Difference limens for formant frequencies of steady-state and consonant-bound vowels[a)]

Paul Mermelstein

Haskins Laboratories, New Haven, Connecticut 06511
(Received 25 April 1977; revised 9 September 1977)

Difference limens (DL) of formant frequencies were measured for two steady-state vowels and the same vowels in symmetric stop-consonant contexts. The stimuli were generated using a computer-programmed synthesizer, and the formant-frequency parameters were adjusted to be steady or symmetric transition functions around the temporal center of the syllable. The DL for the time-varying consonant–vowel–consonant (CVC) stimuli were found to be significantly larger than those for the steady-state vowels. In some cases the DL for the second formant was found to be larger in the direction of expected formant shift due to consonantal coarticulation than in the reverse direction. For CV or VC stimuli the increase in vowel-formant DL is reduced. The difference in DL values in and out of context has, at least partially, an auditory origin. However, the phonetic decoding of the CVC stimuli may also contribute to the loss of vowel-quality information.

PACS numbers: 43.70.Dn, 43.70.Jt

INTRODUCTION

Difference limens (DL) for steady-state vowels reflect the ability of the human auditory system to differentiate complex stimuli having stationary spectral patterns. The speech signal, however, is rarely stationary in its spectral composition for any length of time. The time-varying formant patterns form significant cues for the extraction of phonetic information. It is of interest, therefore, to explore the effects of spectral variation on the just noticeable differences (JND) in formant frequencies. The JND's measured in consonantal context can be expected to be better indicators of formant-frequency discrimination in continuous speech.

Differences in perception of vowels in and out of context have been previously noted (Stevens *et al.*, 1966). Discrimination of changes in vowel quality for vowels presented in isolation has been found to be equally acute within a phoneme region or across a phoneme boundary (Fry *et al.*, 1962). Yet vowels in context tend to be perceived in a categorical fashion; that is, they possess discrimination functions that are characterized by peaks at the phoneme boundaries (Stevens, 1968).

The identification of vowels in context is a complex perceptual task. The effects of context are to introduce different time-dependent formant variations in the vocalic segments. In the presence of such variations, the vowel-category boundaries are displaced as well (Lindblom and Studdert-Kennedy, 1967). Acoustic data on vowel reduction (Stevens and House, 1963) reveal that the characteristic formant frequencies of vowels in isolation are not attained in dynamic speech contexts. The consonantal context always has the effect of shifting F_2 from a value appropriate for the null environment toward a more central position. The shift in F_2 for labial and alveolar stops is generally of the order of 200 Hz; however, for the rounded vowel /u/ it was as much as 350 Hz. It has been suggested that during the perception of vowels in consonantal contexts a corresponding compensation is made for undershoot in vowel articulation (Lindblom and Studdert-Kennedy, 1967; Stevens, 1968). The perceptual significance of such undershoot should be assessed with the aid of DL data on the vowels in context. If the measured shift is equal to the DL value for the same proportional change in the various formants, only 50% of the shifted tokens will be judged different relative to the tokens where the formant frequencies have not been shifted.

Speech coding systems extract a set of time-varying parameters from the speech signal at the source, transmit the values of these parameters, and reconstitute the signal at the destination. The DL's of these parameters represent the maximum permissible transmission errors if the signal is not to be noticeably degraded in transmission. The bandwidth requirements of the encoded signal are therefore directly dependent on the DL's of the parameters. Any significant increase in the DL as the size of the context is increased would suggest that larger signal segments be selected for encoding as a unit. Such procedures offer potential bandwidth savings in the transmission of the individual parameters beyond those due to the inherent limitations on the time variations of these parameters.

In measuring the DL for vowel formants we attempt to exclude explicit labeling of the speech stimuli from the perception process by focusing on discrimination differences irrespective of their origin. Identification of stimuli must be based on discriminable differences; therefore, the DL values are of interest as indicators of human speech perception performance. We have been primarily interested in the relative DL values at various points in the formant-space for vowels in and out of context, rather than in the absolute DL values which are of greater interest in designing speech communication systems (Flanagan, 1955).

Strange *et al.* (1976) report better identification of vowels spoken in a consonantal environment than in iso-

[a)]This paper was presented in part at the 92nd meeting of the Acoustical Society of America, San Diego, CA 16–19 November 1976. [J. Acoust. Soc. Am. 60, S119(A) (1976)].

 0001-4966/78/6302-0572$00.80

lation. They interpret these results as evidence that dynamic acoustic information, distributed over the temporal course of the syllable, is used by the listener to identify vowels. We may ask whether the presence of transitions *per se*, without regard to the dependence of the transition on the vowel, assists in the discrimination of the vowels. When vowels are differentiated on the basis of spectral information alone, the addition of transitions to and from fixed points in acoustic space provides a relatively invariant consonantal frame. The transitions employed are equal in duration and differ only in the vowel formants which they meet. If such limited transitional information is useful for vowel discrimination, then listeners should discriminate synthesized consonant-embedded vowels better than their steady-state counterparts. If such improved discrimination is not found, the more accurate identification of vowels in context must depend on additional factors such as duration associated either with the relatively stationary vowel segment, or with the transitions, or with both. Of course, the model for CVC productions of a vowel segment surrounded by transitions is incomplete. In many productions no stationary segment may be identifiable or, as in diphtongization, the vowel segment is not stationary.

Stevens (1968) reported improved discrimination for isolated vowels over CVC stimuli. The CVC stimuli used in his study, however, varied both in steady-state formant frequency and in duration. The improved discrimination, therefore, may have been due to the additional temporal differences among the CVC stimuli. Our study focuses on discrimination in formant frequencies alone and forms an extension of Stevens' results to different vowels and contexts.

We report in this paper a series of five experiments comparing the DL of vowel formants for the vowels in isolation and in context. Experiment 1 explored DL values in the /ɪ/ vowel region. In experiment 2, the reference formants were those near the boundary between /ɛ/ and /æ/. Experiment 3 compared discrimination of CVC versus V over the $F_1 = 600$ Hz line and covered four different vowel regions. The fourth experiment limited the context to CV and VC syllables. In experiment 5 we attempted to restrict the changes in DL to phonetic processing effects and to eliminate auditory contributions by employing CVC's with nonvarying F_2 frequencies, and allowing the consonantal information to be encoded in the time variations of F_1 and F_3.

I. STIMULI

Steady-state vowels and CVC syllables were generated using a software formant synthesizer modeled after the one presented by Rabiner (1968). The first three formants were adjustable under program control; the fourth and fifth formants were fixed at 3500 and 4500 Hz, respectively. Bandwidth values were fixed at 60, 80, 100, 175, and 281 Hz, respectively. The stimuli were entirely voiced and consisted of steady or changing formant patterns. In experiments 1 and 2 the formant trajectories for the CVC syllables followed symmetric

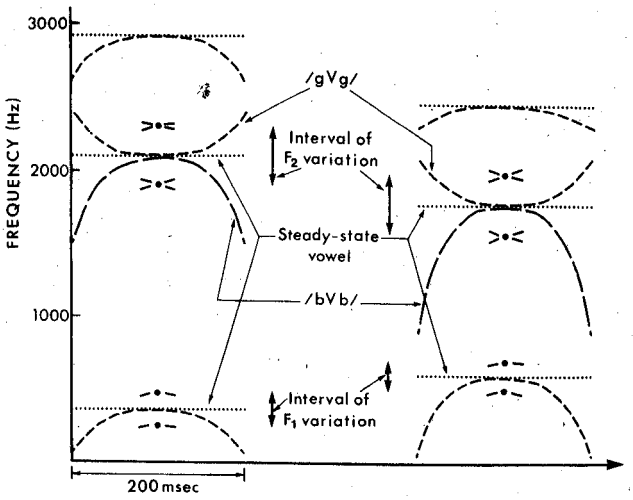

FIG. 1. Formant trajectories for steady-state vowels and consonant–vowel–consonant stimuli. Left—/ɪ/, /bɪb/, /gɪg/ series; right—/ɛ/, −/bɛb/, −/gɛg/ series. Unconnected points at syllable centers indicate ranges of F_1 and F_2 variation.

cubic functions of the time differences from the temporal center of the stimulus (see Fig. 1). In experiment 4 the transitions were linear functions of time. In experiments 3 and 5 they were parabolic functions of time.

The paired stimuli consisted of a standard and a variable counterbalanced to eliminate any order effects. The variable stimuli had central formant frequency values differing in F_1 (± 25-Hz steps) or F_2 (± 50-Hz steps). Differences of one, two, three, and four steps were used to modify each variable up or down to cover the DL range expected on the basis of the data given by Flanagan (1955). Two symmetric consonantal contexts were used—/bVb/ and /gVg/. These differed only in the trajectory of the second formant frequency. All of the stimuli were 200 ms in duration except for an additional shorter steady-state series of 133-ms duration that is discussed further below. The formant frequency variations are illustrated in Fig. 1. The steady-state vowels had a steady fundamental frequency of 125 Hz except for a drop to 100 Hz over the last 50 ms. The fundamental frequency of the CVC stimuli followed the same cubic trajectory as the formant frequencies and covered the same range, namely 125 Hz at the center and 100 Hz at the initial and terminal points. The different fundamental frequency variations chosen for the CVC and V stimuli contributed somewhat to an enhanced naturalness for each group.

In view of Stevens' (1968) results that vowels in consonantal context near the phoneme boundary are discriminated better than vowels with formant values well within the phoneme category, we used two separate standard stimuli as given in Table I. In the first experiment, we used values appropriate for the vowel /ɪ/, in the second, a value near the boundary between /ɛ/ and /æ/ (Peterson and Barney, 1952). Through use of different standards, we attempted to explore what effects, if any, proximity to a phoneme boundary may have on the DL values.

TABLE I. Central and terminal formant frequencies for the standard stimuli of experiments 1 and 2.

			Formant frequencies (Hz)		
Experiment 1	Vowel /ɪ/	(center)	350	2100	2900
	Stop /b/	(terminal)	50	1500	2000
	Stop /g/	(terminal)	50	2400	2600
Experiment 2	Vowel /ɛ/—/ɛ/	(center)	600	1780	2450
	Stop/b/	(terminal)	50	900	2000
	Stop /g/	(terminal)	50	2100	2300

II. SUBJECTS

The subjects were volunteer students from Yale University and the University of Connecticut and were paid $2.00 per hour for their services. Five subjects took part in the first experiment, and six subjects participated in the second experiment. An additional group of five subjects, colleagues and students at Haskins Laboratories, took part in the identification experiments. Further groups of four subjects each acted as listeners in experiments 4 and 5. All subjects were native speakers of English.

III. PROCEDURE

Subjects heard pairs of stimuli 0.4 s apart, separated by 1.6 s of silence from the next pair. They were asked to judge whether each pair consisted of the same or different stimuli. The "same/different" response paradigm was employed to simplify the task and avoid memory-related problems that one encounters with ABX discrimination (Pisoni, 1971; Fujisaki and Kawashima, 1970). Some pairs consisted of the standard presented twice; others consisted of the standard and variable in either order. In experiments 1 and 2 steady-state vowel pairs, and /bVb/ and /gVg/ syllable pairs were all randomized within one grand list. In experiment 3 we employed V and /bVb/ stimuli; in experiment 5 V and /gVg/ stimuli were used. In experiment 4 V, /bV/, and /Vb/ stimulus pairs were randomized together. Each stimulus pair occurred at least five times in each order.

Subjects were allowed to listen to trial pairs of stimuli until they felt comfortable with the quality of the synthetic speech. Roughly ten pairs usually sufficed. Each experiment lasted approximately one-half hour and was divided into two parts by a short rest period. The first experiment contained a total of 336 pairs, the second 360 pairs. Just noticeable differences were determined for each of the three contexts and each individual subject separately.

The fraction of stimuli reported different for any variable increment was adjusted for guessing. We assumed that the probability of a guess that the stimuli are different when in fact they are the same is given by $p(d/\Delta=0)$, the fraction of different responses for the "same" stimulus pair. Following Swets (1964), the apparent probability of discrimination at some formant-frequency difference value Δ is given by

$$p(d/\Delta)=p'(d/\Delta)+p'(d/\Delta=0)[1-p'(d/\Delta)],$$

where $p'(d/\Delta)$ is the true or corrected discrimination probability and $p(d/\Delta=0)$ is the guessing probability. After the data were corrected for guessing, sigmoid curves were fitted to the data (Finney, 1964), and the points of 50% discriminability and their standard errors were estimated.

IV. RESULTS

The DL values for experiment 1 are given in Table II. The direction above or below the center formant frequency in which the first or second formant was moved did not appear to affect the results significantly; therefore, the data are pooled and tabulated according to the formant varied. The mean DL for F_1 variations is 50 Hz for the steady-state vowels and 49 Hz for the consonantal context. For none of the subjects is the difference in F_1 discriminability due to context significant. The mean DL for F_2 variations is 142 Hz for the steady-state vowel, 174 and 199 Hz, respectively, for the /b/ and /g/ contexts. The average increase in DL value is 45 Hz, or 31% over the DL value for the steady-state vowel. All the DL values in consonantal context are higher than the corresponding values for the same subject for the steady-state vowel. Six of the ten differences are significant at the 0.01 level. Significant variations are noted among the data of the various listeners. Note particularly the low DL values for subject 2 and the high values for subject 4 in all contexts.

Is the information provided by the separate formants additive? To explore the effects of simultaneous variation in both the first and the second formant frequencies, variable stimuli were included in the DL tests that were constrained to lie along the line in formant-space given at points such that $\Delta F_2=2F_1$, $\Delta F_2=\pm 25$, 50, 75, and 100 Hz, respectively. The DL values for F_2 under these conditions are also given in Table II for each context.

TABLE II. Difference limen and standard error values (Hz) for listeners to stimuli of experiment 1.

	Subject					
	No. 1	No. 2	No. 3	No. 4	No. 5	Mean
Variation in F_1						
Vowel alone	74±8	36±5	47±6	56±6	39±6	50
/b/ context	61±8	40±5	41±6	57±7	29±4	46
/g/ context	64±6	35±2	46±7	56±6	53±5	51
Variation in F_2						
Vowel alone	166±20	79±15	174±34	164±19	125±14	142
/b/ context	207±46	136±15	222±52	181±12	126±13	174
/g/ context	245±69	109±14	208±38	230±73	206±49	199
$\Delta F_2=2\Delta F_1$						
Variation in F_2						
Vowel alone	66±11	63±11	54±11	76±10	63±14	64
predicted	111	53	83	93	66	81
/b/ context	142±28	66±11	80±14	88±12	63±14	88
predicted	105	69	77	96	53	80
/g/ context	101±21	53±11	80±16	78±19	63±14	75
predicted	114	59	80	101	86	88
Short vowel						
F_1 variation	46±8	30±7	54±10	38±5	42±10	42
F_2 variation	113±22	63±14	120±30	134±25	71±21	100

TABLE III. Difference limen and standard error values (Hz) for listeners to stimuli of experiment 2.

	Subject						
	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	Mean
$+F_1$ variation							
s–s vowel	37±6	29±7	32±4	33±5	39±5	29±4	33
/b/ context	67±7	174±112	41±5	75±6	83±15	52±6	82
/g/ context	83±7	66±9	38±7	81±6	72±13	56±8	66
$-F_1$ variation							
s–s vowel	30±5	40±8	29±5	42±6	37±4	27±4	34
/b/ context	55±5	82±10	39±7	77±8	96±16	85±13	72
/g/ context	41±6	87±17	31±5	88±11	56±8	60±8	61
$+F_2$ variation							
s–s vowel	62±12	115±18[a]	60±9[a]	136±19[a]	107±17[a]	70±12	92
/b/ context	182±19	205±42	114±13	182±19	189±38	202±36	179
/g/ context	161±50	182±35[a]	113±18[a]	149±17	186±37	156±16[a]	158
$-F_2$ variation							
s–s vowel	58±10	35±7	35±7	93±14	59±12	68±11	58
/b/ context	245±80	269±98	125±14	233±52	242±72	261±82	229
/g/ context	180±20	105±15	65±10	151±14	130±19	92±10	120

[a]Significant difference with respect to DL in $-F_2$ direction at the 0.01 level.

Let us assume a simple model that considers F_1 and F_2 to be independent parameters that contribute information to discriminability given by

$$(w_1(\Delta F_1)^2 + w_2(\Delta F_2)^2)^{1/2},$$

where w_1 and w_2 are appropriate weighting factors. This model results in DL regions in formant space composed of elliptical sections. On the basis of DL measurements for F_1 and F_2 variation alone, this model allows prediction of the DL when both parameters are varied simultaneously. The predicted DL values for F_2 under the constraint that $F_2 = 2\Delta F_1$ are also given in Table II. The measured DL values appear to be reasonably well predicted from the DL values for independent formant variation when F_1 and F_2 are not located close to each other.

One hypothesis that may account for the increased DL for the consonant-embedded vowels is that the time integral of the difference between the formant trajectories of the time-varying stimuli differing in a given F_2 value is less than the same difference for the steady-state stimuli. One may argue that if the temporal duration of a given difference is reduced, the discriminability may also be reduced. However, Liang and Chistovich (1971) found no substantial difference in the DL value for steady tones near 1 kHz between duration values of 100 and 200 ms. The duration had to be reduced below 100 ms in order for an increase in DL to be manifested. To assess the duration effects for speechlike stimuli, the DL was also measured for 133-ms-long steady-state vowels around the same point in formant space. No increase in DL value due to the shorter stimulus duration was noted for any subject. In fact, a decrease was measured in the second-formant DL value relative to the longer vowels that is significant at the $p < 0.01$ level for three of five subjects. We know no good explanation for the causes of this reduction in DL values accompanying a reduction in stimulus duration. We may conclude, however, that the decreased discriminability for consonant-embedded vowels is not due to the decreased average duration of the formant differences.

The results of the second experiment are shown in Table III. Here differences were apparent in the DL values determined for variations in the positive and negative directions of the formant frequencies. Therefore, the data for these directions are shown separately. For each of the four directions of variations and for all contexts, the DL's in consonantal context are larger than the DL's for the steady-state vowels. All but one of the 48 differences are significant at the 0.01 level. The average DL for F_1 is 33 Hz for steady-state vowels, 70 Hz in consonantal context. The average DL for F_2 is 75 Hz for steady-state vowels, and 171 Hz in consonantal context. The average increase in each case exceeds 100%. The mean DL data for each vowel and context are summarized in Fig. 2.

An additional experiment (experiment 3) was carried out to test the comparative discriminability of vowels in isolation and in CVC context over larger region of the vowel space. Stimuli were regularly spaced along a line in the space of the first two formant frequencies such that F_1 was constant at 600 Hz, and F_2 ranged from 1 to 2 kHz in 50 Hz increments. This allowed generation of vowels varying through /ɔ, ʌ, ɸ, ɛ, and æ/. The consonant in experiment 3 was always /b/, achieved by using initial and terminal second-formant frequency values ranging from 500 to 1000 Hz in 25 Hz steps. The initial and terminal F_1 frequency was always 50 Hz. The F_3 pattern was unchanged from experiment 2. All formant trajectories were parabolic in shape and had a total duration of 200 ms. The corresponding steady vowels had the same duration.

In Fig. 3 we have plotted the average discrimination by four subjects of pairs of stimuli spaced three steps,

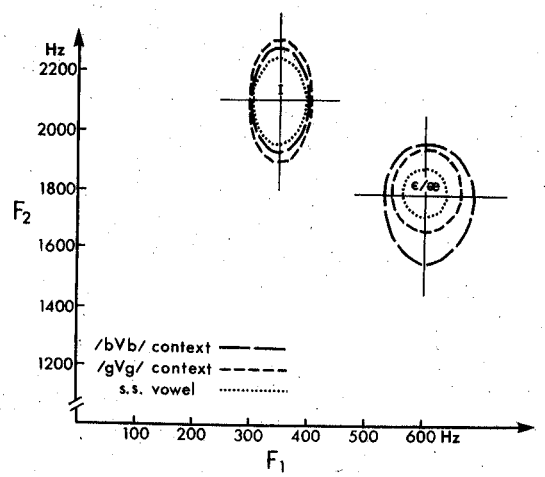

FIG. 2. Just noticeable difference regions for vowels in and out of consonantal context. The results are mean values for five subjects (/ɪ/ group) and six subjects (/ɛ/ group), respectively.

or 150 Hz in the central F_2 value. The short bars indicate the one standard deviation limit in each case. The overall average discrimination is 83 ± 14% for the vowel alone, but only 26 ± 22% for the vowel in the CVC context. The loss in discrimination due to the consonantal context is clearly demonstrated.

Experiment 4 compared the DL values determined for F_2 variations for /bV/ and /Vb/ syllables with the same values for the corresponding vowels. The syllabic stimuli were synthesized with 48-ms linear formant transitions preceding or following 96-ms steady-state segments. All duration values were integral multiples of the fundamental voicing period of 8 ms. The steady-state segments alone constituted the vowel stimuli. Four new subjects acted as listeners in this experiment and heard eight presentations of stimulus pairs differing by 0, ± 50, ± 100, ± 150, and ± 200 Hz from the F_2 of the standard. Results for these positive and negative excursions about the $F_2 = 1780$ value at an F_1 value of 600 Hz are given in Table IV.

Twelve of the 16 CV or VC versus V comparisons yielded positive DL differences significant at the 0.01 level. In most cases even a single prevocalic or postvocalic transition suffices to increase the vowel-formant DL. The differences between DL values between CV and VC syllables are not generally significant. Also, the DL differences remain even when all fundamental frequency variations have been eliminated.

TABLE IV. Difference limen and standard error values for listeners to stimuli of experiment 4.

	Subject				
	No. 1	No. 2	No. 3	No. 4	Mean
$+F_2$ variation					
V	140 ± 16	132 ± 17	93 ± 14	82 ± 13	112
CV	176 ± 11[a]	171 ± 15[a]	139 ± 13[a]	175 ± 15[a]	165
VC	160 ± 16	192 ± 36[a]	169 ± 11[a]	155 ± 18[a]	169
$-F_2$ variation					
V	103 ± 13	164 ± 27	54 ± 8	92 ± 15	103
CV	163 ± 14[a]	93 ± 19	73 ± 8[a]	125 ± 11[a]	114
VC	156 ± 15[a]	131 ± 13	77 ± 8[a]	105 ± 11	117

[a]Difference limen of CV and VC significantly larger than that for corresponding V at the 0.01 level.

TABLE V. Difference limen and standard error values for listeners for vowels and CVC stimuli with stationary F_2 patterns (experiment 5).

	Subject				
	No. 1	No. 2	No. 3	No. 4	Mean
$+F_2$ variation					
Vowel	44 ± 8	47 ± 8	32 ± 7	32 ± 7	39
CVC	101 ± 11[a]	52 ± 10	67 ± 10[a]	43 ± 10	66
$-F_2$ variation					
Vowel	89 ± 10	67 ± 9	42 ± 8	32 ± 7	58
CVC	112 ± 18[a]	138 ± 17[a]	177 ± 28[a]	102 ± 16[a]	132

[a]Difference with respect to corresponding DL for the vowel alone significant at the 0.01 level.

The average magnitude of the DL difference for CV or VC versus V is significantly smaller than the difference measured for similar CVC syllables. Admittedly, different groups of subjects took part in the two experiments and the stimuli employed were not precisely comparable. In experiment 4 we used linear transitions and 100-ms vowels, in experiment 2 we used parabolic transitions and 200-ms vowels. Nevertheless, the results suggest that CV or VC syllables occupy an intermediate position in the contextual hierarchy in that they increase the vowel DL's less than the CVC syllables.

Is there an increase in the DL even when the target formant is stationary and the consonantal information is supplied entirely by the other formants? Experiment 5 compared the perceived DL of F_2 for CVC stimuli with stationary F_2 patterns, but time varying F_1 and F_3, with vowel stimuli having F_1, F_2, and F_3 stationary. The vowel stimuli included a standard with F_1 at 600 Hz, F_2 at 1780 Hz, and F_3 at 2600 Hz; the CVC standard had terminal F_1 and F_3 values of 50 and 2300 Hz, respectively. As before, F_2 variations of ± 50, ± 100, ± 150, and ± 200 Hz were employed. Each subject responded to a total of 160 pairs of stimuli. Table V gives the DL

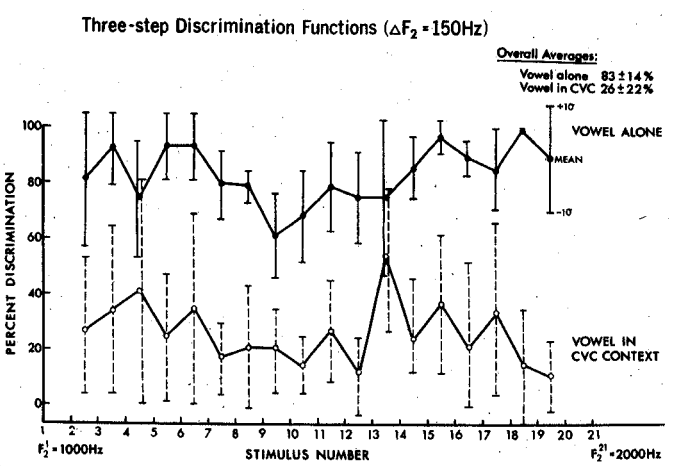

FIG. 3. Three-step discrimination functions for vowels alone and in CVC context.

results obtained. Six of the eight comparisons (two F_2 directions and four subjects) yield significant DL increases for the CVC stimuli. The DL increase for positive F_2 variation was found to be significantly smaller than the DL increase for negative variation. This asymmetry in the DL value may result from consonantal quality differences induced by the increased proximity of F_2 to the terminal F_3 values when F_2 was varied in the positive direction. Although the frequency values were selected so as to leave consonant identity unchanged at /g/ for the entire range of F_2 variation, slight quality changes could not be avoided. Nevertheless, the overall increase in DL in CVC context is reduced compared to the more natural CVC stimuli having time-varying F_2 patterns used in the first four experiments. We attribute this reduction to the elimination of the auditory contribution. The residual DL increase appears to result from the structure information carried by the other formants. Since this information alters the phonetic interpretation of the steady second formant, we tentatively describe the effects as a phonetic contribution to the increased DL.

V. DISCUSSION

Comparison between experiments 1 and 2 reveals a decrease in the steady-state vowel DL for F_1 as the frequency of F_1 increased. Similarly, there was a decrease in the steady-state vowel DL for F_2 with the decreasing frequency of F_2. These agree with the decrease in DL found by Flanagan (1955) when the formant amplitude increases due to the increased proximity of the two formants.

Flanagan obtained a DL in F_1 of 15 Hz at 350 Hz. Our DL result of 50 Hz at 300 Hz appears to be significantly higher. At 600 Hz our result of 33 Hz is comparable to the 25 Hz Flanagan obtained at 500 and 700 Hz. As seen in Table II, the measured values depend greatly on the listener. An additional factor may be the fact that at an F_1 of 300 Hz the second and third harmonics of the 120-Hz fundamental are equally excited. Discrimination of formant differences in such regions may be poorer than when the formant coincides with a maximally excited harmonic.

When we compare the sets of DL values measured in the two experiments, we find that the values in the consonantal context are larger for both F_1 and F_2 in experiment 2. The relative increase in DL may be due to the more extensive formant transitions employed in the second experiment. The lack of a significant difference in the DL for F_1 in experiment 1 in and out of context may be due to the small range of F_1 transition there.

We noted an interesting tendency for asymmetry in the DL values with the direction of variation of F_2. For the steady-state vowels, the data of four of six subjects revealed significantly larger (0.01 level) positive DL values than negative DL values. For the /gVg/ stimuli three of six subjects showed significantly larger DL values in the same direction. For the /bVb/ the data show a trend in the opposite direction, namely larger DL values in the negative F_2 direction. However, since the largest F_2 difference included in the stimuli was 200 Hz, when the DL estimates exceed this value the error of the estimate may also be large. Thus we could not establish the significance of the directional asymmetry for the /bVb/ stimuli. F_2 variations in /b/ context are concave downward, F_2 variations in /g/ context are concave upward. In both cases the larger DL value is in the direction of the terminal F_2 value.

These increases in discriminability might simply be the usual increases observed at a phoneme boundary, but at a boundary that shifts with consonantal context due to perceptual compensation for articulatory undershoot. Thus a shift in the /ɛ/–/æ/ boundary to a point above the central F_2 frequency for the /g/ context, but to a point below the central F_2 frequency for the /b/ context would be in a direction consistent with the findings of Lindblom and Studdert-Kennedy (1967). Moving F_2 toward the perceptually adjusted boundary might then result in a lower DL than moving it away from the boundary. If this hypothesis were correct, the stimuli of the second experiment should yield a clear /ɛ/–/æ/ boundary near the formant values of the standard. To test for this, a vowel identification experiment was run with five additional subjects. (The subjects used in the original discrimination tests of experiment 2 were no longer available.) The stimuli used were those of experiment 2 which varied in the central F_2 value in steady state, in /b/, and in /g/ contexts. Subjects were instructed to identify these stimuli as /ɛ/, /æ/, or "other vowel." Within the limits of the variation among the stimuli, no crossover was observed for any subject between a majority of /æ/ and a majority of /ɛ/ observations along any of the three stimulus continua. The asymmetry in /ɛ/–/æ/ identification can be expressed as

$$(N_{\varepsilon}^{+} - N_{\text{æ}}^{+}) - (N_{\varepsilon}^{-} - N_{\text{æ}}^{-})/(N_{\varepsilon}^{+} + N_{\varepsilon}^{-} + N_{\text{æ}}^{+} + N_{\text{æ}}^{-}),$$

where $N_{\text{æ}}^{+}$ is the number of positive F_2 excusion stimuli identified as /æ/, N_{ε}^{+} the number of positive F_2 excursion stimuli identified as /ɛ/, and $N_{\text{æ}}^{-}$ and N_{ε}^{-} are defined similarly for the negative F_2 excursion stimuli. A perfect boundary at the formant values of the standard would have resulted in a ratio of 1. For no subject did this ratio exceed 0.20. Therefore it is unlikely that, for a fixed stimulus duration of 200 ms, the 400-Hz range of F_2 variation was sufficient to effect significantly the probability of /ɛ/ versus /æ/ judgment. The difference in discriminability between the positive and negative directions of formant variation observed in experiment 2 appears not to be due to perceptual shifts in vowel boundary as a function of consonantal context.

The results of the third experiment indicate that discrimination of the steady-state vowel series is clearly continuous. Two minor peaks in the CVC discrimination curve do appear at 1175 and 1625 Hz, and these roughly correspond to the /ɔ/–/ʌ/ and /ʌ/–/ɛ/ boundaries as observed from the identification of the same stimuli. However, discrimination is not sharply categorical, and the identification curves of the different vowels show significant overlap. Furthermore, discrimination does not vary appreciably with the central F_2 value as that value is changed from 1000 to 2000 Hz. Thus, unlike the case for single formant stimuli, it

does not appear that the critical bandwidth of the auditory system, which increases as we traverse this frequency range, is a limiting factor in establishing the difference-limen value.

VI. CONCLUSIONS

The overall mean DL value in CVC context in the first two experiments encompassing 11 different listeners was 60 Hz for F_1 and 176 Hz for F_2. In experiment 3 a change in F_2 of 150 Hz could be discriminated on the average only 26% of the time. These results question the perceptual significance of most of the F_2 shifts due to consonantal context observed by Stevens and House (1963). Only when the formant shift exceeds the DL can the listener be expected to invoke generative rules to correct for context-induced formant shift. Relatively greater significance must be assigned to the measured F_1 shifts than to the F_2 shifts.

Formant-frequency differences that are discriminated among steady-state vowels are not discriminated in the presence of consonantal transitions. We interpret this result as due to a stage in processing the incoming speech signal where some information concerning formant differences is lost. Two mechanisms may account for this result.

First, we may consider the result as purely auditory in origin. If steady-state vowels are mapped into a unique excitation pattern on the basilar membrane, any time variation of the vowel spectrum can be expected to broaden the excitation pattern and thus add uncertainty to the spectral component of perceived vowel quality.

Tsumura, Sone, and Nimura (1973) carried out experiments on the just noticeable differences between two steady portions of a tone burst separated by rising or falling transitions. The threshold frequency difference increased roughly logarithmically with decreasing transition duration for transitions shorter than 50 ms. For tones near 1 kHz in frequency, the threshold was found to be greater when the frequency transition occurs near the burst onset, than if it occurs near the cutoff. The threshold was generally higher for a rising transition than for a falling transition. Evidently frequency transitions may exert an auditory masking effect on steady tones. The frequency transitions in our stimuli (as often in speech) were significantly larger than those used in the Tsumura study; so that masking effects of significantly greater magnitude may occur in speech than those reported for tones. These effects would tend to increase the just noticeable difference (JND) of the formant frequencies in context above those that hold for steady-state vowels.

Second, the effect may be due to the phonetic decoding that must follow the auditory processing of the CVC stimuli. Liberman, Mattingly, and Turvey (1972) suggest that phonetic decoding of consonants "strips away auditory information" in short-term memory. A similar process may be involved in the decoding of vowels, though perhaps at a slower rate. Discrimination in consonantal context is still quite good despite the fact that the vowels belong to the same phoneme category. Yet it is not as good as that observed for the steady-state vowels. A partial loss of information has taken place.

Experiment 4 was designed to test vowel discriminability in the simplest linguistic context, namely, when the vowel is accompanied by just one prevocalic or postvocalic stop. In comparing the DL value for CV and VC stimuli, we were testing the retention of vowel-quality information as a function of the position of the interstimulus pause with respect to the vowel segment. For CV syllables the pause followed the vowel segment directly; for VC syllables the postvocalic transition intervened. If purely auditory effects such as forward and backward masking were the main contributors to the increase in DL, for the CV syllables forward masking would be expected to predominate and for the VC syllables backward masking would be more important. In each case the 400-ms intersyllabic pause would tend to prevent masking across the stimuli. Extrapolating from results on recognition masking of tones (Massaro, 1973), we would expect backward masking to have more severe effects on the vowel DL than forward masking. Since there was no significant difference between the DL values for VC and CV syllables, forward and backward masking appear unlikely to be the major causes of the increase in DL. Furthermore, a postvocalic pause does not facilitate retention of vowel-quality information as against a postvocalic stop. These results argue for the existence of phonetic processes which incorporate temporally parallel decoding of vocalic and consonantal information. The presence of a consonant, whether it occurs before or after the vowel segment, worsens equally the retention of vowel-quality information. The amount of interference appears to be related to the complexity of the syllabic stimulus. The presence of both a prevocalic and postvocalic consonant affects discrimination more than one consonant alone. We may hypothesize that when the amount of ongoing phonetic processing is increased, a correspondingly higher loss of information regarding vowel quality is incurred. Further research in this direction may result in methods for objective assessment of phonetic complexity.

Comparison of the DL results for vowels in experiments 2, 4, and 5 reveals significant variations over the separate groups of listeners. Therefore we will compare only the relative increases in DL within the various experiments. The average increase in the DL of F_2 for CV and VC stimuli with respect to the corresponding vowel (experiment 4) alone was 34 Hz. The average increase in DL for the CVC with flat F_2 trajectories (experiment 5) was 51 Hz. The largest average increase, 97 Hz, was noted for the CVC with time-varying F_2 patterns. The relative magnitudes of the DL changes obtained allow us to conclude that both auditory effects (time varying versus stationary formant patterns) and phonetic effects (perception of one or more consonants) contribute to increased values for the DL of vowel formants. The explanation in terms of phonetic effects is a tentative hypothesis and further experiments are needed to delimit the environments, both speechlike and nonspeechlike, where such DL increases can be observed.

The improved identification for natural vowels in con-

text over those in isolation obtained by Strange *et al.* (1976) is apparently not due to improved formant frequency discrimination. Our data imply that on the basis of frequency information alone, steady-state vowels are better discriminated than consonant-embedded vowels. However, if temporal information concerning the differences in the transitions were used in the discrimination task in addition to the formant-frequency differences, the consonant-embedded vowels might be better discriminated. Durational information, in particular, was missing from the /e/–/æ/ stimulus series. Such durational information, which is regularly used when interpreting continuous speech (Peterson and Lehiste, 1960; Klatt, 1976), may more than suffice to overcome the reduced frequency discrimination, and thereby result in improved identification of vowels in consonantal context. The results suggest that improved vowel-recognition performance in automatic speech recognition is not to be attained by improved accuracies in formant frequency determination. Rather, contextual and temporal factors must be utilized as well.

At an F_1 value of 600 Hz, the 1–2 Hz-frequency range for F_2 covers four different English vowels, /ɔ/, /ʌ/, /æ/, and /ɛ/. Since the DL for F_2 in consonantal context was nearly 200 Hz, one may argue that no more than five vowels can be reliably differentiated on the basis of spectral information in a 1000-Hz frequency range. One may suspect that it is to overcome this limitation that different vowels with similar formant values are likely to differ in other characteristics such as duration and diphthongization. Liljencrants and Lindblom (1972) discuss a 12-vowel model where the vowels are distributed in F_1–F_2 space on the basis of maximum perceptual contrast. Even for that system, no more than four vowels are located in the above frequency range. It appears that there exist fundamental auditory and phonetic limitations on the density of vowels in formant space, and vowel-rich languages such as English have closely approached those limits.

The results concerning increased DL values in consonantal context appear to have significant implications for speech coding applications. The limited data indicate an increase that frequently exceeds 100% for the DL values in consonantal context. More detailed exploration of the limitations of these results, if any, needs to be carried out. In particular, one must determine whether the same results are obtained throughout the vowel space and in all consonantal contexts. If the results can be generalized to the encoding of all vowels and contexts in syllable-sized units, the resolution need to quantize the formant frequencies of the syllabic peak (the point of maximal spectral stability) may only be half as fine as that expected on the basis of DL values for steady-state vowels. Furthermore, since spectral discrimination at points of greater spectral variation can be expected to be worse than at points of relative spectral stability, discrimination at the syllabic peak appears to impose the tightest requirements.

The DL values cited cannot be applied directly for the independent encoding of formant information for successive short-time segments. Clearly, if on resynthesis independent perturbations of formant-frequency values of a magnitude comparable to the measured DL values were introduced, the result would be unacceptable to the listener. However, where syllable-sized segments are encoded as a unit, in terms of duration and spectral parameters, substantial information-rate savings may be achievable.

ACKNOWLEDGMENTS

Hollis Fitch and Anne Fowler assisted with the preparation of stimuli and the collection of data. Their help in the conduct of the experiments is greatly appreciated. I wish to thank Michael Studdert-Kennedy for numerous penetrating comments on this manuscript. Support for this research was received from the National Institute for Child Health and Development, N.I.H.

Finney, D. J. (1964). *Probit Analysis—A Statistical Treatment of the Sigmoid Response Curve* (Cambridge University, Cambridge).

Flanagan, J. L. (1955). "Difference limen for vowel formant frequency," J. Acoust. Soc. Am. **27**, 613–617.

Fry, D. B., Abramson, A. S., Eimas, P. D., and Liberman, A. M. (1962). "The identification and discrimination of synthetic vowels," Lang. Speech **5**, 171–189.

Fujisaki, H., and Kawashima, T. (1970). "Some experiments on speech perception and a model for the perceptual mechanism," Ann. Rep. Eng. Res. Inst. (University of Tokyo) **29**, 207–214.

Klatt, D. H. (1976). "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," J. Acoust. Soc. Am. **59**, 1208–1221.

Liang, L. C., and Chistovich, L. A. (1971). "Frequency difference limens as a function of tonal duration," Sov. Phys. Acoust. **6**, 75–80.

Liberman, A. M., Mattingly, I. G., and Turvey, M. T. (1972). "Language codes and memory codes," in *Coding Processes in Human Memory*, edited by A. W. Melton and E. Martin (V. H. Winston, Washington, DC).

Liljencrants, J., and Lindblom, B. (1972). "Numerical simulation of vowel quality systems: The role of perceptual contrast," Language **48**, 839–862.

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

Massaro, D. W. (1973). "A comparison of forward and backward recognition masking," J. Exp. Psychol. **100**, 434–436.

Peterson, G. E., and Barney, H. L. (1952). "Control methods used in a study of vowels," J. Acoust. Soc. Am. **24**, 175–184.

Peterson, G. E., and Lehiste, I. (1960). "Duration of syllable nuclei in English," J. Acoust. Soc. Am. **32**, 693–703.

Pisoni, D. B. (1971). "On the nature of categorical perception of speech sounds," Ph.D. thesis (University of Michigan). (*Supplement to Haskins Laboratories Status Report on Speech Research*) (unpublished).

Rabiner, L. R. (1968). "Digital-formant synthesizer for speech-synthesis studies," J. Acoust. Soc. Am. **43**, 822–828.

Stevens, K. N. (1968). "On the relations between speech movements and speech perception," Z. Phon. Sprachwiss. Komm. Forschung **21**, 102–106.

Stevens, K. N., and House, A. S. (1963). "Perturbations of vowel articulations by consonantal context: An acoustical study," J. Speech Hear. Res. **6**, 111–128.

Stevens, K. N., House, A. S., and Paul, A. P. (1966). "Acoustical description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation," J. Acoust. Soc. Am. **40**, 123–132.

Strange, W., Verbrugge, R. R., Shankweiler, D. P., and Edman, T. R. (1976). "Consonant environment specifies vowel identity," J. Acoust. Soc. Am. **60**, 213–224.

Swets, J. A. (1964). *Signal Detection and Recognition by Human Observers* (Wiley, New York).

Tsumura, T., Sone, T., and Nimura, T. (1973). "Auditory detection of frequency transition," J. Acoust. Soc. Am. **53**, 17–25.