Journal of Phonetics (1994) **22**, 205–224

Acoustic and perceptual evidence of a complex relation between F_1 and F_0 in determining vowel height

Maria-Gabriella Di Benedetto*

Department of Information and Communication (INFOCOM), Faculty of Engineering; University of Rome 'La Sapienza', Via Eudossiana 18, 00184 Rome, Italy

Received 1st August 1990, and in revised form 14th November 1991

An acoustic analysis of five vowels of American English [ɪ, ɛ, æ, ɑ, ʌ], spoken by three speakers, showed that in the dimension representing vowel height, individual differences for low vowels were reduced when the vowels are represented by the ($F_1 - F_0$) difference rather than by F_1. However, individual differences for high vowels were increased by the use of the same distance. Results of perceptual experiments using synthetic CVC and one-formant stimuli were in agreement with the observations based on the acoustic analysis. They suggest that the relation between F_1 and F_0 depends on the ranges of both F_0 and F_1 values. A possible interpretation consists in giving to F_0 the role of an anchor point when both F_0 and F_1 are sufficiently high, while the extreme end of the scale would serve as the reference point when F_0 or F_1 assume values in the low frequency 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 (F_1) has been related to vowel height and the second formant frequency (F_2) to vowel backness. Vowels which are coded as being [+high] are characterized by lower F_1 values than vowels which are coded as [−high], and [+back] vowels are characterized by lower F_2 values than [−back] vowels.

However, it is well known that different F_1 and F_2 values can correspond to the same vowel, when this vowel is pronounced by different speakers. Results of several analyses on vowels (Stevens & House, 1963; Lisker, 1984; Di Benedetto 1989*a*) have shown that comparing the vowel areas in the F_1 *vs.* F_2 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 F_1 values similar to those of the vowel β-dispersion area for the other speaker, [α] and [β] being both [−back] or both [+back]. A normalization mechanism must be applied somewhere. For

* This work was carried out while the author was with the Speech Communication Group of the Massachusetts Institute of Technology, Cambridge, MA, U.S.A.

0095-4470/94/030205 + 20 $08.00/0

example, considering a 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.

A previous study (Di Benedetto, 1989*a*, *b*) examined the hypothesis that temporal properties of F_1 patterns distinguish vowels with identical central F_1 values. The present study examines the hypothesis that vowels are normalized according to F_0. It uses acoustical analysis of vowels and 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 F_0 time variations on the perception of vowel height was not investigated. Although F_0 and formant frequencies are associated with different elements of the speech production mechanism, temporal properties of F_0 might be important in determining vowel height, as was already shown for F_1 in the earlier study (Di Benedetto, 1989*a*).

Several other studies showed that F_0 can act as a normalizing factor of formant displacements by influencing the perception of vowel quality (Potter & Steinberg, 1950; Miller, 1953; Fant, Carlson & Granström, 1974; Traünmuller, 1981). Perceptual experiments carried out by Potter & Steinberg (1950) showed that a synthesized [æ], characterized by F_1 and F_2 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 F_0 gave rise to unnatural sounds.

The influence of F_0 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 F_0 (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 F_0 values, but that there was a shift toward higher frequencies of the boundaries in F_1 between different vowel areas when F_0 was higher.

Fant, Carlson & 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 F_1 and F_2 values and similar F_3 values. In their experiment, F_0 was switched from a "male" F_0 of 110 Hz to a "female" F_0 of 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. They found that the increase of the values of F_3 and higher formants shifted the perceptual quality of the stimuli in the [ø] direction. This result is opposite to the effect that they expected, since this change would increase F2′, a parameter which is a combination of F_2, F_3 and F_4 and which they earlier showed to be, together with F_1, sufficient to synthesize Swedish vowels (Carlson, Granström & Fant, 1970), and consequently should favor [e] responses. Fant *et al.* (1974) showed how this shift provokes a loss of spectral energy above F_2 and suggested that F2′ for [ø] might be very close to F_2 for female speakers. Other experiments (Carlson *et al.*, 1970), had shown that F2′ for the male [e] is located halfway between F_2 and F_3. 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 F_0 for the female speaker, and, to a smaller extent, to F_3 and/or F_4 and higher formants.

More recently, Traunmüller (1981) examined the role of intrinsic factors, such as F_1 and F_0, in the determination of perceptual degree of openness. In one of his experiments, one-formant stimuli, in which F_1 and F_0 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 $(F_1 - F_0)$, the difference between F_1 and F_0, using synthetic versions of natural vowels in order to analyze the influence of the higher formants on the perception of the $(F_1 - F_0)$ cue. The general conclusion was that the distance between F_1 and F_0, 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 & Gopal, 1986), the distance between F_1 and F_0 was incorporated together with the distances between F_2 and F_1, F_3 and F_2, F_4 and F_3, and F_4 and F_2, 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 & Lublinskaya, 1979). Syrdal's findings were that high vowels are separated from mid and low vowels by the critical distance of the Bark-transformed $(F_1 - F_0)$ of about 3 Bark and that the $(F_1 - F_0)$ 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 F_1 and F_0, expressed in Bark, is more appropriate than the traditional measurement F_1 to represent the height of vowels, pronounced by different speakers and in several consonantal contexts. The analysis used the five unrounded and non-diphthongal vowels of American English [ɪ, ɛ, æ, ɑ, ʌ], uttered by three speakers. These results are presented in Section 2. Second, the way F_0 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 are described in Section 3. The agreement of the results obtained with the findings of the acoustic analysis and interpretation of the results are discussed in the last section.

2. Acoustic analysis

2.1. *Experimental conditions and procedures*

The aim 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 diphthongal vowels were excluded since, as is well known, rounding and diphthongal quality affect the F_1 values, conflicting with F_1 effects due to height variations. In the American English vowel system, [ɪ, ɛ, æ, ɑ, ʌ] constitute the set of the unrounded and monophthongal vowels. This includes two different subsets, back vowels [ɑ, ʌ] and front vowels [ɪ, ɛ, æ], within each of which vowels can be distinguished on the sole basis of height. Results of a previous analysis (Di Benedetto, 1989*a*) showed that the three front vowels could be separated from the two back vowels in the F_2 dimension with a sufficient degree of accuracy.

An extended description of the experimental conditions and procedures can be found in Di Benedetto (1989*a*) 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 (i.e., syllables such as /bɪd/ were included but those such as /bɪt/ were excluded) leading to 18 different syllables, and pronounced in the frame *The_again*. The /hVd/ and /# Vd/ syllables were also included in the analysis to serve as reference points for comparison with a previous study on vowel reduction (Stevens & House, 1963). All the syllables were uttered by three native speakers of American English (two males and one female). Each syllable type was repeated three times. The corpus obtained included then five vowels in 20 different consonantal contexts in three repetitions, leading to 300 tokens per speaker. Thus, within-speaker formant variability could be acceptably determined on the basis of these data (Di Benedetto 1989*a*). Analysis of between-speaker variability would benefit from an extension to more speakers, but the results of the present study should constitute a good basis for the perceptual investigation.

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

Values for F_1 and F_2 were manually extracted for each vowel by plotting the vowel spectrum (256-point DFT) every 5 ms. These values were systematically compared to those obtained automatically by means of the program KLSPEC developed by Klatt (1984) on the pseudospectrum. The pseudospectrum was obtained by windowing a slice of the signal (for example, 256 samples or 25.6 ms 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 spectrogram-like filters.

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

The temporal sampling point of F_1, F_2 and F_0 was the time at which F_1 reached its maximum, as discussed in Di Benedetto (1989*a*).

2.2. *Results of acoustic measurements*

Table I shows the results of measurements of F_0 for each speaker and each vowel. These data were obtained by averaging the F_0 values of each vowel across the 20 consonantal contexts and the three repetitions. As expected, the highest and most different F_0 was found for the female speaker CR, while F_0 values for the two male speakers JP and KS were lower and more comparable to each other. Table I also shows that F_0 is related to vowel height. This same effect was found in the past by other investigators (Peterson & Barney, 1952; House & Fairbanks, 1953) and may

TABLE I. Averaged overall F_0 values for each speaker and averaged F_0 values for each vowel and speaker

Speaker	Average F_0 (Hz)	F_0 (Hz)				
		Vowel [ɪ]	Vowel [ɛ]	Vowel [æ]	Vowel [ɑ]	Vowel [ʌ]
KS	127	132	126	122	129	124
JP	118	124	117	116	117	115
CR	191	199	192	186	191	186

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 a previous study on the same speech material (Di Benedetto, 1987), it was observed that vowels in voiced consonantal contexts had lower F_1 values than in voiceless consonantal contexts. Here we see that vowels in voiced consonantal contexts also had lower F_0 values than in voiceless consonantal contexts, in agreement with House & Fairbanks (1953), who suggested that the lower F_0 of voiced consonants may have a lowering effect on the F_0 values of adjacent vowels.

Because of the congruence of these two effects, the difference in raw F_1 value between tokens of the same vowel in voiced *vs.* voiceless consonantal contexts was larger than the difference in the F_0 normalized ($F_1 - F_0$) value even when F_1 and F_0 are expressed in hertz (see Fig. 1).

The results of the analysis in the ($F_1 - F_0$) *vs.* F_2 space for speakers KS, JP and CR are presented in Fig. 2(b), 3(b) and 4(b), respectively. Figures 2(a), 3(a) and 4(a) show, for comparison, the representations in the raw F_1 *vs.* F_2 space. The area shown for each vowel is the regular polyhedron which included all 60 data points for the vowel (18 possible stop contexts plus the /hVd/ and /#Vd/ contexts, each in three repetitions). This type of representation, which makes use of the polyhedron rather than of the ellipsis of equiprobability, was proposed in Di Benedetto (1989*a*). It provides a schematic representation without obscuring potentially critical details. Note that as in Syrdal (1985) the ($F_1 - F_0$) values are expressed in the critical tonality scale Bark according to Zwicker & Terhardt's (1980) formula for the critical band rate B:

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

where F is the frequency in kHz. On the other hand, the F_2 values are left in hertz since results of the previous analysis (Di Benedetto, 1989*a*) indicated that no

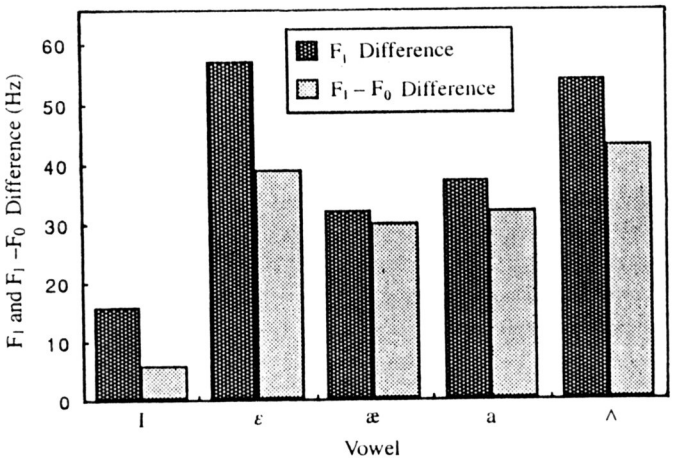


Figure 1. Differences between vowels in voiced *vs.* voiceless consonental contexts in raw F_1 values and in F_0 normalized ($F_1 - F_0$) values for each vowel averaged over the results obtained for the three speakers, the three repetitions and the 20 consonantal contexts.

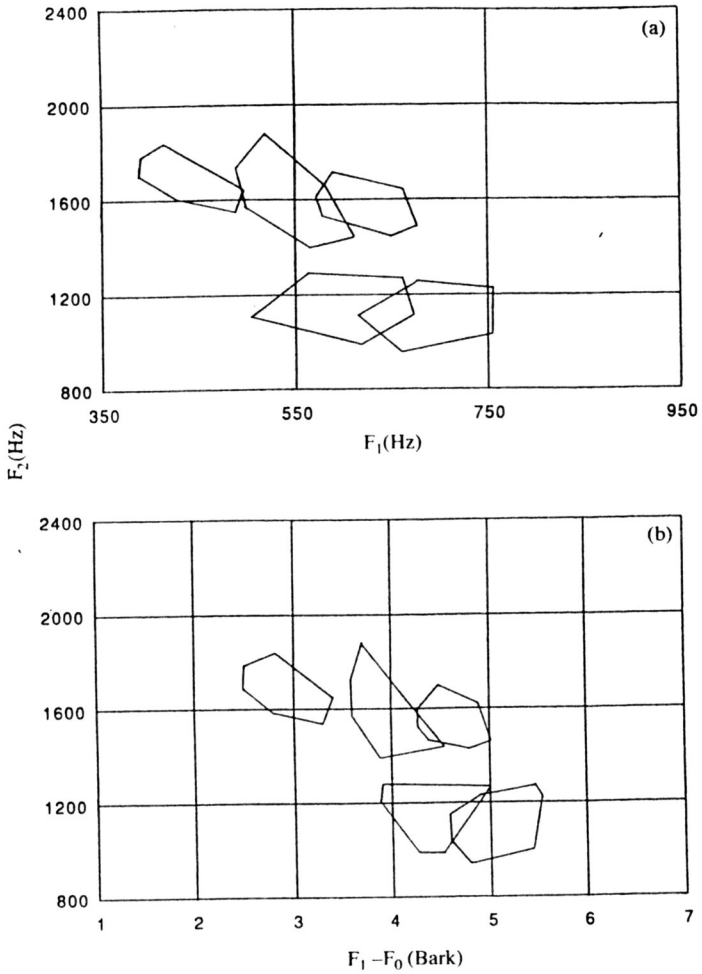


Figure 2. Vowel areas in (a) F_1 *vs.* F_2 space, and (b) $(F_1 - F_0)$ *vs.* F_2 space, for male speaker KS. Each vowel is represented by the regular polyhedron which included all the *F* patterns for that particular vowel (60 tokens for each vowel).

overlap occurred between vowel areas in the F_2 dimension for any single speaker, suggesting that the representation of the dimension of vowel backness by F_2 is sufficiently accurate, at least for the purpose of this study.

Figures 2(b), 3(b) and 4(b) show that overlap still occurred in the $(F_1 - F_0)$ dimension between contiguous vowel areas. This overlap was between [ɑ] and [ʌ] for all speakers, and also between [ɛ] and [æ] for speakers JP and KS. An analysis was carried out to compare the results of the acoustic analysis in the $(F_1 - F_0)$ *vs.* F_2 space to the results of the analysis of the same speech materials in the F_1 *vs.* F_2 space. To this aim, the amount of overlap between contiguous vowel areas in the $(F_1 - F_0)$ dimension was quantified, for each speaker, by determining the straight lines which better separate the sets represented by the $(F_1 - F_0) - F_2$ values of [ɪ] *vs.* [ɛ], [ɛ] *vs.* [æ], and [ɑ] *vs.* [ʌ] in a linear discriminant analysis. The statistical

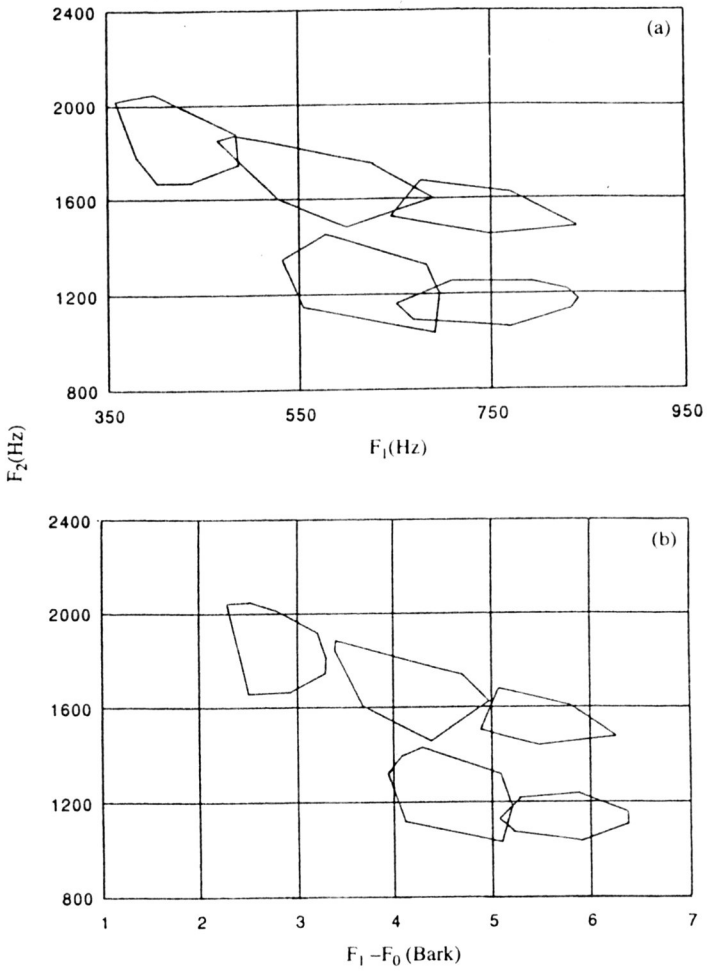


Figure 3. Vowel areas in (a) F_1 *vs.* F_2 space, and (b) $(F_1 - F_0)$ *vs.* F_2 space, for male speaker JP, as in Fig. 2.

distribution of the measurements was assumed 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 (1989*a*), a generalized Euclidean distance (Mahalanobis distance) was computed in order to quantify the distance between the two sets in each pair. 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 an equal amount of spreading or an 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 $(F_1 - F_0)$ *vs.* F_2 space gave a better grouping and better separation of the vowel areas, compared to what was obtained in the F_1 *vs.* F_2 space, although problems of overlap still occurred. Since the differences between vowels in voiced *vs.* voiceless consonantal contexts were smaller

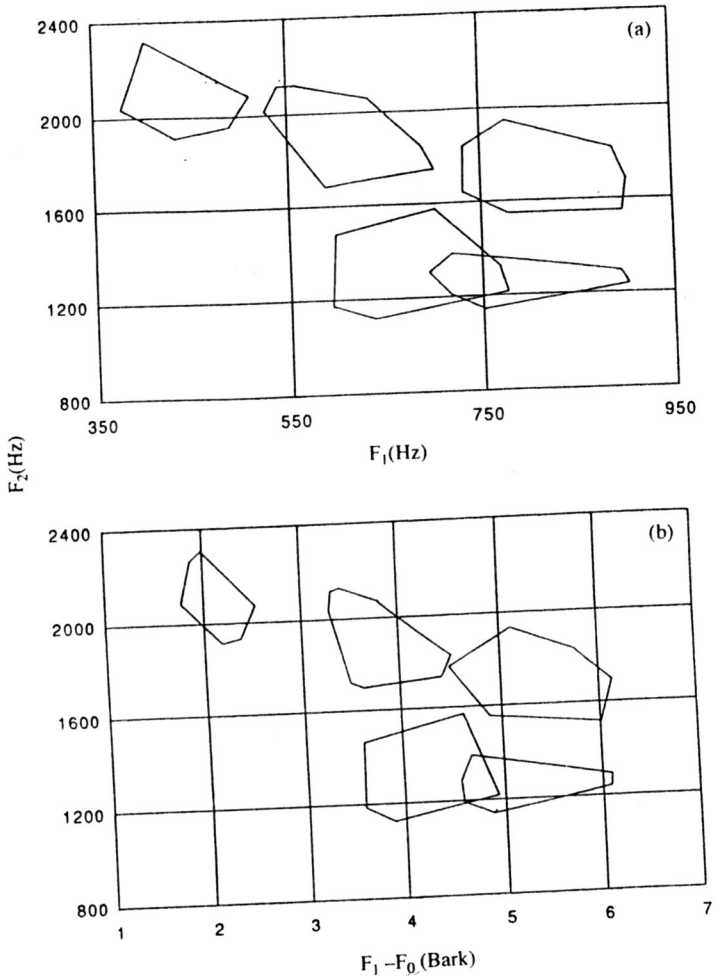


Figure 4. Vowel areas in (a) F_1 *vs.* F_2 space, and (b) $(F_1 - F_0)$ *vs.* F_2 space, for female speaker CR, as in Fig. 2.

in $(F_1 - F_0)$ values than in F_1 values, one of the factors contributing to a better separation of the vowel areas must be that the vowel areas for voiced *vs.* voiceless consonantal contexts were more closely grouped.

In order to compare the interspeaker variations in the F_1 *vs.* F_2 and the $(F_1 - F_0)$ *vs.* F_2 representations, the F_1, $(F_1 - F_0)$ and F_2 values were averaged over all the consonantal contexts and repetitions for each speaker and each vowel. In addition, a different set of $(F_1 - F_0)$ values was computed by applying an end-correction of the Bark scale, as proposed by Traumüller (1981). This low-frequency end-correction raises all frequencies below 150 Hz to 150 Hz. For frequencies between 150 Hz and 200 Hz, the modified frequency is obtained by subtracting from the original value a percentage of the original value normalized to 150 Hz. An identical procedure is used for frequencies between 200 Hz and 250 Hz, except that the original frequency is normalized to 250 Hz. The F_1, $(F_1 - F_0)$, $(F_1 - F_0)$ end-corrected and F_2 values obtained for each speaker and each vowel are presented in Table III. Since the three

TABLE II. Classification rates and Mahalanobis distance values, for the three vowel pairs [ɪ]–[ɛ], [ɛ]–[æ] and [ɑ]–[ʌ], obtained with the F_1 and F_2 values and with the $(F_1 - F_0)$ and F_2 values, for the three speakers

Speakers	Values	Vowel pairs [ɪ] [ɛ]	[ɛ] [æ]	[ɑ] [ʌ]
KS				
Classification rate (%)	F_1 *vs.* F_2	98 100	94 96	92 68
	$(F_1 - F_0)$ *vs.* F_2	100 100	95 100	91 91
Mahalanobis distance	F_1 *vs.* F_2	20	14	3
	$(F_1 - F_0)$ *vs.* F_2	27	20	6
JP				
Classification rate (%)	F_1 *vs.* F_2	100 96	94 98	92 98
	$(F_1 - F_0)$ *vs.* F_2	100 96	96 100	94 100
Mahalanobis distance	F_1 *vs.* F_2	13	14	10
	$(F_1 - F_0)$ *vs.* F_2	16	17	13
CR				
Classification rate (%)	F_1 *vs.* F_2	100 100	100 100	79 90
	$(F_1 - F_0)$ *vs.* F_2	100 100	100 100	85 89
Mahalanobis distance	F_1 *vs.* F_2	24	27	5
	$(F_1 - F_0)$ *vs.* F_2	52	29	6

measurement sets (Hz, Bark and end-corrected Bark values) give incomparable units, an absolute interspeaker variation measure was also computed using the formula:

$$S_{[\text{speaker X,vowel A}]} = \frac{M_{[\text{speaker X,vowel A}]} - M_{[\text{speaker Y,vowel A}]}}{M_{[\text{speaker X,vowel A}]}}$$

where $M_{[\text{speaker X,vowel A}]}$ is either the F_1 value in Hz, or the $(F_1 - F_0)$ value in Bark, or the $(F_1 - F_0)$ value in end-corrected Bark units, for speaker X and vowel A. The S parameter gives an indication of the distance between speaker X and speaker Y for vowel A, relative to vowel A measurements for speaker X. S is adimensional and the F_1 values in incomparable measurement units (Bark *vs.* Hz) can therefore be compared on the basis of the S values. Table IV shows the S values obtained by comparing each speaker with all the others, averaged for each vowel. The inter-speaker variation was reduced using the $(F_1 - F_0)$ parameter for the low front vowel [æ] and the two back vowels [ɑ, ʌ]. A reduction of the inter-speaker variation was not obtained for the vowels [ɪ, ɛ], and in particular for the high vowel [ɪ] a noticeable increase in the inter-speaker variation was observed. When the end-corrected Bark units were used, the results were the following: for [ɪ], the inter-speaker variation was still higher in the $(F_1 - F_0)$ than in the F_1 dimension; a lower inter-speaker variation was obtained with the $(F_1 - F_0)$ parameter for [ɛ, ɑ, ʌ]; in the case of [æ] the results were comparable. Comparing the two $(F_1 - F_0)$ representations, we see a substantially better grouping using the end-correction for [ɪ] and [ɛ] and also a slightly better grouping for [ɑ], but not for [æ] and [ʌ].

TABLE III. F_1, $(F_1 - F_0)$, $(F_1 - F_0)$ end-corrected and F_2 values for each vowel and speaker, averaged over all the consonantal contexts and repetitions

	F_1 values (Hz)				
Speaker	Vowel [ɪ]	Vowel [ɛ]	Vowel [æ]	Vowel [ɑ]	Vowel [ʌ]
KS	432	539	634	703	604
JP	431	571	749	752	622
CR	428	600	818	791	693
	$F_1 - F_0$ values (Bark)				
	Vowel [ɪ]	Vowel [ɛ]	Vowel [æ]	Vowel [ɑ]	Vowel [ʌ]
KS	2.8	3.8	4.7	5.1	4.4
JP	2.9	4.2	5.6	5.6	4.6
CR	2	3.7	5.6	5.2	4.4
	$F_1 - F_0$ end-corrected values (Bark)				
	Vowel [ɪ]	Vowel [ɛ]	Vowel [æ]	Vowel [ɑ]	Vowel [ʌ]
KS	2.6	3.6	4.4	4.9	4.1
JP	2.6	3.8	5.3	5.3	4.3
CR	2.1	3.8	5.7	5.3	4.5
	F_2 values (Hz)				
	Vowel [ɪ]	Vowel [ɛ]	Vowel [ʌ]	Vowel [ɑ]	Vowel [æ]
KS	1690	1612	1576	1167	1169
JP	1837	1646	1551	1166	1242
CR	2118	1933	1665	1260	1296

3. Perceptual experiments

3.1. *Experiment* 1

The aim of this experiment was to investigate the influence of F_0 on the perception of vowel height, using three sets of /dVd/ synthetic syllables. One set consisted of stimuli already used in a previous experiment (Di Benedetto, 1989*b*), to find for

TABLE IV. Average *S* parameter values for each vowel's F_1 values derived from the original Hz, Bark and end-corrected Bark units. The *S* parameter, as defined in the text, is an adimensional parameter which indicates the amount of inter-speaker variation for a given vowel and a given measurement unit

	Vowel				
	[ɪ]	[ɛ]	[æ]	[ɑ]	[ʌ]
S value (Hz)	0.006	0.071	0.171	0.079	0.092
S value (Bark)	0.253	0.085	0.117	0.062	0.029
S value (end-corrected Bark)	0.143	0.036	0.175	0.052	0.062

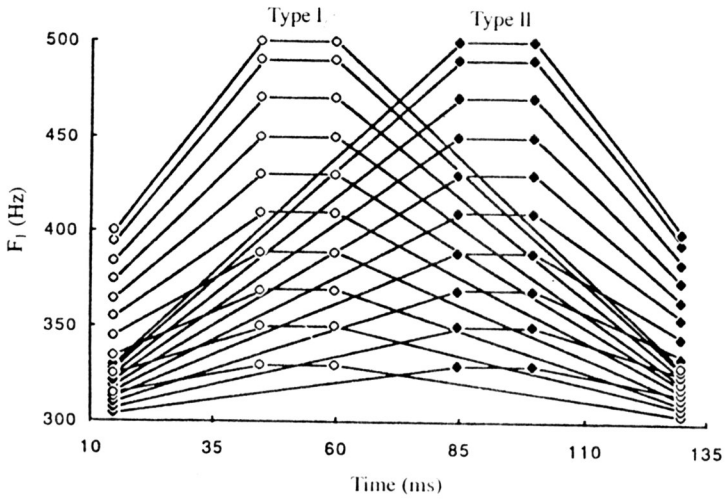
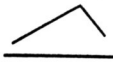


Figure 5. Schematic F_1 trajectories for the stimuli of type I and of type II used in Experiment 1.

different listeners the effect on the perceptual boundary between the high vowels [i, ɪ] and the non-high vowels [e, ɛ] of differences in F_1 onset values and in location of F_1 maximum within the synthetic vowel. Figure 5 shows the F_1 trajectories of these synthetic stimuli, which have two shapes: type I with early F_1 peak, and type II with late F_1 peak. Ten stimuli of each type were synthesized with F_1 maximum values of 330, 350, 370, 390, 410, 430, 450, 470, 490 or 500 Hz. The trajectories of the higher formants and of fundamental frequency (F_0 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 as regards the F_1 trajectory shape and higher formant trajectories, while the maximum fundamental frequency was increased in two steps: 60 Hz and 120 Hz. (All the stimuli described in this and the following sections were synthesized with the Klatt synthesizer (Klatt, 1980, 1984).) Table V summarizes the resulting six continua.

Experiment 1 consisted of two tests: a vowel identification test, and a "boundary" identification test. Subjects were all phonetically trained listeners, native speakers of American English and members of the Speech Communication Group at the Massachusetts Institute of Technology. Four subjects participated in the vowel identification test and two of these and one other subject participated in the boundary identification test. The vowel identification test was first carried out using

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

Trajectory type \ F_0 values (Hz)	125	185	245
	125–I stimuli	185–I stimuli	245–I stimuli
	125–II stimuli	185–II stimuli	245–II stimuli

the 125–I and 125–II continua. It consisted of three parts. In the first part, each 125–I stimulus was presented 10 times. The stimuli were ordered in such a way that each stimulus followed another only once. In this way, the average number of high-vowel 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 the same way using the 125–II stimuli. Then in the third part of the test, stimuli of both types were presented mixed together, in two subsets: the first containing the even-numbered stimuli, and the second containing the odd-numbered stimuli. The stimuli of each subset were ordered to make each stimulus follow another only once. Each stimulus was presented 10 times in the relevent subpart of the mixed part, so that it was presented 20 times over all three parts. In each part of the experiment, the stimuli were separated by a pause of 3 s. At the end of each part, the test was interrupted and the subject could rest for 2–3 minutes. The total test was approximately 45 minutes long. The same procedure was then repeated using the 185–I and –II stimulus continua.

A previous experiment, based on an open-response set (Di Benedetto 1989*b*) showed that American English listeners identified the vowels in the stimuli as [i], [ɪ], [e] or [ɛ]. Therefore, in the present experiment, the subjects were asked to identify the vowel of the synthetic utterances in a closed-set response as one of [i, ɪ, e, ɛ]. In interviews with the subjects after the tests, none of the subjects acknowledged perceiving a different vowel from these four. There were 20 responses per data point.

A "boundary" identification test was then carried out. The three different type I stimulus continua were used. Results of previous experiments presented in Di Benedetto (1989*b*) showed that the type I stimuli were perceived by the American subjects mainly as the tense vowels [i] and [e]. Sequences of the 10 stimuli characterized by the same F_0 (and the same sequences in reverse order) were played to the subjects, who were asked to identify each stimulus as [i] or [e] and to note when their perception of the synthetic vowels changed from [i] to [e] or *vice versa*. Each sequence, in each order, was presented three times.

3.1.1. *Results of the identification test*

Results of the identification test are presented in Fig. 6, for each subject separately. Note that subjects JP and SSH identified type II stimuli as [i] or [ɪ] and never as [e] or [ɛ]. The behaviour of subjects SSH and JP was explained in Di Benedetto (1989*b*) as due to the shape of the F_1 trajectory being more relevant than the F_1 maximum value for these subjects. Figure 6 shows that a change of 60 Hz in F_0 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% cross-over point were computed (see Neter & Wassermann, 1974). The cross-over points, indicated by their value in "stimulus numbers", are presented in Table VI. The difference in cross-over values obtained with 125 Hz and 185 Hz stimuli, in the type I and in the type II sets, was small for all subjects, and in all cases less than one stimulus number.

3.1.2. *Results of the "boundary" identification test*

During a preliminary informal test, the three subjects who participated in this experiment reported that they perceived the vowels of the synthetic utterances as

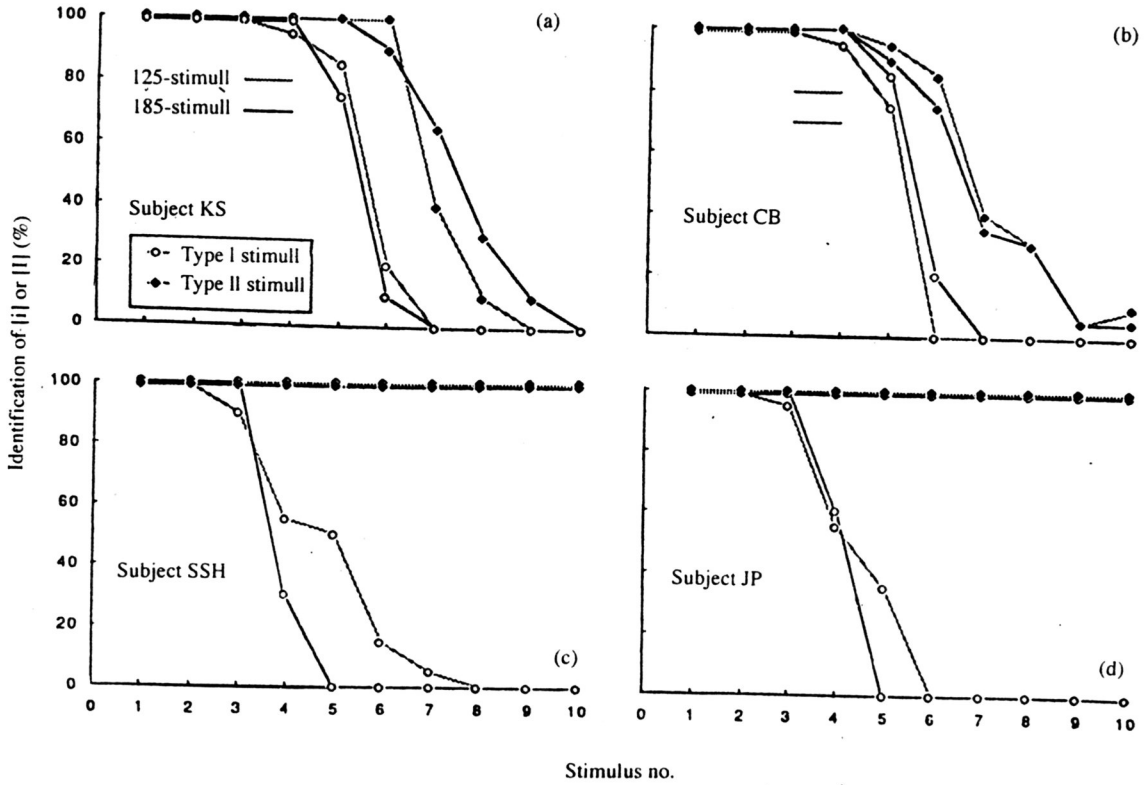


Figure 6. Results of the identification test for the four subjects. Per cent identification as a high vowel ([i] or [ɪ]) as a function of F_1 maximum for the two different F_1 trajectory shapes and two F_0 maxima. Each data point represents 20 responses.

TABLE VI. Cross-over values (in stimulus) for identification functions in Fig. 6 for each subject for each stimulus continuum

Stimuli	Subject's cross-over 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	—	—

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 stimulus numbers, or the last stimulus which was perceived as [e] in the case of sequences ordered according to a descending stimulus number progression. Each data point is the average over six presentations (three in each order) of the stimulus sequence. The figure shows that in the case of the three subjects who particpated in this test, an increase in F_0 from 125 to 185 Hz did not result in a change of the perceptual boundary between [i] and [e], while an increase from 125 Hz 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 F_1 maximum) for subject KS, between one and two stimuli numbers for CB and two stimuli numbers (40 Hz) for RS. The results were consistent across the three

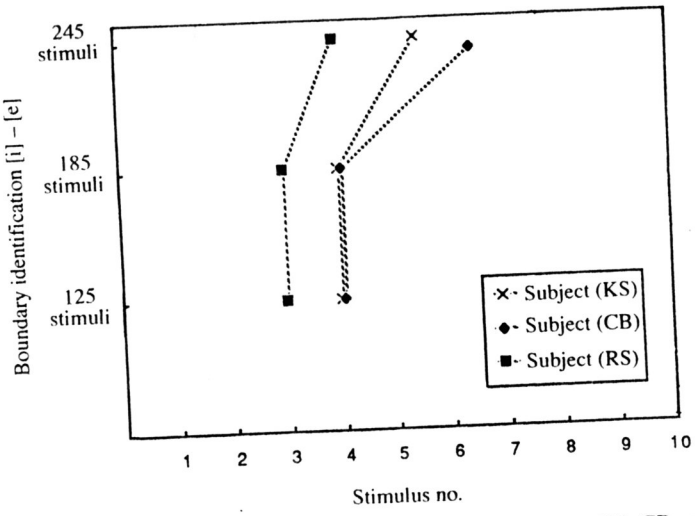


Figure 7. Results of the boundary identification test for subjects KS, CB and RS. Each point indicates the average stimulus at which the identification changes from [i] to [e] (or [e] to [i]), in the three ascending and three descending presentations of the stimulus sequence. Higher stimulus number on the x-axis corresponds to higher F_1.

presentations in each order, and, moreover, no difference was observed in the results obtained with sequences of stimuli with F_1 increasing or in the reverse order. The results of the boundary identification test were in agreement with the results of the identification test for F_0 changes from 125 Hz to 185 Hz and showed, in addition, that when F_0 was increased from 125 Hz to 245 Hz (change of 120 Hz) the perceptual boundary between [i] and [e] shifted towards higher F_1 values (higher stimuli numbers) by about 20–40 Hz which is considerably less than the shift in the value of $(F_1 - F_0)$ between the two stimulus continua. 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

The second experiment used one-formant vowel stimuli with stationary F_0 and F_1, and $F_0 = 125$ Hz, 185 Hz or 245 Hz. Standard stimuli with $F_0 = 125$ Hz were generated with five different F_1 of 300, 350, 400, 500 and 600 Hz. Each of these stimuli was then paired with three one-formant stimuli with $F_0 = 185$ Hz and values of F_1 ranging from the F_1 value of the standard stimulus to the F_1 value that would give the same $(F_1 - F_0)$ value (in hertz) for the comparison and the standard stimulus. For each standard stimulus a series of three pairs was played. For example, the standard stimulus with $F_1 = 300$ Hz was paired with a comparison stimulus with $F_1 = 300$ Hz ($F_0 = 185$ Hz) with a comparison stimulus with $F_1 = 360$ Hz($F_0 = 185$ Hz) and consequently the same $(F_1 - F_0) = 155$ Hz, and with an intermediate comparison stimulus with $F_1 = 330$ Hz ($F_0 = 185$ Hz). Each stimulus pair was played three times. Another set of pairs was produced with the same standard stimuli ($F_0 = 125$ Hz) but paired against comparison stimuli with $F_0 = 245$ Hz. To accommodate the larger F_0 difference, here a series of five pairs was played. For example, the standard stimulus with $F_1 = 500$ Hz was paired with a comparison stimulus with the same $F_1 = 500$ Hz, a comparison stimulus with $F_1 = 620$ Hz and consequently the same $(F_1 - F_0)$ of 375 Hz, and with each of three intermediate comparison stimuli with $F_1 = 530$, 560 and 590 Hz.

Seven subjects participated in this experiment. All were phonetically trained listeners, native speakers of American English, and members of the Speech Communication Group at the Massachusetts Institute of Technology. They were asked to indicate, in each series of three or five pairs, the pair in which the comparison stimuli was most similar to the standard in terms of vowel height. The stimulus chosen did not have to be necessarily identical to the standard stimulus in vowel quality. In fact, the judgements could be based upon different features perceptible to the listener. For example, different matching results could be obtained if different subjects were judging the stimuli for their height *vs.* for their backness. No control experiment was run on this aspect.

Figures 8 and 9 show the results. They plot on the abscissa the F_1 of the standard stimulus (with $F_0 = 125$ Hz), and on the ordinate the average F_1 of comparison stimuli (with $F_0 = 185$ Hz or 245 Hz) identified with that standard stimulus by the different listeners. Each standard stimulus could be identified with any of the three or five comparison stimuli: the one with the same F_1, the one with the same $(F_1 - F_0)$ (in hertz) or the one with an F_1 value intermediate between these two

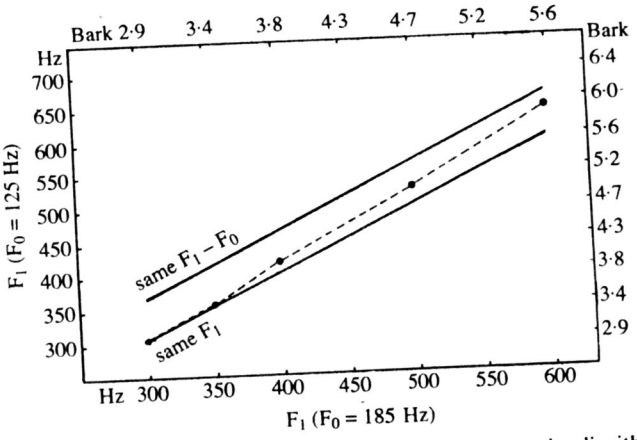


Figure 8. Results of Experiment 2 in the case of comparison stimuli with $F_0 = 185$ Hz. The data points connected with the dashed line are the average F_1 of the comparison stimuli (on the y-axis)identified by the different listeners with that standard stimulus (on the x-axis). The two solid reference lines are for the comparison stimulus with the same F_1 value and the same $(F_1 - F_0)$ values. Axes to the left and bottom show values in Hz; axes to the right and top show Bark equivalents.

values. The figures show reference lines for (i) if no effect of F_0 is observed, and (ii) if listeners based their judgement on equal $F_1 - F_0$.

Figure 8 shows that with low F_0, the F_1 value of the comparison stimulus that best matched the standard corresponded to an exact formant match for the lowest F_1 values (300 Hz and 350 Hz), but increasingly higher values of F_1 for standard stimuli with higher F_1 relative to the standard. Note that when F_1 is high enough (for values higher than 400 Hz, approximately), F_1 is out of the linear part of the Bark scale. Consequently, the $(F_1 - F_0)$ distance expressed in Bark is always lower for comparison stimuli than for standard stimuli when F_1 is in this range, although the difference is small.

Figure 9 shows that the value of the comparison stimulus that best matched the

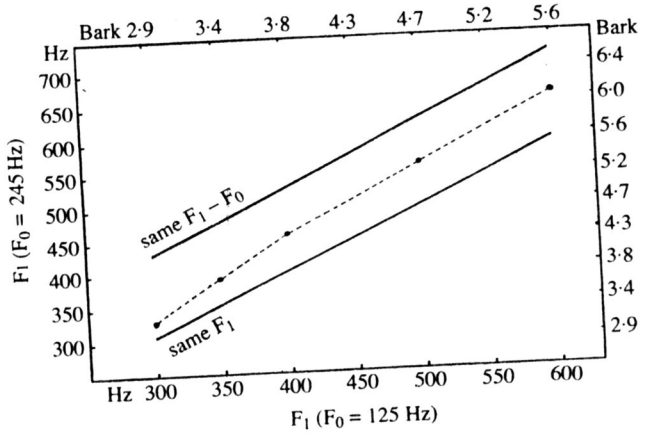


Figure 9. Results of Experiment 2 in the case of comparison stimuli with $F_0 = 245$ Hz. See Fig. 8 for details.

standard in the case of stimuli with $F_0 = 245$ Hz was in general at intermediate values of F_1, between an exact formant match and the value yielding similar $(F_1 - F_0)$ values. Again, however, the F_1 value of the matching comparison stimulus was relatively higher for higher standard F_1 values.

4. Discussion

Results of perceptual experiments showed that the perception of vowel height is related to F_0 as well as to F_1 values. Vowel identification experiments using CVC synthetic stimuli with F_0 of 125 Hz *vs.* 185 Hz did not result in a clear effect on the identification functions. However, in a boundary identification test with similar stimuli, a variation between 125 Hz and 245 Hz did consistently result in different judgements. In a second experiment, one-formant standard stimuli with $F_0 = 125$ Hz and various values of F_1 (300, 350, 400, 500, 600 Hz) were paired with one-formant comparison stimuli in which F_1 could assume three or five different values and F_0 was equal to 185 Hz or 245 Hz. The results of this experiment showed that the value of F_1 for the comparison stimulus most closely identified with the standard was usually intermediate between an exact formant pairing and a pairing yielding similar values of $(F_1 - F_0)$ for comparison and standard stimuli. The pairing was close to the standard's F_1 for low F_1 values and closer to an F_1 value yielding similar $(F_1 - F_0)$ for higher F_1 values. In some cases, in particular when comparison stimuli with $F_0 = 185$ Hz were considered, the pairing reached the same $(F_1 - F_0)$ values (in hertz) for comparison and standard stimuli. In these cases, the $(F_1 - F_0)$ 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 $F_0 = 125$ Hz and $F_0 = 185$ Hz, results of perceptual experiments showed that for low values of F_1, F_0 did not seem to influence the perception of vowel height. Correspondingly, in the acoustic analysis, it was observed that the high vowel [ɪ] area for the male and the female speakers was located at similar values of F_1 (note that the average F_0 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 F_1 was high, a change of F_0 from 125 Hz to 185 Hz influenced the perception of vowel height and that stimuli with different values of F_1 and F_0 but similar $(F_1 - F_0)$ 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 F_1 values in the case of the female speaker, and to similar $(F_1 - F_0)$ values for the female and male speakers.

Thus, there seems to be evidence for a relation between F_1 and F_0 which is not as simple as the $(F_1 - F_0)$ distance proposed by Traunmüller (1981). Traunmüller (1983) observed that the distance between F_1 and F_0 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 F_1 and the lower flank of the configuration, would be independent of F_0 for F_0 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 rule out on the basis of the perceptual experiments presented in this paper that this is a categorical effect dependent on higher level cognitive processes. Nor can it be ruled out that lower level psychoacoustic processes took place in judging the similarity between harmonic simple resonance signals with different F_0. Recent studies by Johnson (1990*a*, *b*) show that perceived speaker identity influences the perceptual normalization, with reference to the /ʊ–u/ vowel continuum, supporting the hypothesis that the information used by the hearers in normalizing vowels is not only directly but also indirectly related to acoustic parameters such as formant values and F_0. If Johnson's observations prove true also in the case of the [i]–[e] continuum, more experiments would be needed to interpret the experiments reported in the present paper.

If a simple auditory effect is assumed, the results of the present study can be interpreted as follows. When F_1 is sufficiently low (as in high vowels) and F_0 also assumes low values (below ~200 Hz) F_1 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 F_1 is high (as in low vowels) and F_0 is sufficiently far from F_1, F_1 may be considered relative to F_0 (not, as previously, to the end of the scale), F_0 being used as an anchor point, and the distance between F_1 and F_0 (in Bark) determines the perception of vowel height. When F_1 is at intermediate values, or the distance between F_1 and F_0 is not large enough, F_1 and F_0 would both intervene in the perceptual process determining vowel height in a relation which would not attribute the same weight to F_1 and F_0. 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 & Kiang (1984), as pointed out by Stevens (personal communication). 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 & Kiang noted that, for all vowels, there was a CF region which was located around F_1 where the harmonics close to F_1 dominated the response spectra. In addition, they observed that this region was flanked on the low-CF side by another region in which the harmonics which were the largest components in the response spectra 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, for high vowels, this region was not distinct. Delgutte & 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" (1984: 872). Therefore, in the case of high vowels, there is a region which corresponds to the F_1 location, which dominates the response spectra, whereas for low vowels, there are two regions which dominate the response spectra, one corresponding to F_1 and the second to either F_0 or to intermediate values between F_1 and F_0. This observation is in keeping with 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 & 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 $F_2 - F_1$ less than 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 F_1 and F_2.

The author thanks Ken Stevens for his continued guidance and support. The helpful comments and suggestions made by Jean-Sylvain Liénard. Maxine Eskenazi and Christophe d'Alessandro on an early draft of this paper 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 & S. Öhman, editors), pp. 143–157. London: Academic Press.

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. (1989*a*) 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. (1989*b*) 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**, 1–19.

Fant, C. G. M., Carlson, R. & Granström, B. (1974) "The [e]–[ø] ambiguity". *Speech Communication Seminar, Stockholm, 1–3 August,* 117–121.

Honda, K. (1983) Relationship between pitch control and vowel articulation. In *Vocal fold physiology: contemporary research and clinical issues* (D. M. Bless & J. H. Abbs, editors), pp. 286–297. San Diego, CA: College-Hill Press.

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.

Johnson, K. (1990*a*) The role of perceived speaker identity in F_0 normalization of vowels, *Journal of the Acoustical Society of America*, **88**(2), 642–654.

Johnson, K. (1990*b*) Contrast and normalization in vowel perception, *Journal of Phonetics*, **18**, 229–254.

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. Speech VAX user's guide* (preliminary version). Cambridge, MA: M.I.T.

Lindblom, B. (1963) On vowel reduction. *Royal Institute of Technology, Stockholm, Speech Transmission Laboratory Report no. 29.*

Lisker, L. (1984) On reconciling monophtongal vowel percepts and continuously varying F patterns, *Haskins Laboratories: Status Report on Speech Research,* SR-79/80, pp. 167–174.

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,* pp. 329–338. Homewood, IL: Irwin.

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. & House, A. S. (1963) Perturbation on vowel articulations by consonantal context: an acoustical study, *Journal of Speech and Hearing Research,* **6**(2), 111–128.

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, *Ph.D. dissertation, University of Stockholm, Sweden.*

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, *Journal of the Acoustical Society of America*, **68**, 1523–1525.