ON VOWEL HEIGHT: ACOUSTIC AND PERCEPTUAL REPRESENTATION BY THE FUNDAMENTAL AND THE FIRST FORMANT FREQUENCY

Maria-Gabriella Di Benedetto (*)

Department of Information and Communication (INFO-COM)
University of Rome 'La Sapienza'- Faculty of Engineering
Via Eudossiana, 18- 00184 Rome, Italy.

ABSTRACT

Acoustic properties of vowels, which can be hypothesized to classify vowels along a dimension of height, are investigated. In particular, vowel representation in the (F1-F0) dimension (F1 and F0 are expressed in Bark) for five vowels of American English is presented and this analysis is compared with the analysis of the same speech materials in the traditional F1 vs F2 space. Results show that individual differences are reduced when the (F1-F0) dimension is used in the case of low vowels while for high and mid vowels the difference in F0 values among speakers is larger than that of F1 values. Perceptual experiments have been carried out using CVC and one-formant synthetic stimuli to examine the influence of F0 on the perception of vowel height. Results are in agreement with the observations on the acoustic analysis and suggest that either F1 and F0 are related in a more complex way than the (F1-F0) Bark-transformed difference or that the Bark scale should be modified at low frequencies.

INTRODUCTION

Traditionally, vowel sounds have been classified along several dimensions: height, backness, tenseness, etc. The formant frequencies of vowels have been widely used as acoustic parameters representative of the different dimensions. For example, it is well known that the first formant frequency (F1) is an acoustic feature related to vowel height and the second formant frequency (F2) to vowel backness.

Syrdal (1985) has introduced the Bark-transformed (F1-F0) distance into a model for the auditory representation of vowels. Syrdal observes that the Bark-transformed (F1-F0) dimension corresponds to a dimension of vowel height. The results of Syrdal's analyses are in agreement with the perceptual results found by Traunmüller (1981). The latter proposes that the prevailing criterion for the perception of vowel height is the distance between F1 and F0 expressed in Bark, when F0 is not between 350 and 400 Hz, approximately.

The present study examines the effectiveness of the (F1-F0) distance to classify vowels according to vowel height. Acoustic analysis of five vowels of American English, in the (F1-F0) vs F2 space (F1 and F0 are expressed in Bark), is presented and compared with the analysis in the F1 vs F2 space. Perceptual experiments which have been carried out, using CVC and one-formant synthetic stimuli, to investigate the influence of F0 in the perception of vowel height are described. The agreement of the results obtained with the findings of the acoustic analysis and their interpretation are then discussed.

ACOUSTIC ANALYSIS

Experimental conditions and procedures

Five vowels of American English [I,ε,æ,*a*,ʌ] are the object of this analysis. In the vowel system of American English, these vowels are characterized by the feature (-round) and by being monophthongal, while the other vowels are all either (+round) or diphthongized. [I,ε,æ] are front vowels and [*a*,ʌ] are back vowels. [I] is (+high), [*a*,æ] are (+low), and [ε,ʌ] are (-high, -low). These vowels are considered in the context of voiced and voiceless stop consonants ([b,d,g,p,t,k]), forming CVC syllables, pronounced in the sentence frame "The __ again". All the combinations between the vowels and the consonants listed are considered, except the non-symmetrical contexts with respect of voicing. In addition, hVd and #Vd syllables are analyzed. Three native speakers of American English, one female and two males, uttered the speech materials. They were asked to pronounce the sentences carefully and clearly. If a mistake occurs, the sentence is repeated. The sentences are pronounced in a random order. The set of syllables is pronounced three times. Thus, three versions of each vowel in each consonantal context are available. The speech materials are recorded in a sound-treated room using high quality equipment. The distance between the microphone and the speaker's mouth is about 20 cm. The recorded materials are then evaluated by a phonetically sophisticated listener. The speech signal is then stored on the MIT-Speech VAX-750. For this purpose, it is low-pass filtered at 4.8 kHz and sampled at 10 kHz.

The speech materials are analyzed using a software program KLSPEC developed by Dennis Klatt (1984). This program computes a 512-point DFT transform of slices of the signal (predifferenced and premultiplied by a Hamming window). The duration of the Hamming window is 30 ms at the sampling rate considered. In addition, fundamental frequency (F0) is determined by collecting frequencies of local maxima occuring below 3000 Hz and judging it to be that frequency (F0) which accounts for most peaks as harmonics. The program KLSPEC also calculates a spectrogram-like spectrum which is obtained by windowing a slice of signal (256 samples) and computing a 256-point DFT. A weighted sum of adjacent DFT sample energy is then computed for each of 128 spectrogram-like filters. Local maxima in this spectrum are most often indicative of the frequency positions of the formants. An interpolation algorithm improves the accuracy over the 40 Hz resolution implied by a 128-sample spectrum over 5 kHz. The spectrogram-like spectrum has been used for the estimation of the formant frequencies of the vowels under analysis. In some cases, in which this algorithm is not successful, the formant frequencies are manually extracted. DFT spectrum slices sampled every 5 ms are plotted and the frequency positions of the formants are evaluated by visual examination of the evolution of the locations of the DFT spectrum peaks in time. The temporal sampling point of F1, F2 and F0 is the time at which F1 reaches its maximum, as discussed in Di Benedetto (1987). The values of F0 and F1 are converted into a critical band tonality scale, according to Zwicker and Terhardt's (1980) mathematical approximation as adopted by Syrdal (1985).

Results of acoustic measurements

As expected, the highest F0 is found for the female speaker (CR) (191 Hz), while F0 for the two male speakers (JP) and (KS) is comparable (118 and 127 Hz, respectively).

The results of the analysis of the vowels [I,ε,æ,*a*,ʌ] for the three speakers considered in the (F1-F0) vs F2 space and in the F1 vs F2 space are extensively described in Di Benedetto (1987). In the present paper, results for only one of the speakers ((KS)) and one of the versions are presented as shown in Fig.1. Figure 1a shows that overlapping occurs in the (F1-F0) dimension only between [*a*] and [ʌ]. In the F1 vs F2 space (Fig.1b) overlapping occurs between [I] and [ε], [ε] and [æ], and [*a*] and [ʌ] while in the (F1- F0) vs F2 space, the [I], [ε] and [æ] areas are well separated. The use of the (F1-F0) dimension seems to improve the distinction between different vowels contiguous along the (F1-F0) dimension, for (KS). The results cbtained for the other versions and speakers (Di Benedetto, 1987) show that similarly an improvement is obtained, in terms of better grouping and separation of the vowel areas in the (F1-F0) dimension, compared to what was obtained in the F1 dimension. However, problems of overlapping still occur between vowel areas of a single speaker in the (F1-F0) dimension. One should note that the differences in (F1-F0) values between vowels in voiced and voiceless consonantal contexts are lower than in F1 values (Di Benedetto, 1987). Consequently, one of the factors which contributes to a better separation of the vowel areas is that in the (F1-F0) dimension the vowel areas are better grouped.

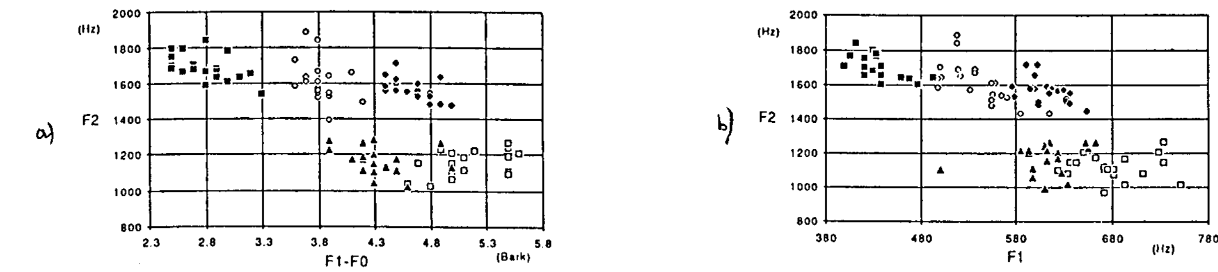

Figure 1: Results of the analysis in the a) (F1-F0) vs F2, and b) in the F1vs F2 spaces of the vowels [I,ε,æ,*a*,ʌ] (speaker (KS)). Each vowel is considered in 20 different consonantal contexts.

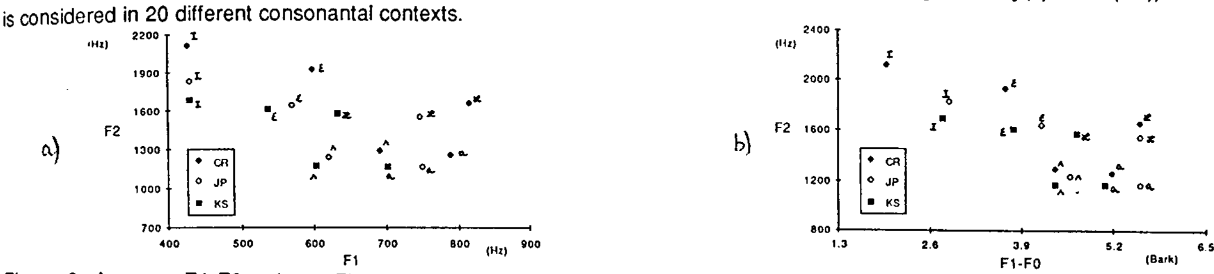

Figure 2: Average F1-F2 values (Fig.2a) and (F1-F0)-F2 values (Fig 2b) of the vowels [I,ε,æ,*a*,ʌ] for speakers (CR), (JP) and (KS). The averaged values are obtained by pooling all the consonantal contexts and versions.

The results of the comparison the vowel areas of the three speakers are summarized in Fig.2. Figure 2a (2b) shows for speaker (CR) (full losangues), (JP) (open losangues) and (KS) (full squares) the F1 and F2 values ((F1-F0) and F2 values) for each vowel, averaged over all the consonantal contexts and versions. The comparison of Fig.2a and Fig.2b shows that the difference in the representation of vowels for different speakers is reduced using the (F1-F0) parameter for the low and front vowel [æ] and the two back vowels [*a*,ʌ]. For the mid vowel [ε], in the (F1-F0) dimension the [ε]-area of the female speaker (CR) is shifted to lower values than those characterizing the [ε]-area of (KS) and (JP) and this effect is even more accentuated in the case of the vowel [I].

A comparison of these results with the analysis of American English vowels by Peterson (1961) has been carried out. It is noticed on Peterson's data that the difference in F1 values between male and female speakers, depends upon the range of F1 values. In particular, it is observed that this difference for high vowels is much smaller than for non-high vowels, and this difference increases when F1 increases. This result confirms what is observed in the present study.

Syrdal (1985) reports the Bark-difference means for ten vowels of American English on the Texas Instruments data base which consists of vowels in hVC words pronounced in isolation by 52 men, 51 women and 51 children. The data reported by Syrdal confirm that the difference in (F1-F0) values between male and female speakers depends on the height of the vowel considered. Note that both Peterson's and Syrdal's results are based on vowels pronounced in hVd or hVC words while the vowels of the present study are considered in several consonantal contexts.

In conclusion, acoustic analysis of the five vowels [I,ε,æ,*a*,ʌ] has shown that, in the dimension representing vowel height, individual differences for low vowels are reduced when the vowels are represented by the difference (F1-F0) rather than by F1. For high and mid vowels, on the other hand, a smaller shift in the F1 dimension would be needed to correct the differences in F0.

PERCEPTUAL EXPERIMENTS

All the stimuli used in the experiment described were synthesized with the Klatt synthesizer (1980, 1984).

Experiment1

Description. The aim of this experiment is to investigate the influence of F0 on the perception of vowel height, using dVd synthetic syllables. One set of stimuli is characterized by F0=125 Hz (125-stimuli) while the two other sets of stimuli consists of stimuli which are identical to the previous ones as regards F1 and higher formant, while F0 of the stimuli of this experiment is increased in two steps: 60 Hz (185-stimuli) and 120 Hz (245-stimuli). Each set consists of 10 stimuli characterized by different values of F1 maximum, ranging from 300 Hz (stimulus #1) to 500 Hz (stimulus # 10) in steps of 30 Hz. Experiment 1 consists of two phases: a vowel identification test and a "boundary" identification test.

In the first phase the test was carried out on four american subjects. The subjects were all non-naive listeners, native speakers of American English and members of the Speech Communication Group at the Massachusetts Institute of Technology. They all name English as their best language.The stimuli used are the 125- and185-stimuli. The subjects were asked to identify the vowel of the synthetic utterances as [*i*,*e*] as justified by the results of a previous experiment (Di Benedetto, 1987).

In the second phase, 125-stimuli, 185-stimuli and 245-stimuli were used. Sequences of stimuli (and the same sequences in reverse order) characterized by the same F0 were played to the subjects who were asked to declare when their perception of the synthetic vowels presented changed from [*i*] to [*e*] or viceversa. Each sequence, in each order, was presented three times. Three subjects participated in this test.The subjects' description is identical to that of the subjects who participated in the vowel identification experiment

Results. Results of the identification test are presented for each subject, separately, in Fig.3 which shows that a change of 60 Hz in F0 does not result in a clear effect on the identification functions for any of the subjects who participated in the test. The three subjects who participated in the "boundary" identification test reported that they perceived the vowels of the synthetic utterances as either [*i*] or [*e*], as they were instructed. Figure 4 shows the results for each of the three subjects and indicates the first stimulus which is perceived as [*e*], when the sequences presented are ordered with ascending stimuli number, or the last stimulus which is perceived as [*e*], in the case of sequences ordered according to a descending stimulus number progression. Figure 4 shows that, in the case of the three subjects who participated in this test, an increase in F0 from 125 to 185 Hz does not result in a change of the perceptual boundary between [*i*] and [*e*], while a variation in F0 from 125 to 245 Hz does result in a consistent shift in this boundary. No difference was observed in the results obtained with sequences of stimuli with F1 increasing or in reverse order.

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

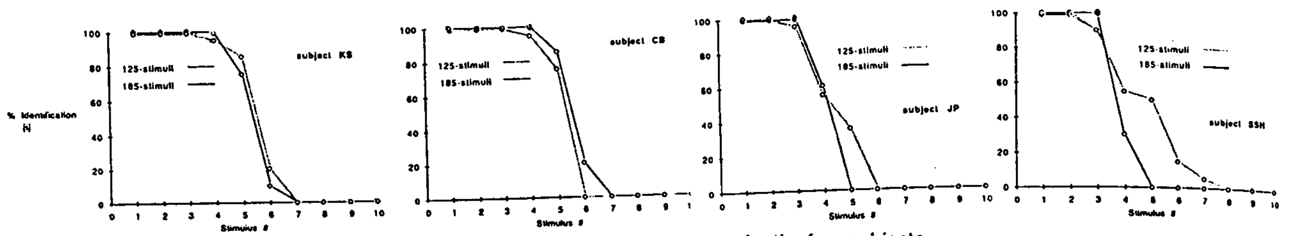

Figure 3: Results of the identification test for the four subjects.

Experiment 2

Description. The aim of this experiment is to investigate the influence of F0 in the perception of vowel height, using one-formant stimuli. Various one-formant stimuli with F0=125 Hz, 185 Hz or 245 Hz were generated. The one-formant stimuli with F0=125 Hz were characterized by five values of the formant (F1) (300, 350, 400, 500, 600 Hz). Each of these stimuli was matched against 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 for comparison and standard stimuli. Each pair was played three times. The same procedure was repeated with the same standard stimuli (F0=125 Hz) but the comparison stimuli were characterized by F0=245 Hz.Seven subjects participated in this experiment.They were non-naive listeners, native speakers of American English, and members of the Speech Communication Group at the Massachusetts Institute of Technology. They all named English as their best language. They were asked to indicate which pair of stimuli was most similar in terms of vowel height.

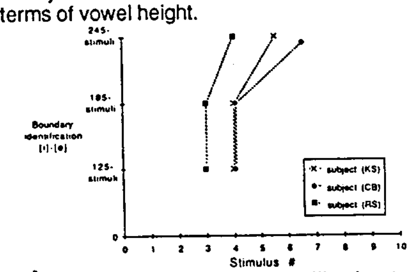

Figure 4: Results of the boundary identification test for the three subjects. Each dot on the figure (of different shape for each subject) indicates the stimulus at which the identification changes from [i] to [e], in the case of the three stimuli F0 types.

Results. Figures 5 and 6 show the results obtained in experiment 2. Figure 5 shows on the abscissa the standard stimuli (with F0=125 Hz) identified by the F1 maximum value, and on the ordinate the comparison stimuli (with F0=185 Hz) which are matched against the standard stimuli. As shown on Fig.5 each standard stimulus is matched against three comparison stimuli: one with the same F1, one with the same (F1-F0) (in Hertz) and one with a F1 value intermediate between the same F1 and the same (F1-F0). For each standard stimulus, Fig.5 shows the value of F1 for best match in the case of each subject individually (1° column: subject (MA), 2° column: subject (TC), etc, as shown on the figure). A full (open) symbol indicates that the corresponding comparison stimulus was never (always) chosen as stimulus for best match by the subject. Partially open symbols indicate the percentage of times that this particular stimulus was chosen for best match. Figure 6 is similar to Fig.5 but indicates the results of the test in the case of the comparison stimuli with F0=245 Hz. In this case, each standard stimulus can be matched against five comparison stimuli: one with the same F1, one with the same (F1-F0) and three with intermediate values of F1, between the same F1 and the same (F1-F0). As in Fig.5, the value of F1 for best match is indicated by partial or complete blanking of the corresponding symbol, for each subject. Figure 5 shows that the F1 value for best match, in the case of stimuli with F0=185 Hz, corresponds to an exact formant match for low F1 values (300 and 350 Hz). For other values of F1 the match is in general between an exact formant match and values of F1 leading to similar (F1-F0) values. Note that in the case of the highest F1 value for the standard stimuli (600 Hz) the match is similar to (F1-F0) for subjects (TC) and (CH) and is close to this value for (KS). One should note that when F1 is high enough (for values higher than 400 Hz, approximately) F1 is out of 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.

Figure 6 shows that the value for best match, in the case of stimuli with F0=245 Hz is 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), the match is almost in all cases with stimuli characterized by F1=330 Hz corresponding to the first intermediate step. For values of F1 in the middle range (350, 400 and 500 Hz) the match shifts to stimuli with intermediate F1 values higher than in the case of standard stimuli with F1=300 Hz, with increasing F1 of the standard stimuli. Note, in fact, that for standard stimuli with F1=350 Hz the match is in general against comparison stimuli with F1=410 Hz and that for standard stimuli with F1=400 Hz or F1=500 Hz, the match is in general against comparison stimuli with F1=460~490 Hz and F1=560~590 Hz, respectively. The case of standard stimuli with F1=600 Hz is similar to the case of F1=400 Hz and F1=500 Hz, but note that for one subject (CH) the match is partially against stimuli with F1 values (720 Hz) leading to similar (F1-F0) values for comparison and standard stimuli.

DISCUSSION

Results of the perceptual experiments have shown that the influence of F0 in the perception of vowel height is related to F0 and F1-values. In particular, vowel identification experiments using CVC synthetic stimuli, have shown that an increase in F0 from 125 to 185 Hz does not result in a clear effect on the identification functions, while a variation from 125 to 245 Hz does result in consistently different judgements. A second experiment has been described, in which one-formant stimuli with F0=125 Hz and various values of F1 (300, 350, 400, 500, 600 Hz) were matched against one-formant stimuli in which F1 was adjustable and F0 equal to 185 or 245 Hz. Results show that the value of F1 for best match was usually between an exact formant match and a match yielding similar values of (F1-F0) for comparison and standard stimuli. The match was close to F1 for low F1 values and approached in general similar (F1-F0) values for higher F1. In some cases, in particular when comparison stimuli with F0=185 Hz were considered, the match reached the same (F1-F0) values (in Hertz) for comparison and standard stimuli. It has been noticed that in these cases, the (F1-F0) values expressed in Bark are lower for comparison stimuli than for standard stimuli.

The results of the perceptual experiments presented are in agreement with the observations on the acoustic analysis. The results of the acoustic analysis have shown 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. For high vowels, the shift in the F1 dimension to account for differences in (F1-F0) increases the acoustic variability of the same vowel among speakers. For mid vowels, an intermediate effect is observed. In these cases (high and mid vowels), it has been observed that a smaller shift in the F1 dimension would be needed to correct the differences in F0. In the perceptual experiments, stimuli with three different values of F0 (125, 185 and 245 Hz) have been used. The average F0 value of the female speaker considered in the acoustic analysis is ~ 190 Hz and of the male speakers ~120-130 Hz, as previously mentioned. The results of the perceptual experiments for F0=125 Hz and F0=185 Hz, have shown that for low values of F1, F0 does not seem to influence the perception of vowel height. Correspondingly, one should note that it has been observed that the vowel area of the high vowel [I] for the male and the female speakers is located at similar values of F1. Experiment 2 has shown that when F1 is high, a change of F0 from 125 to 185 Hz influences the perception of vowel height and that stimuli with different values of F1 and F0 but similar (F1-F0) values are perceived as similar in terms of

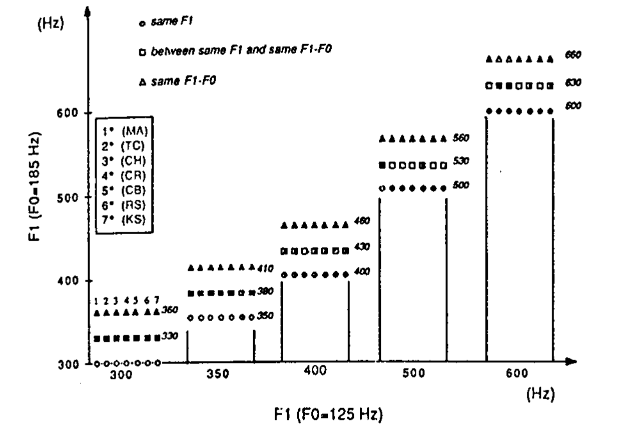

Figure 5: Results of experiment 2 for comparison stimuli with F0=185 Hz.

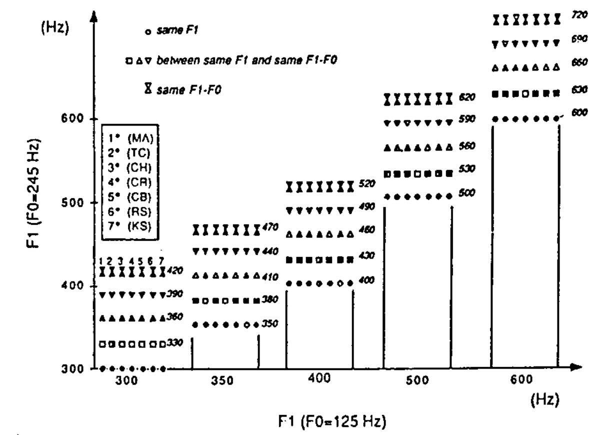

Figure 6: Results of experiment 2 for comparison stimuli with F0=245 Hz.

vowel height. Correspondingly, the acoustic analysis has indicated that the location of the [æ]-area corresponds to higher F1 values in the case of the female speaker, and to similar (F1-F0) values for the female and male speakers.

The interpretation of the results obtained can be given as follows. When F1 is sufficiently low (as in high vowels) and F0 assumes also 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 ananchor 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) is determinant in 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 have 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 F0=125 Hz. Delgutte and Kiang note that for all vowels, there is a CF region which is located around F1 (F1 region) where the harmonics close to F1 dominate the response spectra. In addition, they observe that this region is flanked on the low-CF by another region in which the harmonics close to CF are the largest components in the response spectra. These harmonics correspond to the fundamental frequency or to intermediate values between F1 and F0. For low vowels, this region extends up to about 400 Hz while on the contrary, for high vowels, this region is not distinct. Delgutte and Kiang observe that "...the open-close dimension of phonetics correlates with both the position of the F1 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 for low F0 values, F1 determines the perception of vowel height when F1 is low (high vowels), whereas if F1 is high (low vowels) F0 influences vowel height perception. Unfortunately, Delgutte and Kiang do not present results in the case of higher values of F0. Consequently, the results of the present study in the case of higher values of F0 cannot be interpreted on the same basis. We want to point out that the perception of vowels with F1 and F2 closer than 3.5 Bark could be based on one equivalent formant located in a position intermediate between the two formants, according to the categorical perceptual effect SCG (Spectral Center of Gravity) found by Chistovich et al. (1979). It could be hypothesized then that this one formant is relevant, in the cases of vowels with F2-F1 < 3.5 Bark, to vowel height perception.

This aspect of the problem is not addressed in the present study. We want to suggest that our interpretation of the relation between F1 and F0 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.

REFERENCES

Chistovich, L.A., Sheikin, R.L. & Lublinskaya, V.V. (1979) "Centres of gravity and spectral peaks as the determinants of vowel quality", in: B.Lindblom and S.Öhman, eds., *Frontiers of Speech Communication Research* , Academic Press, London, 143-157.

Delgutte, B. & Kiang, N.Y.S. (1984) "Speech coding in the auditory nerve: I. vowel-like sounds", *J. Acoust. Soc. Am.* 75 no.3, 866-878.

Di Benedetto, M.G. (1987) "An acoustical and perceptual study on vowel height", Ph.D. thesis, University of Rome, Italy.

Fant, C.G.M. (1975) "Non-uniform vowel normalization", *STL-QPSR* 2-3.

Klatt, D.H. (1980) "Software for cascade/parallel formant synthesizer", *J. Acoust. Soc. Am.* 67 no.3, 971-995.

Klatt, D.H. (1984) "M.I.T. SpeechVAX user's guide", preliminary version.

Peterson, G.E. (1961) "Parameters of vowel quality", *J. Speech and Hearing Res.* 4, 10-29.

Stevens, K.N. (1985) Personal communication.

Syrdal, A.K. (1985) "Aspects of a model of the auditory representation of American English vowels", *Speech Communication* 4, 121-135.

Traunmüller, H. (1981) "Perceptual dimension of openness in vowels", *J. Acoust. Soc. Am.* 69 no.5, 1465-1475.

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.