XI. SPEECH COMMUNICATION*

Prof. M. Halle†	Dr. T. T. Sandel	M. H. L. Hecker
Prof. K. N. Stevens	Jane B. Arnold	J. M. Heinz
Prof. J. B. Dennis	P. T. Brady	D. L. Hogan
Dr. A. S. House	O. Fujimura‡	A. P. Paul
	H. Fujisaki	

A. SOME SYNTHESIS EXPERIMENTS ON STOP CONSONANTS IN THE INITIAL POSITION

As a result of studies at Haskins Laboratories (1), and elsewhere (2, 3, 4), it has been recognized that formant transitions constitute important cues for the recognition of initial voiced stop consonants. Some generally accepted rules concerning the formant transitions associated with different consonants have emerged from these studies, but there are considerable areas of disagreement. We have carried out a series of speech-synthesis experiments that are designed to supplement the results obtained previously and to study certain features that have not been examined in detail heretofore.

Two different speech synthesizers have been used in the present experiments: a dynamic analog of the vocal tract (called DAVO), and a terminal analog (called POVO). The characteristics of the synthetic speech produced by DAVO are controlled in terms of the articulatory configuration (5). The control system of this synthesizer specifies precisely the time course of the articulatory changes, as well as the time-variant characteristics of the source excitation (buzz and/or noise); that is, the amplitudes, the fundamental frequency of the buzz, and the amount of amplitude modulation of the noise by the buzz. The same control system is used for POVO, but in this case the formant transitions are specified directly and the formant frequencies are changed in a piecewise-linear fashion. POVO can be excited by the same accurately controlled sources that are used to excite DAVO. The naturalness of the sounds generated by these synthesizers is generally very good when the variables are appropriately controlled, and it is often possible to generate stimuli that are judged almost unanimously to be members of a given phoneme. Consequently, we are in a position to evaluate not only the so-called primary cues, but also the secondary cues whose effects might be obscured if stimuli of poor quality were used (6).

Before performing the detailed synthesis experiments, we made a series of informal evaluations of the sounds produced by POVO with a buzz source. We found that the rules

---

*This work was supported in part by the U.S. Air Force Command and Control Development Division under Contract AF19(604)-6102; and in part by National Science Foundation.

†On leave, 1960-61, as Guggenheim Fellow at the Center for Study of the Behavioral Sciences, Stanford University.

‡On leave from the Research Institute of Communication Science, University of Electro-Communications, Tokyo, Japan.

provided by the Haskins group for the most favorable transitions of the second and third formants were generally appropriate, but the identifiability or the naturalness of the sounds was quite different for different stop-vowel combinations. Some syllables were ambiguous or not acceptable as the intended sound on an absolute-judgment basis, even when the control parameters were adjusted to give stimuli of optimum quality. The parameters that were controlled were the starting points and the duration of the linear shift of the lower three formants; the onset characteristics and the changes in frequency of the buzz source were also controlled. (In some cases more acceptable sounds could be generated with DAVO.) Synchronization of the time clock (5) to the buzz source was effective in improving the quality of initial voiced stops. Use of a very fast shift of formant frequencies was sometimes necessary, particularly for the study of the most favorable transition rate for a syllable /b/ + vowel. In this connection, we modified the output stage of the POVO circuits so that no appreciable thump would affect the output waveform even in a 10-msec transition (7).

a. Voiced Stop Followed by /ɛ/

The informal tests discussed above had demonstrated that reasonably satisfactory versions of the voiced stop consonants /b/, /d/, and /g/ can be obtained with both POVO and DAVO when these consonants are combined with the vowel /ɛ/. Synthetic syllables with this vowel context were selected, therefore, as test materials for detailed studies of the influence of formant transition rate and of the effect of introducing an initial noise burst on the identifiability of voiced stop consonants. Stimuli for formal listening tests were prepared for POVO and DAVO.

For the sounds generated by POVO, appropriate starting frequencies for the formants at the beginning of the transition were determined by listening to the POVO output. The values finally selected and employed in the formal tests are given in Table XI-1. The formant frequencies for the vowel were selected similarly by informal listening; the previously reported results for natural utterances (8) were taken into consideration.

For DAVO an articulatory configuration appropriate for the vowel /ɛ/ was determined by methods that took into account data from previous experiments with DAVO (5) and provided a perceptual quality, formant frequencies, and a vocal-tract geometry consistent with known data for the vowel. The articulatory configurations for the consonants (the starting points of the transition) were determined in similar fashion. Measured formant frequencies for these DAVO configurations are listed in Table XI-1. The frequencies of the first formants for the consonants were rather high because a sufficiently narrow constriction could not be attained with the continuously controlled circuit elements.

In each of two similar experiments, one with POVO and another with DAVO, 96 test items were recorded in random order on a magnetic tape. Four formant transition rates for transition durations of 10, 20, 30, and 40 msec were provided for each of the

Table XI-1. Formant frequencies for the consonants (the starting points of the transitions) and for the vowel /ɛ/ for the stimuli used in the tests of identification of voiced stops followed by /ɛ/.

	Formant Frequencies (cps)	b	d	g	ɛ
POVO	F_1	185	185	200	490
	F_2	1150	1750	2100	1750
	F_3	2200	2700	2250	2300
	F_4	3300	3300	3300	3300
DAVO	F_1	240	240	270	485
	F_2	1160	1740	2060	1725
	F_3	2200	2750	2260	2300
	F_4	3200	3680	3420	3170

syllables /bɛ/, /dɛ/, and /gɛ/. For POVO, half of the /g/-stimuli were produced with a short burst of noise at the onset of the syllable (g_1), and the rest were produced without the noise (g_0). The noise burst was generated by mixing a short burst of white noise with the buzz signal at the input to the POVO circuit. For DAVO, a similar noise burst was applied at the glottis for half of the g-stimuli, and at the articulatory constriction for half of the d-stimuli. In one test (a forced-judgment test), 6 subjects were asked to identify each of the stimuli as one of the three syllables; and in another test (an unforced-judgment test), as one of five possibilities, /bɛ/, /dɛ/, /gɛ/, /ɛ/ or "something else." The results are given in Table XI-2, and from these data the following conclusions have been drawn.

(a) For all of the three stops practically perfect identification could be obtained without the use of the noise burst, when the transition time was appropriate. This was true for sounds generated both by POVO and by DAVO.

(b) The most favorable transition duration was relatively short for /b/, long for /g/, and intermediate for /d/. (Preliminary tests showed that a longer duration (as long as 60 msec) was acceptable for /gɛ/.)

(c) There is considerable difference in the score when the forced judgment is compared with the unforced judgment. In the POVO experiment, for example, a 10-msec transition for /b/ and a 20-msec transition for /d/ gave rise to almost perfect identification (92 per cent and 94 per cent, respectively) when the subjects were forced to choose one from among /b/, /d/, and /g/, whereas for the unforced judgment the scores were

62 per cent and 73 per cent, respectively.

(d) In the DAVO experiment, a marked effect of the noise burst on the recognizability was observed. When the transition rate was relatively high, the presence of the burst helped to separate /gɛ/ from /ɛ/. For a slower transition, it was helpful in distinguishing /dɛ/ from both /gɛ/ and /ɛ/. The use of a noise source thus expanded the range of the appropriate formant-transition rates.

Table XI-2. Correct judgment score for the stimuli /bɛ/, /dɛ/, and /gɛ/ produced by POVO and DAVO. The subscript 1 to the consonants indicates the presence of the noise burst; the subscript o, its absence. The consonants without subscripts were produced without noise. Numbers represent the percentages of correct judgments out of 48 votes when there is no subscript, and out of 24 votes when there is a subscript, for the forced-judgment test. Numbers in parentheses represent the results of the unforced-judgment test.

POVO

Duration of Formant Transition	b	d	g_1	g_o
10 msec	92 (62)	44 (12)	50 (25)	54 (0)
20 msec	100 (100)	94 (73)	88 (58)	88 (62)
30 msec	100 (100)	98 (100)	100 (92)	100 (100)
40 msec	98 (100)	100 (100)	100 (96)	100 (96)

DAVO

Duration of Formant Transition	b	d_1	d_o	g_1	g_o
10 msec	100 (100)	83 (83)	79 (79)	96 (100)	79 (54)
20 msec	100 (100)	92 (100)	100 (92)	100 (100)	92 (83)
30 msec	100 (92)	100 (92)	88 (67)	100 (100)	100 (92)
40 msec	100 (92)	88 (83)	92 (67)	100 (100)	100 (96)

Considerable difference can be observed in the preferred transition time when the score obtained for DAVO is compared with that for POVO. For DAVO, the subjects always preferred faster transitions and, in particular, they could identify /b/ with a very fast transition (10-msec duration) without difficulty. The difference may be partially ascribable to a smoothing circuit (with a 5-msec time constant) which was

employed for the control signal of DAVO, but not for POVO. Another possible explanation is that the nonlinear transition of the formants (at the acoustic level), which depends on the particular configuration and is automatically given by DAVO, can have a perceptual effect that is significantly different from that of the linear formant transitions given by POVO.

b. Cues for /gɑ/

In the informal tests described above the syllable /gɑ/ was one of the stop + vowel combinations that could not be synthesized successfully by appropriate selection of a linear formant transition. Further informal evaluations of this syllable as generated by the POVO synthesizer indicated that the quality of the output could be improved considerably by modifying the syllable in the following manner: (a) by mixing a short burst of noise with the buzz excitation at the initial portion of the syllable; (b) by using a higher starting point for the first formant, for example, 450 cps instead of 200 cps; and (c) by delaying the start of the formant transition, thereby leaving a short buzz-excited portion with stationary formants.

Table XI-3. Formant frequencies for the consonants and for the vowel /ɑ/ in the tests of cues for the syllable /gɑ/.

Formant Frequencies (cps)	b	d	g	g'	ɑ
F_1	200	200	200	450	750
F_2	800	1800	1500	1500	1100
F_3	2000	2800	1800	1800	2500
F_4	3300	3300	3300	3300	3300

Table XI-4. Percentages of votes indicating preferences for the stimulus of the row over that of the column. For the stimuli represented by g_1 and g_2 the starting point of the first formant is at 225 cps; for g_3 and g_4, at 450 cps. For g_1 and g_3, the formant transition starts at the same time as the buzz onset; for g_2 and g_4 it starts 10 msec later. (See Fig. XI-1.)

	g_1	g_2	g_3
g_2	31		
g_3	85	94	
g_4	94	92	85

In order to evaluate the effects of these features, we have conducted two listening tests. In the first experiment, the effect of these features in combination was ascertained by an absolute-judgment test. Syllables /bɑ/, /dɑ/, and /gɑ/ were produced by POVO by a procedure similar to that described above for the synthesis of /bɛ/, /dɛ/, and /gɛ/. The formant frequencies for the consonants (the starting points), as well as for the vowel, are given in Table XI-3. Using the same second- and third-formant frequencies for /g/, we synthesized another /gɑ/ in which the three additional features were included. This second /gɑ/, which will be called the g'-syllable, was mixed with the three other syllables /bɑ/, /dɑ/, and /gɑ/ in randomized order to make a recorded tape for the listening test. The time structures of the stimuli, as specified by the control signals, are shown in Fig. XI-1. The control signals for b, d, and g are compared

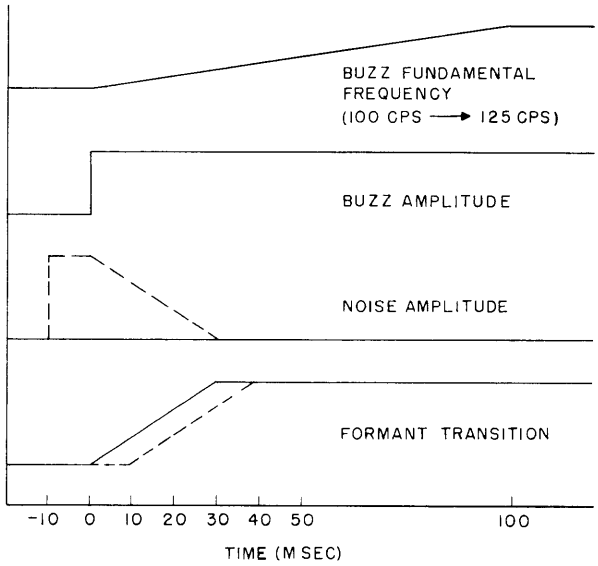

Fig. XI-1. Forms of the control signals for the fundamental frequency of buzz source, buzz amplitude, noise amplitude, and formant frequencies. Solid lines are for the stimuli b, d, and g; broken lines, for the stimulus g'.

with those for g'. The recorded tape contained a total of 40 items, 10 for each of the four syllables. Eleven subjects listened to the stimuli, and they were forced to choose one from among /bɑ/, /dɑ/, and /gɑ/. The results were that b was judged as /b/ by 99 per cent of 110 votes, and as /d/ by 1 per cent (one vote); d was judged correctly without any error (110 unanimous votes); g was judged as /d/ by 79 per cent, as /g/ by 19 per cent, and as /b/ by 2 per cent; g' was judged as /g/ by 99 per cent and as /d/ by 1 per cent (one vote).

The judgments of the g-items varied considerably from subject to subject, the number of /g/ responses ranging from 7 to 0 votes out of 10. For most subjects, however,

it is clear that the shift from /dɑ/ to /gɑ/ was caused by the three additional features.

In order to estimate the relative importance of the three cues, a second experiment was conducted. By preliminary listening, it was found that an identifiable /gɑ/ was not obtained unless the first feature – the presence of noise – was included. Therefore, the remaining four combinations of presence or absence of the last two features were compared in a paired comparison test, with the noise always employed. Twenty-four pairs consisting of 12 different pair items were recorded in random order, and the subjects were asked to judge which of the pair was more like an utterance of /gɑ/. Every pair item was presented twice for one judgment. Eight subjects participated in the test, and five of them repeated the same test five months later. All of these judgments are summarized in Table XI-4. The designations of the four stimuli, g_1, g_2, g_3, and g_4, are explained in Table XI-4. A total of 52 judgments was made for each of the 6 combinations.

Versions g_3 and g_4 were preferred to either g_1 or g_2 in 91 per cent of the judgments, and hence the higher location of the starting frequency of the first formant definitely improves the quality. The delay of the start of the formant transition was preferred when the other condition (concerning the first formant) was favorable (85 per cent preference for g_4 over g_3). The delay feature alone, on the other hand, does not seem to improve the sound. The stimulus designated as g_4 in this experiment is approximately identical to the g' in the previous absolute-judgment test.

It is interesting to note that a noise of 10-msec duration preceding the voice onset was necessary to make an unambiguous /gɑ/. This duration of unvoiced hiss is often observed in spectrograms of natural utterances of /g/ in the initial position in an isolated syllable. It is plausible that the reaction (in the low-frequency region) of the vocal tract to the vocal cords is particularly appreciable in the case of the velar stop because the cavity behind the occlusion is very small. An appreciable disturbance or a delay of the onset of regular vibration of the vocal cords can be expected as a result of the occlusion or its sudden release.

A higher location of the starting point of the first-formant transition can be a cue for velar stops as opposed to dental or labial stops. The locus of the first formant may be significantly higher for /g/ than for /b/ or /d/ because the cavity behind the occlusion of /g/ is substantially smaller than that for other stops, and possibly because the glottis is more open for /g/ than for /b/ or /d/. A value as high as 450 cps, however, does not necessarily represent the actual resonant frequency for the g-configuration. It is quite possible that the initial part of the rapid transition is overlooked in the perception of the natural utterance. In view of this, the preference for the delayed formant transition is particularly interesting. In the spectrograms of natural utterances of /gɑ/, a characteristic separation of the consecutively located second and third formants takes place (2, 3), but this separation starts very gradually and apparently only at a later stage

relative to the start of the first-formant transition. This characteristic is caused by a nonlinear relation between the resonant frequencies and the articulatory dimensions. Thus the entire pattern of the formant transition can be simulated effectively by the use of the additional features, if the initial part of the rapidly changing first formant is overlooked.

c. Effect of a Fundamental Frequency Inflection on Tense-Lax Opposition

This portion of the report describes a study of the cues for tense-lax opposition of stop consonants in the initial position, with the DAVO synthesizer used. The test items consisted of the consonant pair g-k in combination with the vowel /ɛ/. Test stimuli were prepared with 10 different buzz onset times (in steps of 10 msec), and the presence or absence of a fundamental frequency inflection was also controlled. The fundamental frequency, when it had an inflection, was shifted linearly from 70 cps to 100 cps during the initial 150-msec period of the voicing. A white-noise source was applied at the glottis input of the electrical vocal tract during the initial portion of the syllable. The noise started abruptly (with an onset time of 1 msec), and then immediately decayed linearly to zero in 50 msec. The vocal-tract configurations for the consonant and the vowel were similar to those used in the previous experiment, and the g-configuration was shifted toward the ɛ-configuration with a 40-msec transition time. For all test stimuli, the start of the noise and the configuration transition occurred simultaneously. The buzz started abruptly (with a 1-msec ramp) and the inflection started at the same time. In this experiment the ramp starting times were not synchronized with the buzz waveform.

Eighty test items containing 4 each of 20 different stimuli were recorded in a randomized order. The subjects were asked to decide whether the stimulus sounded like /gɛ/ or like /kɛ/. Ten American subjects and three Japanese subjects participated in the listening test. The results for two subjects who gave particularly interesting responses are shown in Fig. XI-2. Both subjects P. B. and H. F. showed the same complete (4 votes) jump from g to k in one step, that is, with a 10-msec shift of the buzz onset time, when there was no fundamental frequency inflection. When the inflection was present, this boundary shifted invariably to a later onset time for all subjects, but the curve of the stimuli with inflection present appeared considerably different from subject to subject. Subject P. B. did not give any g-responses if the onset time was 50 msec, or more, after the burst of noise; H. F., on the other hand, was reluctant to give a k-vote, even with an unvoiced period as long as 80 msec, but he had no problem when there was no inflection. This subject (H. F.) is Japanese, but other Japanese subjects did not show this marked behavior. This tendency, however, was often seen in American subjects, although usually less markedly. The average for 10 American subjects is shown in Fig. XI-3.

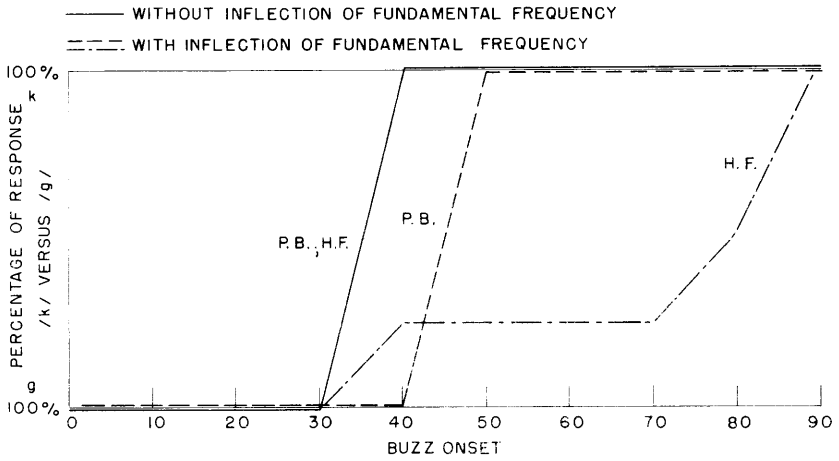

Fig. XI-2. Responses /kɛ/ versus /gɛ/ as a function of the buzz onset time for two conditions of inflection of the fundamental frequency. Curves are shown for two subjects, P. B. and H. F. The ordinate represents the percentage of k-judgments (as opposed to g-judgments).

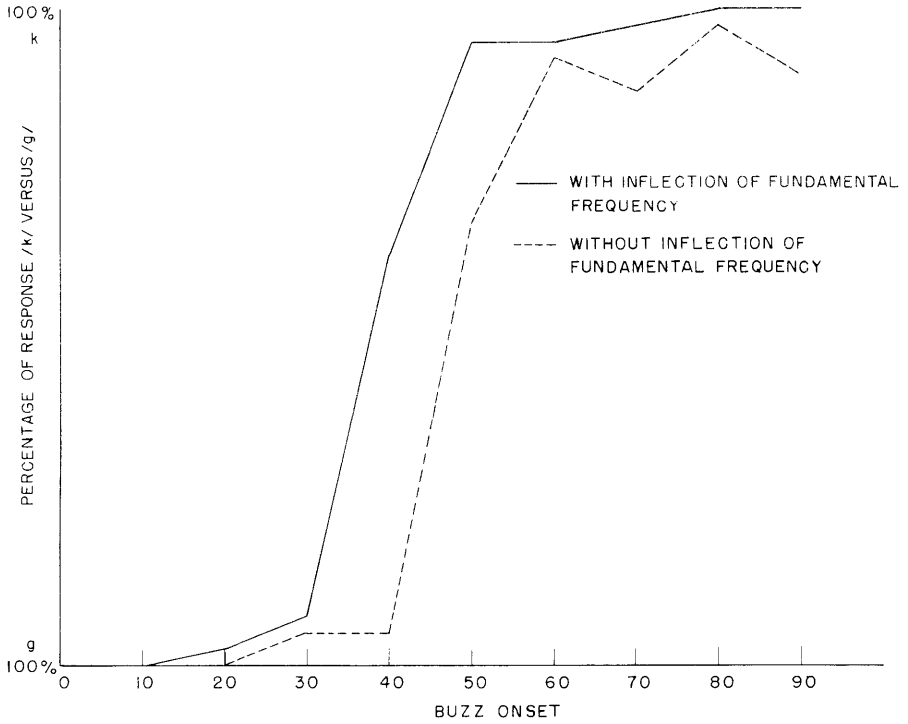

Fig. XI-3. Averaged responses of 10 subjects for the same experiment as in Fig. XI-2.

It would appear that an inflection of the fundamental frequency, as used in this experiment, is not favorable for the naturalness of syllables consisting of a tense stop + a vowel, whereas it is favorable for voiced stops. Also, the judgment can possibly be switched from tense to lax just by changing the inflection, if the buzz onset is in the boundary region. For the initial stops of English, the primary cue of tense versus lax distinction is considered as a delay of buzz onset time associated with a certain amount of unvoiced hiss, while the inflection in the fundamental frequency of voice constitutes a secondary cue.

A similar experiment has been conducted for the alveolar stops combined with the vowel /i/. In this experiment the noise was inserted at the articulatory constriction, and the ramp initiation was synchronized with the buzz. The range of buzz onset times was decreased, and 10 judgments were collected from each subject for each condition of the test stimuli. The result was very similar to that of the experiment just described, except that the shift from /di/ to /ti/ occurred at an earlier value of the buzz onset time.

O. Fujimura

References

1. A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. J. Gerstman, The role of consonant-vowel transitions in the perception of stop and nasal consonants, Psychol. Monogr. Vol. 68, No. 8, pp. 1-13, 1954; K. S. Harris, H. S. Hoffman, A. M. Liberman, P. C. Delattre, and F. S. Cooper, Effect of third-formant transition on the perception of the voiced stop consonants, J. Acoust. Soc. Am. 30, 122-126 (1958); H. S. Hoffman, Study of some cues in the perception of the voiced stop consonants, J. Acoust. Soc. Am. 30, 1035-1041 (1958); P. C. Delattre, A. M. Liberman, and F. S. Cooper, Acoustic loci and transitional cues for consonants, J. Acoust. Soc. Am. 27, 769-773 (1955).

2. M. Halle, G. W. Hughes, and J. P. Radley, Acoustic properties of stop consonants, J. Acoust. Soc. Am. 29, 107-116 (1952).

3. E. Fischer-Jørgensen, Acoustic analysis of stop consonants, Miscellanea Phonetica 2, 42-59 (1954).

4. C. G. M. Fant, Acoustic Theory of Speech Production (Mouton and Company, The Hague, 1960).

5. G. Rosen, Dynamic Analog Speech Synthesizer, Technical Report 353, Research Laboratory of Electronics, M.I.T., Feb. 10, 1960.

6. A. S. House, Formant bandwidth and vowel preference, Quarterly Progress Report No. 56, Research Laboratory of Electronics, M.I.T., Jan. 15, 1960, pp. 172-173.

7. Final Report on Research on Speech Synthesis, Report AFCRC-TR-60-101, December 1959.

8. G. E. Peterson and L. L. Barney, Control methods used in a study of the vowels, J. Acoust. Soc. Am. 24, 175-184 (1952).