Frequency and time variations of the first formant: Properties relevant to the perception of vowel height

Maria-Gabriella Di Benedetto[a)]
Speech Communication Group, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 20 May 1987; accepted for publication 4 January 1989)

Perceptual experiments using consonant–vowel–consonant (CVC) syllables were carried out to examine the perceptual relevance of the first formant frequency ($F1$) trajectory in the perception of high vowels versus nonhigh vowels. Results show that stimuli characterized by a higher onset frequency and $F1$ maximum at the beginning of the vocalic portion are perceived as lower vowels than stimuli with a lower $F1$ onset frequency and $F1$ maximum toward the end of the vocalic portion. These findings are in agreement with the hypothesis, based on the acoustic analyses of Di Benedetto (1989), that stimuli with higher $F1$ onset frequencies and $F1$ maximum at the beginning of the vocalic portion characterize lower vowels. Results are similar for native speakers of different languages, leading to a suggestion that this phenomenon may have either an articulatory or an auditory basis. Possible interpretations based on an overshoot hypothesis or a formant time average theory were investigated through an additional perceptual experiment. Results of this last experiment agree with a weighted average time formant theory.

PACS numbers: 43.71.Es, 43.70.Fq

INTRODUCTION

The aim of the perceptual experiments presented in this paper was to investigate the perceptual relevance of the first formant frequency ($F1$) trajectory in the identification of high vowels versus nonhigh vowels. In the experiments, synthetic dVd syllables were used in which the vowel represented the high front vowels [i] and [ɪ] and the nonhigh front vowels [e] and [ɛ]. Durational measurements of the vowels [ɪ] and [ɛ] showed (Di Benedetto, 1987) that these vowels have similar duration, although [ɛ] is usually longer than [ɪ]. Durational measurements in other languages, for example Italian (Ferrero *et al.*, 1975), showed that [e] and [ɛ] have very similar durations and are slightly longer than [i]. It was hypothesized that these vowels could be synthesized, using for the stimuli the same total duration of the vocalic portion with appropriate values of $F1$ maximum.

In experiment 1, which will be described in Sec. I, stimuli characterized by two different $F1$ trajectories (higher $F1$ onset frequency and $F1$ maximum towards the beginning or lower $F1$ onset frequency and $F1$ maximum towards the end of the vocalic portion) were used. Identification tests were carried out on subjects of three different languages: American English, Italian, and Japanese. The motivation for the inclusion of American English, Italian, and Japanese listeners was that the vowel system of American English includes the four vowels [i,ɪ,e,ɛ], the Italian vowel system only three of these vowels [i,e,ɛ], and the Japanese vowel system only two [i,ɛ]; supposedly, the perceptual boundary between [i] and [e] should change in American English and Italian listeners responses since, as the vowel [ɪ] is not present in the Italian vowel system, the vowels [i] and [e] are contiguous in the $F1$ dimension. The same observation can be made for the boundary between [i] and [ɛ] in American English and Italian compared to Japanese, for which these two vowels are contiguous in the $F1$ dimension. Stimuli used in experiment 2, described in Sec. II, were shorter in duration than those used in experiment 1. The effect of duration of the vocalic portion was observed on the identification curves obtained for the same subjects who participated in experiment 1. Results showed that two different hypotheses could account for them: perceptual overshoot or a formant time averaging process. These hypotheses were investigated in experiment 3, which will be described in Sec. III.

The results of the three experiments will be discussed in Sec. IV. They will be compared with those found in previous studies that have dealt with the perception of vowels or of stimuli characterized by time-varying properties (Stevens, 1959; Huang, 1985; Brady *et al.*, 1961; Lindblom and Studdert-Kennedy, 1967). An interpretation of the results of experiments 1 and 2, in the light of the results of experiment 3, will be given. A weighted formant time average theory will be proposed for describing the strategy used in processing $F1$ in the case of the distinction between high vowels [i] and [ɪ] versus nonhigh vowels [e] and [ɛ]. A link will be made between the acoustic analysis data described in Di Benedetto (1989) and the perception data. Given the similarity of the results for subjects of different languages, it will be suggested that the phenomenon observed may have either articulatory or auditory bases. A discussion of the agreement of the results of the present study with previous investigations on vowel specification in terms of dynamic rather than static properties (Strange *et al.*, 1976; Strange and Gottfried, 1980; Gottfried and Strange, 1980; Strange *et al.*, 1983; Rakerd *et al.*, 1984; Verbugge and Rakerd, 1986) will follow. The

a) Present address: Department of Information and Communication (INFOCOM), Faculty of Engineering, University of Rome La Sapienza, Via Eudossiana, 18, 00184 Rome, Italy.

 0001-4966/89/070067-11$00.80

aspects of the perceptual data obtained in the present study that fit in this larger context will be outlined.

I. EXPERIMENT 1: ROLE OF THE *F*1 TRAJECTORY

The aim of experiment 1 was to investigate the perceptual relevance of the *F*1 trajectory using dVd synthetic syllables. The experiment consisted of two parts: a preliminary experiment and an extended experiment. Both were identification tests. The aim of the preliminary experiment was to identify the set of vowels, belonging to the vowel system of each subjects' language, to which the vowels of the stimuli were matched by the listeners. In the preliminary experiment, no information concerning the inventory of vowel qualities was given to the subject participating in the experiment. The stimuli were described to the subjects as being dVd synthetic syllables, and subjects were asked to identify the vowel in the stimuli as any vowel system of their language. On the basis of the results of the preliminary experiment, the subjects participating in the extended experiment were asked to identify the vowel in the stimuli as one of the vowels belonging to the set of vowels identified in the preliminary experiment consisting of vowels belonging to the vowel system of the subject's language.

A. Subjects

Five subjects participated in the preliminary experiment. None of the subjects had any knowledge of the purpose of the experiment. Two of these subjects were native speakers of Italian, two were native speakers of American English, and one was a native speaker of Japanese. The Italian subjects were both naive listeners without a good knowledge of any other language. They both named Italian as their best language. The American subjects were phonetically trained listeners and monolingual. They were both members of the Speech Communication Group of MIT, living in the Massachusetts area. The Japanese subject was a phonetically trained listener. At the time of the experiment, he had been living in Cambridge, MA for a few weeks, and his knowledge of American English was rather limited. He named Japanese as his first language.

Seven subjects participated in the extended experiment. None of the subjects had any knowledge about the purpose of the experiment. Four subjects were native speakers of American English, two of Italian, and one of Japanese. The American subjects were all phonetically trained listeners and were members of the Speech Communication Group at MIT. Only one of the American subjects served as a subject in the preliminary experiment. None of the American subjects had profound knowledge of other languages and they all lived in Cambridge, MA. The Italian subjects were both naive listeners. None of the Italian subjects served as a subject in the preliminary experiment. One of the Italian subjects had good knowledge of French but named Italian as his first language; the other subject did not have any knowledge of any other language. The Japanese subject was the one who served as a subject in the preliminary experiment.

It will be shown in the analysis of the results of this experiment that the presence of a limited number of subjects per language could be justified by the fact that no important differences across subjects were observed. However, the presence of only one Japanese subject is somewhat problematic, and the results obtained for this speaker should be considered to be only preliminary.

B. Stimuli

All the stimuli consisted of synthetic dVd syllables and were synthesized with the Klatt synthesizer. This cascade/parallel formant synthesizer has been extensively described by Klatt (1980, 1984). The time where *F*1 reached its maximum and the *F*1 onset frequency were the parameters by which two stimuli having the same *F*1 maximum differed. Figure 1 shows the *F*1 trajectories of the stimuli used. As shown in Fig. 1, this trajectory can have two shapes; depending on the shape, the stimuli are identified as type I or type II. The duration of the stimuli was 115 ms and the duration of the steady-state interval was 15 ms. In type I stimuli, the duration of the onglide was 30 ms and the duration of the offglide was 70 ms. In type II stimuli, the durations were reversed: 70 ms for the onglide and 30 ms for the offglide. The vocalic portions of all stimuli were preceded by an aperiodic portion, whose duration was 15 ms, corresponding to the [d] release burst.

Ten stimuli of type I and ten stimuli of type II were synthesized, each stimulus being characterized by a different *F*1 maximum value (330, 350, 370, 390, 410, 430, 450, 470, 490, 500 Hz) and different *F*1 onset and offset values, which are listed in Table I. The trajectories of the formants higher than *F*1 and of the fundamental frequency were identical for both stimulus types and symmetrical around the center of the vowels. The values of the fundamental frequency and of the formants above *F*1 are indicated in Table II.

One should note that, in the *F*1 vs *F*2 space representation, no distinction could be made between two stimuli having the same *F*1 maximum but different *F*1 trajectory shapes. This would occur either if the sampling time of the trajectories were the time where *F*1 reaches its maximum or if the sampling point were at the middle of the vowel, and if the representation used were more generally in the *F*1 vs *F*2 vs *Fx* space.

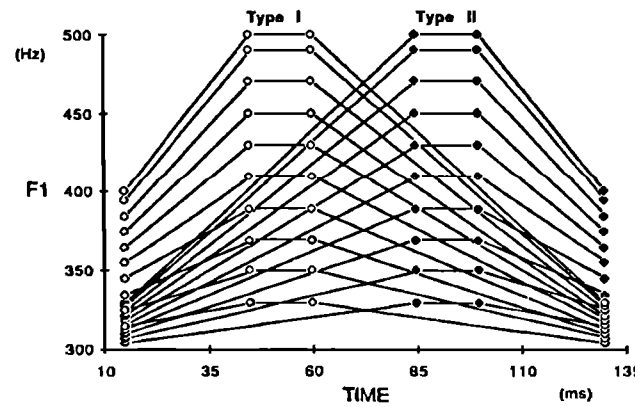

FIG. 1. Schematic *F*1 trajectories for type I and type II stimuli, used in experiment 1. The *F*1 trajectories start at time = 15 ms, as there was a previous aperiodic portion corresponding to the [d] release burst.

TABLE I. The $F1$ onset, $F1$ maximum, and $F1$ offset values (in hertz) for stimuli of types I and II, used in experiment 1.

Stimulus number	Stimulus type	$F1$ onset value (Hz)	$F1$ maximum value (Hz)	$F1$ offset value (Hz)
1	1	304	330	315
1	2	315	330	304
2	1	325	350	307
2	2	307	350	325
3	1	335	370	310
3	2	310	370	335
4	1	345	390	313
4	2	313	390	345
5	1	355	410	316
5	2	316	410	355
6	1	365	430	319
6	2	319	430	365
7	1	375	450	322
7	2	322	450	375
8	1	385	470	325
8	2	325	470	385
9	1	395	490	328
9	2	328	490	395
10	1	400	500	330
10	2	330	500	400

C. Procedure

In the preliminary experiment, type I and type II stimuli were presented in two phases. In the first phase, the ten type I stimuli were first randomized and then presented in three sets of ten stimuli. Each set consisted of one repetition of each of ten type I stimuli; thus, each type I stimulus was presented three times. In each set, the stimuli were spaced by a pause of 3 s and the sets of stimuli were spaced by a pause of 7 s. In the second phase, an identical procedure was used, with type II stimuli.

The extended experiment consisted of three phases. In a first phase, only type I stimuli and in a second phase only type II stimuli were presented to the listeners. In a third phase, type I and type II stimuli were both presented to the listeners. In the first phase, each type I stimulus was presented ten times. The ten type I stimuli were ordered in such a way that all possible combinations of two consecutive stimuli appeared; the responses given by the listeners for each stimulus could then be supposed to be independent of the stimulus preceding it. The second phase was organized as the first one, but the stimuli used were of type II. In the third phase, the ten stimuli of type I and the ten stimuli of type II were divided into two sets of ten stimuli each. Each set sampled the range from near the lowest to near the highest $F1$ values. The stimuli of each set were ordered to obtain, as previously, all possible combinations of two consecutive stimuli. Each stimulus was then presented 20 times. In each phase, the stimuli were spaced by a pause of 3 s. At the end of each part, the test was interrupted and the subject could rest for a few minutes. These pauses did not last more than 2–3 min. The test lasted approximately 40 min. The response format consisted of an answer sheet, printed with as many lines as those of the played-back stimuli, on which the subjects had to write consecutively the vowel they perceived (in phonetic symbols for the phonetically trained listeners and in standard orthography for the naive listeners).

TABLE II. Fundamental frequency and higher formats than $F1$ values (in hertz) for stimuli of types I and II, used in experiment 1.

Time (ms)	$F0$ value (Hz)	$F2$ value (Hz)	$F3$ value (Hz)	$F4$ value (Hz)	$F5$ value (Hz)
15	120	2593	3500	4000	4500
70	125	2800	3500	4000	4500
75	125	2800	3500	4000	4500
130	120	2593	3500	4000	4500

The results obtained for each subject in the three phases were averaged since the identification functions obtained were identical in the first and third phases for stimuli of type I, and in the second and third phases for type II stimuli. The results were represented by the identification curves in the plane identified on the abscissa by the stimulus number (see Table I) and on the ordinate by the percent of identification of the vowel (or vowels), each time specified. A logistic curve fitting the data and the 50% crossover point were computed. The crossover value represents the point at which the identification changes from the vowel (or vowels) specified to a different vowel. The logistic curve was found according to the procedure proposed by Neter and Wassermann (1974) and it represents the probability of correct response (psychometric function).

D. Results

In the preliminary experiment, the vowels of the synthetic stimuli were identified as [i], [ɪ], [e] or [ɛ] by the American subjects, as [i], [e], or [ɛ] by the Italian subjects, and as [i] or [ɛ] by the Japanese subject. The vowel [e] was characterized by the American subjects "as the nondiphthongized [e^{y}]" or "as the first part of the diphthong [e^{y}]." Note that the vowel [ɪ] does not belong to the Italian and the Japanese vowel systems, and that the vowel [e] does not belong to the Japanese vowel system.

Based on the results of the preliminary experiment, the American subjects were asked, in the extended experiment, to identify the vowel of the synthetic syllables as one vowel of the set [i,ɪ,e,ɛ]. In any case, none of the subjects reported hearing a vowel different from these four.

Results of the extended experiment obtained for each American subject individually are presented in Fig. 2, which shows that all type II stimuli were identified by the subjects SSH and JP as [i] or [ɪ] and never as [e] or [ɛ], while the two other subjects' identification curves for type II stimuli cross the 50% line. For type I stimuli, similar identification curves were found for all subjects. A possible explanation of this result was that, for SSH and JP, the shape of the $F1$ trajectory was more perceptually relevant than the $F1$ maximum value. Thus, even when $F1$ was in the high-value range, these subjects based their judgment of the stimuli on the fact that the $F1$ trajectory had that particular shape. Average results were then obtained for type I stimuli by considering the four subjects' responses [Fig. 3(a)], and for type II stimuli using only those of KS and CB [Fig. 3(b)]. Consider

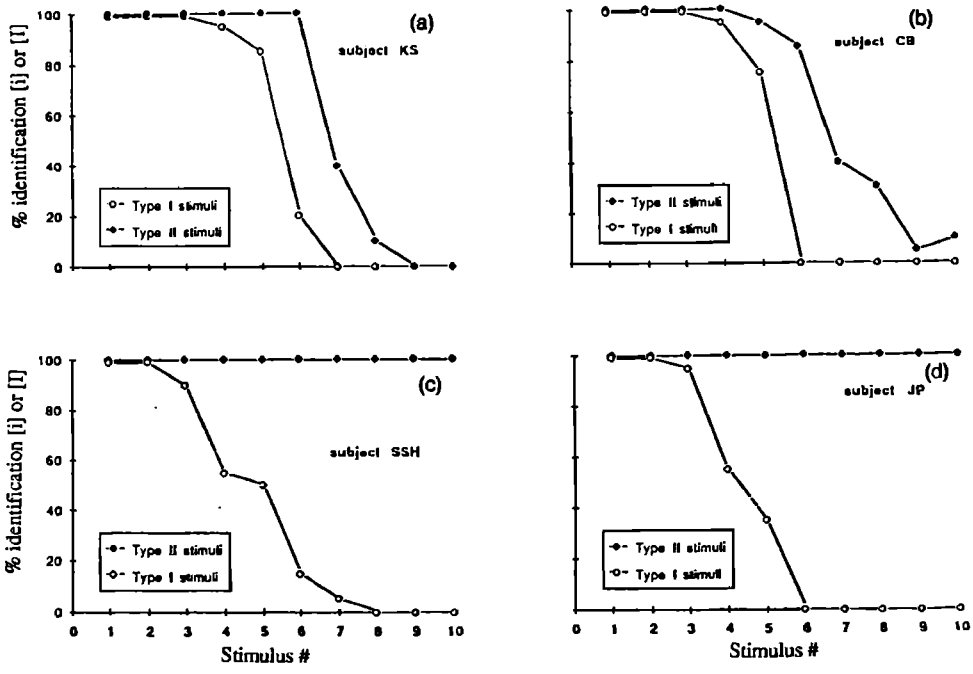

FIG. 2. Results of the extended experiment (identification curves) for (a) KS, (b) CB, (c) SSH, and (d) JP, for type I and type II stimuli.

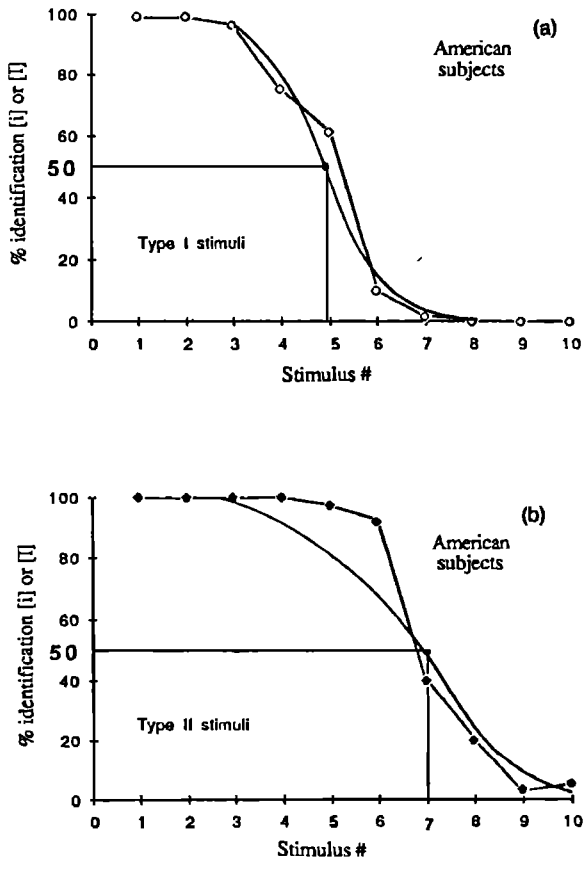

FIG. 3. Average results of the extended experiment in terms of the percent identification of the high vowels [i] and [ɪ] versus the nonhigh vowels [e] and [ɛ], for (a) type I stimuli, and (b) type II stimuli, for the American subjects. The results obtained for type II stimuli, (b), represent an average of the results of two subjects (KS and CB) while the results obtained for type I stimuli (a), represent an average of the results of the four subjects (KS, CB, JP, and SSH).

that, nevertheless, if the results of JP and SSH were taken into account in computing averaged responses for type II stimuli, the crossover point would be at a higher value than that indicated in Fig. 3(b). Figure 3 shows that a lower crossover value for type I than for type II stimuli was found. The difference was about two stimulus numbers or about 40 Hz.

Based on the results of the preliminary experiment, the Italian subjects were asked to identify the vowels of the synthetic syllables as [i], [e], or [ɛ]. In any case, none of the subjects reported hearing a different vowel from these three. Figure 4 shows the identification curves of [i] and the logistic curves for type I [Fig. 4(a)] and type II stimuli [Fig. 4(b)], obtained by averaging the results of the two Italian subjects who participated in this experiment. As observed for the American subjects, there was a difference in the crossover values for type I and type II stimuli. The difference was about 1.3 stimulus numbers. This difference was slightly lower than the one found for the American subjects, but in the same direction. Results of this experiment for each listener individually, reported in Di Benedetto (1987), were similar to the average results (for subject MA the crossover values were 4.8 for type I and 6.3 for type II stimuli, and for subject ZB the crossover values were 4.5 for type I and 5.6 for type II stimuli).

The vowel system of Japanese does not include the vowels [ɪ] and [e]. Consequently, the Japanese subject who participated in this experiment was asked to identify the vowel of the synthetic syllables presented as [i] or [ɛ]. Note that, in any case, the subject who participated in this experiment reported that he did not perceive vowels other than [i] and [ɛ]. The results of the experiment in the case of the Japanese subject, presented in Fig. 5, show that, similarly to what was found for American and Italian subjects, there was a differ-

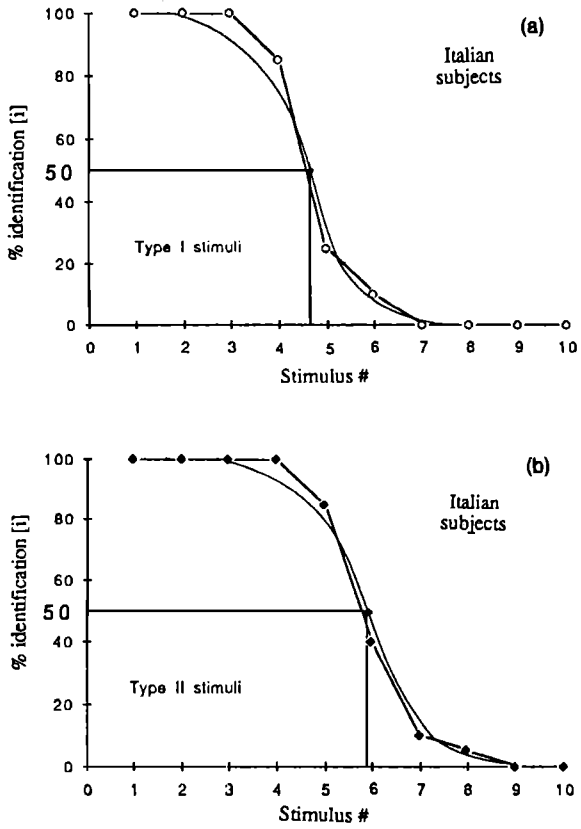

FIG. 4. Average results of the extended experiment in terms of the percent identification of the high vowel [i] versus the nonhigh vowels [e] and [ɛ], for (a) type I stimuli and (b) type II stimuli, for the Italian subjects.

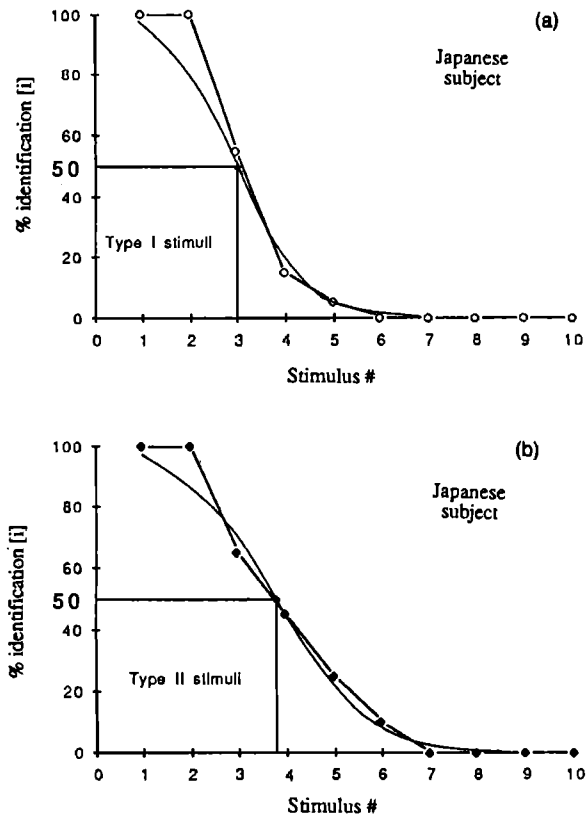

FIG. 5. Results of the extended experiment in terms of the percent identification of the high vowel [i] versus the nonhigh vowel [ɛ] for (a) type I stimuli and (b) type II stimuli, for the Japanese subject.

ence in the crossover values for type I and type II stimuli. This difference was less than in the previous cases (about 0.8 stimulus number) and in the same direction.

One should again consider that the number of subjects was small and that there was variation among the results obtained by American subjects. However, all subjects, American, Italian, and Japanese, showed a tendency to associate synthetic vowels characterized by higher $F1$ onset frequency and $F1$ maximum at the beginning of the vocalic portion (type I stimuli) with lower vowels than synthetic vowels characterized by a lower $F1$ onset value and $F1$ maximum towards the end of the vocalic portion (type II stimuli). All the subjects participating in experiment 1 (preliminary and extended) declared that they were not disturbed by the perception of the consonants and that they always perceived the consonants of the synthetic utterances as [d].

II. EXPERIMENT 2: ROLE OF DURATION

The aim of experiment 2 was to investigate whether shortening the duration of the stimuli used in experiment 1 would lead to different listeners' responses. In fact, by shortening the duration of the stimuli, the $F1$ onset frequencies could be kept the same in the two experiments, but the time where $F1$ reached its maximum relative to the total vowel duration was changed. Experiment 2 was carried out to gain insight on how the relative $F1$ onglide duration with respect to total duration influenced vowel perception.

A. Subjects

The subjects were the same as those of the extended experiment 1.

B. Stimuli

The stimuli used in experiment 2 differed from those of experiment 1 in terms of duration. The duration of the stimuli was 95 ms (20 ms shorter in duration than the stimuli of experiment 1), and the duration of the steady-state interval of the vowel was 15 ms for all stimuli. The $F1$ onset and offset frequencies were kept the same as in experiment 1. In type I stimuli, the duration of the onglide was 30 ms and the duration of the offglide 50 ms. In type II stimuli, the durations were reversed: 50 ms for the onglide duration and 30 for the offglide. Consequently, for type I stimuli, the shape of the $F1$ onglide was the same in experiments 1 and 2 and the $F1$ offglide was shorter in experiment 2, while for stimuli of type II the $F1$ onglide was shorter in experiment 2 and the shape of the $F1$ offglide was the same in the two experiments.

C. Procedure

The organization and the method of representing the results of experiment 2 were the same as those of the extended experiment 1. The subjects were asked to identify the vowels of the synthetic syllables as vowels belonging to the same sets of experiment 1, i.e., [i,ɪ,e,ɛ] for American subjects, [i,e,ɛ] for Italian subjects, and [i,ɛ] for the Japanese subject.

D. Results

Figure 6 shows on the same plot the results of experiments 1 and 2 for type I and type II stimuli in the case of American subjects. The results for type I stimuli represent an average of the responses of the four subjects, and the ones for type II stimuli only of KS and CB, because JP and SSH never identified those stimuli as [e] or [ɛ]. Figure 6 shows that, in experiment 2, the crossover values are higher by about 0.5 stimulus number for type I stimuli and by about 1 stimulus number for type II stimuli than in experiment 1. Note that durational measurements on the vowels [i] and [ɛ] have shown that [i] is, in general, shorter in duration than [ɛ] (Di Benedetto, 1987).

Figure 7 shows on the same plot the results obtained in experiments 1 and 2 for type I and type II stimuli for the Italian subjects. Similar to the results of the American subjects, the crossover values were higher in experiment 2 by about 0.8 stimulus number for type I stimuli and by about 1 stimulus number for type II stimuli than they were in experiment 1. Note that, in Italian, the vowel [i] is, in general, shorter than [e] and [ɛ] (Ferrero *et al.*, 1975).

Figure 8 shows on the same plot the results of experiments 1 and 2 for the Japanese subject. As for the American and Italian subjects, the Japanese subject tended to associate stimuli of experiment 2 with a higher vowel than stimuli of

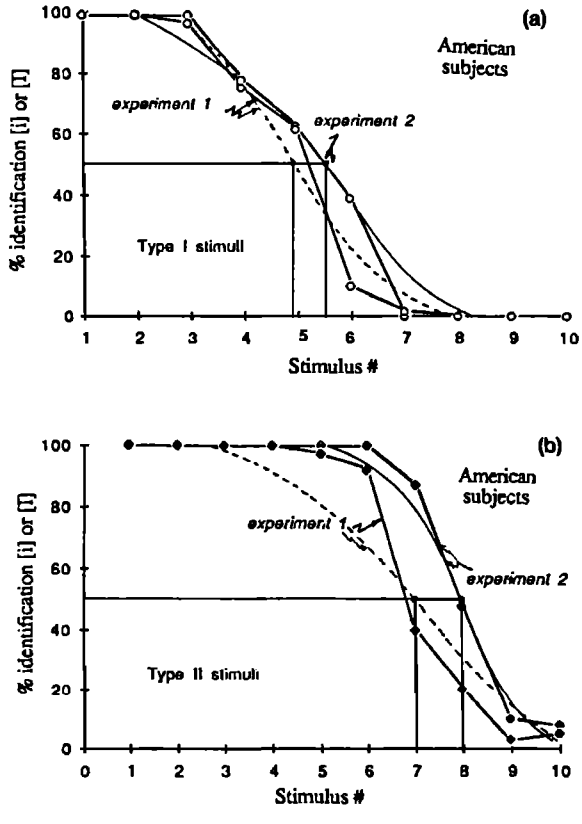

FIG. 6. Average results of experiments 1 and 2 in terms of the percent identification of the high vowels [i] and [ɪ] versus the nonhigh vowels [e] and [ɛ], for (a) type I stimuli and (b) type II stimuli, for the American subjects. The results obtained for type II stimuli, (b), represent an average of the results of two subjects (KS and CB), while the results obtained for type I stimuli, represent an average of the results of the four subjects (KS, CB, JP, and SSH).

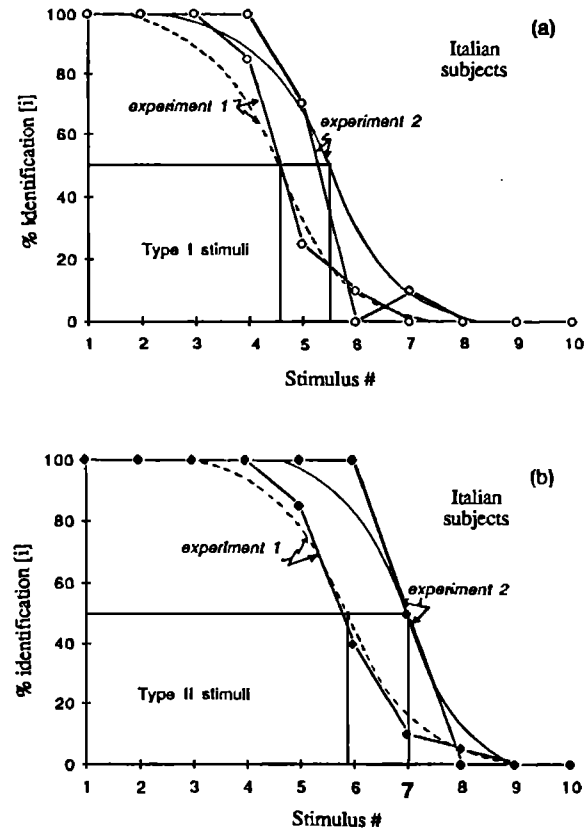

FIG. 7. Average results of experiments 1 and 2 in terms of the percent identification of the high vowel [i] versus the nonhigh vowels [e] and [ɛ] for (a) type I stimuli and (b) type II stimuli, for the Italian subjects.

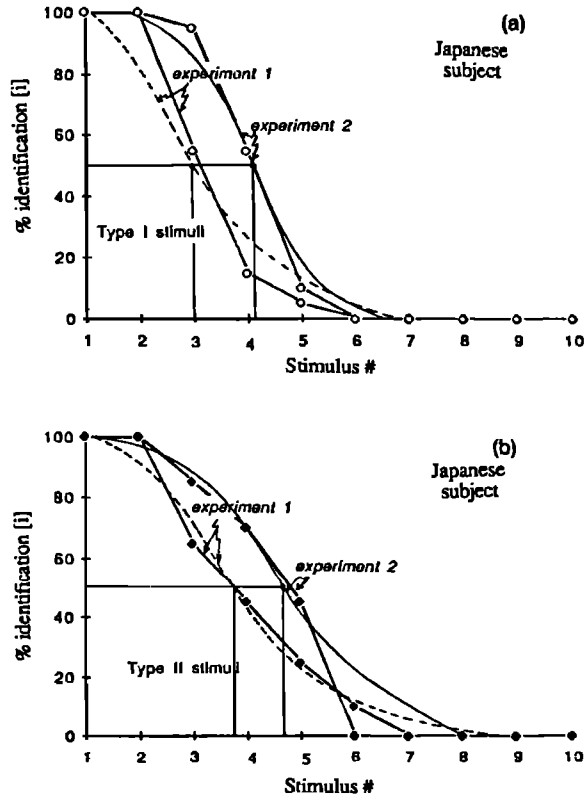

FIG. 8. Results of experiments 1 and 2 in terms of the percent identification of the high vowel [i] versus the nonhigh vowel [ɛ] for (a) type I stimuli and (b) type II stimuli, for the Japanese subject.

experiment 1 (shift of the crossover towards higher values by about 1 stimulus number for type I and 0.9 stimulus number for type II stimuli). Note that, in Japanese, [i] is usually shorter than [ɛ] (Klatt, 1985).

Note that, for the American English and the Italian subjects, the differences in the crossover values in experiments 1 and 2 were larger for type II than for type I stimuli. The shift found in the crossover values for type I stimuli in experiments 1 and 2 was expected since, for these stimuli, the onglide duration was unchanged while the offglide duration was shortened. The onglide duration relative to the total vowel duration was then larger for type I stimuli in experiment 2 than in experiment 1, making these stimuli resemble more, in terms of relative onglide duration, type II stimuli; this could justify the results that type I stimuli in experiment 2 were perceived as higher vowels than in experiment 1. The shifts found in the crossover values for type II stimuli in experiments 1 and 2 were expected to be in the opposite direction. Since for these stimuli the onglide duration was shortened and the offglide duration was unchanged, the relative onglide duration with respect to the total vowel duration was smaller in experiment 2 than in experiment 1, making these stimuli resemble more, in terms of relative onglide duration, type I stimuli. These stimuli should have then been perceived as lower vowels in experiment 2 than in experiment 1. It thus appears that, for type II stimuli in experiment 2, the overall stimulus duration overrode the expected effect of shorter onglide, which would have presumably produced more [e]–[ɛ] responses, since the onglide duration of these stimuli (50 ms) was intermediate between the onglide duration of type I and type II stimuli of experiment 1 (30 and 70 ms, respectively). Note that, as already pointed out, the high vowels are shorter than the nonhigh vowels under consideration in the three languages considered.

III. EXPERIMENT 3: FORMANT TIME AVERAGE VERSUS OVERSHOOT

The aim of experiment 3 was to investigate whether either a formant time average theory or an overshoot hypothesis based on perceptual extrapolation of the $F1$ trajectory was appropriate for interpreting the results obtained in experiments 1 and 2. On the basis of a formant time average hypothesis, the perceived $F1$ frequency should correspond to a frequency that is included in the set of frequencies swept by the $F1$ trajectory, while on the basis of an overshoot hypothesis, the perceived $F1$ could correspond to higher frequency values with respect to those of the $F1$ trajectory.

A. Subjects

Three subjects participated in this experiment. Two of these subjects (CH and RS) were native speakers of American English. They were phonetically trained listeners and named American English as their best language. They were both members of the Speech Communication Group at MIT. One of the subjects (SM) was a native speaker of Italian with some knowledge of American English, but he named Italian as his first language. This subject was a naive listener. None of the subjects had participated in experiments 1 and 2. American English and Italian subjects were considered in order to compare the results obtained in experiments 1 and 2 with native speakers of the same languages.

B. Stimuli

Seven stimuli were considered. They all consisted of dVd synthetic syllables of the kind used in experiment 1. The $F1$ trajectories of these stimuli are shown in Fig. 9. The trajectories for formants above $F1$ and the fundamental frequency contour were identical to those of the stimuli of experiment 1. A more detailed description of the stimuli is given in Table III. Stimuli A–D were chosen in order to gain insight on how stimuli characterized by a very rapid onset and various offglide durations were perceived. The responses to those stimuli could be compared to the ones corresponding to stimuli E and F, which were characterized, on the contrary, by a long onglide duration (note that E and F have the same onglide slope).

C. Procedure

Experiment 3 consisted of two phases. In a first phase, identification tests were carried out. The stimuli were randomized and presented to the listeners. Each stimulus occurred eight times. The subjects were asked to identify the vowel of each stimulus as any vowel in the vowel system of their language. In a second phase, matching tests were carried out. The subjects were asked to listen to pairs of stimuli (one of the stimuli in each pair was stimulus A and the other stimulus in the pair was stimulus B, C, D, E, or F) and to grade their similarity in three steps: very similar (VS), similar (S), nonsimilar (NS). Each stimulus pair was presented ten times.

D. Results

Results of the identification test obtained from the three subjects showed that the vowels of the synthetic utterances were identified as [ɪ] or [e] by the American subjects and [i] or [e] by the Italian subject. The American subjects defined the vowel [e] as the nondiphtongized [e^{γ}]. The two American subjects declared, in addition, that in some cases they perceived a vowel which would be in between [e] and [ɛ] in terms of height. If forced, they would classify these vowels as [e], but they would attribute to it an [ɛ]-like color. We will call the identification of these particular cases as [e]–[ɛ] identifications.

The results for subjects RS, CH, and SM are shown in Fig. 10(a)–(c), respectively. This figure shows for each speaker individually the percentage of identification of [i], [ɪ], [e], or [e]–[ɛ] for each stimulus identified by the same letter as in Fig. 9. Figure 10 shows that the three subjects identified stimuli A–D as [e] or [e]–[ɛ] sounds. In particular, subject CH identified stimulus A as [e]–[ɛ] in all cases. Stimuli E and F were, on the contrary, generally identified as [ɪ] by the American subjects and as [i] 90% of the time by the Italian subject (100% for stimulus E and 80% for stimulus F).

Results of the matching test are shown in Fig. 11. Re-

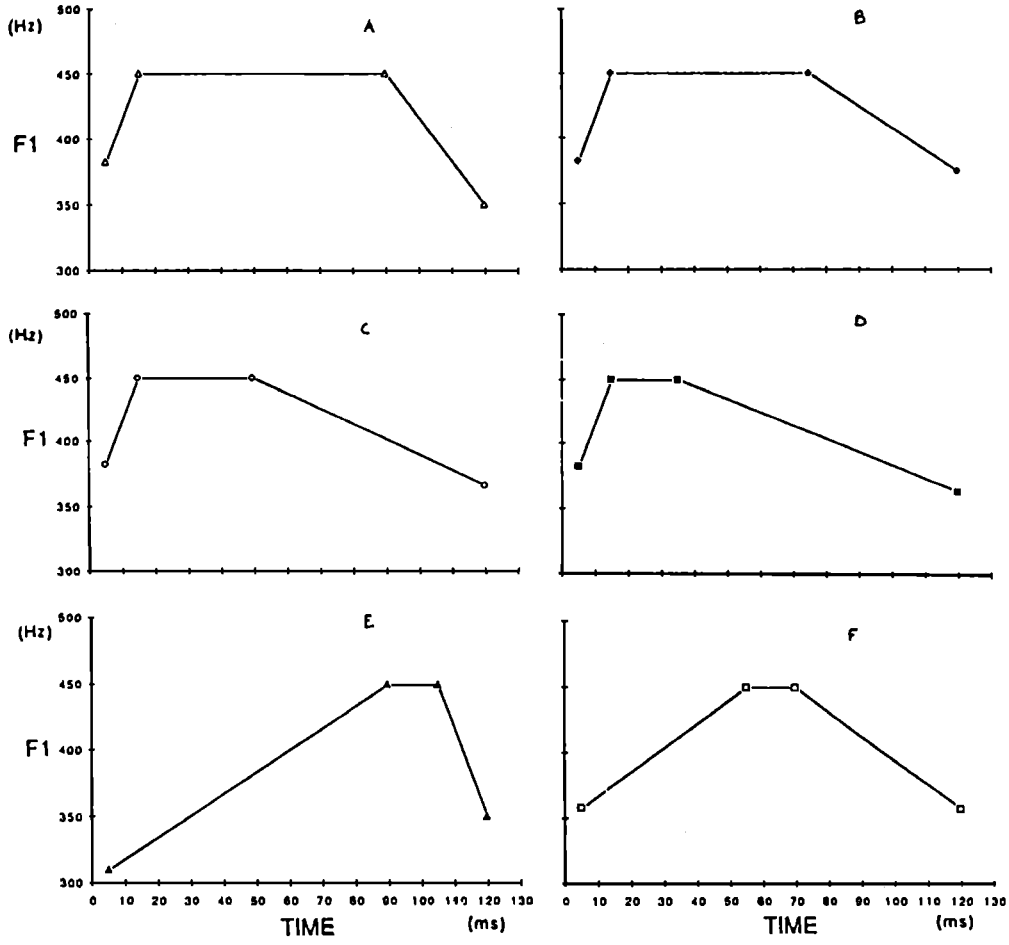

FIG. 9. Schematic $F1$ trajectories of the stimuli used in experiment 3.

sults of each subject were identical (there was no variability across subjects) to the average results. Figure 11 shows, on the abscissa, the stimulus pairs (the reference stimulus is always stimulus A in all pairs), represented by the same letters as in Fig. 9, and, on the ordinate, the judgment given by the subjects. Figure 11 shows that, for the three subjects, the results of the matching test were in agreement with the results of the identification test. The subjects perceived stimuli A–D as similar vowels, while stimuli E and F were perceived as different vowels from stimulus A. Subject CH declared, as in the identification test, to perceive stimulus A as a vowel somewhat more open than the vowels of stimuli B–D. None of the subjects reported to be disturbed by the per-

TABLE III. Descriptive characteristic values for the stimuli used in experiment 3. $F1$ onset, $F1$ maximum, and $F1$ offset values (in hertz) and onglide, steady-state, and offglide durations (in ms) are listed.

Stimulus name	$F1$ onset value (Hz)	$F1$ max value (Hz)	$F1$ offset value (Hz)	Onglide dur. (ms)	Steady-state dur. (ms)	Offglide dur. (ms)
A	383	450	375	10	65	40
B	383	450	375	10	60	45
C	383	450	375	10	35	70
D	383	450	375	10	20	85
E	283	450	358	85	15	15
F	358	450	358	50	15	50

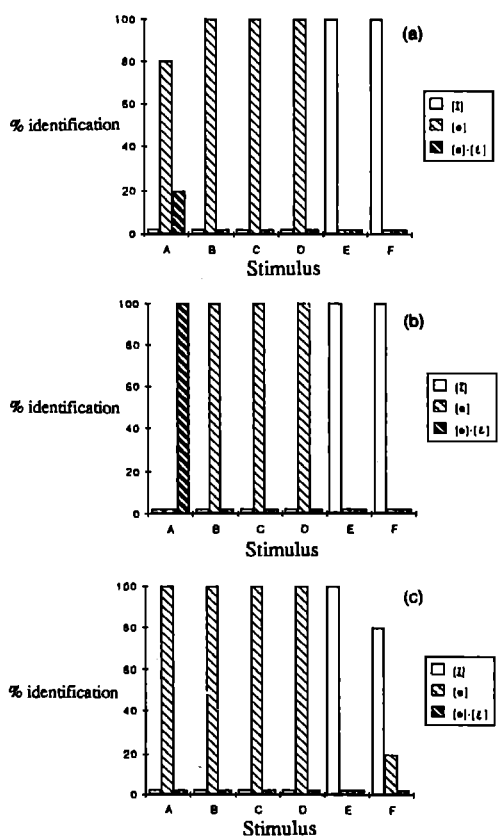

FIG. 10. Results of the identification test of experiment 3 for subjects (a) RS, (b) CH, and (c) SM.

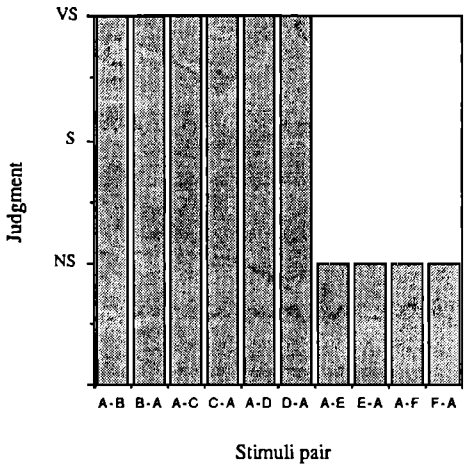

FIG. 11. Results of the matching test of experiment 3.

ception of the consonants of the stimuli. They all reported to perceive the consonants as [d].

Results of experiment 3 show that stimuli characterized by a higher $F1$ onset frequency and a shorter $F1$ onglide duration (stimuli A–D) were perceived as lower vowels than stimuli with a lower $F1$ onset frequency and a longer $F1$ onglide duration (stimuli E and F). Among stimuli A–D, stimulus A was perceived as the more open one, showing that a shorter offglide duration and a longer steady-state results in the perception of a lower vowel. Stimuli E and F were perceived as similar, although F was perceived as more open than E by subject SM. The results obtained for the Italian subject (F perceived as [e] part of the time) could be justified by the fact that stimuli F had a longer $F1$ offglide and a shorter $F1$ onglide. Note, in addition, that E had a lower $F1$ onset value than F.

IV. INTERPRETATION AND CONCLUSION

Perceptual experiments 1 and 2, which were carried out on American, Italian, and Japanese subjects, showed that stimuli that were characterized by a higher $F1$ onset frequency and $F1$ maximum at the beginning of the vocalic portion were perceived as lower vowels than stimuli in which the $F1$ maximum was reached towards the end of the vocalic portion and the $F1$ onset value was lower. This result is in agreement with the observation based on the analysis of the $F1$ trajectories (Di Benedetto, 1989) that, when two different vowels such as [ɪ] and [ɛ] have the same $F1$ maximum, then $F1$ starts higher and reaches its maximum earlier in the lower vowel. The results of experiments 1 and 2 show evidence that supports the hypothesis that the $F1$ onset frequency and the temporal location of the $F1$ maximum are relevant to the perceptual identification of the high vowels [i] or [ɪ] versus the nonhigh vowels [e] or [ɛ]. The acoustic analysis data presented in Di Benedetto (1989) showed that the variations of the $F1$ onset frequency and of the relative $F1$ onglide duration with respect to total vowel duration did not contribute to the variation of a single parameter, represented by the onglide slope of the $F1$ trajectory (this parameter was called "$F1$ speed"). The results of the present study were in agreement with this observation; their interpretation in terms of an association between steeper onglides and the perception of lower vowels could be misleading. In fact, since stimuli of types I and II with the same $F1$ maximum and identified as [e]–[ɛ] vs [i]–[ɪ] differed both in the $F1$ onset frequency values and in the $F1$ onglide duration, the $F1$ speed values for these stimuli could be similar. Consider, for example, stimulus #6 of experiment 1, which was most often identified as a high vowel if of type II and as a low vowel if of type I by the American English listeners; the difference of the $F1$ speed values between stimulus #6 of type I and stimulus #6 of type II, expressed in degrees, was very small (about 6 deg).

In addition, experiment 2, in which all the stimuli considered were 20 ms shorter than those of experiment 1, showed that the duration of the synthetic vowel was a factor that influenced vowel perception. When shorter stimuli were considered, the difference in the perception of type I and type II was maintained, but shorter stimuli were identified by all subjects as higher vowels. It was noticed that stimuli of type I, which had in experiment 2 a shorter offglide than in experiment 2, were perceived as higher vowels, in agreement with the results of experiment 1. On the contrary, stimuli of type II, which would have been expected to produce more [e]–[ɛ] responses, due to the shorter onglide, were perceived as higher vowels. The interpretation was that, for type II stimuli in experiment 2, the overall duration overrode the expected effects of shorter onglide. The results suggest, in addition, that the listener waits for the end of the vocalic portion before identifying it. In fact, it was observed that two stimuli having the same $F1$ maximum and being characterized by the same initial transition of $F1$ could be perceived as two different vowels due to the different offglides. A similar effect, leading to the same suggestion, could be observed in Huang's data (1985).

The possibility of interpreting the data of experiments 1 and 2 on the basis of a formant average theory or of perceptual extrapolation of the trajectories was investigated. Stevens (1959) suggested that the perception of a nonhigh vowel is determined by some time-average values of formant frequencies, including the transitions. Huang (1985,1986) investigated the role of duration and the effect of the formant trajectory shape on the perception of the tense/lax distinction in General American English. Huang's findings were that a formant time average process for $F1$ could account for the results obtained in the case of nonhigh vowels. Results of other perceptual studies suggested perceptual overshoot in vowel perception. Brady *et al.* (1961) found a consistent tendency of the subjects to place the resonant frequency of a time-constant resonance stimulus near the terminal value of a time-varying resonance stimulus. This tendency was stronger when the resonant frequency increased than when it decreased, and was greater when the change in the resonant frequency was more abrupt. Lindblom and Studdert-Kennedy (1967) showed that duration of the vowel stimuli and the direction and rate of the second formant transitions influenced the determination of vowel identity; they hypothesized that articulatory activities (in the production of vowels) are characterized by undershoot and that vowel perception is characterized by overshoot. However, the hypothesis

of $F2$ perceptual overshoot was made without ruling out the possibility that the effect observed could be based on some top–down processing phenomenon (Lindblom *et al.*, 1967).

The results of experiments 1 and 2 suggest that a simple average value of the $F1$ trajectory is not appropriate. In fact, two stimuli characterized by two different values of $F1$ maximum, one of type I and the other of type II, have also the same average value of $F1$, but they could be perceived, for certain values of $F1$ maximum, as different vowels. The formant time average should be a weighted formant time average value. To take into account the results of experiments 1 and 2, the weighting function should attribute more importance (more weight) to the first part of the $F1$ trajectory.

An overshoot hypothesis could also take into account the results of experiments 1 and 2. It was found that four different stimuli characterized by different trajectories and different durations, as shown in Fig. 12(a), were perceived as similar sounds. The listener could extrapolate a new value of a new parameter, as shown in Fig. 12(b) obtained by taking into account the whole $F1$ trajectory on the basis of an extrapolation curve and by sampling this curve at the end of the vocalic portion. Figure 12(b) shows the $F1$ trajectory extrapolation curves for each of the four stimuli specified and the sampling instants located at the end of the vocalic portion. The four synthetic vowels characterized by the four different $F1$ trajectories of Fig. 12(a) could then be characterized by similar values of this new auditory parameter. One should note that it was not possible to assert whether this hypothesis was in agreement or not with the results of Brady *et al.* (1961) and Lindblom and Studdert-Kennedy (1967), since in these studies observations were made on frequencies that characterized $F2$, and there is no evidence that, in the perception of vowels, $F1$ and $F2$ are processed in the same way. The hypothesis that $F1$ and $F2$ could be perceived differently is not unreasonable, as $F1$ is within the lower frequency range of the auditory system in which temporal (synchronous firing rate) coding of stimulus frequency occurs, while at higher frequencies (e.g., $F2$ range), temporal coding breaks down and the coding of frequency is primarily spatial. In addition, one should note that the stimuli considered in the Brady *et al.* study were particularly short and characterized by either a rising or a falling variation of the resonant frequency. The case studied in this paper in which the $F1$ trajectories were characterized by both transitions, the initial rising and the final falling, is different. This may lead to different perceptual judgments. However, even if $F1$ and $F2$ were perceived on the basis of the same auditory mechanism, the results of Brady *et al.* could not explain the results of experiments 1 and 2 that listeners perceived type I stimuli as lower vowels. Note, in fact, that, as observed previously, the initial transition was not significantly more abrupt in type I than in type II stimuli (for stimuli perceived as similar vowels), as the $F1$ onset and $F1$ onglide duration both varied.

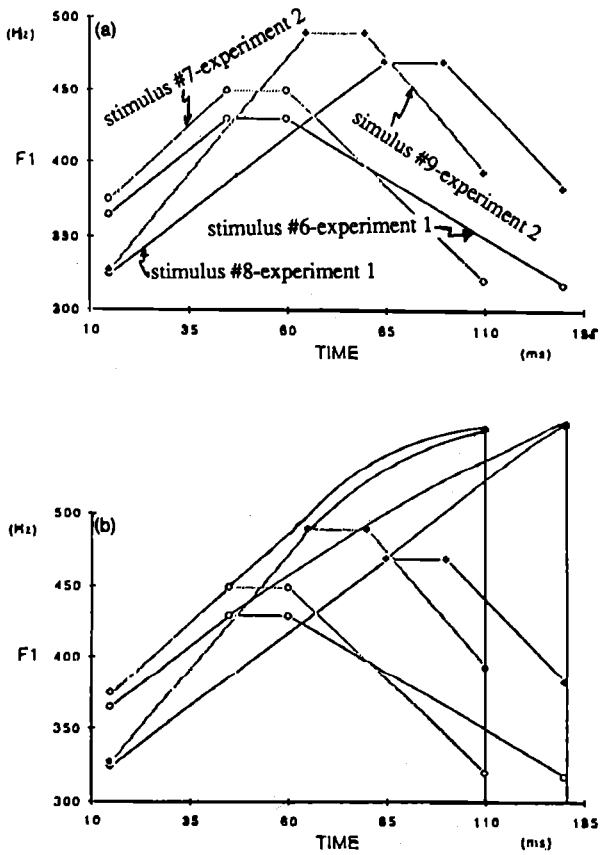

FIG. 12. (a) Schematic $F1$ trajectories of four different stimuli that have been perceived to be similar. Specifically, stimulus #6 and #8 of experiment 1, and stimuli #7 and #9 of experiment 2. (b) Interpretation of the results of experiments 1 and 2 on the basis of an overshoot hypothesis.

The results of experiment 3 showed that stimuli which were characterized by a very short $F1$ onglide and a higher $F1$ onset frequency (stimuli A–D) were perceived as lower vowels than stimuli with a longer $F1$ onglide and a lower $F1$ onset frequency (stimuli E and F). If the fast transition of stimuli A–D was considered as part of the vowel, this result would agree with an overshoot hypothesis. If the fast transition were considered as part of the consonant, they would agree with an average time formant theory. We suggest that this second hypothesis is more reasonable, as such a rapid change in $F1$ (10 ms) implies a rapid spectrum change, and that this property characterizes segments of speech having the feature [+ consonantal] (Stevens, 1980, 1989). As far as the $F1$ transition of the vowel is concerned, we believe that one can consider that there is no change in frequency of $F1$ in stimuli A–D. In addition, the results of experiment 3 give additional support for the hypothesis that the entire $F1$ trajectory is taken into account by the perceptual mechanism that processes it. In fact, stimulus A was perceived as a somewhat more open vowel than stimulus B. The finding that the entire $F1$ trajectory affects vowel judgment would be a good additional argument against the overshoot hypothesis if one assumes that an overshoot mechanism considers the onglide shape (onglide duration and $F1$ onset frequency) to build an extrapolation curve and then uses total duration to determine the sampling point, ignoring the effects of the offglide shape. However, one could imagine that, on the contrary, the overshoot mechanism uses also the offglide trajectory to correct "the direction" pointed by the extrapolation curve.

That shorter stimuli were perceived as higher vowels may reflect a process learned by the subjects. In fact, it was

observed that, in the three languages studied, high vowels are characterized by shorter durations than nonhigh vowels. However, if this phenomenon were language universal, it could be argued that there is some evidence for innate articulatory or auditory bases. In the same way, the similar results observed for listeners of different languages concerning the $F1$ temporal and spectral properties effects could lead to a suggestion that this phenomenon may have either an articulatory or an auditory basis; a possible biomechanical explanation was suggested in Di Benedetto (1989) by relating it to anticipatory coarticulation effects on vowels dependent on postvocalic place of articulation.

The perceptual data presented in this study and in the acoustic analysis described in Di Benedetto (1989) support an hypothesis advanced by Strange *et al.* (1976) that vowels as well as consonants might not be fully characterized by target frequencies alone. Strange *et al.* (1976) and Strange and Gottfried (1980) found that vowels in isolation were more poorly identified than in [p–p] environments. This effect was also observed when the consonant frame varied unpredictably (in this case, the vowels were considered in the context of stop consonants). Strange's conclusions were that, from a perceptual point of view, the formant contours (formant frequency variations and total vowel duration) were much more relevant than the formant patterns sampled in a single temporal cross section. In another study, Gottfried and Strange (1980) confirmed that vowels in the context of stop consonants were better perceived than in isolation. They ruled out the hypothesis that phonological explanations could account for the results, confirming that the effect was due to the richer acoustic information contained in vowels in consonantal contexts. The only exceptions were the syllables with [g] initial or final. No explanation based on the results of the present study could be found to justify this finding. Strange *et al.* (1983), Rakerd *et al.* (1984), and Verbrugge and Rakerd (1986) confirmed the general results. Strange *et al.* and Rakerd *et al.* found that the advantage of having the vowel in consonantal context was greater for open vowels (except for [ɔ]) than for closed vowels. The experiments of the present study dealt with the perception of high vowels versus nonhigh vowels, but no low vowels were considered; thus, they only confirm that, for vowels characterized by the features [+ high] or [+ high, − low], spectral and temporal changes of $F1$ might provide information to the listener about vowel identity.

We identified two properties, one associated with the $F1$ onset frequency and the other with the relative $F1$ onglide duration with respect to total vowel duration, which appeared to be acoustically (Di Benedetto, 1989) and perceptually relevant. This research should be considered as a first attempt of specifying acoustic invariants of vowels, which would account for spectral and temporal variations of $F1$. In future studies, in which temporal and spectral properties will not be confounded, we might be able to determine whether the temporal or the spectral characteristics might account alone for the effects observed.

ACKNOWLEDGMENTS

I am grateful to Paolo Mandarini for his generous support and advice, and I wish to express my deep gratitude to Ken Stevens for his unique help, criticism, and guide throughout this study.

Brady, P. T., House, A. S., and Stevens, K. N. (**1961**). "Perception of sounds characterized by a rapidly changing resonant frequency," J. Acoust. Soc. Am. **33**, 1357–1362.

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

Di Benedetto, M. G. (**1989**). "Vowel representation: Some observations on temporal and spectral properties of the first formant frequency," J. Acoust. Soc. Am. **86**, 55–66.

Ferrero, F. E., Magno-Caldognetto, E., Vagges, K., and Lavagnoli, C. (**1975**). "Some acoustic characteristics of the Italian vowels," Eighth International Congress of Phonetic Sciences, Leeds, England.

Gottfried, T. L., and Strange, W. (**1980**). "Identification of coarticulated vowels," J. Acoust. Soc. Am. **68**, 1626–1635.

Huang, C. B. (**1985**), "Perceptual correlates of the tense/lax distinction in general American English," Master's thesis, MIT, Cambridge, MA.

Huang, C. B. (**1986**). "The effect of formant trajectory and spectral shape on the tense/lax distinction in American vowels," IEEE-ICASSP, Tokyo, Japan.

Klatt, D. H. (**1980**). "Software for cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

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

Klatt, D. H. (**1985**). Personal communication.

Lindblom, B., and Studdert-Kennedy, M. (**1967**). "On the role of formant transitions in vowel recognition," J. Acoust. Soc. Am. **42**, 830–843.

Lindblom, B., Heinz, J. M., and Lindquist, J. (**1967**). "Patterns of residual masking for sounds with speech-like characteristics," R. Inst. Technol. Stockholm, Q. Prog. Stat. Rep. **2-3**.

Neter, J., and Wassermann, W. (**1974**). *Applied Linear Statistical Models* (Irwin, Momewood, IL), pp. 329–338.

Rakerd, B., Verbrugge R. R., and Shankweiler, D. P. (**1984**). "Monitoring for vowels in isolation and in consonantal context," J. Acoust. Soc. Am. **76**, 27–31.

Stevens, K. N. (**1959**). "The role of duration in vowel identification," Q. Prog. Rep. **52**, Res. Lab. Electron. MIT, Cambridge, MA.

Stevens, K. N. (**1980**). "Acoustic correlates of some phonetic categories," J. Acoust. Soc. Am. **68**, 836–842.

Stevens, K. N. (**in press**). *Acoustic Phonetics.*

Strange, W., Verbrugge, R. R., Shankweiler, D. P., and Edman, T. R. (**1976**). "Consonant environment specifies vowel identity," J. Acoust. Soc. Am. **60**, 213–222.

Strange, W., and Gottfried, T. L. (**1980**). "Task variables in the study of vowel perception," J. Acoust. Soc. Am. **68**, 1622–1625.

Strange, W., Jenkins, J. J., and Johnson, T. L. (**1983**). "Dynamic specification of coarticulated vowels," J. Acoust. Soc. Am. **74**, 695–705.

Verbrugge, R. R., and Rakerd, B. (**1986**). "Evidence of talker-independent information for vowels," Lang. and Speech **29**, Part 1, 39–57.