Reprinted from THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA, Vol. 30, No. 11, 1035–1041, November, 1958

Study of Some Cues in the Perception of the Voiced Stop Consonants*

HOWARD S. HOFFMAN
The Pennsylvania State University, University Park, Pennsylvania
(Received August 6, 1958)

Previous research involving synthetic speech reveals that both the second- and the third-formant transitions play a role in the perception of the voiced stops, |b|, |d|, and |g|. The present experiment examined an additional cue (burst frequency), repeated a portion of the previous research, and collected more information about how the cues act in various combinations.

Synthetic speech sounds containing one cue, all possible combinations of two cues, and all possible combinations of three cues were tested on a large group of listeners.

Burst frequency was found to act as a cue for the perception of the voiced stops in much the same manner as this variable affects the perception of the unvoiced stops. To the extent that the present experiment overlapped previous research, the two sets of findings were in very close agreement. When cues were combined, they shared in the control of perception in such a way that the contribution of any one cue was largely independent of the nature and the number of the other cues present in the sound. The individual cues seemed to have directional properties somewhat like vectors. The addition of cues produced effects somewhat like the addition of vectors.

INTRODUCTION

IN most areas of perception the search for cues begins with an inspection of the stimulus and generally culminates with a series of experiments in which the investigator carefully modifies the separate aspects of the stimulus and determines how his operations have affected perception. In the domain of speech, however, the necessary stimulus control has, until recently, been unavailable. Human speech, because of the limitations of the vocal mechanisms, was hardly suitable. The logical alternative was machine-produced speech, but speech synthesizers suitable for research purposes have only been developed within the past ten years.

One of the more versatile of the currently available speech synthesizers, the pattern playback,[1] generates speech by converting hand-painted spectrograms into sound. The experiment to be reported derives from and is an extension of the large body of research which has been undertaken by workers at the Haskins Laboratories, where the pattern playback was designed and built. It constitutes an attempt to extend one segment of the research, namely, that portion which has been concerned with the cues to the distinctions among the voiced stop consonants.

The patterns shown in Fig. 1 illustrate one of the cues which, in previous investigations, had been found to be important for the distinctions among |b|, |d|, and |g|.[2,3] The variable aspect of the patterns is the extent and direction of the frequency shift (transition) of the second formant. These hand-painted spectrograms were, of course, schematic in the extreme; yet, when they were converted to sound on the pattern playback, most listeners identified each as a syllable in which a voiced stop was followed by the vowel |æ|. When the transition of the second formant (hereafter referred to as the transition of F2) was rising, the consonant was heard as |b|. When the transition of F2 was either straight or fell slightly, the consonant was heard as |d|. When the F2 transition fell from higher points on the frequency scale, the consonant was heard as |g|.

Although other research indicated that the transition of F2 is a very nearly sufficient cue for the distinctions among the voiced stops, more recent experiments by Harris, Hoffman, Liberman, Delattre, and Cooper have revealed that the transition of a third formant is also important.[4] This cue was investigated by determining how perception was affected when a third (higher) formant with various transitions was added to each of the patterns that were seen in Fig. 1.

In general, the data appeared to be consistent with the notion that the separate acoustic events (the transition of F2 and the transition of F3) were making relatively independent contributions to perception.

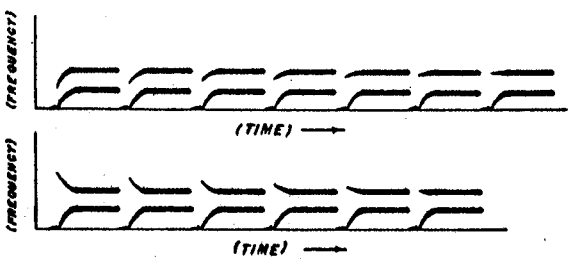

FIG. 1. Patterns used in previous investigations of the transition of the second formant.

* This research was supported by the Haskins Laboratories and is based upon a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the University of Connecticut, 1957.

[1] Cooper, Liberman, and Borst, Proc. Natl. Acad. Sci. U. S. 37, 318–325 (1951).

[2] Delattre, Liberman, and Cooper, J. Acoust. Soc. Am. 27, 769–773 (1955).

[3] Liberman, Delattre, Cooper, and Gerstman, Psychol. Monogr. 68, No. 8 (Whole No. 379), (1954).

[4] Harris, Hoffman, Liberman, Delattre, and Cooper, J. Acoust. Soc. Am. 30, 122–126 (1958).

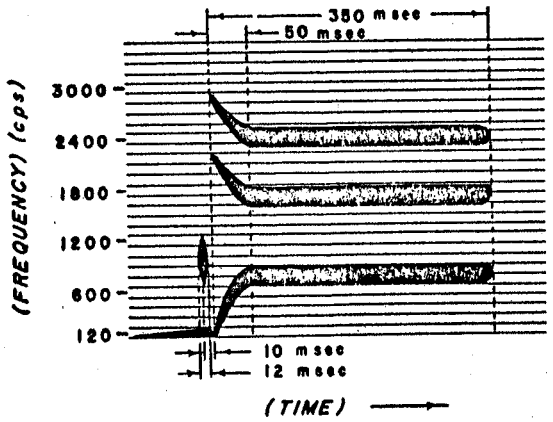

FIG. 2. One of 336 patterns used in the present investigation. This pattern illustrates the temporal arrangement of the burst and the transitions, and also illustrates the frequencies at which the formants centered.

When cues for the same phoneme were combined, the sound continued to be heard as that phoneme. When, on the other hand, cues for different phonemes were combined the sound was heard as one phoneme by some of the listeners and as the other phoneme by the rest of the listeners.

The present study was designed to extend the scope of this research by investigating a third cue for the same distinctions as are given by the transitions of F2 and F3. This cue, the frequency position of a short burst of noise, although not previously studied in connection with the voiced stops, was one of the first of the acoustic variables that were found to be important for the distinctions among |p|, |t|, and |k|, the unvoiced counterparts of |b|, |d|, and |g|.[5]

On the strength of those findings and because bursts are often seen in front of transitions in spectrograms of the voiced stops, it seemed possible that burst frequency might also function as a cue to the distinctions among |b|, |d|, and |g|. One purpose of the present experiment was to explore this possibility.

A second purpose of the present investigation was to determine how bursts and transitions share in the control of perception. More generally the study was designed to provide information relevant to the question of how cues combine. Obviously, the problem of cue combination is of fundamental importance in any perceptual process, or, indeed, in any behavioral process in which the response is determined by more than one cue element.

METHOD

The feasibility of this experiment hinged in part upon the question of whether or not bursts, as synthesized on the pattern playback, could be included when sounds were intended to be heard as realistic voiced stops. The possibility existed that the burst might not combine with the transitions or, if it did, that it would act as a cue for unvoicing and cause the sounds to be heard as |p|, |t|, and |k| rather than |b|, |d|, and |g|. This question was resolved affirmatively through exploratory work on both the pattern playback and the sound spectrograph.

In order to satisfy the purpose of this experiment and, at the same time, keep the number of patterns within a reasonable limit, it was necessary to restrict the study to a single vowel. The American |æ| was chosen since this vowel had been used in previous research on the transitions.[4]

Figure 2 illustrates one of the patterns used in this study. As can be seen in this figure, the burst precedes the onset of the transitions by 10 msec. This is approxi-

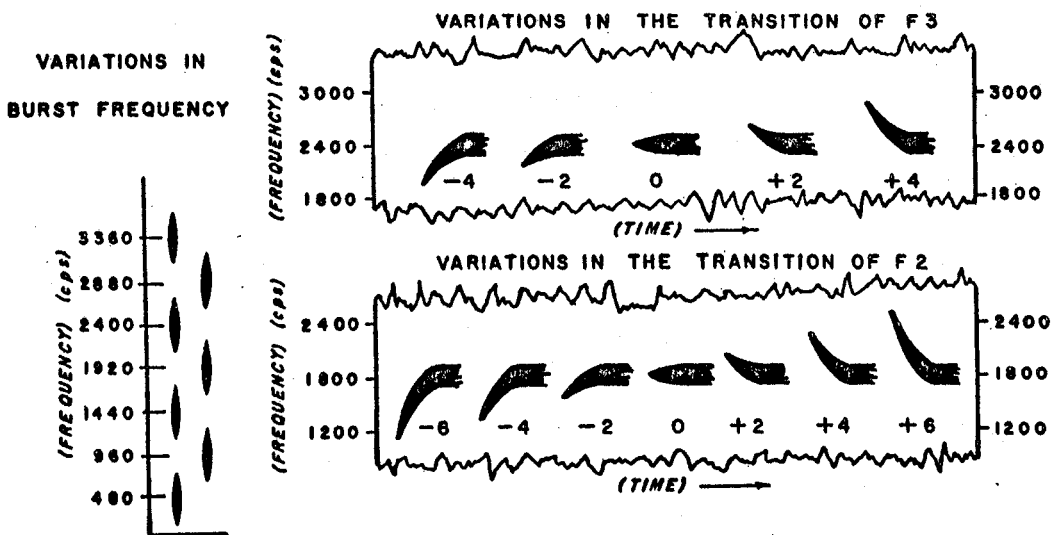

FIG. 3. Schematic diagram showing variations in burst frequency, the transition of the second formant, and the transition of the third formant.

[5] Liberman, Delattre, and Cooper, Am. J. Psychol. 65, 497–516 (1952).

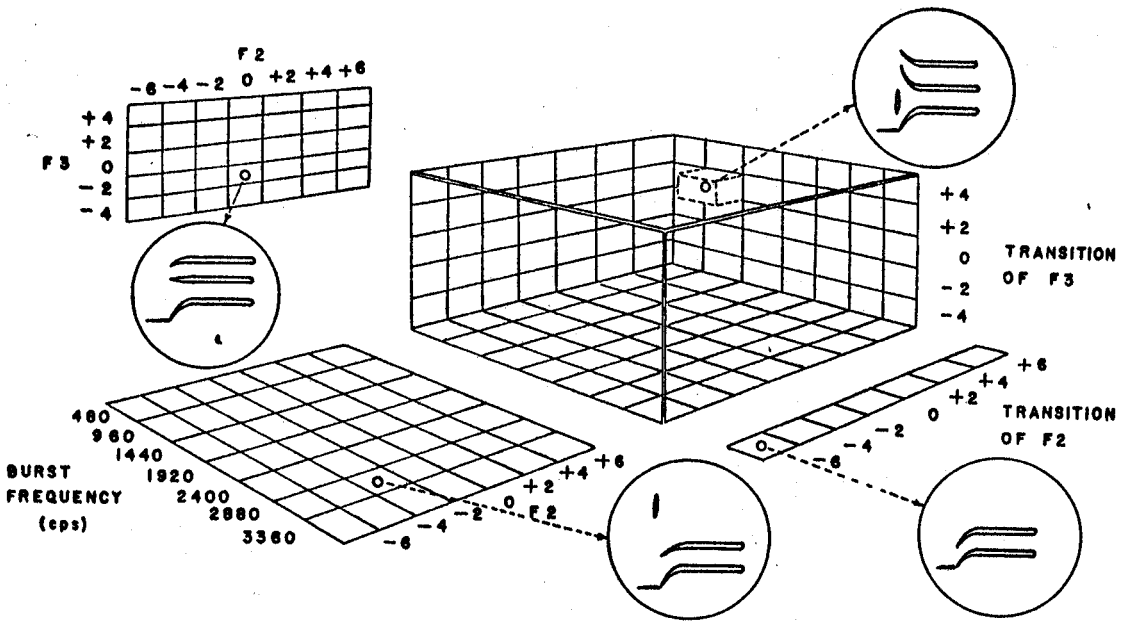

FIG. 4. The patterns that were used in the present investigation. Each of the acoustic variables is represented by a dimension, and each of the patterns is represented by a cell within the spaces formed by combining the dimensions. The circles indicate one of each of the four kinds of patterns that were used.

grams of actual speech. The duration of the burst and the duration of the transitions (12 msec and 50 msec, respectively) are also approximations to actual speech. The steady-state portions of the formants have a duration of 300 msec and, with the exception of F1, were set so as to approximate the frequencies reported by Peterson and Barney in their study of the vowels of American English.[6] In order to synthesize this vowel on the pattern playback, it was necessary to place F1 at 780 cps.[7] The transition of the first formant and the bar of low-frequency acoustic energy which precedes it approximate the kinds of acoustic events which speakers typically produce when they articulate the voiced stops. These acoustic features tend to act as cues for voicing but have little, if any, effect upon the distinctions among |b|, |d|, and |g|

The pattern shown in Fig. 2 represents one combination of the three acoustic variables (burst frequency, the extent and direction of the F2 transition, and the extent and direction of the F3 transition). In all, there were 245 such combinations, each of which represented a different arrangement of one of seven F2 transitions, one of five F3 transitions, and one of seven burst frequencies. Figure 3 shows the way in which the bursts and the transitions were varied. The transitions in this figure have been specified by indicating whether they are rising or falling (− or +) and by reporting the number of harmonics of 120 cps that are covered in the frequency excursion. As seen in Fig. 3, the F2 transition was varied from −6 to +6 in seven steps of 240 cps. The transition of F3, on the other hand, was varied from −4 to +4 in five steps of 240 cps. The bursts used in this study have been specified by indicating the frequency position of their centers. As can be seen in Fig. 3, the burst could occupy any one of seven positions. The low burst centered at 480 cps, while the highest burst centered at 3360 cps.

In addition to the 245 patterns that were formed by combining all three cues, three other series of patterns were tested: (1) a series of patterns which contained bursts and F2's but not F3's—there were 49 such patterns representing all possible combinations of the seven second-formant transitions and the seven bursts that were illustrated in Fig. 3; (2) a series of patterns that contained F2's and F3's but not bursts—there were 35 such patterns representing all possible combinations of the seven second-formant transitions and the five third-formant transitions that were seen in Fig. 3; (3) a series of two-formant patterns which represented the seven second-formant transitions that were seen in Fig. 3.

The patterns that were used in this study are illustrated in a second way in Fig. 4. In this figure, each of the acoustic variables is represented by a dimension and each of the patterns is represented by a cell within the spaces formed by combining these dimensions. According to this representation, the acoustic parameters for a given pattern are specified by the coordinates of its position within the stimulus space.

The three-dimensional stimulus space is filled by the 245 patterns which represent all possible combinations

[6] G. E. Peterson and H. L. Barney, J. Acoust. Soc. Am. 24, 175–184 (1952).

[7] Peterson and Barney report that in |æ| F1 tends to center at 660 cps.

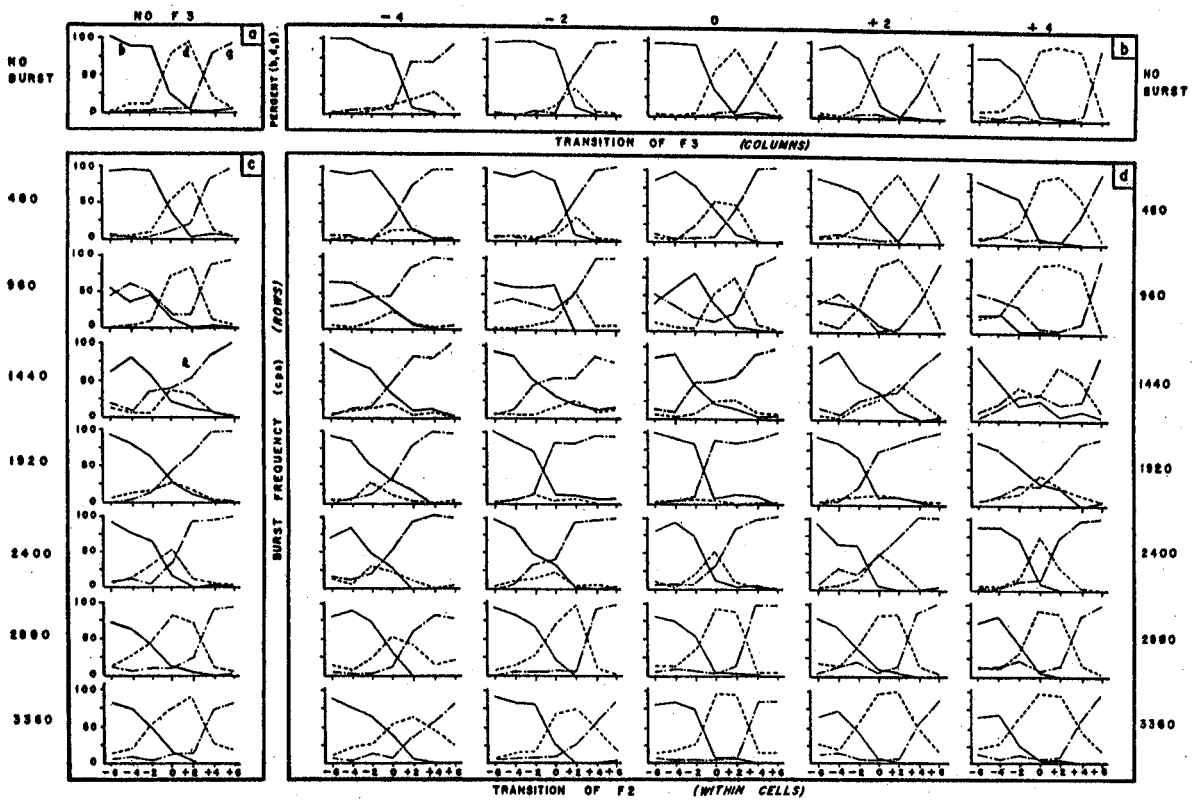

FIG. 5. Judgments of the 336 patterns used in the present investigation. Each subplot represents the responses to seven patterns. The second-formant transition is indicated along the abscissa of the subplot (cell). For a given second-formant transition, the ordinates to the three response curves indicate the percentage of times the pattern was judged as |b|, as |d|, and as |g|. The third-formant transition in the pattern is indicated at the head of each column of subplots. The burst in the pattern is indicated alongside of each row of subplots. (These response curves are based upon 52 judgments per pattern.)

of the three acoustic variables. The two-dimensional stimulus space off the left-hand face of the solid is filled by the 35 patterns containing second- and third-formant transitions, but not bursts. The two-dimensional space off the bottom face of the solid is filled by the 49 patterns containing second-formant transitions and bursts, but not third-formant transitions. The cells along the stimulus continuum near the lower right-hand edge of the solid represent the seven basic two-formant patterns. One example of each of the four kinds of patterns has been represented in the insets in Fig. 4.

Preparation of the Sounds

The entire series of patterns (336 in all) was converted to sound on the pattern playback and recorded on magnetic tape. By cutting and splicing the magnetic tape, the sounds were arranged in random order. The tape was then rerecorded in such a way that each stimulus would appear and then be repeated after an interval of one second. There was an interval of six seconds between successive pairs of stimuli (i.e., a sound and its repetition). The purpose of these operations was to permit the listener to hear each sound twice before making an identification and also to provide him with sufficient time to record his decision.

This large tape was then divided into four smaller tapes, each of which contained 84 stimuli.

Subjects and Instructions

Twenty-six paid volunteers, recruited from psychology courses at the University of Connecticut, judged the sounds used in this study. Each subject attended four experimental sessions. On the first session, they judged two of the four tapes, on the second session they judged the other two, on the third and fourth sessions subjects judged the four tapes for the second time.

The subjects were told that each stimulus (and its repetition) would be a synthetically produced syllable consisting of a voiced stop followed by the vowel |æ|. They were instructed to identify only the consonant and to limit their responses to |b|, |d|, and |g|. Listeners were urged to make an identification of every stimulus even though in some cases they might find it necessary to guess.

In all, fifty-two judgments were obtained for each of the 336 stimuli (two identifications from each of twenty-six listeners).

RESULTS

Since an inspection of the results showed essentially no difference between the first and second sets of judgments, these data have been combined in reporting the findings.

Figure 5 shows the responses to all 336 of the sounds that were used in this experiment. In this figure each subplot represents the responses to seven different acoustic patterns. For a given pattern, the second-formant transition is indicated along the abscissa of the subplot, the third-formant transition is indicated at the head of the column of subplots and the burst is indicated alongside of the row of subplots. Within a given subplot, the ordinate to the three response curves reveals the percentage of times that each sound was judged as |b|, as |d|, and as |g|.

Figure 5(a) shows the responses to the patterns that contained second-formant transitions but neither bursts nor third formants. As revealed by these response curves, the extent and direction of the second-formant transition exercises considerable control over the judgments of the listeners. When F2 was negative, the sound was judged as |b|. When F2 was straight or fell slightly, the sound was judged as |d|. When F2 was either +4 or +6 the sound was judged as |g|.[8] Since, in these patterns, the F2 transition was the only cue for the distinctions among |b|, |d|, and |g|, one may interpret the three response curves as representing, rather directly, the cue properties of the several second-formant transitions. The cue properties of the other variables (the burst and the transition of F3) are not, however, represented so directly in Fig. 5. Rather, the effects of these cues must be inferred by determining how the response curves shifted and changed in contour when these cues were added to the two-formant patterns.[9]

One way to carry out such an analysis is to compare visually the several subplots in Fig. 5. An alternative, though somewhat less sensitive technique, involves adding across variables. In Fig. 6 the latter technique has been used to generate what amounts to a very nearly direct representation of the effects of the bursts and the transitions of F3.

First, the percentage of |b|, of |d|, and of |g| responses for the entire series of seven two-formant patterns was determined. The next step was to determine the way in which these percentages changed when various other cues were added to the series. Thus, in Fig. 6, the ordinates at an F3 transition of −4 in the subplot labeled *F3 added to F2 alone* indicate that when a −4F3 was added to each of the two-formant patterns, the entire series yielded 9% more |b| responses, 9% more |g| responses, and 18% less |d| responses. The ordinates in the other subplots were derived in a like manner. For the subplot labeled *Burst added to F2 plus F3*, the first step was to obtain the percentage of |b|, of |d|, and of |g| responses from the entire series of 35 patterns that contained F2 and F3 but not bursts. These values were then compared to the percentages that were obtained when a given burst was added to each of the 35 three-formant patterns.

It can be seen in Fig. 6 that, regardless of whether F3's were added to F2's or to F2's plus bursts, the negative F3's enhanced |b| and |g| at the expense of |d| while the positive F3's enhanced |d| at the expense of both |b| and |g|. Bursts, when added to the three-formant patterns, have the same over-all effects as when these cues are added to the two-formant patterns. The low burst has comparatively little effect; the 960 burst enhances |g| at the expense of |b|; the 1440, 1920, and 2400 bursts enhances |g| primarily at the expense of |d|; and the high burst enhances |d| primarily at the expense of |b|

It would seem, that except for |b|, burst frequency serves as a cue for the perception of the voiced stops in much the same manner as this variable affects the perception of the unvoiced stops. In the earlier experiment on |p|, |t|, and |k|,[5] the high bursts were found to act as cues for |t|. In the present experiment (where all of the sounds were voiced) these same high bursts were found to act as cues for |d| (the voiced counterpart of |t|). In the earlier study it had been found that bursts near the second formant of the vowel acted as cues for |k|. In the present experiment, bursts near the second formant of the vowel acted as cues for |g| (the voiced counterpart of |k|). One difference between the results of the present study and those of previous research was that the present study did not reveal a burst which acted as a cue for |b| (the voiced counterpart of |p|). It may be noted, however, that the sample of burst frequencies was rather coarse. It is possible, therefore, that a finer sampling of the lower frequencies would have revealed such a cue.

INTERPRETATION AND DISCUSSION

In Fig. 6 it was seen that the addition of bursts to the three-formant patterns changed perception in the same way as when bursts were added to the two-formant patterns. Similarly, F3's added to F2 plus bursts changed perception in the same way as when third formants were added to the series of patterns which contained only F2's. This suggests the general conclusion that the cues had effects which were independent of each other. One can also see this independence by examining the more or less raw data of Fig. 5. If cues have independent effects, one would expect that the best cue for a given phoneme would

[8] These findings correspond in detail to the results of previous research on the transition of F2.

[9] Ideally, bursts and F3's would have been tested in isolation before combining them with F2's. As it turns out, however, neither F3's nor bursts yield speech-like sounds when presented in the absence of an F2. Since even a straight F2 acts as a cue, there seems no way of directly isolating the effects of the other cues.

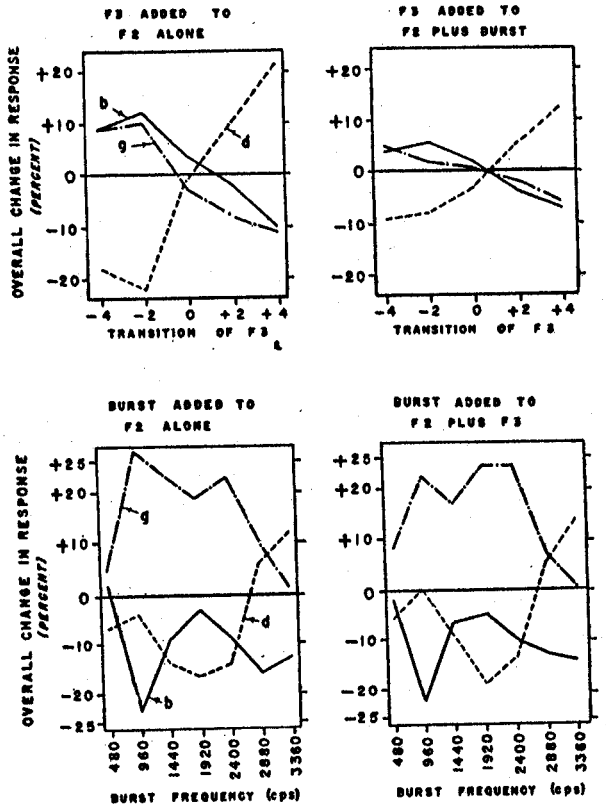

FIG. 6. The effects of the bursts and the transition of F3. This figure shows the over-all change in response resulting from the addition of a given cue to a given series of patterns.

remain the best regardless of the other cues in the patterns.

Consider, for example, the F2 cue for |d|. As seen in Fig. 5, the modes for the several response curves for |d| tend always to center at about the same value of F2. Apparently the best F2 cue for |d| tended to remain the best regardless of the nature or the number of the other cues in the patterns. Since an examination of the response curves for |b| and for |g| reveals further support for the interpretation of independence, it seems reasonable to suppose that the cues have, in fact, retained their separate effects when they are combined in various ways.[10]

This finding may be set against a background formed by the research of George Miller and Patricia Nicely.[11] These investigators took account of the confusions that occurred when various spoken consonants were heard against increasing amounts of noise. They found that listeners would misidentify the sounds with regard to place of articulation long before they would begin to misidentify the sounds with reference to manner of articulation. In general, their results led them to the conclusion that spoken consonants contain separate cues for place and manner of articulation and that these classes of cues have largely independent effects in perception.

Apparently the perceptual responses to speech sounds are complex in that they can vary along any of several dimensions. In general, these response dimensions seem to correspond to the dimensions of the articulations involved in the production of the sounds. The cues to the perception of the phonemes enable the listener simultaneously to fix his response along each of these dimensions.

Miller and Nicely's experiment suggests that the separate dimensions of the perceptual response are controlled by separate classes of cues and that, in general, the cues which control one dimension of the response do not affect the other dimensions. The results of the present study suggest a tentative answer to one of the next logical questions, namely, can the several cues which control the *same* dimension of the response also have effects which are independent of each other? As indicated in our previous discussion, the answer seems to be yes.

Although the cues appear to have independent effects, the results of this study also suggest that the cue properties of a given stimulus element are sometimes relatively complex. Typically, when we speak of a cue for a given response we mean that the stimulus tends to evoke that response at the expense of all other responses. Although this is often the case (for example, as seen in Fig. 6 the +4F3 enhances |d| at the expense of both |b| and |g|), there are some cases in which the cue tends to elicit a given phoneme at the expense of only one of the other competing phonemes. For example, as seen in Fig. 6, the 960 burst enhances |g| at the expense of |b|; the 1920 burst, on the other hand, produces almost the same degree of enhancement in |g|, but does so primarily at the expense of |d|. It is as if the cues have directional properties, somewhat like vectors. The two different cues elicit the same response, but do so by pulling from entirely different directions.

[10] One possible exception to the concept of independence would seem to be the negative F3 transition. As seen in Fig. 5, these transitions tend to enhance |b| when the transition of F2 is negative, and they tend to enhance |g| when combined with a positive F2 transition. In view of the fact that this case is unique with respect to the rest of our data, it seemed appropriate to examine the possibility that the effect might be explained without giving up the concept of the cues making independent contributions to perception.

When negative transitions of F3 are added to stimuli which tend to yield |b|, we get an enhancement in |b|. On the other hand, when these same negative F3 transitions are added to stimuli which tend to yield |g|, we get an enhancement in |g|. The possibility exists that the negative F3 transitions act as cues for *both* |b| and |g|, but when combined with a cue for |b|, the |b| tendencies summate so as to overwhelm the |g|. Similarly, in combination with a cue for |g|, the |g| tendencies summate so as to overwhelm the |b|. According to this explanation, one would expect that if a negative F3 transition is added to a cue which has little or no differential effects for either |b| or |g|, we should observe an enhancement in both |b| and |g|. When we examine the data we find that this does, in fact, happen—although the effects are small.

[11] G. A. Miller and P. E. Nicely, J. Acoust. Soc. Am. 27, 338–352 (1955).

On this basis one might suspect that in combining cues we are performing an operation which is very much like the addition of vectors. In general, this is what we seem to find. If, for example, a cue for |d| is added to a pattern which yields |g|, a sound is obtained which leads to equivocation between |g| and |d|. It is as if we have added two opposing vectors and so produced a resultant which points to something in between. If, on the other hand, the same cue for |d| is added to a sound in which there is already equivocation between |g| and |d|, the confusions are resolved in favor of |d|. It is as if the |d| components of the two vectors summate to push the response all the way to |d|.

Considerations such as these suggest that on a qualitative level, at least, some sort of vector model is useful. One hopes, however, that eventually it will be possible to develop a model which will handle the quantitative aspects of these results. It seems clear from the present study that such a model will involve a set of assumptions to account for the vectorlike properties of the separate cues. It is also clear, however, that if one is to formulate such a model it will be necessary to have more information than is now available. In particular it would be important to find out what happens to perception when strong cues for the same phoneme are combined. In the present study most of the two-formant patterns yeilded responses near the 100% point. As a result, the response measure (listener agreement) could show little or no increase when a second or third cue for the same phoneme was added to the two-formant pattern. Perhaps by using some auxiliary response measure (latency, for example) it will be possible to overcome this difficulty.