

where F is the frequency in kHz.

2. Acoustic analysis

2.1. Experimental conditions and procedures

The interest of the present study was to analyze vowels which could be distinguished only on the basis of their height and not of any other property. Rounded or diphthongized American English vowels were excluded from the present analysis since, as well known, roundness and diphthongization affect the F1 values, conflicting with F1 effects due to height variations. In the American English vowel system, the vowels left are [I,ɛ,æ,a,ʌ] which constitute the set of the unrounded and non diphthongized vowels of American English. These vowels can be distinguished on the basis of their height and of their backness. If one considers two different sets, back vowels [a,ʌ] and front vowels [I,ɛ,æ], it can be observed that, within each set, vowels can be distinguished on the sole basis of height. Results of a previous analysis (Di Benedetto, 1989a) showed that front vowels could be separated from back vowels in the F2 dimension, with a sufficient degree of accuracy. Therefore, the five vowels [I,ɛ,æ,a,ʌ] were taken into consideration in the present study.

An extended description of the experimental conditions and procedures can be found in Di Benedetto (1989a) and will therefore be only briefly reported here.

The vowels under study were considered in the context of the voiced and voiceless stops [b,d,g,p,t,k] forming CVC syllables, symmetric with respect to voicing (syllables such as bId were included but such as bIt were excluded) leading to 18 different syllables, and pronounced in the sentence frame "The___ again". The hVd and #Vd syllables were also included in the analysis and could serve as "reference points" for comparison with previous studies on vowel reduction (Stevens and House, 1963; Lindblom, 1963). All the syllables were uttered by three speakers (two males and one female), native speakers of American English. Each sentence was repeated three times. The corpus obtained included then 5 vowels uttered by three speakers in 20 different consonantal contexts each one in three repetitions, leading to 900 vowels. The

considerable size of this corpus limited the extension to other speakers or other repetitions. Due to the fair amount of consonantal contexts and repetitions available for each vowel, within-speaker formant variability could be acceptably determined on the basis of these data (Di Benedetto 1989a). Between-speaker variability analysis would benefit from an extension to more speakers, although the ultimate goal of this investigation is finding some evidence for a universal effect, and not specifying a statistically based satisfactory measurement of some physical phenomenon. Although limited to three speakers, the results of the present study may constitute a first basis for further investigation.

The recorded materials were evaluated by a phonetically sophisticated listener. All the syllables were judged to be good samples of the phonemes considered.

The attention was focused on F1 and F2 estimation, which were manually extracted for each vowel by plotting the vowel spectrum (256-point DFT), every 5 msec. These values were systematically compared to those obtained automatically by means of the program KLSPEC developed by Dennis Klatt (1984) on the pseudospectrum. The pseudospectrum was obtained by windowing a slice of signal (for example 256 samples or 25.6 msec at 10 kHz) and computing a 256-point DFT. An approximation to the filter set used in a broadband spectrogram display was then obtained by forming a weighted sum of adjacent DFT sample energies for each of the 129 spectrogramlike filters.

Estimation of the fundamental frequency was obtained by measuring the harmonics in a narrow-band spectrum.

The temporal sampling point of F1, F2 and F0 was the time at which F1 reached its maximum, as discussed in Di Benedetto (1989a).

2.2 Results of acoustic measurements

Table I shows the results of measurements of F0 for each speaker and each vowel. These data were obtained by averaging the F0 values of each vowel in the 20 consonantal contexts and in the 3 repetitions. As expected, the highest F0 was found for

the female speaker (CR), while F0 for the two male speakers (JP) and (KS) was comparable. Table I also shows that F0 is related to vowel height. This same effect was found in the past by other investigators (Peterson and Barney, 1952; House and Fairbanks, 1953) and may be explained in terms of a mechanical connection between forward movement of the tongue root (giving rise to changes in tongue height) and movements of the hyoid bone and thyroid cartilage leading to variation in vocal-fold tension and in fundamental frequency of the vowel (Honda, 1983).

In addition, in a previous study on the same speech material (Di Benedetto, 1987), it was observed that vowels in voiced consonantal contexts had lower F1 values than in voiceless consonantal contexts. Moreover, vowels in voiced consonantal contexts also had lower F0 values than in voiceless consonantal contexts, in agreement with House and Fairbanks who suggested that a possible explanation to this effect was that the F0 of the surrounding voiced consonants was lower than that of the vowels and may have a lowering effect on the F0 values of vowels.

Consequently to these two effects, the difference in F1 values between vowels in voiced and voiceless consonantal contexts was higher than in (F1-F0) values (F1 and F0 expressed in Hertz), for all vowels, as shown in Figure 1.

The results of the analysis in the (F1-F0) vs F2 space for speakers (KS), (JP) and (CR) are presented in Fig. 2b, 3b, and 4b, respectively. Figures 2a, 3a, and 4a show, for comparison, the representations in the F1 vs F2 space for each speaker. The areas shown represent the regular polyhedra which included all the F patterns of each vowel. This type of representation, which makes use of the polyhedra rather than of the ellipses of equiprobability, was proposed in Di Benedetto (1989a). It provides a schematic representation without obscuring critical details which might have no statistical significance but be representative of an important property. The F patterns of each vowel in each consonantal context and in the three repetitions, leading to 60 points for each vowel (18 possible stop contexts plus the hVd and #Vd contexts, three repetitions), were included in each area. Note that the (F1-F0) values were expressed in Bark and the F2 values were in Hertz since results of a previous analysis (Di Benedetto, 1989a) indicated

that no overlapping occurred between vowel areas in the F2 dimension, for any single speaker, suggesting that the representation of the dimension of vowel backness by F2 is sufficiently accurate, at least for the purpose of this study. Details on the location of the vowel areas for the vowels [I], [ε], [æ], [a], [ʌ] can be found in Di Benedetto (1987).

Figures 2b, 3b and 4b show that overlapping occurred in the (F1-F0) dimension between contiguous vowel areas. This overlap was mainly between [a] and [ʌ], and for speakers (JP) and (KS) also between [ε] and [æ]. An analysis was carried out to compare the results of the acoustic analysis in the (F1-F0) vs F2 space to the results of the analysis of the same speech materials in the F1 vs F2 space. To this aim, the amount of overlapping between contiguous vowel areas in the (F1-F0) dimension, for each speaker, was quantified, by determining the straight lines which better separate the sets represented by the (F1-F0)-F2 values of [I]-[ε], [ε]-[æ], and [a]-[ʌ] (linear discriminant analysis). The statistical distribution of the measurements was supposed to be Gaussian and the covariance matrix was hypothesized to be similar for the measurements of the vowels in each pair. In addition, as proposed in Di Benedetto (1989a), a generalized Euclidean distance (Mahalanobis distance) was computed in order to quantify the distance between two sets. The Mahalanobis distance is equal to the distance between the means of two sets, divided by the amount of spreading in each set, and is a dimensionless parameter. For equal amount of spreading or equal euclidean distance between the means of two sets, a higher value of the Mahalanobis distance corresponds to a better separation of the two sets. The classification rates and Mahalanobis distances obtained with the two representations under comparison are shown in Table II. The results reported in Table II show that an improvement was obtained in the (F1-F0) vs F2 space, in terms of better grouping and better separation of the vowel areas, compared to what was obtained in the F1 vs F2 space, although problems of overlapping still occurred between vowel areas of a single speaker in the (F1-F0) dimension. One should note that, as observed earlier, the differences in (F1-F0) values between vowels in voiced and voiceless consonantal contexts were smaller than in F1 values. Consequently, one of the factors which

contributed to a better separation of the vowel areas was that in the (F1-F0) dimension the vowel areas for voiced and voiceless consonantal contexts were better grouped.

The F1, (F1-F0), and F2 values were averaged over all the consonantal contexts and repetitions, for each speaker and each vowel. In addition, a different set of (F1-F0) values was computed by applying an end-correction of the Bark scale, as proposed by Traunmüller (1981) and applied by Syrdal and Gopal (1986). This low-frequency end correction implies that all frequencies below 150 Hz be raised to 150 Hz. For frequencies between 150 and 200 Hz, the modified frequency is obtained by subtracting a percentage of the original value normalized to 150 Hz from the original value. An identical procedure is used for frequencies between 200 and 250 Hz, except that the original frequency is normalized to 250 Hz. The F1, (F1-F0), (F1-F0) end-corrected, and F2 values obtained for each speaker and each vowel are presented in Table III. These values show that the difference in the representation of vowels for different speakers was reduced using the (F1-F0) parameter for the low and front vowel [æ] and the two back vowels [a, ʌ]. For the mid vowel [ɛ], in the (F1-F0) dimension the [ɛ]-area of the female speaker (CR) was shifted to lower values than those characterizing the [ɛ]-area of (KS) and (JP) and this effect is even more evident for [ɪ]. The results presented in Table III were compared by considering the maximum variations in F1, (F1-F0), and (F1-F0) end-corrected values between the speakers, for each vowel. The maximum variation values quantify the differences between speakers in the representations of each vowel. The values obtained are presented in Table IV. Note that better grouping was obtained, using the end correction for [ɪ] and [ɛ] and also slightly for [a], but not for [æ] and [ʌ]. The values of Table IV also confirm the observations made above on the comparison between F1 vs F2 and (F1-F0) vs F2 representations.

3. Perceptual experiments

In this part of the study, an analysis was carried out to observe whether perceptually F0 and F1 could be related in characterizing vowel height. In all experiments, the stimuli considered were synthesized with the Klatt synthesizer. This cascade/parallel formant synthesizer has been extensively described by Klatt (1980, 1984).

3.1 Experiment 1

The aim of this experiment was to investigate the influence of F0 on the perception of vowel height, using dVd synthetic syllables.

One set of stimuli consisted of stimuli already used in a previous experiment (Di Benedetto, 1989b). The results of this previous experiment made it possible to find, for different listeners, the perceptual boundary between the high vowels [i, I] and the non high vowels [e, ε] on the sole basis of differences in F1 onset values and F1 maximum location in the synthetic vowel. In the present experiment, it was possible then, considering the same listeners, to examine whether a change in F0 values would affect the identifications of the vowels in the stimuli and shift the perceptual boundaries previously found.

The description of the first set of stimuli, previously given in Di Benedetto (1989b), will only briefly be reported here. These stimuli were obtained on the basis of an analysis of real speech. The F1 maximum location and F1 onset were the only parameters by which two stimuli having the same F1 maximum differed. Figure 5 shows the F1 trajectories of the synthetic stimuli, which can have two shapes; depending on the shape, the stimuli were identified as type I or type II. Ten stimuli of type I and ten stimuli of type II were synthesized, each stimulus being characterized by a different F1 maximum value (330, 350, 370, 390, 410, 430, 450, 470, 490, 500 Hz). The trajectories of the higher formants and of fundamental frequency (F0 maximum was 125

Hz) were identical for both stimuli types and were symmetrical around the center of the vowels.

In the second and third sets, the stimuli were identical to the previous ones as regards the F1 trajectory shape and higher formant trajectories, while the fundamental frequency was increased in two steps: 60 Hz and 120 Hz.

The following stimuli were then used: ten stimuli of type I and ten stimuli of type II with F0 maximum at 125 Hz (125-I and 125-II stimuli); ten stimuli of type I and ten stimuli of type II with F0 maximum at 185 Hz (185-I and 185-II stimuli); and ten stimuli of type I and ten stimuli of type II with F0 maximum at 245 Hz (245-I and 245-II stimuli). Table V shows an overview of the stimuli considered in the present experiment.

Five subjects participated in Experiment 1. They were all phonetically trained listeners, native speakers of American English and members of the Speech Communication Group at the Massachusetts Institute of Technology.

Experiment 1 consisted of two tests: a vowel identification test and a "boundary" identification test. Four subjects participated in the vowel identification test. The stimuli used in this test were the 125-I and -II and the 185-I and -II stimuli. The test was first carried out using 125-I and II stimuli. It consisted of three parts. In the first part, only 125-I stimuli, while in the second part only 125-II stimuli, were presented to the listeners. In the third part, 125-I and 125-II stimuli were combined. In the first part of the test, each 125-I stimulus was presented ten times. The 125-I stimuli were ordered in such a way that each stimulus followed another only once. In this way, the average number of responses given by the listeners for each stimulus could be supposed to be independent of the stimulus preceding it. The second part of the experiment was organized as the first part with 125-II stimuli. In the third part of the test, stimuli of both types were presented. The ten 125-I stimuli and the ten 125-II stimuli were divided in two sets of ten stimuli each (stimuli having an even number in one set and stimuli having an odd number in the other). The stimuli of each set were ordered to make each stimulus follow another only once. In all, each stimulus was heard 20 times. In each part of the experiment, the stimuli were spaced by a pause of three seconds. At the end of each part, the test was interrupted

and the subject could rest for a few minutes. The pauses did not last more than 2-3 minutes. The test was approximately 45 minutes long.

The same procedure was then repeated using 185-I and -II stimuli.

A previous experiment, based on an open-response set (Di Benedetto 1989b) showed that American-English listeners identified the vowels in the stimuli as [i, I, e, ε]. Therefore, in the present experiment, the subjects were asked to identify the vowel of the synthetic utterances as [i, I, e, ε]. None of the subjects declared to perceive a different vowel from these four. There were 20 responses per data point.

A "boundary" identification test was then carried out. 125-I stimuli, 185-I stimuli and 245-I stimuli were used. Results of previous experiments presented in Di Benedetto (1989b) showed that type I stimuli were mainly perceived by the American subjects as the tense vowels [i, e]. Sequences of stimuli (and the same sequences in reverse order) characterized by the same F_0 were played to the subjects who were asked to declare when their perception of the synthetic vowels changed from [i] to [e] or viceversa. Each sequence, in each order, was presented three times. The perceptual boundaries between stimuli identified as [e] or [i], obtained for the 125-I stimuli, 185-I stimuli and the 245-I stimuli, could then be compared, in order to observe the effect of different values of F_0 on the perception of vowel height. Three subjects participated in this test. Two of these subjects (KS) and (CB) also participated in the vowel identification test.

3.1.1 Results of the identification test

Results of the identification test are presented in Fig.6, for each subject separately, showing the identification curves in respect to an abscissa of the stimulus number (#1 corresponds to F_1 maximum=330 Hz and #10 to F_1 maximum=500 Hz) and an ordinate of the percent of identification of the vowel specified. Note that subjects (JP) and (SSH) identified type II stimuli as [i] or [I] and never as [e] or [ε]. The behaviour of subjects SSH and JP was explained in Di Benedetto (1989b) as due to the shape of the F_1 trajectory being more relevant than the F_1 maximum value; Even if F_1

was high, the subjects based their judgement on the fact that the F1 trajectory had that particular shape. Figure 6 shows that a change of 60 Hz in F0 did not result in a clear effect on the identification functions. In order to quantify this observation, a logistic curve fitting the data and the 50% crossover point were computed. The logistic curve was found according to the procedure proposed by Neter and Wassermann (1974) and it represents the probability of correct response (psychometric function). The logistic curves are not shown on the figure to keep the representation as clear as possible. The crossover points, indicated by their value in "stimulus numbers", for each curve, are presented in Table VI. The difference in crossover values obtained with 125- and 185-stimuli, in the type I and in the type II set, was small for all subjects, and in all cases less than one stimulus number. The highest difference (0.8) was obtained for subject (SSH) with type I stimuli. Note that, in this case, the lower crossover value was obtained with 185-stimuli while if the (F1-F0) effect was occurring, the crossover should have shifted to higher values, favouring [i] and [I] responses. /

3.1.2. Results of the "boundary" identification test

During a pre-informal test, the three subjects who participated in this test reported that they perceived the vowels of the synthetic utterances as either [i] or [e]. In the "boundary" identification test they were asked then to identify the vowels in the stimuli as [i] or [e]. Figure 7 shows the results for each of the three subjects. This figure indicates the stimulus at which the identification changes from [i] to [e], and specifically the first stimulus which was perceived as [e], when the sequences presented were ordered with ascending stimuli number, or the last stimulus which was perceived as [e], in the case of sequences ordered according to a descending stimulus number progression. Figure 7 shows that, in the case of the three subjects who participated in this test, an increase in F0 from 125 to 185 Hz did not result in a change of the perceptual boundary between [i] and [e], while a variation in F0 from 125 to 245 Hz did result in a consistent shift in this boundary. The shift in the perceptual boundary between [i] and [e] was one stimulus

number (corresponding to a change of 20 Hz in the F1 maximum) for (KS), between one and two stimulus numbers for (CB), and two stimulus numbers (corresponding to a change of 40 Hz in the F1 maximum value) for the other subject (RS). No difference was observed in the results obtained with sequences of stimuli with F1 increasing or in reverse order. The results of the boundary identification test were in agreement with the results of the identification test for F0 changes from 125 to 185 Hz and showed, in addition, that when F0 was increased from 125 to 245 Hz (change of 120 Hz) the perceptual boundary between [i] and [e] shifted towards higher stimulus numbers by about 20-40 Hz (considerably less than F1-F0 shifts). One should note that the boundary for (KS) and (CB) in the "boundary" identification test was lower than in the identification test. No explanation was found for this effect.

3.2 Experiment 2

In the second experiment, the acoustic information contained in the stimuli was reduced to F1 and F0, by removing the stabilizing factors F2, F3 and F4 of experiment 1; One-formant stimuli were used.

Various one-formant vowel stimuli with F0=125 Hz, 185 Hz or 245 Hz, stationary in F0 and F1 were generated. The one-formant stimuli with F0=125 Hz were generated with five different F1 (300, 350, 400, 500, 600 Hz). Each of these stimuli was paired with one-formant stimuli with F0=185 Hz and values of F1 ranging from the F1 value of the standard stimulus to the F1 value that would give the same F1-F0 (in Hertz) for the comparison and the standard stimulus. Each stimulus pair was played three times. The same procedure was repeated with the same standard stimuli (F0=125 Hz) but the stimuli against which they were paired were characterized by F0=245 Hz.

Seven subjects participated in this experiment. The seven subjects were all phonetically trained listeners, native speakers of American English, and all members of the Speech Communication Group at the Massachusetts Institute of Technology. They were asked to indicate which pair of stimuli was most similar in terms of vowel height.

Note that, as pointed out by Stevens (1986), the stimuli paired by the listeners did not have to be necessarily identical in vowel quality. In fact, the judgement on the stimuli could be based upon different features perceptible to the listener. For example, different matching results could be obtained if the subjects were judging the stimuli for their height or for their backness. No control experiment was run on this aspect.

Figures 8 and 9 show the results obtained in experiment 2. Figure 8 shows on the abscissa the standard stimuli (with $F_0=125$ Hz) identified by the F_1 value in Hertz (bottom axis) and Bark (top axis), and on the ordinate the comparison stimuli (with $F_0=185$ Hz) identified by the F_1 values in Hertz (left axis) and Bark (right axis). As shown on Fig.8, each standard stimulus could be paired with three comparison stimuli: one with the same F_1 , one with the same (F_1-F_0) (in Hertz) and one with a F_1 value intermediate between the same F_1 and the same (F_1-F_0) . For example, the standard stimulus with $F_1=300$ Hz could be paired with a comparison stimulus with $F_1=360$ Hz and consequently the same $(F_1-F_0)=155$ Hz, or a comparison stimulus with $F_1=330$ Hz. For each standard stimulus, Fig.8 shows the value of F_1 for best pairing for all subjects. The circled numbers indicate the percentage of times that each comparison stimulus was chosen for best pairing by all subjects.

Figure 9 is similar to Fig.8 but indicates the results of the test in the case of the comparison stimuli with $F_0=245$ Hz. In this case, each standard stimulus could be paired with five comparison stimuli: one with the same F_1 , one with the same (F_1-F_0) and three with intermediate values of F_1 , between the same F_1 and the same (F_1-F_0) . For example, the standard stimulus with $F_1=500$ Hz ($F_0=125$ Hz) could be paired with a comparison stimulus with $F_1=500$ Hz ($F_0=185$ Hz), a comparison stimulus with $F_1=620$ Hz (and consequently the same $(F_1-F_0)=375$ Hz), and three comparison stimuli with $F_1=530, 560, 590$ Hz.

Figure 8 shows that the F_1 value for best pairing, in the case of stimuli with $F_0=185$ Hz, corresponded to an exact formant match for low F_1 values (300 and 350 Hz). For other values of F_1 the match was in general between an exact formant match and values of F_1 leading to similar (F_1-F_0) values. Note that in the case of the highest F_1

value for the standard stimuli ($F1=600$ Hz) the match was similar to $(F1-F0)$ and close to this value for some of the subjects. One should note that when $F1$ is high enough (for values higher than 400 Hz, approximately) $F1$ is out of the linear Bark range. Consequently, the $(F1-F0)$ distance expressed in Bark is always lower for comparison stimuli than for standard stimuli when $F1$ is in this range although the difference is small.

Figure 9 shows that the value for best pairing, in the case of stimuli with $F0=245$ Hz was in general at intermediate values of $F1$, between an exact formant match and values leading to similar $(F1-F0)$ values for comparison and standard stimuli. In the case of the lowest values of $F1$ for standard stimuli ($F1=300$ Hz, $F0=125$ Hz), the match was made in almost all cases with stimuli characterized by $F1=330$ Hz ($F0=245$ Hz) corresponding to the first intermediate step. For values of $F1$ in the middle range ($F1=350$, 400 and 500 Hz) the match shifted to stimuli with intermediate $F1$ values higher than in the case of standard stimuli with $F1=300$ Hz. Standard stimuli with $F1=350$ Hz ($F0=125$ Hz) were generally paired with comparison stimuli with $F1=410$ Hz ($F0=245$ Hz) and for standard stimuli with $F1=400$ Hz or $F1=500$ Hz ($F0=125$ Hz), the pairing was generally with comparison stimuli with $F1=460\sim490$ Hz ($F0=245$ Hz) and $F1=560\sim590$ Hz ($F0=245$ Hz), respectively. The case of standard stimuli with $F1=600$ Hz ($F0=125$ Hz) is similar to the case of $F1=400$ Hz and $F1=500$ Hz, but note that few responses were paired with stimuli with $F1=720$ Hz ($F0=245$ Hz) leading to similar $(F1-F0)$ values for comparison and standard stimuli.

4. Discussion

Results of perceptual experiments showed that the perception of vowel height is related to F0 values and F1 values.

In particular, vowel identification experiments, using CVC synthetic stimuli, showed that an increase in F0 from 125 to 185 Hz did not result in a clear effect on the identification functions. In a boundary identification experiment, a variation from 125 to 245 Hz did consistently result in different judgements. In a second experiment, one-formant stimuli with F0=125 Hz and various values of F1 (300, 350, 400, 500, 600 Hz) were paired with one-formant stimuli in which F1 could assume 3 to 5 different values and F0 equal to 185 or 245 Hz. The results of this experiment showed that the value of F1 for best pairing was usually between an exact formant pairing and a pairing yielding similar values of (F1-F0) for comparison and standard stimuli. The pairing was close to F1 for low F1 values and tended to be closer to similar (F1-F0) values for higher F1. In some cases, in particular when comparison stimuli with F0=185 Hz were considered, the pairing reached the same (F1-F0) values (in Hertz) for comparison and standard stimuli. In these cases, the (F1-F0) values expressed in Bark were lower for comparison stimuli than for standard stimuli.

The results of the perceptual experiments were in agreement with the results of the acoustic analysis presented in section 2. In particular, in the cases of F0=125 Hz and F0=185 Hz, results of perceptual experiments showed that for low values of F1, F0 did not seem to influence the perception of vowel height. Correspondingly, in the acoustic analysis, it was observed that the high vowel [I]-area for the male and the female speakers was located at similar values of F1 (note that the average F0 value of the female speaker was ~ 190 Hz and of the male speakers ~120-130 Hz). The results of a second perceptual experiment showed that when F1 was high, a change of F0 from 125 to 185 Hz influenced the perception of vowel height and that stimuli with different values of F1 and F0 but similar (F1-F0) values were perceived as similar in terms of vowel height. Correspondingly, the acoustic analysis indicated that the location of the low vowel [æ]-

area corresponded to higher F1 values in the case of the female speaker, and to similar (F1-F0) values for the female and male speakers.

There seems to be evidence for a relation between F1 and F0 which is not as simple as the (F1-F0) distance proposed by Traunmüller (1981). Traunmüller (1983) observed that the distance between F1 and F0 is not strictly invariant in vowels with similar perceived height. Traunmüller proposed, on the basis of a hypothesis of spectral integration over a range of 3 Bark, that, in the representation of high vowels, the distance between the peak of the auditory spectral representation, which is shaped by F1 and the lower flank of the configuration would be independent of F0 for F0 around 150 Hz, because the lower flank of the configuration would be the origin. The results of the present study agree with Traunmüller's observation.

It is not possible to exclude that the effect observed on the basis of the perceptual experiments presented in this paper is not an auditory effect but a categorical effect dependent on higher level cognitive processes or that, on the contrary, lower level psychoacoustic processes took place in judging the similarity between harmonic simple resonance signals with different F0.

If the hypothesis of an auditory effect is made, the interpretation of the results of the present study can be given as follows. When F1 is sufficiently low (as in high vowels) and F0 also assumes low values (below ~200 Hz) F1 may be considered, by the perceptual mechanism which processes it, relative to the extreme end of the scale (the end of the scale is used as an anchor point) and is then the most relevant factor in vowel height perception. When F1 is high (as in low vowels) and F0 is sufficiently far from F1, F1 may be considered relative to F0 (not, as previously, to the end of the scale), F0 being used as an anchor point, and the distance between F1 and F0 (in Bark) determines the perception of vowel height. When F1 is at intermediate values, or the distance between F1 and F0 is not large enough, F1 and F0 would both intervene in the perceptual process determining vowel height in a relation which would not attribute the same weight to F1 and F0.

This interpretation would imply a non-uniform vowel normalization in agreement with Fant's study (1975).

This hypothesis finds support in results of physiological experiments carried out by Delgutte and Kiang (1984), as pointed out by Stevens (1985). These investigators observed the location of the largest components in the discrete Fourier transforms of period histograms obtained from auditory-nerve fibers with various values of the characteristic frequency (CF). The stimuli were steady-state two formant stimuli with $F_0=125$ Hz. Delgutte and Kiang noted that, for all vowels, there was a CF region which was located around F_1 (F_1 region) where the harmonics close to F_1 dominated the response spectra. In addition, they observed that this region was flanked on the low-CF by another region in which the harmonics close to CF were the largest components in the response spectra. These harmonics corresponded to the fundamental frequency or to intermediate values between F_1 and F_0 . For low vowels, this region extended up to about 400 Hz while on the contrary, for high vowels, this region was not distinct. Delgutte and Kiang observed that "...the open-close dimension of phonetics correlates with both the position of the F_1 region along the CF dimension and with the extent of the low-CF region". This observation could justify the results of the present study that F_1 alone determines the perception of vowel height when F_1 is low (high vowels), whereas if F_1 is high (low vowels) F_0 influences vowel height perception.

Unfortunately, Delgutte and Kiang did not present results in the case of higher values of F_0 . Consequently, the results of the present study in the case of higher values of F_0 cannot be interpreted on the same basis.

In the introduction, we have mentioned the categorical perceptual effect SCG (Spectral Center of Gravity) found by Chistovich et al. (1979). We want to point out that the perception of vowels with F_1 and F_2 closer than 3.5 Bark could be based on one equivalent formant located in an intermediate position between the two formants, according to the SCG theory. It could then be hypothesized that this one formant is relevant, in the cases of vowels with $F_2-F_1 < 3.5$ Bark, to vowel height perception. We want to suggest that our interpretation of the relation between F_1 and F_0 in the perception

of vowel height is appropriate in the case of front vowels, but that for back vowels additional factors could be relevant, such as, according to the SCG theory, the relative amplitudes of F1 and F2.

An analysis of temporal variations of F0 is in progress. The object of future studies will be to analyze the relative temporal variations of F1 and F0 and to attempt to relate these properties to vowel height.

Acknowledgements

The author wishes to express her indebtedness to Paolo Mandarini and Ken Stevens for their continued support and help. The helpful comments and suggestions made by Jean-Sylvain Liénard, Maxine Eskenazi and Christophe d'Alessandro are gratefully acknowledged.

References

- Carlson, R., Granström, B. & Fant, C.G.M. (1970) Some studies concerning perception of isolated vowels, *STL-QPSR* 2-3, 19-46.
- Chistovich, L.A., Sheikin, R.L. & Lublinskaya, V.V. (1979) Centres of gravity and spectral peaks as the determinants of vowel quality. In *Frontiers of Speech Communication Research* (B.Lindblom and S.Öhman, editors), pp.143-157. Academic Press.
- Chowmsky, N., and Halle, M. (1968) *The sound pattern of English* (Harper and Row, New York).
- Delgutte, B. & Kiang, N.Y.S. (1984) Speech coding in the auditory nerve: I. vowel-like sounds, *Journal of the Acoustical Society of America* 75(3), 866-878.
- Di Benedetto, M.G. (1987) An acoustical and perceptual study on vowel height, Ph.D. thesis, University of Rome 'La Sapienza', Italy.
- Di Benedetto, M.G. (1989a) Vowel representation: Some observations on temporal and spectral properties of the first formant frequency, *Journal of the Acoustical Society of America* 86(1), 55-66.
- Di Benedetto, M.G. (1989b) Frequency and time variations of the first formant: Properties relevant to the perception of vowel height, *Journal of the Acoustical Society of America* 86(1), 67-77.
- Fant, C.G.M. (1975) Non-uniform vowel normalization, *STL-QPSR* 2-3.

- Fant, C.G.M., Carlson, R. & Granström, B. (1974) "The [e] - [ø] ambiguity", Speech Communication Seminar, Stockholm, Aug. 1-3, 117-121.
- Honda, H. (1983) Relationship between pitch control and vowel articulation. In *Vocal Fold Physiology: contemporary research and clinical issues* (Diane M. Bless & James H. Abbs, editors), pp.286-297
- House, A.S. & Fairbanks, G. (1953) The influence of consonant environment upon the secondary acoustical characteristics of vowels, *Journal of the Acoustical Society of America* **25**(1), 105-113.
- Klatt, D.H. (1980) Software for cascade/parallel formant synthesizer, *Journal of the Acoustical Society of America* **67**(3), 971-995.
- Klatt, D.H. (1984) M.I.T. SpeechVAX user's guide, preliminary version.
- Lisker, L. (1984) On reconciling monophthongal vowel percepts and continuously varying F patterns, *Haskins Lab., stat. Rep. Speech Res.* SR 79/80.
- Miller, R.L. (1953) Auditory tests with synthetic vowels, *Journal of the Acoustical Society of America* **25**(1), 114-121.
- Neter, J. & Wassermann, W. (1974) *Applied linear statistical models* (Irwin, Momewood, IL), pp.329-338.
- Peterson, G.E. (1961) Parameters of vowel quality, *Journal of Speech and Hearing Research* **4**, 10-29.

- Peterson, G.E. & Barney, H.L. (1952) Control methods used in a study of the vowels, *Journal of the Acoustical Society of America* **24**(2), 175-184.
- Potter, R.K. & Steinberg, J.C. (1950) Toward the specification of speech, *Journal of the Acoustical Society of America* **22**(6), 807-820.
- Stevens, K.N. (1985) Personal communication.
- Stevens, K.N. (1986) Personal communication.
- Syrdal, A.K. (1985) Aspects of a model of the auditory representation of American English vowels, *Speech Communication* **4**, 121-135.
- Syrdal, A.K. & Gopal, H.S. (1986) A perceptual model of vowel recognition based on the auditory representation of American English vowels, *Journal of the Acoustical Society of America* **79**(4), 1086-1100.
- Traunmüller, H. (1981) Perceptual dimension of openness in vowels, *Journal of the Acoustical Society of America* **69**(5), 1465-1475.
- Traunmüller, H. (1983) On vowels: perception of spectral features, related aspects of production and sociophonetic dimensions, University of Stockholm, PhD Dissertation.
- Wakita, H. (1977) Normalization of vowels by vocal-tract length and its application to vowel identification, *IEEE Transactions on Acoustics Speech and Signal Processing* **25**(2), 183-192

Zwicker, E. & Terhardt, E. (1980) "Analytical expressions for critical band rate and critical bandwidth as a function of frequency", J. Acoust. Soc. Am. 68, 1523-1525.

Figure captions

Figure 1: Differences in F1 and (F1-F0) values for vowels in voiced and voiceless consonantal contexts, for each vowel, averaged over the results obtained for the three speakers, the three repetitions, and the 20 consonantal contexts.

Figure 2: Vowel areas in the a) F1 vs F2 space, and b) (F1-F0) vs F2 space, for speaker (KS), male speaker. Each vowel is represented by the regular and convex polyhedron which included all the F patterns of that particular vowel (60 patterns for each vowel).

Figure 3: Vowel areas in the a) F1 vs F2 space, and b) (F1-F0) vs F2 space, for speaker (JP), male speaker. Each vowel is represented by the regular and convex polyhedron which included all the F patterns of that particular vowel (60 patterns for each vowel).

Figure 4: Vowel areas in the a) F1 vs F2 space, and b) (F1-F0) vs F2 space, for speaker (CR), female speaker. Each vowel is represented by the regular and convex polyhedron which included all the F patterns of that particular vowel (60 patterns for each vowel).

Figure 5: Schematic F1 trajectories for the stimuli of type I and of type II, used in experiment 1.

Figure 6: Results of the identification test using 185-I and 185-II stimuli, compared with the results obtained with 125-I and 125-II stimuli, in the case of a) subject (KS), b) subject (CB), c) subject SSH, and d) subject (JP). For an overview of the stimuli see Table IV.

Figure 7: Results of the boundary identification test for subjects (KS), (CB), and (RS). Each dot on the figure (of a different shape for each subject) indicates the stimulus at which the identification changes from [i] to [e], in the case of the three stimuli F0 types: 125-I, 185-I, and 245-I stimuli. For an overview of the stimuli see Table V.

Figure 8: Results of experiment 2 in the case of comparison stimuli with F0=185 Hz.

Figure 9: Results of experiment 2 in the case of comparison stimuli with $F_0=245$ Hz

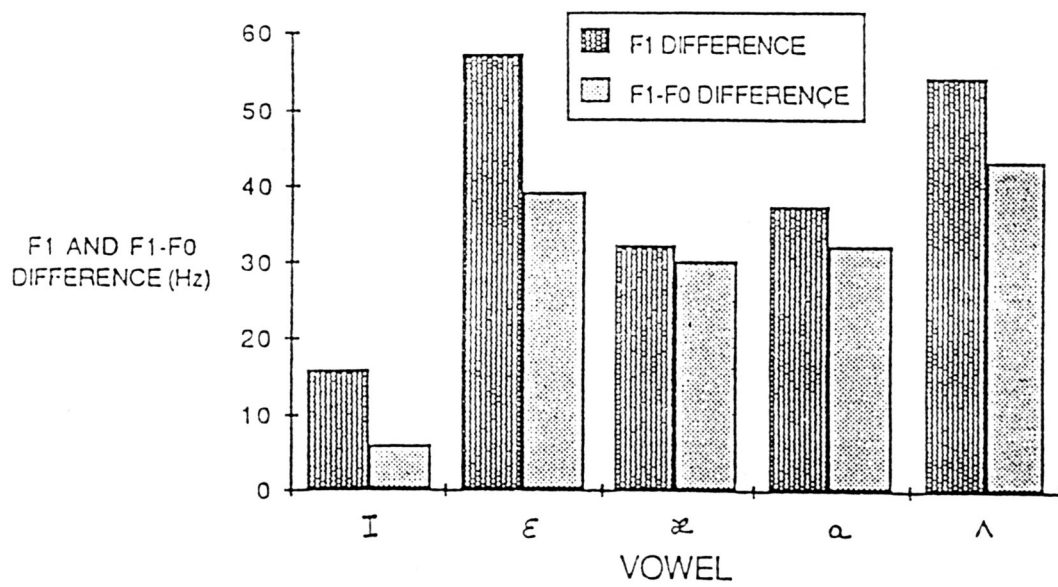
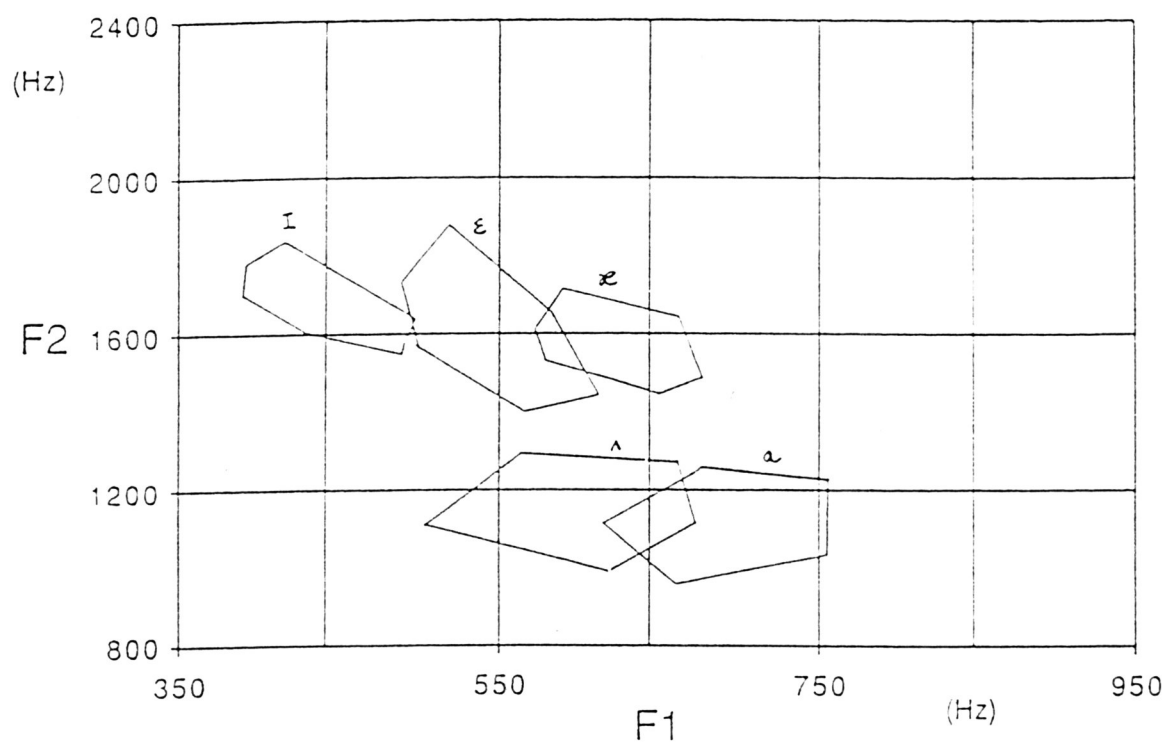
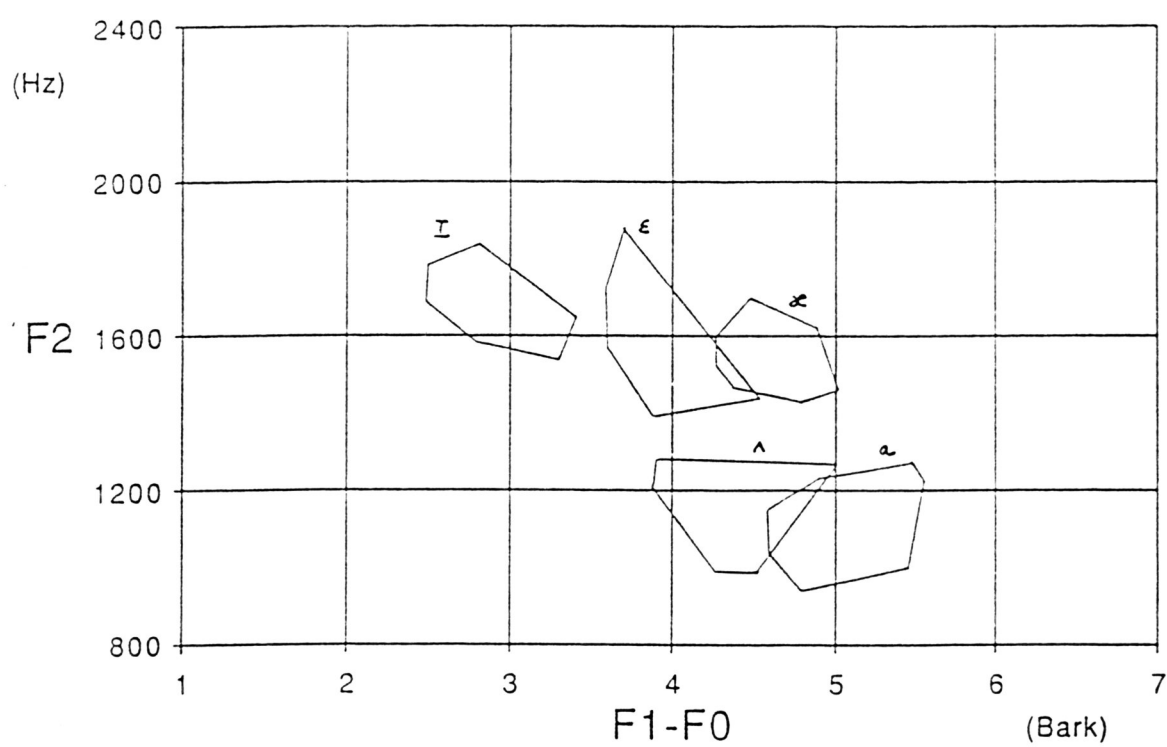


Figure 1

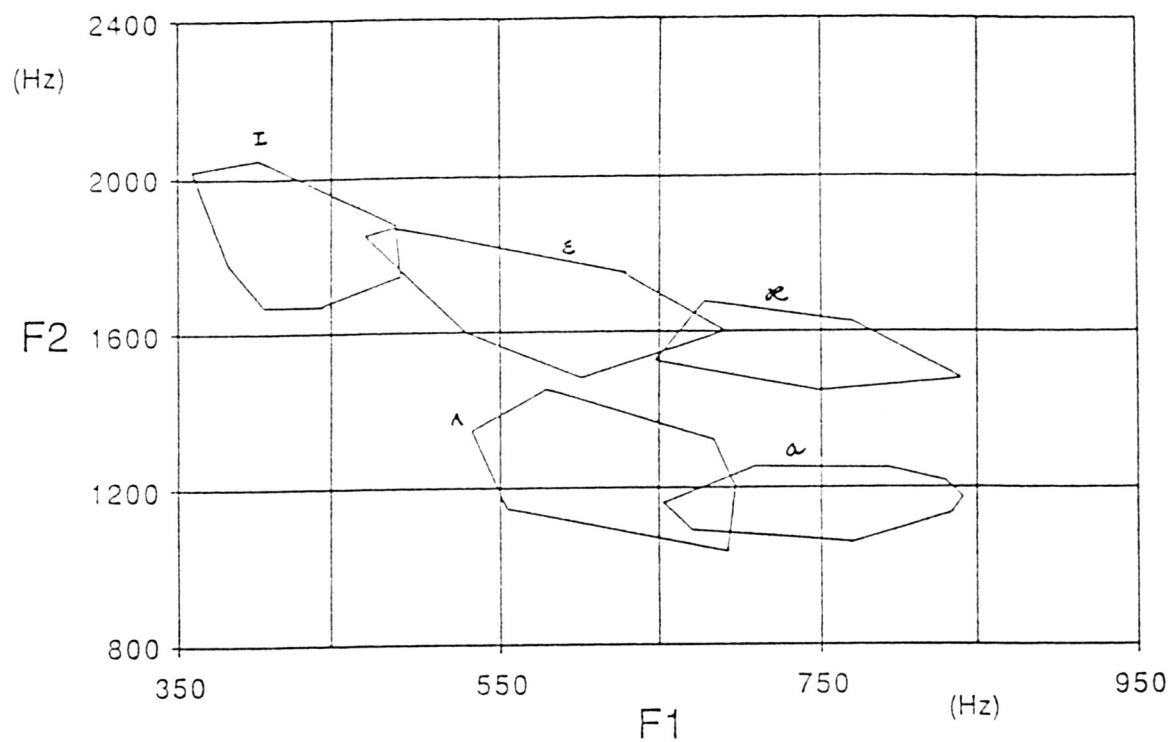


a)

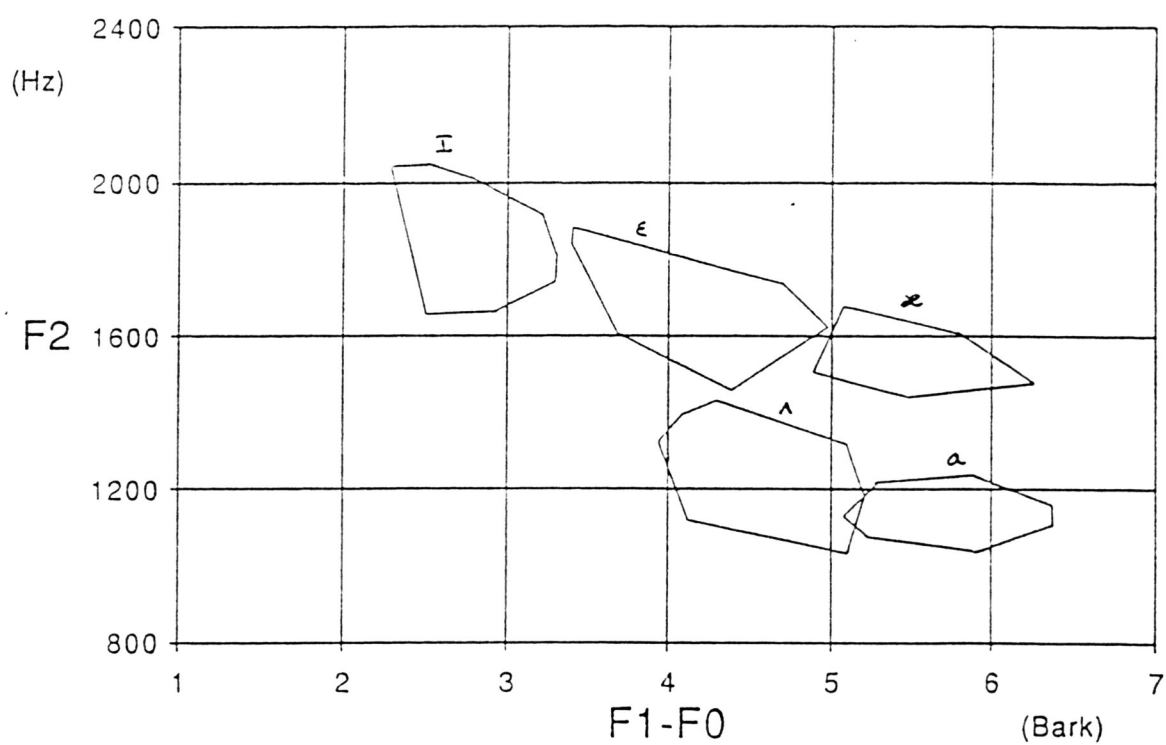


b)

Figure 2

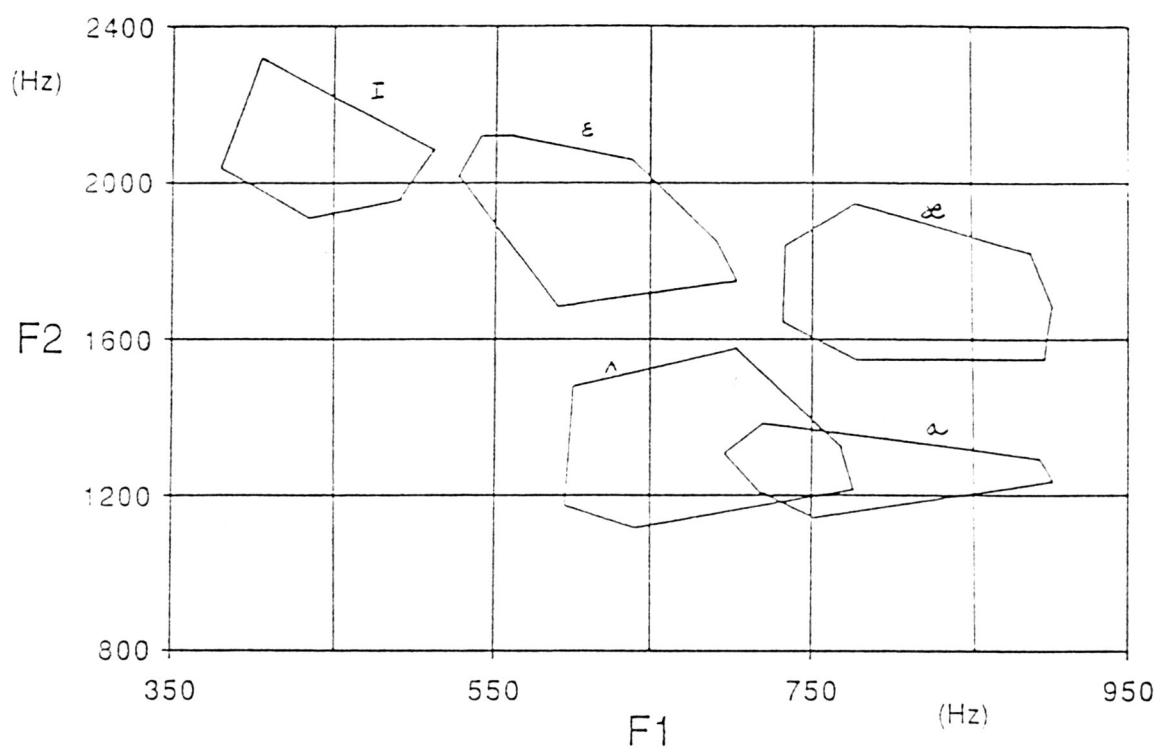


a)

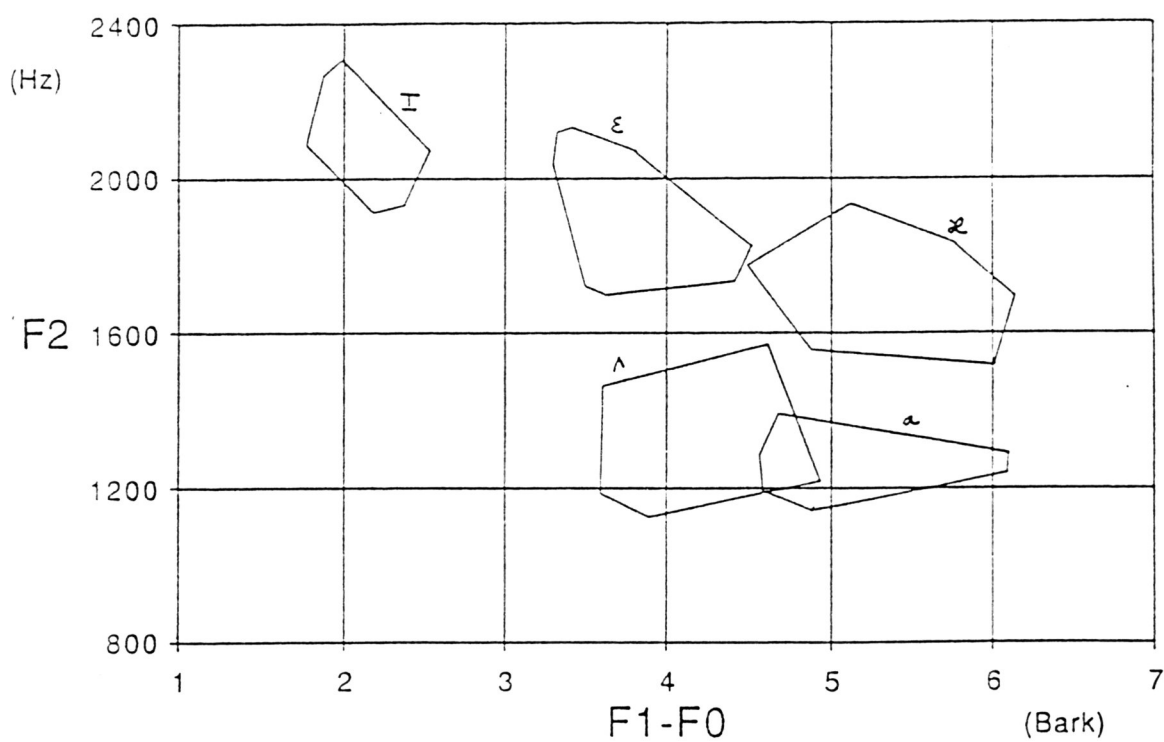


b)

Figure 3



a)



b)

Figure 4

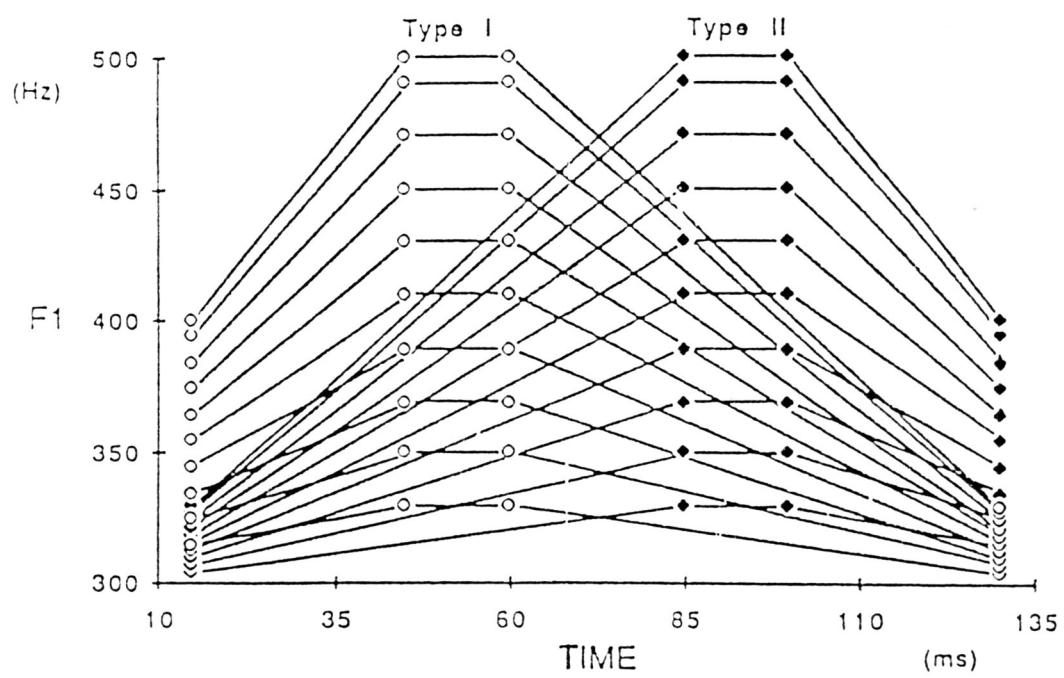


Figure 5

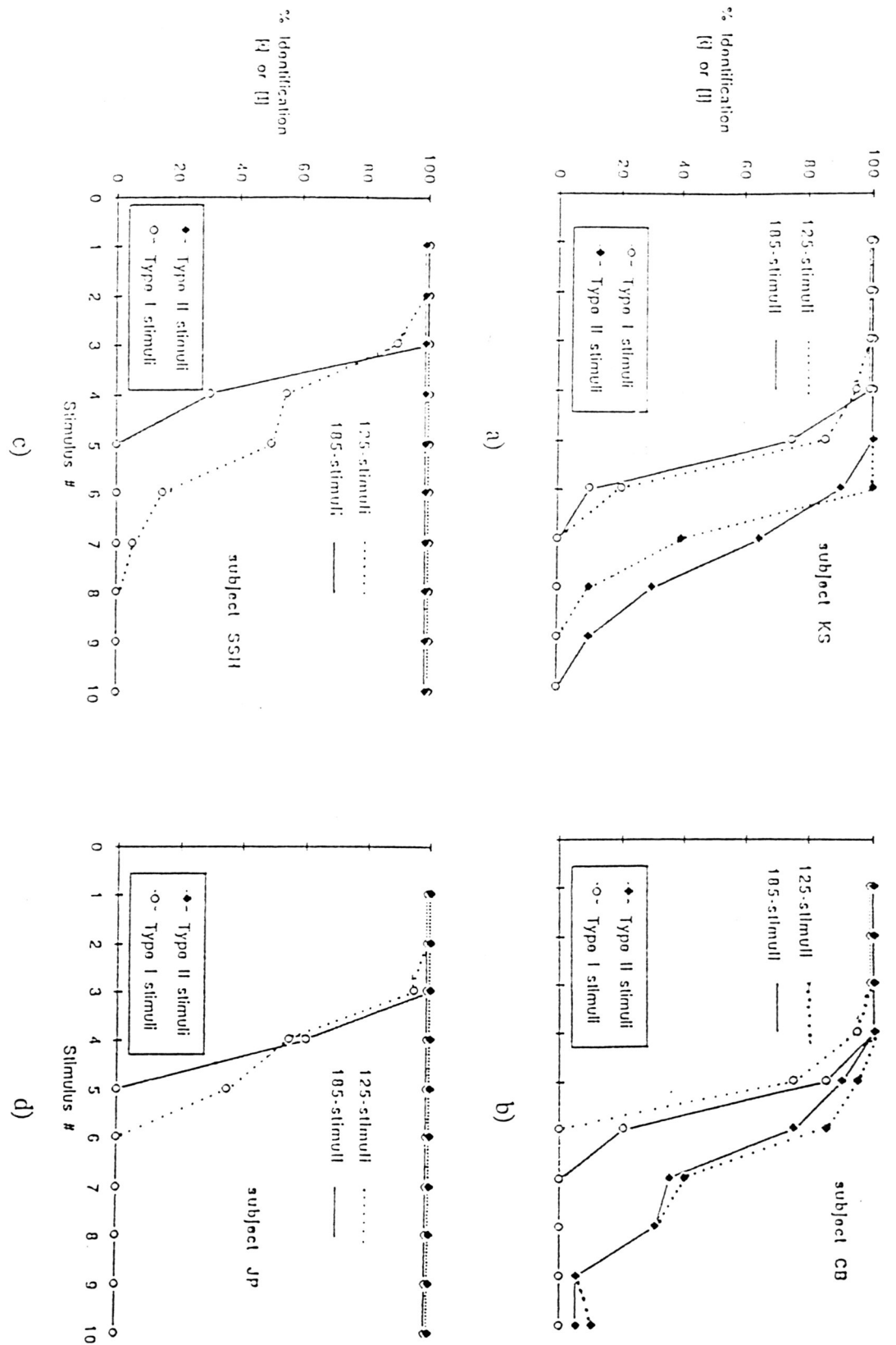


Figure 6

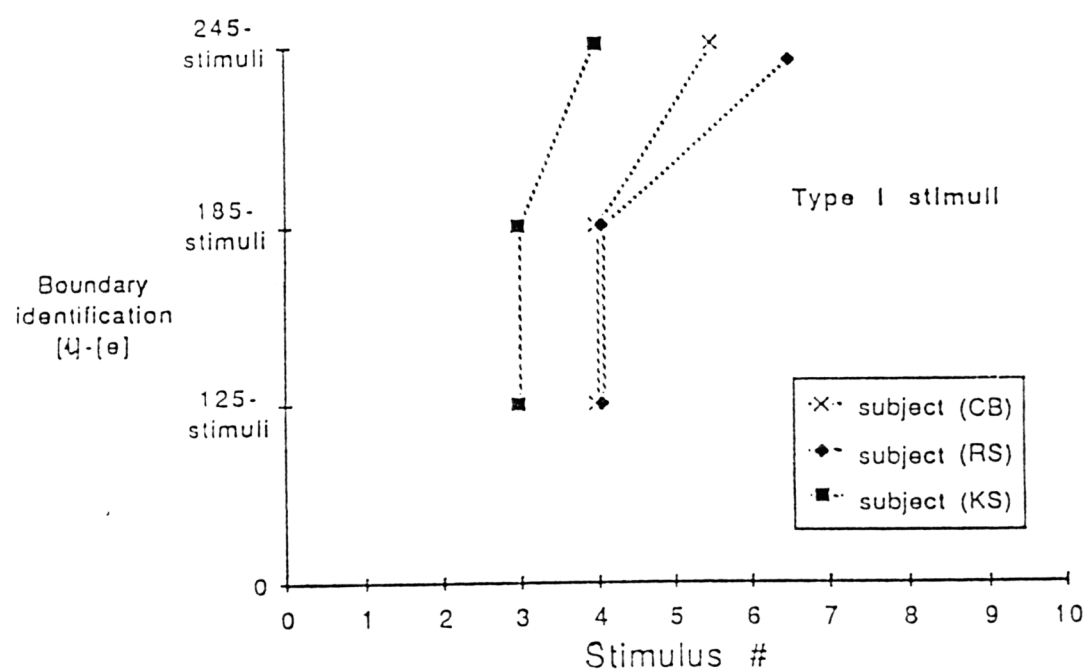


Figure 7

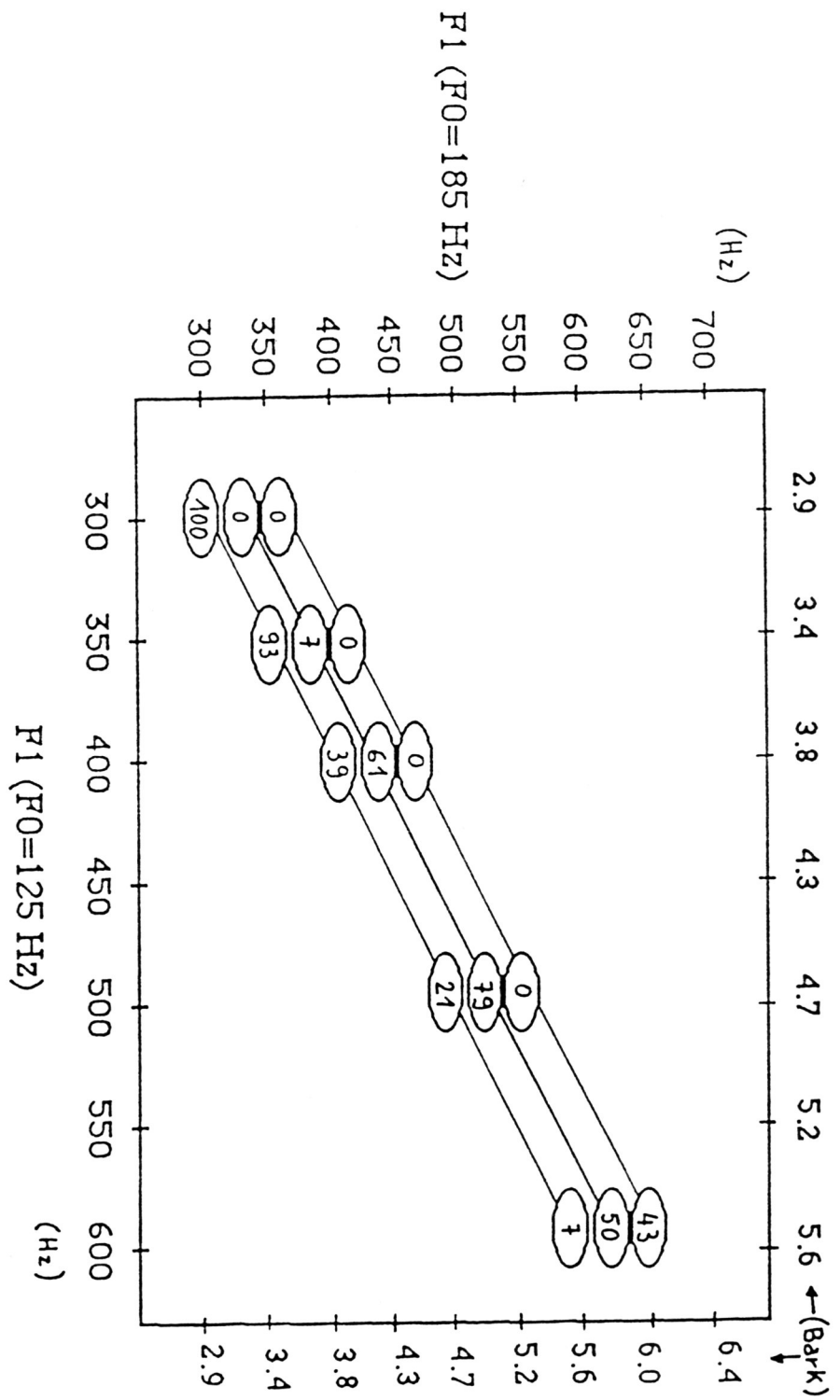


Figure 8

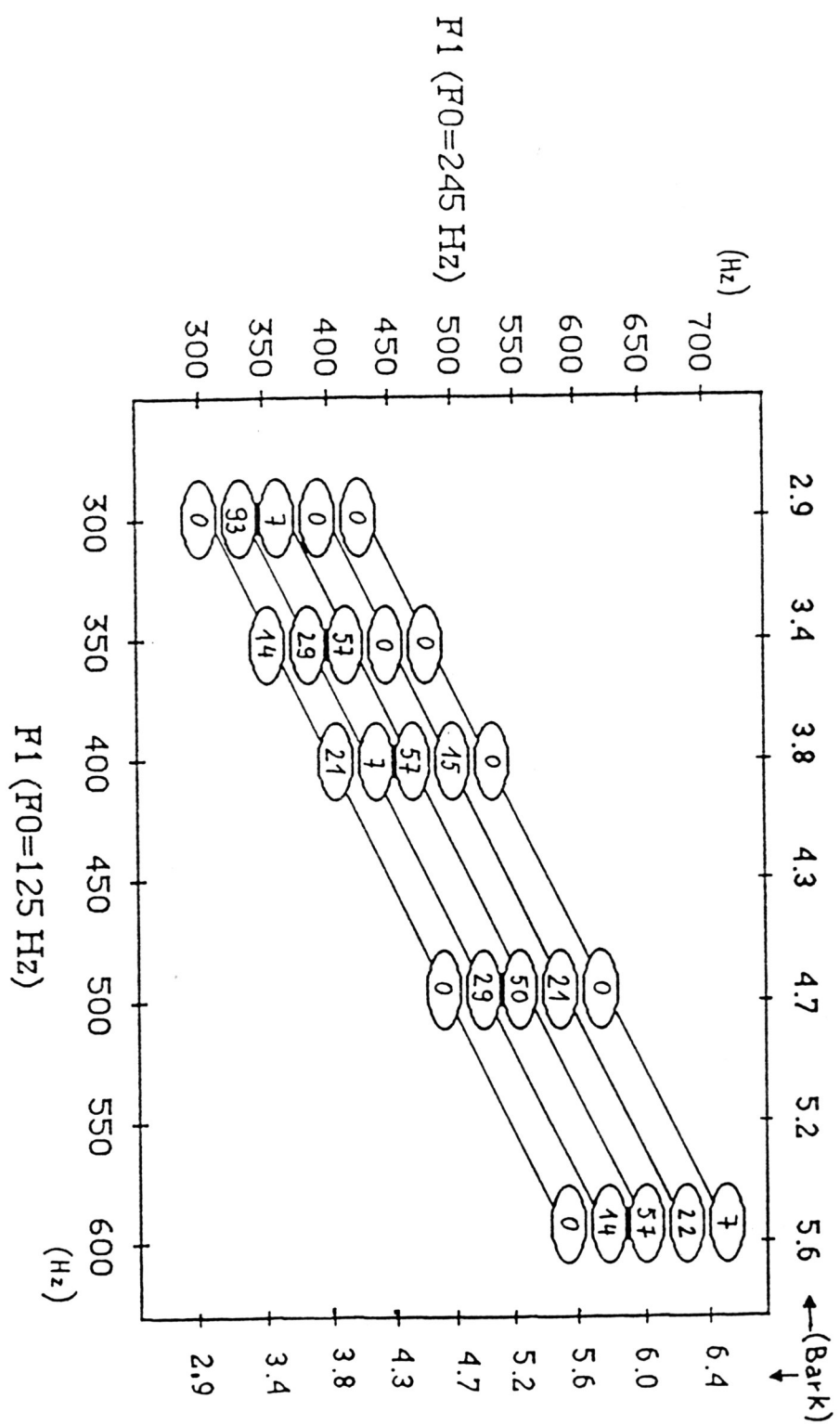


Figure 9

Table I. Averaged F0 values for each speaker and averaged F0 values for each vowel and speaker.

speaker	average	F0 (Hz)	F0 (Hz)	F0 (Hz)	F0 (Hz)	F0 (Hz)
	F0 (Hz)	vowel [ɪ]	vowel [ɛ]	vowel [æ]	vowel [a]	vowel [ʌ]
KS	127	132	126	122	129	124
JP	118	124	117	116	117	115
CR	191	199	192	186	191	186

Table II. Classification rates and Mahalanobis distance values, for the three vowel pairs [I]-[ε], [ε]-[æ], and [a]-[ʌ], obtained with the F1 and F2 values and with the (F1-F0) and F2 values, for the three speakers.

speaker			Vowel pairs					
			[I]	[ε]	[ε]	[æ]	[a]	[ʌ]
KS	classification rate %	F1 vs F2	98	100	94	96	92	68
		(F1-F0) vs F2	100	100	95	100	91	91
	Mahalanobis distance	F1 vs F2	20		14		3	
		(F1-F0) vs F2	27		20		6	
JP	classification rate %	F1 vs F2	100	96	94	98	92	98
		(F1-F0) vs F2	100	96	96	100	94	100
	Mahalanobis distance	F1 vs F2	13		14		10	
		(F1-F0) vs F2	16		17		13	
CR	classification rate %	F1 vs F2	100	100	100	100	79	90
		(F1-F0) vs F2	100	100	100	100	85	89
	Mahalanobis distance	F1 vs F2	24		27		5	
		(F1-F0) vs F2	52		29		6	

Table III. F1, (F1-F0), (F1-F0) end-corrected, and F2 values for each vowel and speaker, averaged over all the consonantal contexts and repetitions.

speaker	F1 values (Hz)				
	vowel [I]	vowel [ε]	vowel [æ]	vowel [a]	vowel [ʌ]
KS	432	539	604	703	634
JP	431	571	622	752	749
CR	428	600	693	791	818
	F1-F0 values (Bark)				
	vowel [I]	vowel [ε]	vowel [æ]	vowel [a]	vowel [ʌ]
KS	2.8	3.8	4.4	5.1	4.7
JP	2.9	4.2	4.6	5.6	5.6
CR	2	3.7	4.4	5.2	5.6
	F1-F0 end-corrected values (Bark)				
	vowel [I]	vowel [ε]	vowel [æ]	vowel [a]	vowel [ʌ]
KS	2.6	3.6	4.1	4.9	4.4
JP	2.6	3.8	4.3	5.3	5.3
CR	2.1	3.8	4.5	5.3	5.7
	F2 values (Hz)				
	vowel [I]	vowel [ε]	vowel [æ]	vowel [a]	vowel [ʌ]
KS	1690	1612	1169	1167	1576
JP	1837	1646	1242	1166	1551
CR	2118	1933	1296	1260	1665

Table IV. Maximum differences in the F1, (F1-F0), and (F1-F0) end-corrected values, for each vowel, between the three speakers (KS), (JP), and (CR).

	Vowel				
	[ɪ]	[ɛ]	[æ]	[a]	[ʌ]
maximum F1 variation (Hz)	4	61	184	88	89
maximum (F1-F0) variation (Bark)	0.9	0.5	0.9	0.5	0.2
" " " " end-cor.	0.5	0.2	1.3	0.4	0.4

Table V. Overview of the synthetic stimuli used in perceptual experiment I.

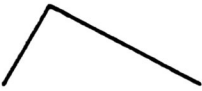

<div>F0 values</div> <div>trajectory type</div>	125	185	245
	125-I stimuli	185-I stimuli	245-I stimuli
	125-II stimuli	185-II stimuli	245-II stimuli

Table VI. Crossover values for type I and type II stimuli in the case of 125- and 185-stimuli for all subjects.

subject	Crossover values			
	KS	CB	SSH	JP
type I 125-stimuli	5.5	5.6	4.5	4.3
type I 185-stimuli	5.3	5.5	3.7	4.1
type II 125-stimuli	6.8	7.0	-	-
type II 185-stimuli	7.4	6.8	-	-