XVIII. SPEECH COMMUNICATION

Prof. M. Halle	Jane B. Arnold	E. H. Lenneberg
Prof. K. N. Stevens	P. T. Brady	C. I. Malme
Dr. T. T. Sandel	O. Fujimura	F. Poza
G. W. Hughes	J. M. Heinz	G. Rosen

RESEARCH OBJECTIVES

The process of communication normally involves a code, a transmitter and receiver, a channel, and a signal. In speech communication the code is the language in which the utterance is spoken; the transmitter and the receiver are the speaker and the listener, respectively; the channel is normally the ambient air of the human environment; and the signal is the acoustic wave produced by the speaker and received by the listener. In the cases that are best known to communication engineers the code and the properties of the transmitter and of the receiver are completely specified. In speech communication our knowledge of these factors is still fragmentary, and it is therefore necessary to gather more information about the factors that determine the physical signal; that is, about the nature of language and the properties of speakers and listeners.

The main problems which we are now pursuing in our research are: (a) What are the capabilities of the human vocal apparatus as a source of speech signals? What limitations are imposed on the signal by the restricted control that man has over his vocal organs? (b) How does man perceive sounds, in general, and speech-like sounds, in particular? What are the limitations of the human organism in dealing with acoustic stimuli that serve as vehicles for the identification of messages? (c) What are the physical properties of utterances in various languages? In particular, what are the properties that serve to distinguish utterances that are different from each other? (d) What is the nature of human language, in general, and of various national languages, in particular? How does the structure of the language determine the physical properties of utterances?

Part of our present research effort is devoted to the development of apparatus for speech synthesis and analysis, as well as to the exploration of new techniques that involve the use of high-speed digital computers. Also within our sphere of interest are several practical applications of research on speech communication, including devices for automatic speech recognition and for the efficient coding of speech information.

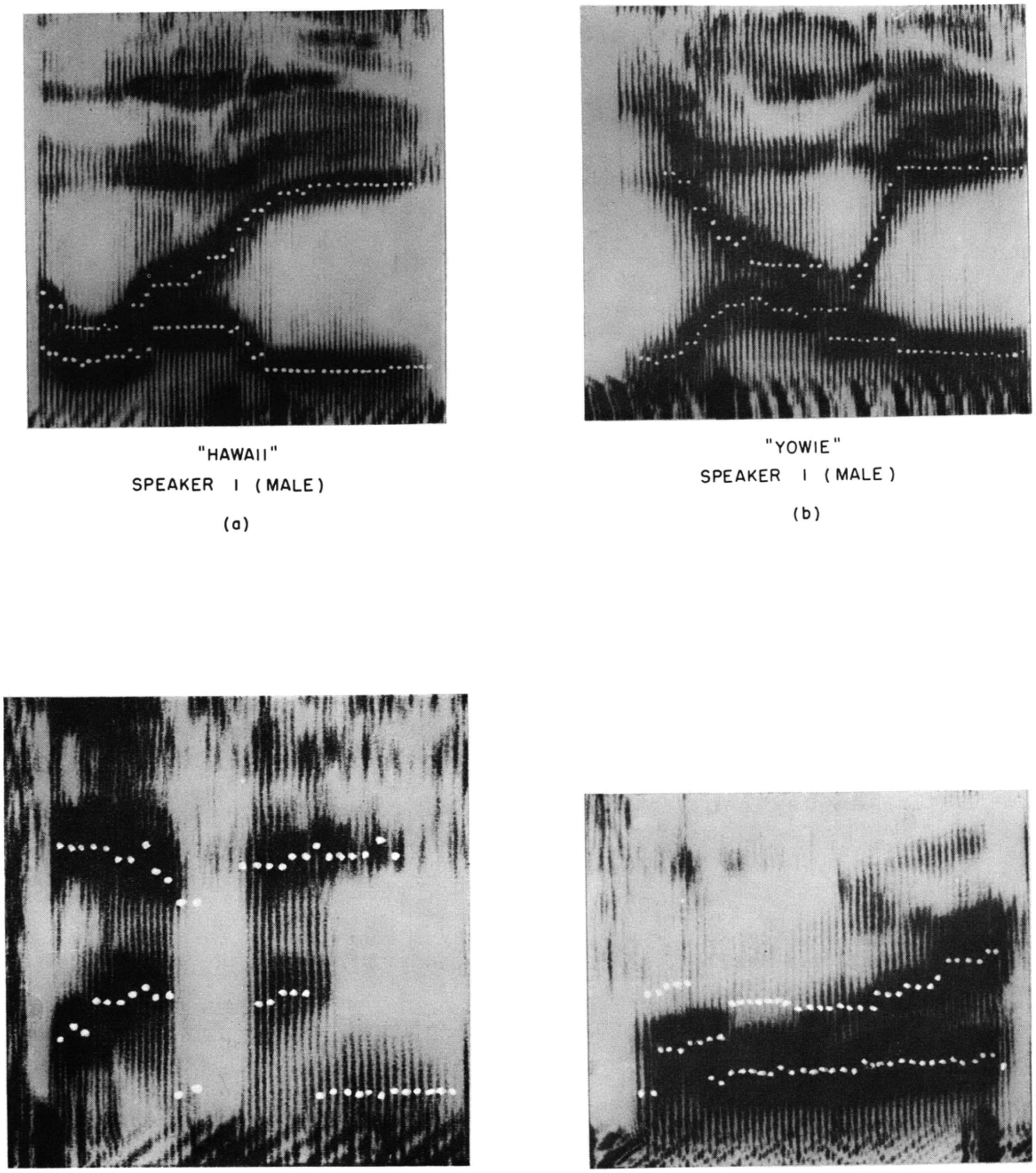
M. Halle, K. N. Stevens

A. SPEECH ANALYSIS WITH THE AID OF A DIGITAL COMPUTER*

We are investigating the extraction of the information-bearing elements of speech, using a large-scale digital computer as a research tool. Because of the speed, reliability, and flexibility of digital computers, they lend themselves well to speech research problems that require the collection and reduction of large amounts of data for obtaining statistically significant results. Also, many hypotheses can be tested, proven, modified, or rejected without the necessity of constructing equipment.

A real-time speech input system has been devised to function with the Whirlwind I computer (1). The rectified and smoothed output of each of 35 bandpass filters (2) covering the frequency range 115 cps to 10,000 cps is sampled 180 times per second. An Epsco Datrac performs the analog-to-digital conversion. Spectral and amplitude

*This work was supported in part by National Science Foundation.



"HAWAII"
SPEAKER 1 (MALE)
(a)

"YOWIE"
SPEAKER 1 (MALE)
(b)

"SEVEN"
SPEAKER 8 (FEMALE)
(c)

"FOUR"
SPEAKER 3 (MALE)
(d)

Fig. XVIII-1. Output of formant-tracking portion of the computer program plotted directly (white dots) on sonagrams of the input utterances.

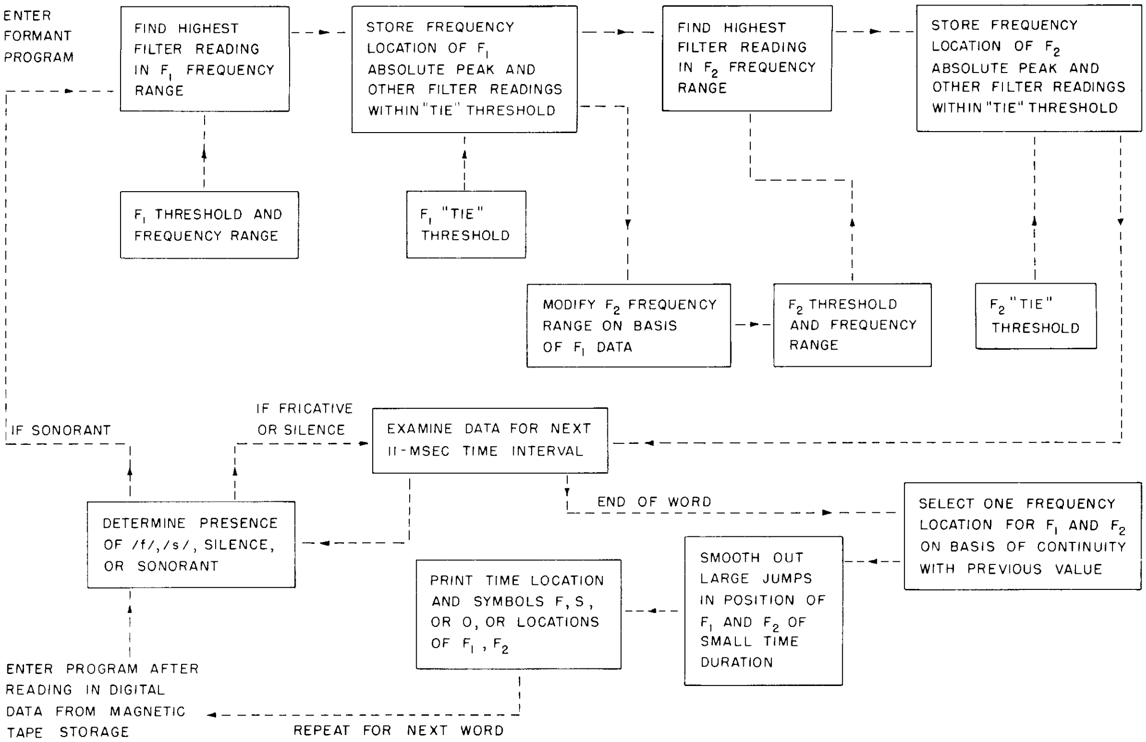

ENTER FORMANT PROGRAM
FIND HIGHEST FILTER READING IN F₁ FREQUENCY RANGE
STORE FREQUENCY LOCATION OF F₁ ABSOLUTE PEAK AND OTHER FILTER READINGS WITHIN "TIE" THRESHOLD
FIND HIGHEST FILTER READING IN F₂ FREQUENCY RANGE
STORE FREQUENCY LOCATION OF F₂ ABSOLUTE PEAK AND OTHER FILTER READINGS WITHIN "TIE" THRESHOLD
F₁ THRESHOLD AND FREQUENCY RANGE
F₁ "TIE" THRESHOLD
MODIFY F₂ FREQUENCY RANGE ON BASIS OF F₁ DATA
F₂ THRESHOLD AND FREQUENCY RANGE
F₂ "TIE" THRESHOLD
IF SONORANT
IF FRICATIVE OR SILENCE
EXAMINE DATA FOR NEXT 11-MSEC TIME INTERVAL
END OF WORD
SELECT ONE FREQUENCY LOCATION FOR F₁ AND F₂ ON BASIS OF CONTINUITY WITH PREVIOUS VALUE
DETERMINE PRESENCE OF /f/,/s/, SILENCE, OR SONORANT
SMOOTH OUT LARGE JUMPS IN POSITION OF F₁ AND F₂ OF SMALL TIME DURATION
PRINT TIME LOCATION AND SYMBOLS F, S, OR O, OR LOCATIONS OF F₁, F₂
ENTER PROGRAM AFTER READING IN DIGITAL DATA FROM MAGNETIC TAPE STORAGE
REPEAT FOR NEXT WORD

Fig. XVIII-2. Block diagram of computer program.

information at the rate of 11 bits every 154 μsec is then stored in the computer memory for analysis and/or semipermanent storage on magnetic tape.

The Whirlwind I computer has been programmed to perform the following operations:

1. Locate fricative sounds in a word and classify them as /f/ or /s/. The class /ʃ/ will be added shortly. As yet, no attempt has been made to include the voiced-voiceless distinction.

2. Locate the frequencies of the two lowest vocal tract resonances (track formants 1 and 2) during the sonorant portion of an utterance.

3. Calculate the over-all level of sound either from the level channel included in the input, as the linear sum, or the sum of the squares of the amplitudes of any preset number of filters.

In the speech data processed thus far, all fricatives were correctly located and classified into one of the two available categories. On the basis of level and formant position discontinuities some success was achieved in locating boundaries between sonorant consonants and vowels, notably of the nasal-vowel type.

Results from the formant-tracking program are best displayed by plotting the data output from the computer directly on a sonagram of the analyzed utterance, as in Fig. XVIII-1. Figure XVIII-1a shows good performance; Figs. XVIII-1b and c, average performance. Figure XVIII-1d shows the poorest performance that we have encountered. Figure XVIII-2 gives an outline in block diagram form of the logical operations performed by the program.

On the basis of data obtained from the computer, word classes, separated by a system of distinctive differences analogous in some ways to the system of distinctive features of Jakobson, Fant, and Halle (3) for phoneme separation, will be set up. At present, the features of an utterance that we are able to extract are too gross to allow phoneme separation in all instances. The work involving defining and tracking these features is a necessary prerequisite to further study of phoneme identification.

G. W. Hughes

References

1. J. W. Forgie and G. W. Hughes, A real-time speech input system for a digital computer, J. Acoust. Soc. Am. 30, 668 (July 1958).
2. G. W. Hughes and J.-P. A. Radley, Set of highpass and lowpass filters, Quarterly Progress Report, Research Laboratory of Electronics, M.I.T., April 15, 1955, p.74.
3. R. Jakobson, C. G. M. Fant, and M. Halle, Preliminaries to speech analysis: distinctive features and their correlates, Technical Report No. 13, Acoustics Laboratory, M.I.T., May 1952.

B. THE ROLE OF DURATION IN VOWEL IDENTIFICATION*

Measurements of the spectra of vowels indicate that these sounds can be distinguished from each other reasonably well on the basis of the frequencies (F_1 and F_2) of the first two formants (1). There is, however, some overlap in the frequency ranges of these formants for some vowels, and it is apparent, therefore, that cues other than F_1 and F_2 may play a role in vowel identification. The experiment described here examines the role of duration in the identification of synthetic stressed vowels in nonsense syllables.

The stimuli used in the experiment are of the type schematized in Fig. XVIII-3. The stimuli were generated by a resonance-analog synthesizer, and the excitations and transfer function of the synthesizer were controlled by a device designed by Rosen, which has been discussed previously (2). The beginning and end of the vowel were always characterized by transitions: The first formant started at approximately 180 cps and moved to the first-formant frequency of the vowel in 40 msec; the second formant started at approximately 1800 cps and moved to the second-formant frequency of the vowel in a piece-wise linear fashion, as shown in Fig. XVIII-3. The transitions at the end of the vowel were similar to the initial transitions. A burst of noise with energy concentrated

*This work was supported in part by National Science Foundation.

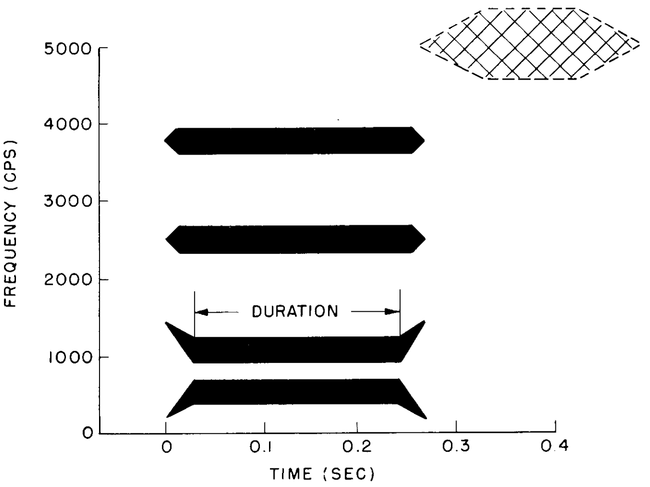


Fig. XVIII-3. Schematized intensity-frequency-time pattern for the test stimuli. The crosshatched area represents a noise burst. The stimulus represents a syllable with initial consonant /d/ and final consonant /s/.

in the vicinity of 5000 cps was generated after the vowel, with appropriate rise and decay times. The synthetic syllables could be readily identified as beginning with /d/ and ending with /s/.

The variables in the experiment were the duration and formant frequencies of the steady vowel portion of the synthetic syllable. The duration, defined in the manner shown in Fig. XVIII-3, could be varied through five values: 25, 50, 100, 200, and 400 msec. Formant frequencies appropriate to back and front vowels were examined in two different tests. In each test, seven different combinations of formant frequencies were used. The particular selection of values of formant frequencies was based on previous data from measurements of formant frequencies for male voices. The values of F_1 and F_2 for front vowels covered the ranges from 250 to 700 cps and from 2300 to 1700 cps, respectively. In back vowels the corresponding ranges were 300 to 730 cps and 900 to 1170 cps, respectively.

The stimuli for each test were tape-recorded in random order, each stimulus being presented three times throughout the test. The test recordings were presented to a group of 10 listeners through a high-quality loud-speaker system. In the test involving syllables with front vowels, the subjects were asked to identify each vowel as one of /i, I, ϵ, æ/. The permissible responses for the back vowel sequence were /u, U, Λ, ɑ/.

The main features of the results for front and back vowels are summarized in Fig. XVIII-4a and b. The seven combinations of F_1 and F_2 in each case are shown on the abscissa. The crosshatched areas depict the values of formant frequency and duration that elicited the various vowel responses at the 75 per cent level or higher. The

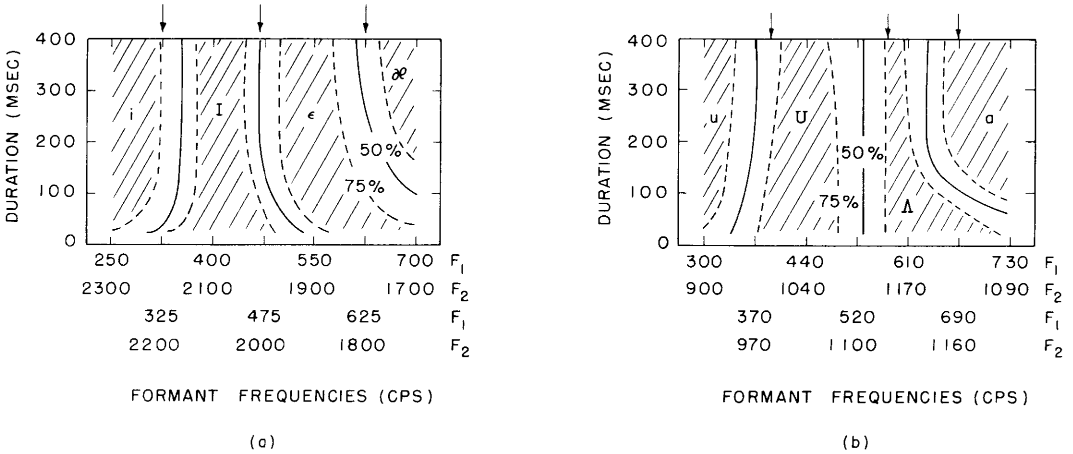


Fig. XVIII-4. Results of vowel identification tests: (a) front vowel series, and (b) back vowel series. Within the crosshatched areas, for each vowel indicated, the response level was higher than 75 per cent; solid lines represent the 50 per cent response level. The arrows at the top of each figure represent the values of formant frequencies corresponding to the boundaries between vowels reported by Peterson and Barney.

solid lines represent the boundaries corresponding to 50 per cent response.

For durations of less than approximately 100 msec, vowels are judged to be /ɛ/ or /Λ/ even though the formant frequencies are in ranges that are usually considered appropriate for /æ/ or /ɑ/, respectively. (The arrows at the top of each figure indicate the values of formant frequencies judged by Peterson and Barney (1) to be the boundaries between the vowels as uttered in a particular consonant context by a number of talkers.) The /i-I/ and /u-U/ distinctions are less influenced by duration. For an extremely short vowel in which the length of the steady portion approaches zero, there is a tendency for /i/ to shift to /I/ and /u/ to /U/.

At least two hypotheses might be suggested as explanations for these results. It is possible that because the vowels /i/ and /u/ are at extreme points on the vowel triangle they serve as anchor points that are reasonably stable and rather easily identifiable. On the other hand, /æ/ and /ɑ/ occupy points in the vowel triangle that are less well defined and reproducible. Thus a listener may tend to rely more on duration and less on spectrum in making /æ-ɛ/ and /ɑ-Λ/ distinctions.

A second possible hypothesis is that the perception of a short vowel is determined not by the frequencies reached by the formants but by some time-average values of formant frequencies for the duration of the vowel, including the transitions. The distances from the /d/ (or /s/) formant loci to the vowel formants are considerably greater for /æ/ and /ɑ/ than they are for /i/ and /u/. Thus as the /æ/ becomes shorter the time-average value of F_1 shifts to lower frequencies, and we might expect a shift to /ɛ/.

Likewise, we would expect a change from /ɑ/ to /Λ/ but not from /u/ to /U/. This hypothesis suggests that the perception of very short vowels depends upon the loci of consonants adjacent to them. Different results might be expected for the bilabial consonants, for example.

We expect to carry out further experiments on vowel identification, particularly with very short vowels in various consonant contexts.

K. N. Stevens

References

1. G. E. Peterson and H. L. Barney, Control methods used in a study of the vowels, J. Acoust. Soc. Am. 24, 175-184 (1952).
2. G. Rosen, Dynamic analog speech synthesizer, Quarterly Progress Report, Research Laboratory of Electronics, M.I.T., April 15, 1958, p. 102.

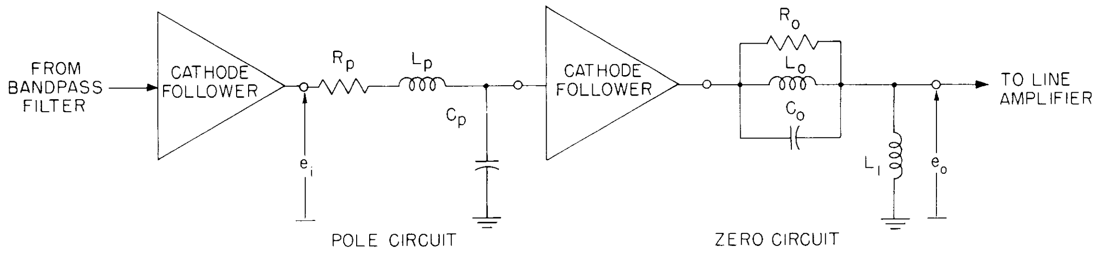
C. DETECTABILITY OF SMALL IRREGULARITIES IN A BROADBAND NOISE SPECTRUM*

The detectability of small changes in the energy content of a narrow frequency band in a white-noise spectrum was measured by using standard psychoacoustical testing techniques. The test stimuli were obtained by passing a white noise (uniform spectrum from 100 cps to 10 kc) through a filter with a resonance and antiresonance. The filter circuit that was used and its transfer function are shown in Fig. XVIII-5. This simple pole-zero circuit was used instead of commercially available bandpass filters because its response is more closely related to the frequency response of physical systems.

The test stimuli were composed of the noise spectra shown in Fig. XVIII-6. Six tests, each involving the presentation of 6 different noise spectra, were conducted. In each test the flat noise spectrum was considered as a standard against which the other spectra were compared. The flat spectrum was produced by tuning the circuits to the desired frequency and adjusting the damping of each to give a Q of 10. In the s-plane, the pole and zero would therefore be coincident. A spectrum notch was produced by increasing the damping of the pole circuit; the zero circuit was left at the reference position (Q = 10). A spectrum peak was produced by interchanging the roles of the pole and zero circuits. The bandwidth and center frequency of the resulting irregularities were thus specified in terms of the positions of the poles and zeros in the s-plane plot.

The test stimuli were presented to the subjects by means of an Ampex tape recorder and Permoflux PDR-8 earphones. Each recorded test contained 100 stimulus groups in standard "ABX" presentation. A stimulus group consisted of three noise bursts, each

*This work was supported in part by the U.S. Navy (Office of Naval Research) under Contract Nonr-1841(42). Reproduction in whole or in part is permitted for any purpose of the United States Government.



$$\frac{E_o}{E_i} = \frac{\left(S^2 + \frac{1}{R_oC_o} S + \frac{1}{L_oC_o}\right)}{L_p C_p\left(S^2 + \frac{R_p}{L_p} S + \frac{1}{L_p C_p}\right)\left(S^2 + \frac{1}{R_oC_o} S + \frac{1}{L_oC_o} + \frac{1}{L_1C_o}\right)}$$

NOTE: $L_1 << L_o$ TO PLACE THE SECOND POLE WELL OUT OF THE RANGE OF INTEREST

Fig. XVIII-5. Spectrum-shaping circuit and transfer function.

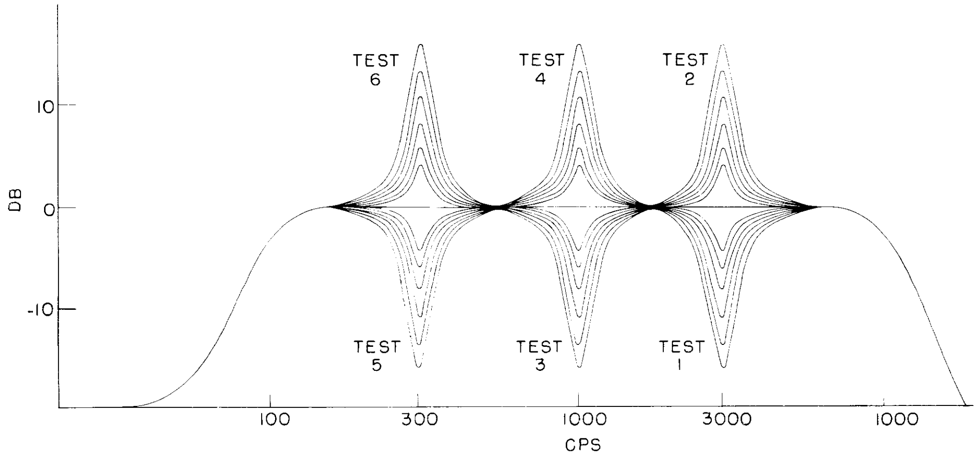


Fig. XVIII-6. Spectra of test stimuli.

of 500-msec duration with a sound-time fraction of one-half. In the ABX test, subjects were asked whether the last stimulus sounded more like the first or more like the second, and were given 5 seconds in which to record their decision. Five subjects gave a total of 80 judgments on each spectrum irregularity.

The experimental results are shown in Table XVIII-1. The magnitude of a simple peak or notch in a white-noise spectrum detectable by an average listener 75 per cent of the time was obtained by averaging the results for each subject. The damping factor of the irregularity detected by each subject 75 per cent of the time was determined. The

Table XVIII-1

Characteristics of Simple Irregularities in a White-Noise Spectrum for Aural Detection 75 Per Cent of the Time.

POLE

Resonant Frequency, f_o (cps)	300	1000	3000
Minimum Effective Damping Factor, σ (sec^{-1})	109	650	1500
Standard Deviation, (sec^{-1})	±7.7	±150	±220
Effective Bandwidth, BW (cps)	35	210	480
Peak Height (db)	9	4.2	5.2
Effective Q	8.6	4.8	6.3

Note: $$\sigma = \frac{\sigma_r}{[1 - (2/K^2)]^{1/2}}$$

where $K = \sigma_m/\sigma_r$

Also note: $BW = \sigma/\pi$; peak height $= 20 \log_{10} K$; $Q = \omega_o/2\sigma$

ZERO

Resonant Frequency, f_o (cps)	300	1000	3000
Minimum Effective Damping Factor, σ (sec^{-1})	98	420	1100
Standard Deviation (sec^{-1})	±3.6	±45	±100
Effective Bandwidth, BW (cps)	31	130	350
Notch Depth (db)	14.3	6.6	8.8
Effective Q	9.6	7.5	8.5

mean value for the 5 subjects was calculated and is shown along with its standard deviation in Table XVIII-1. Information on the peak height (or notch depth) and on the bandwidth of a just detectable irregularity is of greater interest than information on the effective pole or zero location, and hence these values are also tabulated.

It is interesting to note that it is possible to have a spectrum notch as deep as 14 db with a bandwidth of 31 cps at a center frequency of 300 cps before it becomes detectable 75 per cent of the time. At each frequency that we investigated a spectrum peak was more detectable than a spectrum notch of the same magnitude. Of the three center frequencies used, an irregularity was detected most readily at 1000 cps and least readily at 300 cps.

Figure XVIII-7 is an s-plane plot of the center frequencies and damping factors of

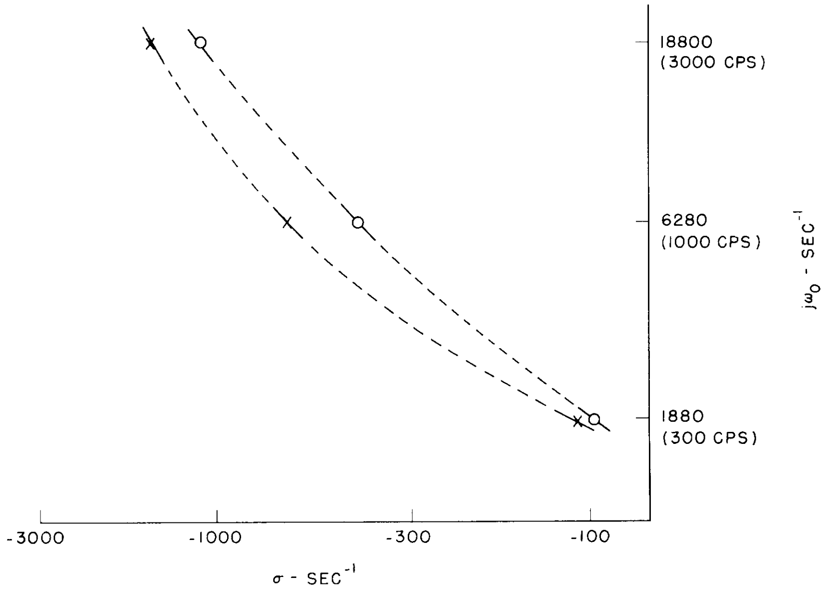


Fig. XVIII-7. s-plane plot of effective poles or zeros detectable 75 per cent of the time.

the just detectable spectrum irregularities listed in Table XVIII-1. The dashed lines indicate the possible location of just detectable irregularities at other center frequencies. Of course, additional information is needed to verify this prediction.

One possible use for a plot of this kind would be to determine the maximum deviation from flatness that can be tolerated in a transducer response curve. Other applications may be of interest in the field of physiological acoustics.

Further investigations are planned for checking the predicted curves of Fig.XVIII-7 and for determining the masking effect of the noise that is present in the spectrum but not at frequencies in the neighborhood of the irregularity. It is quite possible that this "marginal" noise exerts a considerable masking influence on the detection of a spectrum irregularity. The magnitude of a just detectable irregularity may have to be specified in terms of the frequency boundaries of the spectrum in which it is embedded.

C. I. Malme

D. DYNAMIC ANALOG SPEECH SYNTHESIZER*

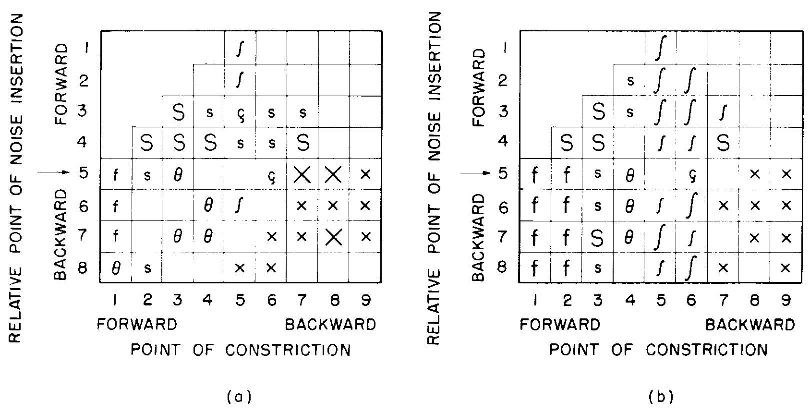
A study of the generation of fricatives by the dynamic analog synthesizer has been started. Synthetic fricative consonants were presented in a consonant-vowel context always with the vowel /ɑ/. The articulatory configuration for the fricatives was a

*This research was supported in part by the U.S. Air Force (Air Force Cambridge Research Center, Air Research and Development Command)under Contract AF19(604)-2061.

uniform tube, 4.5 cm^2 in cross-section area, with a single constriction, and tapered near the glottis. White noise was inserted as a series voltage source at various points along the tube. The effects of moving the constriction and the point of noise insertion were studied.

The analog is composed of sections of different length and configuration so that a definition of "constant degree of constriction" is required. It is defined as constant acoustic mass, that is, constant ratio of length to cross-section area. Two degrees of constriction were used: a loose constriction with an acoustic mass equal to that of a cylinder 1 cm in length and 0.4 cm^2 in area, and a tight constriction with an acoustic mass equal to that of a cylinder 1 cm by 0.13 cm^2.

The stimuli were generated with identical timing patterns for buzz amplitude, buzz frequency, noise amplitude, and articulatory change. The stimuli included many sounds resembling the velar fricatives /ç/ and /χ/ which do not occur in English. The subjects were therefore given some training in these sounds by a human speaker, and all scored 100 per cent in control tests on those sounds. A control test was given before each test of synthetic stimuli and another control test was given afterwards. One test of sounds generated by the analog was recorded for each degree of constriction. The subjects were instructed to make one of the responses /f, θ, s, ʃ , ç, χ/.



(a)

	1	2	3	4	5	6	7	8	9
1					ʃ				
2					ʃ				
3			S	s	ç	s	s		
4		S	S	S	s	s	S		
→ 5	f	s	θ			ç	X	X	x
6	f			θ	ʃ		x	x	x
7	f		θ	θ		x	x	X	x
8	θ	s			x	x			

(b)

	1	2	3	4	5	6	7	8	9
1					ʃ				
2				s	ʃ	ʃ			
3			S	s	ʃ	ʃ	ʃ		
4		S	S		ʃ	ʃ	S		
→ 5	f	f	s	θ		ç		x	x
6	f	f	s	θ	ʃ	ʃ	x	x	x
7	f	f	S	θ	ʃ	ʃ		x	x
8	f	f	s		ʃ	ʃ	x		x

Fig. XVIII-8. Response maps for fricatives made with: (a) a loose constriction; (b) a tight constriction.

The responses for the loose and tight constrictions are summarized in Fig. XVIII-8a and b. Small symbols are used for response levels between 50 and 75 per cent, and large symbols for levels above 75 per cent. All stimuli in the rows indicated by arrows were made with noise inserted at the constriction or just ahead of the constriction, with the choice depending on the section configuration.

It is seen that various fricatives occur in the correct order as the constriction moves from front to back. For example, most /f/ responses were obtained for stimuli generated with a frontal constriction, whereas /χ/ responses were obtained with constrictions at, or posterior to, velar locations. A good /ʃ/ requires a tight constriction, whereas a good /χ/ requires a loose constriction. A highly intelligible /θ/ has not yet been generated.

G. Rosen