in Recent Advances and Applications of Speech Recognition

International Workshop, Rome (Italy), 1986.

TIME-VARYING PROPERTIES OF THE FIRST FORMANT FREQUENCY: AN ACOUSTICAL AND PERCEPTUAL STUDY.

Maria-Gabriella Di Benedetto

Department of Information and Communication (INFOCOM) Faculty of Engineering, Rome, Italy and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139,USA.

Keywords.first formant trajectory,time-varying properties of F1,vowel perception.

Summary. Acoustic analysis of the vocalic portion of CVC nonsense syllables spoken by three speakers shows that ambiguities between vowels, for each speaker, occur, primarily in the F1 dimension, if the vowels are represented by the values of F1 and F2 sampled at the time where F1 reaches its maximum. The examination of the F1 trajectories shows in particular that if two vowels such as /I/ and /ɛ/ have the same maximum F1, then F1 for /ɛ/ reaches its maximum earlier. Perceptual experiments have been carried out to examine the perceptual importance of the F1 trajectory shape, using synthetic CVC syllables. Results are in agreement with the hypothesis that stimuli for which F1 reaches its maximum earlier are perceived as lower vowels. Results are similar for subjects of different languages, leading to a suggestion that this phenomenon can be explained on an auditory basis.

I. INTRODUCTION

A few studies have been carried out to examine the perceptual importance of time-varying properties of F1. Stevens [1] has examined the interaction of formant frequency and duration and has proposed that time-average values of formant frequencies could determine the perception of a vowel. Huang ([2] and [3]), has examined whether the acoustic properties, which have been found in previous studies to be related to the tense/lax distinction in American-English, have an effect in the perception of tense-lax vowel pairs. In particular, the interaction between formant trajectory and formant frequency is investigated. Huang' s findings are that vowels are not fully characterized by their steady-state formant positions and that there is evidence for the existence of perceptual averaging in F1 in non-high vowels. Di Benedetto [4] has reported preliminary results of an investigation on the time-varying properties of F1, in particular on the perceptual importance of the F1 trajectory shape.

In section II, a traditional acoustic analysis of five american-english vowels, /I,ɛ,ae,ɑ,ʌ/, is presented. Problems that arise using a classical vowel representation are pointed out. In addition, properties of the F1 trajectory, in particular where F1 reaches its maximum value, are observed. It is noticed that if confusion between two vowels such as /I/ and /ɛ/ occurs, F1 reaches its maximum earlier for the vowel /ɛ/ than for the vowel /I/, that is for the lower vowel.

An investigation on the perceptual importance of the F1 trajectory shape, described in section III, is then conducted. Identification tests in which the stimuli are dVd synthetic syllables are carried out. Those stimuli differ only in the F1 trajectory shape, which can be of two types: one characterized by having the F1 maximum at the beginning of the vowel (type I) and the other at the end of the vowel (type II). Results of the tests are given and are similar for listeners native of three different languages. The difference in the identification functions obtained for type I and type II stimuli show that type I stimuli are more associated to vowels characterized by a higher F1 than type II stimuli. These results are in agreement with the observations on the F1 trajectory shape

described in section II. Interpretations and conclusions are given in section IV.

II. ACOUSTIC ANALYSIS

Speech materials. Five vowels of american english /I,ɛ,ae,ɑ,ʌ/, are the object of this analysis. They are considered in the context of stop consonants voiced and voiceless /b,d,g,p,t,k/, forming nonsense CVC syllables, pronounced in the sentence frame "The ___ again". All the possible combinations between the five vowels and the six consonants listed above are considered, with the exclusion of non-symmetrical contexts with respect of voicing. In addition, hVd and Vd syllables are analyzed. Three version of each syllable are available.

Speakers and recording conditions. Three speakers native of American-English, one female (CR) and two males ((KS) and (JP)), utter the speech materials. The speakers are asked to pronounce the sentences carefully and clearly.

The speech materials are recorded in a sound treated room on a Nakamichi Model LX-5 cassette deck using a Crown model D-75 amplifier. The microphone used is the dynamic Altec Model 684A. The speech signal is then stored on the MIT-Speech VAX-750. For this purpose, it is first low-pass filtered at 4.8 kHz and then sampled at 10 kHz. The low-pass filter used is a TTE Model J97E-kOhm passive low-pass anti-aliasing filter. The a/d convertion is obtained by means of a AD-11k 8-channel (differential) 12-bit plus-or-minus 5 volt a/d converter.

Measurement procedures. The speech materials are analyzed by means of the program KLSPEC developed by Dennis Klatt on the Speech-VAX [5]. This program allows visualization of the waveform as well as of 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, a spectrogram-like spectrum is available, which is obtained by windowing a slice of signal (256 samples) and computing a 256-point DFT. A weighted sum of adjacent DFT sample energies is then computed for each of 128 spectrogram-like filters.

The use of the spectrogram-like spectrum has been found to be most useful for the estimation of the formant frequencies of the vowels under analysis. In fact, the location of the maxima of the spectrogram-like spectrum gives a very good indication 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. In some cases, in which formant tracking results in being particularly doubtful and this algorithm is not successful, DFT spectrum slices each 5 ms are plotted and the frequency positions of the formants are evaluated by visual examination of the evolution of the DFT spectrum peaks locations in time.

The choice of the spectrogram-like spectrum for the estimation of formant frequency positions is extensively justified in [6].

In this analysis vowels are represented in a traditional way, by a point in the F1 vs F2 plane. The choice of the instant of time where to sample the values of F1 and F2 is of interest. In the present study, it has been chosen to sample F1 and F2 at the time where F1 reaches its maximum. This choice is extensively motivated in [6].

Results. The results of this analysis are presented for each speaker individually and for one of the versions of the vowels. Results corresponding to the other versions and detailed discussion can be found in [6].

Fig. 1 shows the results of the analysis for (KS). This diagram shows that overlapping occurs in the F1 dimension. Observe the /I-ɛ/ and /ɛ-ae/ boundary regions and the non-neglectable overlap between /ɑ/ and /ʌ/. Notice that no overlapping occurs in the F2 dimension.

The results of the analysis for (JP) are presented in Fig. 2. From this diagram, one can see that problems occur especially at the boundaries between /a/ and /ʌ/, /ɛ/ and /ae/. Note that /ɛ/ in the /g-g/ context has a very low F1. No overlapping occurs between the front and back vowels under study.

The results of the analysis for (CR) are presented in Fig. 3 which

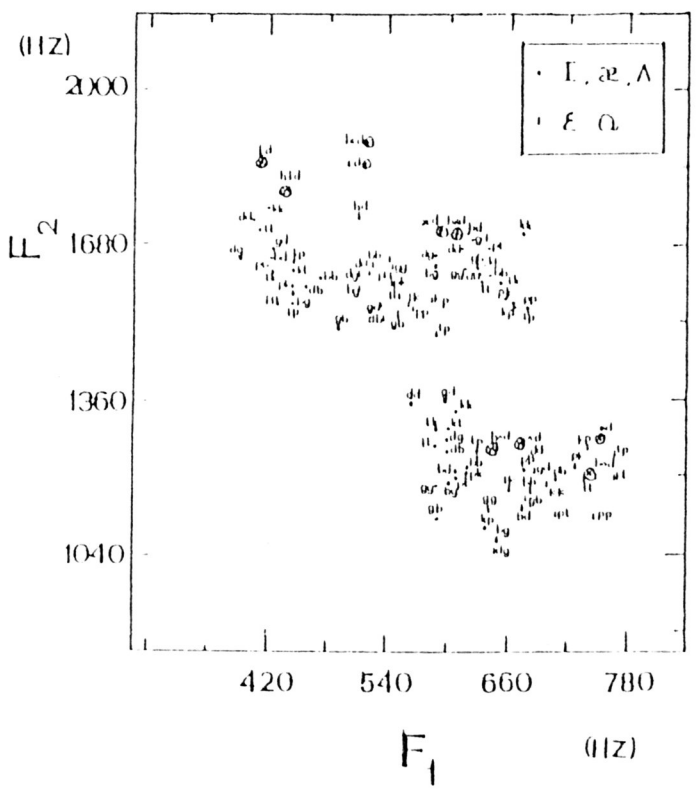

Fig.1 Analysis in the F1 vs F2 plane of the vowels /I,ɛ,ae,ɑ,ʌ/ (speaker (KS)). Each vowel is considered in 20 different consonantal contexts.

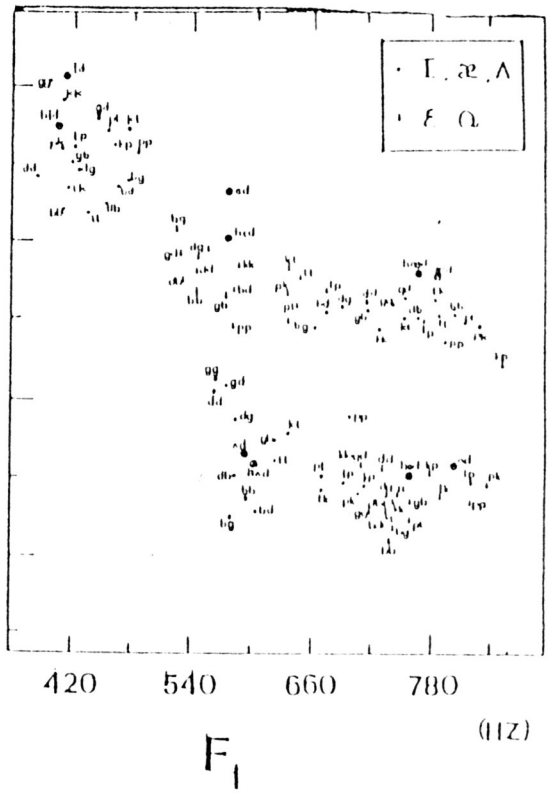

Fig.2 Analysis in the F1 vs F2 plane of the vowels /I,ɛ,ae,ɑ,ʌ/ (speaker (JP)). Each vowel is considered in 20 different consonantal contexts.

that, in this case, problems occur only in the F1 dimension between /ɑ/ and /ʌ/. The three front vowels, /I,ɛ,ae/, do not overlap. No problems of overlapping occur in the F2 dimension.

The results of the present study show that overlapping occurs even though measurements on vowels pronounced by a single speaker are considered. In addition, it is found that for all speakers confusion occurs only in the F1 dimension. The back vowels /ɑ,ʌ/ are in all cases well distinct from the front vowels /I,e,ae/. Confusion between the two back vowels /ɑ/ and /ʌ/ is present for all speakers. Overlapping occurs between /ɛ/ and /ae/ for (JP) and (KS), and problems at the boundary between /I/ and /ɛ/ occur for (KS).

F1 trajectories. Particular attention is given to the confusion area between /ɛ/ and /I/ that has been found in the acoustic analysis of the vowels pronounced by speaker (KS). Fig. 4 shows the F1 trajectories of the vowels for which confusion occurs. As one can see, F1 reaches its maximum earlier in the lower vowel (/ɛ/).

III. PERCEPTUAL EXPERIMENTS

Stimuli in experiment 1. All the stimuli considered were synthesized with the Klatt synthesizer. This cascade/parallel formant synthesizer is described by Klatt [7]. The stimuli consist of dVd synthetic syllables. The duration of the stimuli is 115 ms for all stimuli. The F1 trajectory is the only parameter in which those stimuli differ. This trajectory can have two shapes (stimuli of type I or of type II). The duration of the steady-state is 15 ms for all stimuli. In type I (type II) stimuli, the duration of the onglide is 15 (85) ms and the duration of the offglide is 85 (15) ms.

The higher formant trajectories and the fundamental frequency are identical for both stimuli types and symmetrical around the center of the vowels. Ten stimuli of each type are used, characterized by the F1 maximum value (330, 350, 370, 390, 410, 430, 450, 470, 490, 500 Hz).

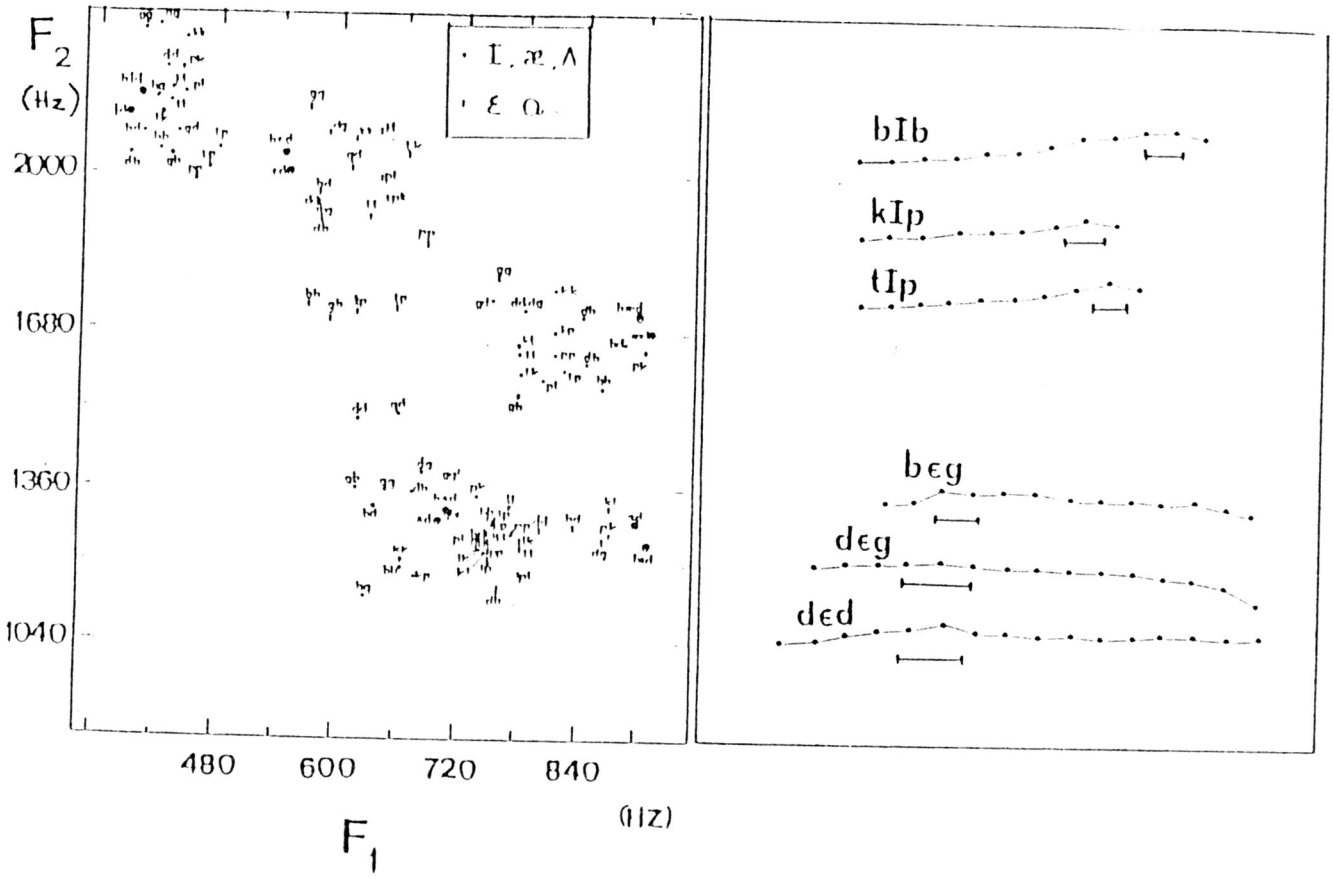

Fig.3 Analysis in the F1 vs F2 plane of the vowels /I,ɛ,ae,ɑ,ʌ/ (speaker (CR)). Each vowel is considered in 20 different consonantal contexts.

Fig.4 F1 trajectories of the vowels /I/ and /ɛ/ for speaker (KS).

Notice that two stimuli, one of each type, having the same F1 maximum would be represented by the same values of F1 and F2 if those were sampled at the time where F1 reaches its maximum (or at the middle of the vowel). This would also be the case if the representation used was more generally in the F1 vs F2 vs Fx space or if a time-average of F1 was considered.

Stimuli in experiment 2. The stimuli of experiment 2 differ from those used in experiment 1 by their duration, which is equal to 75 ms. The duration of the steady-state is 15 ms. For type I (type II) stimuli the duration of the onglide is 15 (45) ms and the duration of the offglide is 45 (15) ms. The higher formant trajectories and the fundamental frequency are identical for both stimuli types and symmetrical around the center of the vowels. Ten stimuli of each type are used, characterized by the same F1 maximum values as stimuli of experiment 1.

Subjects in experiment 1 and experiment 2. Seven subjects participate to these experiments. Four subjects are native speakers of American-English and non-naive listeners. They belong to the Speech Group of M.I.T. None of them has profound knowledge of other languages. They all live in the Cambridge (MA) area. Two subjects are native speakers of Italian and naive listeners. One of those subjects has good knowledge of French but names Italian as his first language. The other subject has no knowledge of any other language. One subject is native of Japanese and non-naive listener. At the time of the tests, he had been living in the Cambridge area for few weeks. He had poor knowledge of American-English and named Japanese as his first language.

Experiment procedures. Experiment 1 and 2 consist of three parts. In the first part, the ten type I stimuli are presented 10 times. Their order is such that each

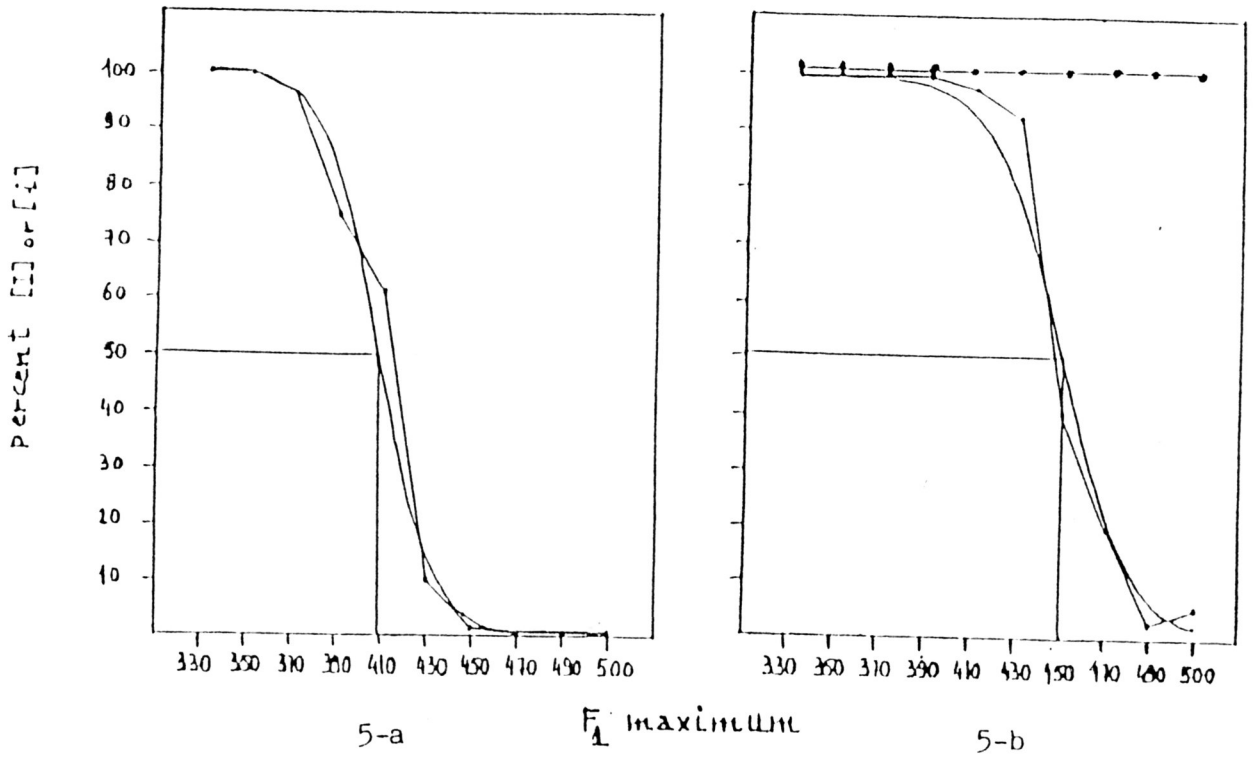

Fig.5 Experiment 1- identification curves for type I (5-a) and type II (5-a) stimuli- american subjects

stimulus follows once each stimulus. By this way, one expects to obtain a judgment on each stimulus independently of the immediately preceding stimulus. The second part is similar to the first, considering type II stimuli. In the third part both stimuli types are presented. The stimuli are grouped in two sets of ten stimuli (five of each type). In each group, the stimuli order is such that each stimulus follows once each stimulus. In each test, each stimulus is presented then 20 times. Each test is approximately 45 minutes long.

A preliminary experiment to experiments 1 and 2, in which only three repetitions of each stimulus were presented, has shown that, when listened by american, italian and french subjects, the vowels of the stimuli are associated to four different vowels, /i,I,e,ɛ/. In particular, italian and french listeners associated them to the /i,e,ɛ/, as /I/ does not belong to the vowel system of those languages.

In experiment 1 and 2, american subjects are asked to identify the vowel of the stimuli as /i/, /I/, /e/, or /ɛ/, italian subjects as /i/, /e/ or /ɛ/, and japanese subjects as /i/ or /ɛ/ (note in fact that the vowels /I/ and /e/ do not belong to the vowel system of Japanese).

Representation of the results. The diagrams representing the results of experiments 1 and 2 show in abscissa the stimuli identified by the F1 maximum value and in ordinate the percent of identification of "i-like" sounds (/i/ and /I/) vs "e-like" sounds (/e/ and /ɛ/). On the same diagram a logistic curve, fitting the data, is shown, and the 50% cross-over point is indicated. This point represents the stimulus number at which the identification changes. The logistic curve is fitted as shown by Neter and Wassermann [8].

It is important to note that the identification curves obtained have in all cases a very regular shape and never cross more than once the 50% line. In most cases the crossovers found on the logistic and on the identification curves coincide. However, the use of the logistic curve is a tool for estimating the crossover points in all cases consistently.

Results of experiment 1. Fig. 5-a and Fig. 5-b represent the identification functions obtained for type I and type II stimuli respectively, in the case of american subjects. Fig. 5-b shows that two of the subjects never classified type

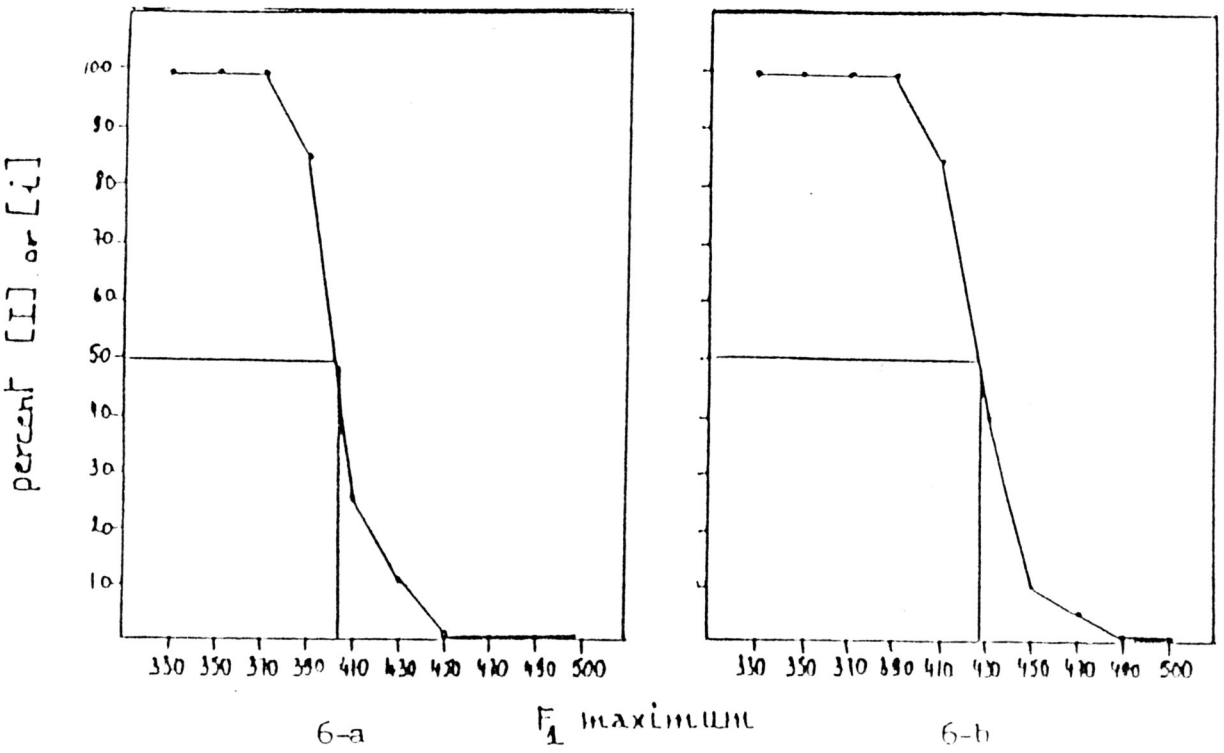

Fig.6 Experiment 1- identification curves for type I (6-a) and type II (6-b) stimuli- italian subjects

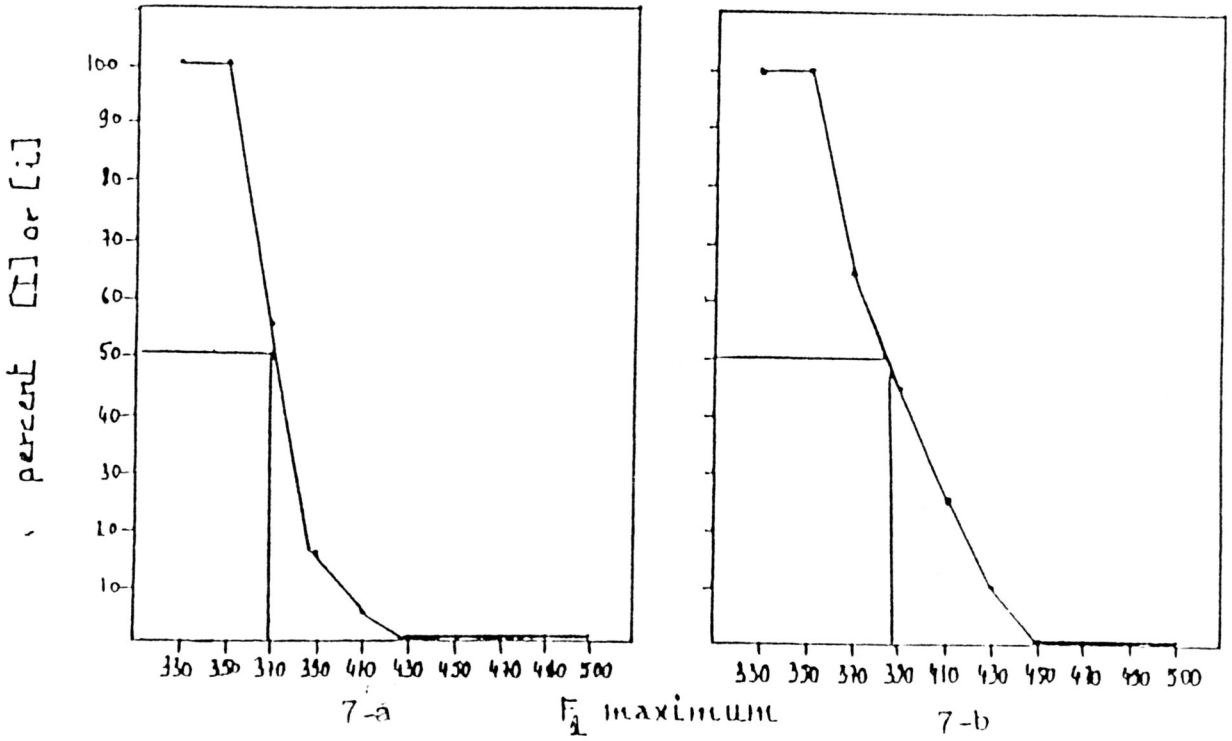

Fig.7 Experiment 1- identification curves for type I (7-a) and type II (7-b) stimuli- japanese subject

II stimuli as "ɛ-like" sounds. In addition, Fig. 5-a and Fig. 5-b show that for type I stimuli a lower crossover than for type II stimuli is found (~ 40 Hz). Type I stimuli are therefore considered more as "ɛ-like" sounds than type II stimuli.

Very similar results are obtained for the subjects of the two other languages considered. Fig. 6 shows the results obtained with italian subjects. One can see that there is a difference in the crossovers for type I and type II stimuli (~ 25 Hz), the higher crossover corresponding to type II stimuli. Fig. 7 shows the results obtained with the japanese subject. Note from Fig. 7 the difference in the crossovers between type I and type II stimuli (~20 Hz).

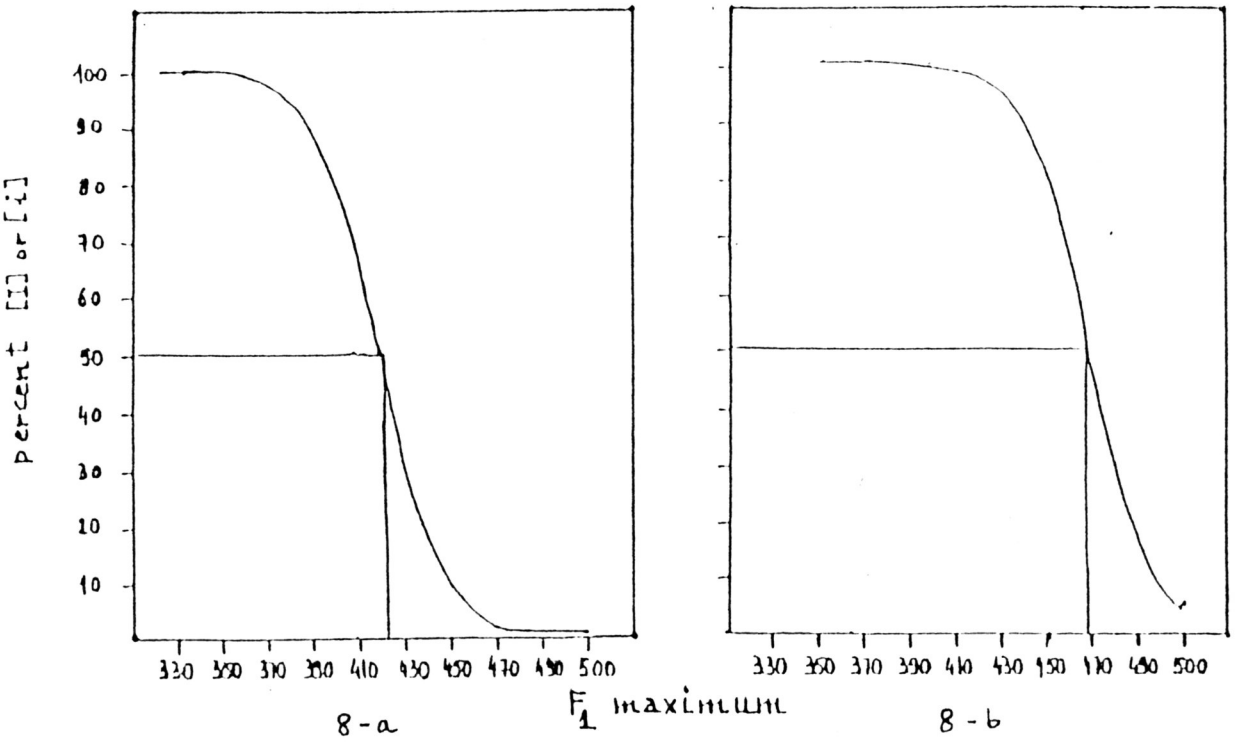

Fig.8 Experiment 2- logistic curves for type I (8-a) and type II (8-b) stimuli- american subjects

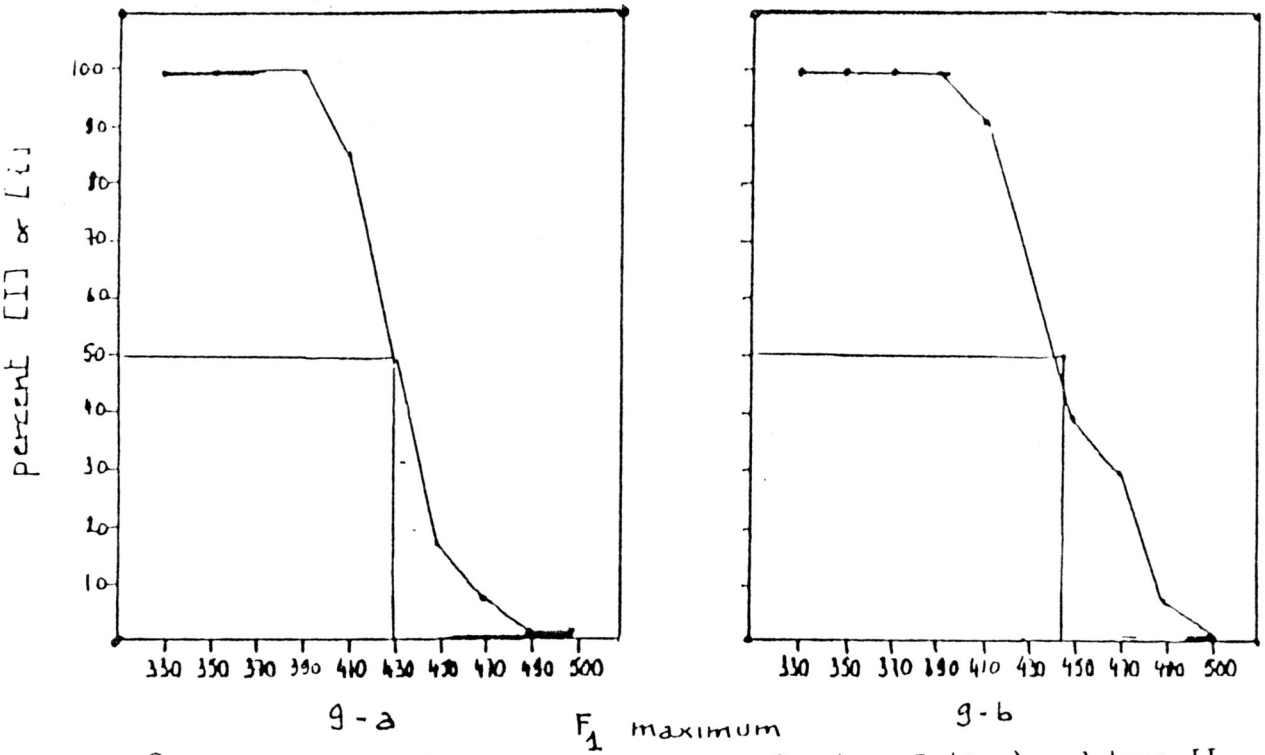

Fig.9 Experiment 2 - identification curves for type I (9-a) and type II (9-b) stimuli- italian subjects

Results of experiment 2. Fig. 8, Fig. 9 and Fig. 10 show the results obtained for american, italian and japanese subjects respectively. One observes that the crossovers obtained with experiment 1 stimuli are in all cases lower (~20 Hz) than with experiment 2 stimuli. In addition, a similar difference to that found in experiment 1 is observed in the crossovers for type I and type II stimuli.

IV. INTERPRETATIONS AND CONCLUSIONS

The results of the perceptual experiments presented in section III show that

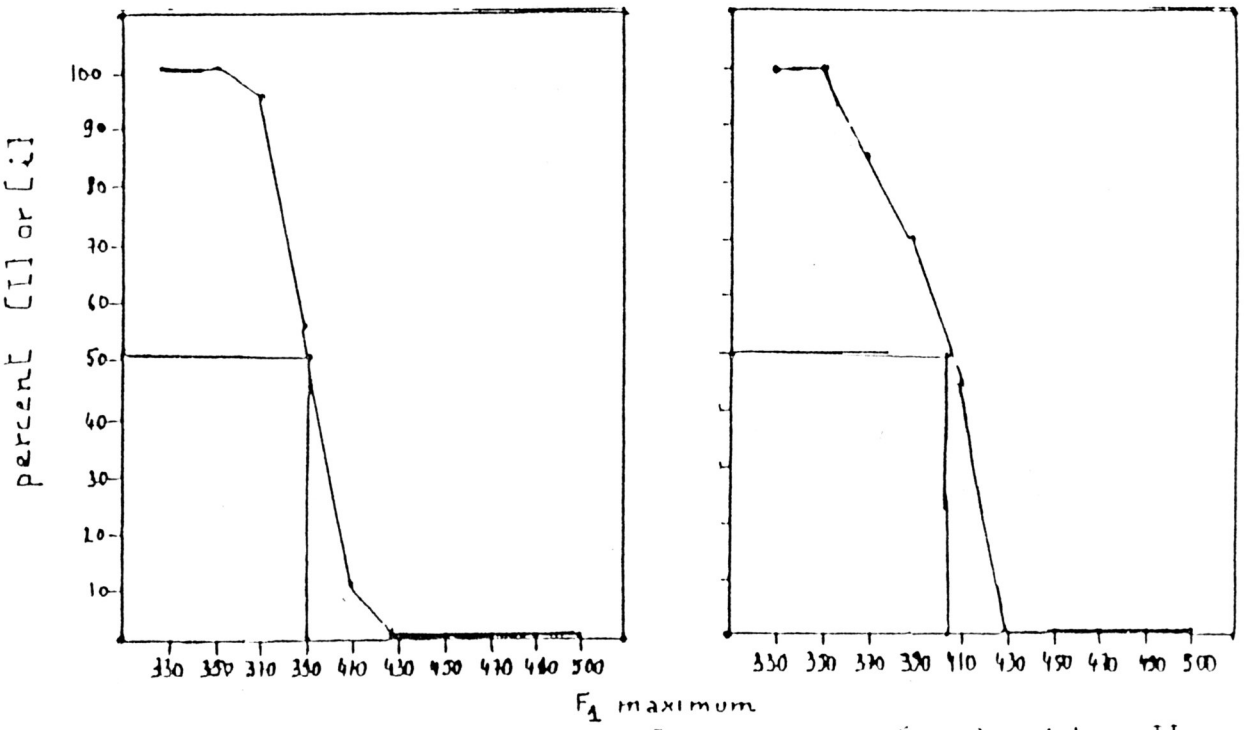

Fig.10Experiment 2 - Identification curves for type I(10-a) and type II (10-b) stimuli- Japanese subject

stimuli characterized by having F1 maximum earlier in the vowel are perceived as lower vowels than stimuli having F1 maximum later in the vowel. These results are in agreement with what observed in section II on the acoustic analysis data: if two different vowels such as /I/ and /ɛ/ have the same F1 maximum, F1 reaches its maximum earlier for the lower vowel. One would tend to conclude then that vowels cannot be represented accurately either by F1 maximum or by the F1 value at the middle point of the vowel, and they can neither be represented satisfactorily by the area (or non weighted time-average) under the F1 trajectory.In addition, it has been shown in experiment 2 that duration influences the perception of the stimuli. Note that in the three languages considered, /i/ and /I/ are shorter than /e/ and /ɛ/ ([9], [10]): the subjects may use their experience when they listen to the shorter stimuli and associate them more to "i-like" sounds. The dependence upon duration can be interpreted in terms of where in the vowel the decision on the vowel is taken and concluding that one waits the end of the vowel before identifying it.

The results are independent of language and this leads to a suggestion that they can have an auditory basis. The actual knowledge of the auditory system does not allow to derive a precise and unique justification to the results obtained. However, some auditory mechanisms, such as short-term adaptation, are not intuitively in disagreement with the results of the present study.

Two hypothesis could account of the results obtained. The first hypothesis is that some extrapolation procedure could occur. The extrapolation curve could be sampled at the end of the vowel, leading to similar values of some new parameter. The second hypothesis is that a weighted time-average process of F1 could occur. The weighting curve should attribute more weight to the first part of the F1 trajectory than to the second one, in order to account of the results obtained in the present study.

Additional perceptual experiments have been carried out to investigate these two hypotheses. Their description is not reported in this paper but can be found in [6]. They tend to be in agreement with the hypothesis of perceptual weighted average F1 process, in agreement with Stevens' and Huang' s findings.

The results of the perceptual experiments presented in this paper do not include observations on the /i/-/I/ and the /e/-/ɛ/ distinction. A complete description of these results can be found in [6].

REFERENCES

1. K.N. Stevens, "The role of duration in vowel identification," Quarterly Progress Report 52, Research Laboratory of Electronics, M.I.T.,1959.
2. C.B. Huang, "Perceptual correlates of the tense/lax distinction in general American-English," Master' s thesis, M.I.T., 1985.
3. C.B.Huang, "The effect of formant trajectory and spectral shape on the tense/lax distinction in american vowels," Proc. IEEE Int. Conf. on Acoust. Speech and Sig. Proc., Tokyo, Japan, April 8-11, 1986.
4. M.G. Di Benedetto, "Relevance of time-varying properties of the first formant frequency in vowel representation," J.A.S.A., Suppl. 1, 78, pp. s81, 1985.
5. D.H. Klatt, "M.I.T. Speechvax user's guide," preliminary version, 1984.
6. M.G. Di Benedetto, "On the role of time-varying properties of the first formant frequency, and of fundamental frequency, in vowel understanding: ar acoustical and perceptual study," Ph.D dissertation, under preparation.
7. D.H. Klatt, "Software for a cascade/parallel formant synthesizer," J.A.S.A. 67(3), PP. 971-995, 1980.
8. J. Neter and W. Wasserman, "Applied linear statistical models," R.D. Irwin, Inc., Homewood, IL 60430, pp. 329-338, 1974.
9. A.S. House, "On vowel duration in english," J.A.S.A., Vol.33, pp. 1174-1178, 1961.
10. D.H. Klatt, personal communication, 1985.