Experiments Relating to the Perception of Formants

JOHN MORTON AND ALAN CARPENTER

Medical Research Council, Applied Psychology Research Unit, Cambridge, England

(Received 23 November 1962)

Experiments are described which suggest that, for human perception, the information concerning the location of a formantlike complex sound is contained in the two most prominent harmonics. This result is limited to the condition where adjacent harmonics are more widely separated than the width of a critical band.

INTRODUCTION

THE method of production of human speech leads to the presence of three or four formants below 3 kc, each of which is usually indicated in the frequency spectrum by an amplitude peak. It is customary for low fundamental frequencies to draw an envelope round the line spectrum and obtain the formant positions from the amplitude peaks, a procedure illustrated in Fig. 1. Automatic speech analyzers which employ "peak pickers" operate by identifying and tracking these peaks, and assigning them as formants.

It does not follow, however, that every formant in each vowel spoken by any person will be characterized by such a peak in the spectrum. Ladefoged[1] has pointed out the difficulty of identifying all the formants from line spectra, especially with the high-back vowels. The absence of a peak corresponding to a formant could be due to a high value of damping in the resonance system. Alternatively, if the half-power points of two adjacent formants overlap, the resulting theoretical envelope will contain a single peak. A third possibility occurs with a particular combination of fundamental frequency and two fairly close formants, as with normal-back vowels, which may produce a line spectrum that contains only a single peak, although there are two formants present. With the line spectrum in Fig. 2(a), we may draw two alternative envelopes. The first, as in Fig. 2(b), would result from our knowledge of the existence of two formants (though it would require a complex mathematical analysis or analysis-by-synthesis technique to deduce it). The second envelope, as in Fig. 2(c), would be the one we would draw in the absence of such knowledge. It is not known whether the analysis of speech sounds in the human auditory system takes account of such knowledge.

Delattre *et al.*[2] found that they could synthesize cardinal vowels with a single "formant," which, for the back vowels, were as identifiable as the two-formant vowels. (The "formants" used were not whole-spectrum formants, but a group of up to 4 harmonics with amplitudes suitably arranged to provide a peak at the required formant position.) The preferred single-formant

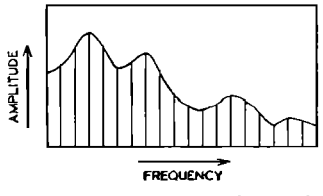


FIG. 1. An illustration of the correspondence between peaks in the spectrum of a vowel and the formant positions.

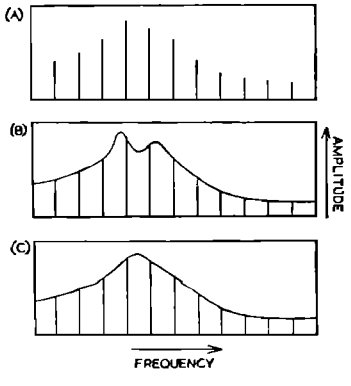
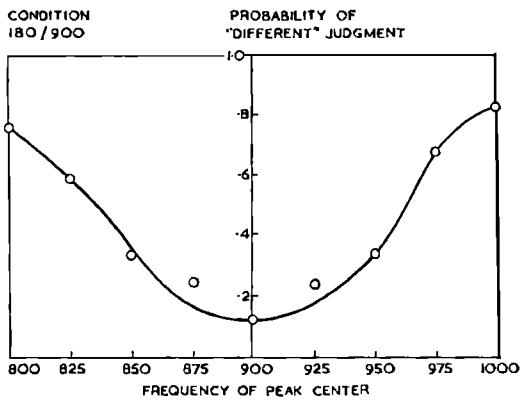


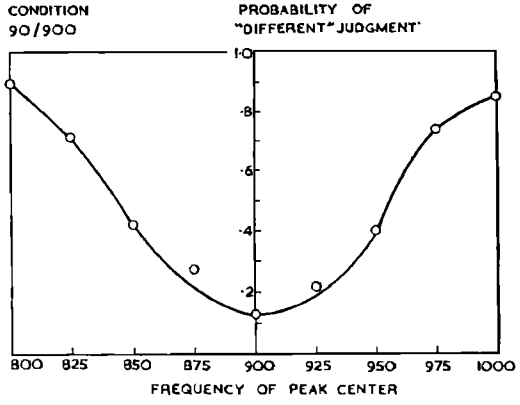
FIG. 2. (a) A line spectrum of a complex harmonic sound. (b) The envelope that would be drawn on the assumption that the origin of the sound contained two formant-producing systems. (c) The envelope that would be drawn around the single peak.

[1] P. Ladefoged, "The Perception of Vowel Sounds," a thesis submitted for the degree of Ph.D. at the University of Edinburgh (1959).

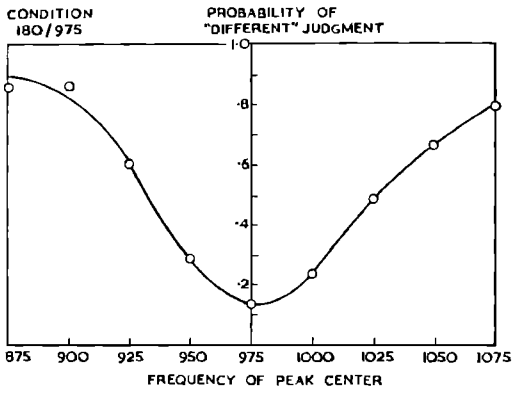
[2] P. Delattre, A. M. Liberman, F. S. Cooper, and L. Gerstman, Word 8, 195–210 (1952).



(a)



(b)



(c)

FIG. 3. The effect of varying the fundamental frequency and the frequency of the standard peak position upon the probability of a 'different' judgment. Each point is derived from 10 observations made by each of 20 subjects.

positions were in general between the positions of the two formants in the preferred synthetic two-formant vowels, in the way that the single peak in Fig. 2(c) is between the positions of the two peaks in Fig. 2(b). It is possible then that a single formant with its single peak could be analyzed as if it were a pair of formants. This is an alternative statement of Delattre *et al.*'s "assumption that the ear effectively 'averages' two formants which are relatively close together (as is the case for the back vowels), and receives from them an over-all quality roughly equivalent to that which would be produced by an intermediate formant" (p. 209).

It is implicit in the above analysis that human perception of speech does depend on the presence and identification of peaks in the frequency spectrum. However, two previous experiments[3,4] have shown that it is not necessary to have peaks in the frequency spectrum of a complex harmonic sound for the sound to be categorized consistently as to its vowel color.

In spite of this, it is likely that when peaks are present they will be important; yet little is known about how they are perceived. Difference limens for formant frequencies using full synthetic vowel sounds containing four formants have been measured by Flanagan.[5] If such stimuli were analyzed by the auditory system in terms of the whole spectrum, then the description given of the sounds in terms of formant positions would be sufficient. However, since his results revealed wide variations in the difference limens, particularly when two formants were close together, it is likely that for explanatory purposes a more detailed level of description of the stimuli is required.

The present experiments examine certain properties of peaks, and also investigate the hypothesis that the amplitudes of the two most prominent harmonic components provide sufficient information for the localization of peaks in human perception.

EXPERIMENT I. DIFFERENCE LIMENS (DL's) FOR PEAKS

This experiment was designed to measure the DL's for complex harmonic sounds containing a single peak. The stimuli were produced by passing the output of a periodic pulse source through a filter. This filter was a Wien bridge network with positive feedback and had a symmetrical response asymptotic to 6 dB per octave on each side, with a half-power bandwidth of 200 cps.

[3] A. Carpenter and J. Morton, "Perception of Vowel Colour in Formantless Complex Sounds," Language and Speech 5, 205–214 (1962).

[4] J. Morton and A. Carpenter, "Judgement of the Vowel Colour of Natural and Artificial Sounds," Language and Speech 5, 190–204 (1962).

[5] J. L. Flanagan, J. Acoust. Soc. Am. 27, 613–617 (1955).

TABLE I. Difference limens for peaks.

Condition	−DL	+DL
180/900	63.5	62.5
90/900	55.0	55.0
189/975	49.0	52.5

DL's were measured under three conditions employing pulse frequencies of 180 or 90 cps and peak frequencies of the standard stimulus of 900 or 975 cps. The conditions are termed 180/900, 90/900, and 180/975.

A comparison of conditions 180/900 and 90/900 would demonstrate the effect of varying the fundamental frequency; i.e., of adding more harmonics to the sound, and so providing, in theory, more information as to the precise position of the peak. The 180/900 and 180/975 conditions were taken as limiting cases. In the first case, a harmonic falls on the formant peak; in the other case, the peak falls almost symmetrically between two harmonics.

Test Procedure

Items were recorded on magnetic tape. Each presentation consisted of two sounds of 0.4-sec duration separated by a 0.6-sec gap. Subjects were instructed to judge whether the quality of the two sounds in a pair was the same or different. One of the sounds was the standard (S), and the other was either a repeat of S or one of the variations (V). The frequency variations used for the peak position of the filter were ±25, ±50, ±75, and ±100 cps from the standard in each condition.

The test started with 10 pairs of items which were treated as practice items, followed by 90 randomized test pairs, of which ten pairs were identical (SS), and 80 pairs involved physical differences (SV or VS). Thus, each subject judged each pair of sounds 10 times. The pairs of sounds followed each other at approximately 6-sec intervals. The three conditions were recorded on separate tapes.

The subjects, 20 experimentally naive men between 17 and 21 years of age, heard the three tapes on separate days. They each recorded their responses as "S" or "D" on a score sheet.

The stimuli were reproduced through a Vortexion tape recorder and a Quad Electrostatic loudspeaker in a normally damped room. The sound-pressure level was approximately 70 dB *re* 0.0002 dyn/cm² and was constant for all samples.

Results

For each stimulus, the observed probability p of a "different" response was obtained from the 200 responses to that stimulus. The values of p and the resulting curves (which were fitted by eye) are shown in Fig. 3 for the three conditions. The DL's were taken as the $p=0.5$ points on the curves and are given in Table I. A note on the treatment of the results is given in the appendix.

Discussion

The values for the DL are somewhat larger than the equivalent values given by Flanagan (−20 and +50 cps for a standard of 1000 cps). However, the filters used in his experiment had a bandwidth of 150 cps, as opposed to 200 cps in the present experiment, and on the present hypothesis such a difference would be expected. This point will be taken up below.

The threshold in condition 90/900 is lower than in condition 180/900, as was predicted. However, this cannot be due simply to the presence of more harmonics enabling the position of the peak to be defined more precisely (in some gestalt sense), since the threshold in condition 180/975 is lower still. It might be noted that the missing condition (90/975) was used, but on analyzing the stimuli after the experiment, it was discovered that half of them had been incorrectly recorded. The results for the other half were, however, substantially the same as for condition 90/900, as might be expected.

For the two conditions with 180 cps fundamental, the levels of the prominent harmonics in the standard stimuli and in the stimuli which would correspond to the threshold conditions were measured. These values are shown in Table II. It will be seen that there is a high degree of consistency between the figures in the last column. These figures are the sums of the differences in intensity, regardless of sign, of the two most prominent harmonics between the standard and threshold.

If information concerning the other harmonics is being used in the discrimination, then this value, c.4.8 dB, should be lower than the difference threshold of two tones alone when one is increased and the other decreased in intensity. As no work appears to have been done on the perception of such sounds, Experiment II was designed.

TABLE II. Relative amplitude of the harmonics in the standard and threshold stimuli harmonic frequencies (in cps).

	Position of peak	540	720	900	1080	1260	S[a]
Condition 180/900							
Standard	900 cps	−16.4	− 8.2	0	−6.5	−12.0	
−Threshold	836.5	−13.9	− 4.6	−1.4	−8.8	−13.7	
Difference	63.5	+ 2.5	+ 3.6[b]	−1.4[b]	−2.3	− 1.7	5.0
+Threshold	962.5	−19.0	−11.3	−1.6	−3.7	−10.4	
Difference	62.5	− 2.6	− 3.1	−1.6[b]	+2.8[b]	+ 1.6	4.4
Condition 180/975							
Standard	975	−19.6	−12.0	−2.2	−3.0	−10.1	
−Threshold	926	−17.3	− 9.1	−0.2	−5.7	−11.6	
Difference	49	+ 2.3	+ 2.9	+2.0[b]	−2.7[b]	− 1.6	4.7
+Threshold	1027.5	−20.1	−14.2	−5.1	−1.0	− 8.5	
Difference	52.5	− 0.5	− 2.2	−2.9[b]	+2.0[b]	+ 1.6	4.9

[a] S is the sum of the modulus differences in intensity of the prominent harmonics.
[b] Most prominent harmonics.

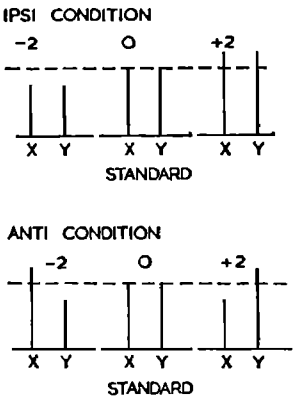


FIG. 4. An illustration of the kind of stimuli used in experiment II. *X* and *Y* represent pure tones. The numbers ±2 refer to a change in intensity of the tones in the variable stimuli relative to the standard.

EXPERIMENT II. DIFFERENCE LIMENS FOR TWO TONE STIMULI

The stimuli consisted of two pure tones, x and y, where x is the lower in frequency. We were primarily interested in the difference threshold when one of these tones was increased in amplitude and the other was decreased in amplitude. However, from the work on critical bands,[6,7] it seemed likely that the result would depend upon the separation of the two tones in frequency. The implication of the work for the present experiment is that a pair of pure tones might be inseparable as far as detection of energy changes is concerned when they are less than a certain "critical" frequency apart. Accordingly, we used three conditions, *P*, *Q*, and *R*, employing frequency separation of the two tones of 60, 180, and 300 cps, respectively, centered on 930 cps. (See Table III.) The 60-cps separation would leave the two tones within the same critical band; the 180-cps separation would be marginal (depending upon which of the widely different measures of critical bandwidth was applicable); and the 300-cps separation would leave the tones unambiguously separate.

The design of the experiment was identical to that of experiment I, subjects being presented with a pair of stimuli, each consisting of two pure tones. Each pair included a standard (S), where the two tones were equal in intensity, and either a repeat of S or one of the variations (V). For each of the conditions there were two subconditions: (a) *Ipsi* condition (I)—where, for the variable stimuli, both tones were increased or decreased in intensity together by the same amount relative to the standard; (b) *Anti* condition (A)—where, for the variable stimuli, one of the tones was increased in intensity and the other was decreased in intensity by the same amount relative to the standard.

TABLE III. Pure tones used in experiment II.

Condition	Frequency of x (cps)	Frequency of y (cps)	Separation (cps)
P	900	960	60
Q	840	1020	180
R	780	1080	300

TABLE IV. DL's (in dB relative to standard) for two-tone stimulus.

Condition (separation of tones)	*Ipsi* −DL	*Ipsi* +DL	*Anti* −DL	*Anti* +DL
P (60 cps)	2.15	2.25	3.25	3.4
Q (180 cps)	2.1	2.4	2.4	2.4
R (300 cps)	2.0	2.6	2.35	2.5

The total energy change is greater in the *Ipsi* than in the corresponding *Anti* condition. If a pair of tones were in the same critical band and there was summation of the energy in the two tones, then the threshold in the *Anti* condition should be higher than that in the *Ipsi* condition. If the tones were in different critical bands and could be treated separately by the auditory mechanism, then the thresholds in the *Ipsi* and *Anti* conditions should be the same.

The changes of intensity used for the variable stimuli were ±1, ±2, ±3, and ±4 dB. Equivalent stimuli in the two conditions are illustrated in Fig. 4.

Six tapes were made of the six conditions and were played to a total of 12 young men under the same conditions, as in the previous experiment.

Results

Figure 5 shows the results as plotted using the same procedure as in experiment I. The $p=0.5$ points on the curves were taken as thresholds; these values are given in Table IV. Only the condition *P* (60-cps separation) *Anti* condition yields a threshold significantly greater than the others, indicating that if the above analysis is correct, then the critical bandwidth at 930 cps is between 60 and 180 cps. This range encompasses all the previous estimates of this measure, from the estimate of 65 cps obtained by Fletcher[8] and Schafer *et al.*[9] to that of 160 cps made by Zwicker *et al.*[6] and by Greenwood.[10]

In experiment I, it was found that at the thresholds for formant frequencies, the sum of the modulus differences in intensity of the two dominant harmonics ranged from 4.4 to 5.0 dB, with a mean of 4.75 dB. The equivalent condition (QA) in experiment II, where two tones, 180 cps apart, undergo intensity changes in opposite directions, yields a threshold of 2.4 dB. That is equivalent to a sum change of 4.8 dB. The correspondence of these two figures is in accord with the original hypothesis that the amplitudes of the two most prominent harmonics are sufficient to discriminate between two different formant positions.

[6] E. Zwicker, G. Flottorp, and S. S. Stevens, J. Acoust. Soc. Am. **29**, 548–559 (1957).
[7] B. Scharf, J. Acoust. Soc. Am. **33**, 503–511 (1961).
[8] H. Fletcher, *Speech and Hearing in Communication* (D. Van Nostrand Company, Inc., New York, 1953).
[9] T. H. Schafer, R. S. Gales, C. Shewmaker, and P. O. Thompson, J. Acoust. Soc. Am. **22**, 490–496 (1950).
[10] D. D. Greenwood, J. Acoust. Soc. Am. **33**, 484–502 (1961).

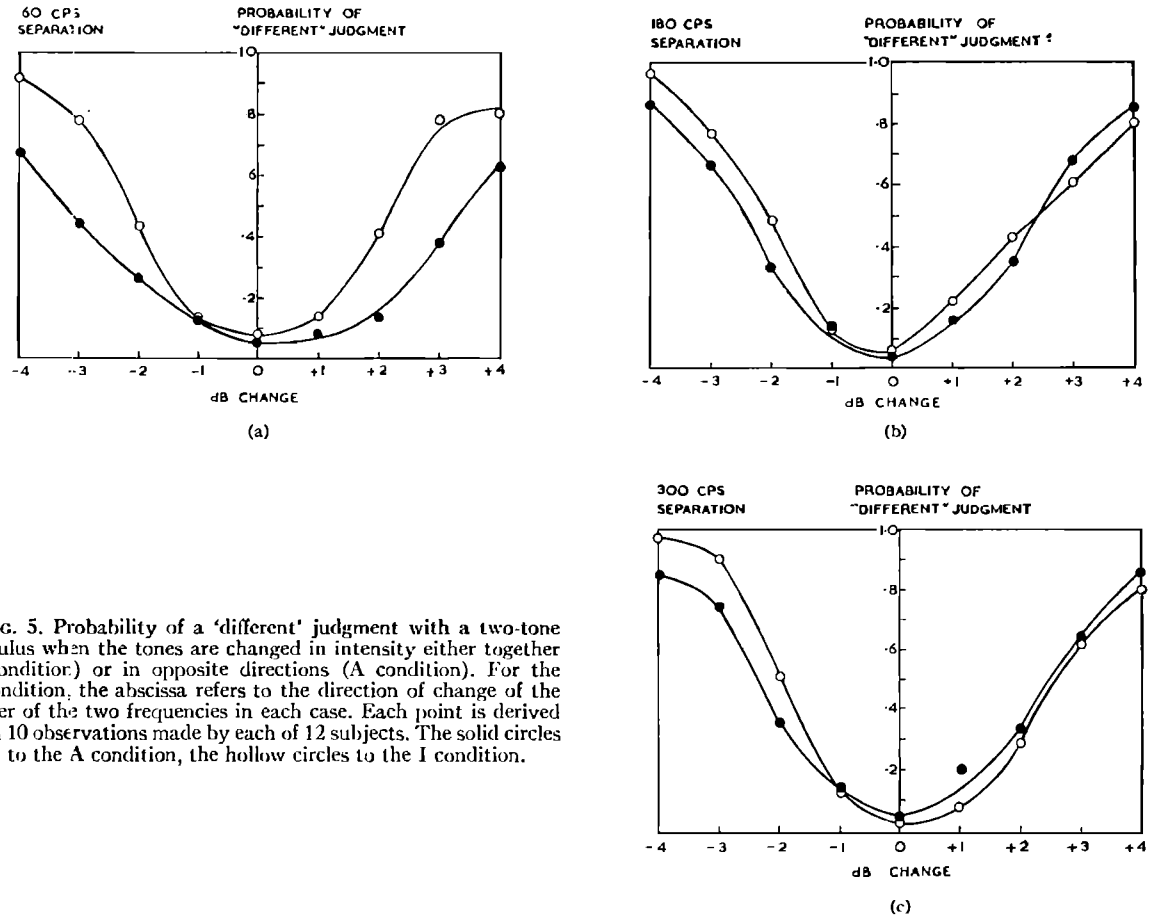


FIG. 5. Probability of a 'different' judgment with a two-tone stimulus when the tones are changed in intensity either together (I condition) or in opposite directions (A condition). For the A condition, the abscissa refers to the direction of change of the higher of the two frequencies in each case. Each point is derived from 10 observations made by each of 12 subjects. The solid circles refer to the A condition, the hollow circles to the I condition.

General Discussion

These results suggest a series of experiments in order to investigate more closely the effects of having a fundamental frequency low enough for adjacent harmonics to fall within the same critical band. The results for condition 90/900 in experiment I do not fit in directly with the other results when individual harmonics are examined, as can be seen in Table V. The sums of the changes of the two most prominent harmonics at threshold in this condition were much lower than the values in the other two conditions (c.f. Table II). If the changes in the *three* most prominent harmonics are summed, we obtain values at threshold of 6.4 and 6.0 dB for the $-ve$ and $+ve$ thresholds, respectively; and these figures are too high. If, on the other hand, we consider that basically the auditory mechanism deals with the two prominent harmonics only, but that, owing to the low fundamental frequency, some or all of the energy in the adjacent harmonics is included, then agreement may be found. As an example, we can suppose that for the negative threshold all the energy in harmonics *a* and *b* (in Table V) is summed, as is that in harmonics *c* and *d*. For the positive threshold, *b* is taken with *c* and *d* with *e*. The resulting changes in intensity at threshold compared with the standard are given in Table VI. For the lower threshold, the figure 4.98 dB corresponds with the values found in the 180-cps conditions of experiment I. The value at the higher threshold, 3.8 dB, however, is still too low.

To proceed further with this kind of analysis, it would clearly be necessary to have more detailed information as to the shape, extent, and mode of operation of the internal filters which lead to the critical-band phenomena, or to produce an alternative model to explain this phenomenon. Where there is no summation of the components of a complex stimulus, our results do suggest that our perception of a peaked sound is limited

TABLE V. Relative amplitude of the harmonics in the standard and threshold stimuli of condition 90/900.

	Position of peak	Harmonic frequencies in cps a 720	b 810	c 900	d 990	e 1080	S^a
Standard	900	− 8.2	−3.4	0	−2.6	−6.5	
−Threshold	845	− 3.5	−0.5	1.2	−4.9	8.6	
Difference	55	+ 4.7	+2.9[b]	−1.2[b]	−2.3	−2.1	4.1
+Threshold	955	−11.0	−6.1	−1.2	−0.5	−4.1	
Difference	55	− 2.8	−2.7	−1.2[b]	+2.1[b]	+2.4	3.3

[a] Sum of modulus differences in intensity of the two most prominent harmonics.
[b] Prominent harmonics.

TABLE VI. The resulting changes in intensity at threshold in condition 90/900 assuming that adjacent harmonics are treated together.[a]

	Harmonic (see Table V)	Relative intensity in the standard stimulus	Relative intensity at threshold	Intensity difference at threshold	Sum of modulus differences
Lower threshold	a	−8.2	−3.5		
	b	−3.4	−0.5		
	(a+b)	−2.16	+1.27	+3.43	
	c	0	−1.2		4.98
	d	−2.6	−4.9		
	(c+d)	+1.90	+0.35	−1.55	
Upper threshold	b	−3.4	−6.1		
	c	0	−1.2		
	(b+c)	+1.64	+0.02	−1.64	
	d	−2.6	−0.5		3.81
	e	−6.5	−4.1		
	(d+e)	−1.11	+1.08	+2.19	

[a] All intensities are measured relative to that of the 900-cps component in the standard stimulus.

by our perception of the leading harmonics as far as discrimination is concerned. (It is trivial to remark that the other components contribute to the constant *quality* of the sound; this appears to be a separate matter.)

We would expect to find a lower threshold with peaks of a smaller bandwidth, as did Flanagan,[5] since the same shift in peak position would produce a larger change in the intensity of the harmonics near to the peak. He explained the wide variations he found in threshold at different standard frequencies in terms of interaction between the spectral peaks of adjacent formants. We would expect that an explanation in terms of the changes in amplitude of individual harmonics would be superior predictively.

APPENDIX

We have employed the simplest treatment of results in this preliminary experiment. If the data are given a "correction for guessing," the results remain substantially the same. If we consider them in the light of signal-detection theory, the following points arise:

1. In experiment I the false positive rates for the group were identical for the three conditions (13%), and the variation within subjects was small. It is considered that the comparison of conditions 180/900 and 180/975 remains valid.

2. The false positive rates in experiment II were much lower, ranging from 7.5 to 3.3%. When d' values are computed for the whole group (it is realized that this is a primitive concession to detection theory, but it remains a useful indication), the results for condition *X-Anti* remain dramatically different from those in the other conditions.

3. The comparison made between the results of experiment I and the *X-Anti* condition of experiment II is strengthened in favor of the hypothesis by the differences in false-alarm rates.