X. SPEECH ANALYSIS[*]

Prof. M. Halle
A. I. Engel

G. W. Hughes

H. J. Jacobsen
F. Poza

A. VOWEL IDENTIFIER

Most vowel identifiers constructed in the past were designed on the principle of "pattern matching"; that is, for each vowel a pattern of some set of measurable parameters, such as a particular energy density spectrum, frequency location of the formants, axis crossing density, and so on, was chosen to represent the average or "ideal" set for that vowel. These patterns were stored in the apparatus, and the results of measurements on an unknown input were compared with the patterns in order to obtain the best match, which was interpreted as the uttered vowel. Implicit in these schemes is the assumption that with each utterance of a vowel, the speaker is trying to hit a complicated target pattern of such measurable parameters. This assumption has never been proven, nor is it logically the only possible one.

An alternative model, the distinctive feature model, stresses the fact that identification is possible so long as the speaker differentiates each vowel from all other vowels in his language in some consistent manner. Since with optimal coding it is possible to distinguish 2^n different vowels by means of n binary features, a distinctive feature model is evidently very economical. It also obviates the necessity of assuming the stringent requirements of accuracy in producing the vowels that are demanded in the "pattern matching" hypothesis. As long as the relevant features are properly produced, identification will be possible, regardless of other measurable features of the pattern. The description of these relevant features in terms of electrical measurements may be quite complex, but the feature "pattern" associated with each vowel is simply stated in binary terms of presence or absence of the pertinent features. Thus this approach eliminates the storage of complex parameter patterns, but may entail an involved measurement procedure to extract information in terms of distinctive features.

Since all properties of the acoustic stimulus except those serving to distinguish different words in the language are ignored as carrying more information about the speaker, the linguistic and extra-linguistic context, and so forth, than about the vowel that is spoken, identification should be relatively independent of the speaker provided his dialect possesses the selected features.

Table X-1(a) gives a summary of the results of a theoretical analysis of a set of American English vowels on the basis of the four pertinent distinctive features. If the indicated binary decision is made for each feature, every vowel shown will be uniquely specified. Note that in four cases specification can be made on the basis of only three

[*] Supported in part by National Science Foundation.

Table X-1. Distinctive Feature Analysis of Vowels.

(a)

Feature	/i/	/ɪ/	/e/	/ɛ/	/æ̦/	/æ/	/a/	/ʌ/	/o/	/ɔ/	/u/	/ʊ/
1. Tense-Lax	T	L	T	L	T	L	T	L	T	L	T	L
2. Grave-Acute	A	A	A	A	A	A	G	G	G	G	G	G
3. Compact-Noncompact	N	N	N	N	C	C	C	C	N	N	N	N
4. Diffuse-Nondiffuse	D	D	N	N	R	R	R	R	N	N	D	D

R = Redundant (feature not necessary to distinguish this phoneme)

(b)

2nd formant above 1400 cps = Acute }
2nd formant below 1400 cps = Grave } feature 2
1st formant above 700 cps = Compact }
1st formant below 700 cps = Noncompact } feature 3
1st formant 150-400 cps = Noncompact + Diffuse
1st formant 400-700 cps = Noncompact + Nondiffuse

features. Thus, if the acoustical correlates of these features can be found, and the appropriate measurements instrumented, an "unknown" vowel input can be classified mechanically.

It has been shown that frequency position of the first three formants is sufficient to specify completely the spectral envelope of a vowel sound. Therefore, this parameter was chosen as the one in terms of which the features indicated in Table X-1(a) were formulated. A preliminary investigation of approximately 60 detailed energy density spectra of the vowels of Table X-1 (spoken by several speakers in word context) indicated that the simple formulation given in Table X-1(b) was sufficient as far as features 2, 3, and 4 were concerned, since the feature compact never occurs with diffuse-nondiffuse. To simplify the construction of the present vowel identifier, feature 1 (tense-lax) was ignored, thus allowing only a six-category classification, hereafter designated as /i/, /e/, /æ/, /a/, /o/, and /u/ (the tense member of the pair coalesced with the lax one as the result of omitting feature 1).

DESCRIPTION OF IDENTIFIER

A device was built for the purpose of exploring the possibilities of the model outlined in the preceding section rather than of producing a finished piece of equipment. This preliminary vowel identifier locates the first two vowel formants relative to the six frequency regions given in Table X-1(b) and carries out the logical procedure required by the feature analysis given in Table X-1. The device consists of three main parts: a filter set, a pulse-position modulator, and a relay selector circuit.

The output of the microphone (into which the vowel that is to be identified is spoken) is amplified and fed into a set of six fixed bandpass filters. These filters are of the cascaded, m-derived type with very sharp cutoff characteristics (120 db/octave skirt slope). The frequency bands of the filters that enable location of formant position [consistent with Table X-1(b)] are given in Table X-2. The small discrepancies in frequency band limits can be ascribed to the finite skirt slopes and to the fact that the closest available filters were used. Table X-2 also shows which two of the given set of filters pass maximum energy for each of the six vowel categories.

The outputs of the six filters are fed into a pulse-position modulator which locates the bands of highest energy and second-highest energy. The details of this section of the vowel identifier are described in the appendix. In essence, this component transforms the voltage-level information emanating from the filters into the time domain. Thus the filter channel with the highest output produces a pulse that comes first in time during a fixed sampling period; the channel with the next highest output produces a pulse that follows the first with a delay proportional to the difference in level between the two channels, and so forth, for each of the six channels.

The relay selector circuit accepts these pulses spaced in time, and on the basis of

Table X-2. Filter Frequencies.

Filter Channel	Frequency Range
1	165-365 cps
2	365-645
3	645-800
4	800-1100
5	990-1370
6	1370-2400

i(ɪ)	e(ɛ)	æ̦(æ)	a(ʌ)	o(ɔ)	u(ʊ)	Vowel category
1, 6	2, 6	3,6 or 4,6	3,5 or 4,5	2, 5	1, 5	Highest energy output channels

Table X-3. Logical Connections.

Filter of highest energy	Relays so connected that:
1	2, 3, 4 cannot be second
2	1, 3, 4 cannot be second
3	1, 2, 4 cannot be second
4	1, 2, 3 cannot be second
5	6 cannot be second
6 (never occurs)	5 cannot be second

The first two filters that have been selected as having highest energy output subject to the constraints above uniquely determine vowel indication.

which came first and second, causes an indicator to flash under one of the six vowel classifications in accordance with the combinations indicated in Table X-2. Simple interconnection of thyratrons and relays would have accomplished this selection but for the fact that formant peaks are often broad enough to cause adjacent channels to fire first and second in order. In any actual vowel input, adjacent formants are not this close (except for channels 4 and 5); therefore certain disenabling connections were made. These are given in Table X-3. Thus, for example, an actual sequence of pulses might be 2 - 1 - 5 - 6, which would be identified as 2 - 5 or as belonging to the vowel class /o/.

EVALUATION OF PERFORMANCE

This preliminary version of a vowel identifier based on distinctive feature principles was tested by having many speakers speak vowels into the machine and by noting its operation. The errors of the machine can be ascribed, in most cases, to limitations on obtaining fine formant frequency discrimination when as in the present case only six filter channels are used. Dialectal variations of the vowels /æ / and /o/ accounted for a great many additional errors. Although, theoretically, the system is independent of the quality and pitch of the voice, for extremely high-pitched female voices errors do occur frequently. It is interesting to note that whispered vowels are identified reasonably well. In general, most errors are caused by locating the first formant in a band adjacent to the correct one. It is felt that the use of more channels would help make first formant frequency location more precise. A series of tests is being planned that will present short recorded segments of vowels for identification by both a group of listeners and the machine, in order to obtain better performance evaluation.

Further tests on the response of the vowel identifier were made with a vowel resonance synthesizer that was built by Gunnar Fant. This machine (called OVE by Fant) allows manual control of the circuits that simulate the first three resonances (formants) of the vocal tract. They are connected in cascade and driven by a buzz source of variable pitch which simulates the larynx. The output is very close to real speech waveforms of vowel sounds, but unlike natural speech it can be precisely controlled by the experimenter. The frequency location of formants 1(F1) and 2(F2) is controlled by moving a pointer over a calibrated plot of F1 versus F2. Additional controls adjust F3 and the fundamental frequency.

We present the results of feeding the output of OVE into the vowel identifier in Figs. X-1 and X-2, where lines are plotted on the F1-F2 plane representing boundaries between the six vowel categories into which the identifier is forced to place all inputs. F3 was held fixed at 3000 cps. Figure X-1 shows the category boundaries that are obtained by using the filter cutoff frequencies given in Table X-2 and a fundamental pitch frequency of 120 cps from OVE. Figure X-2 shows these boundaries under the

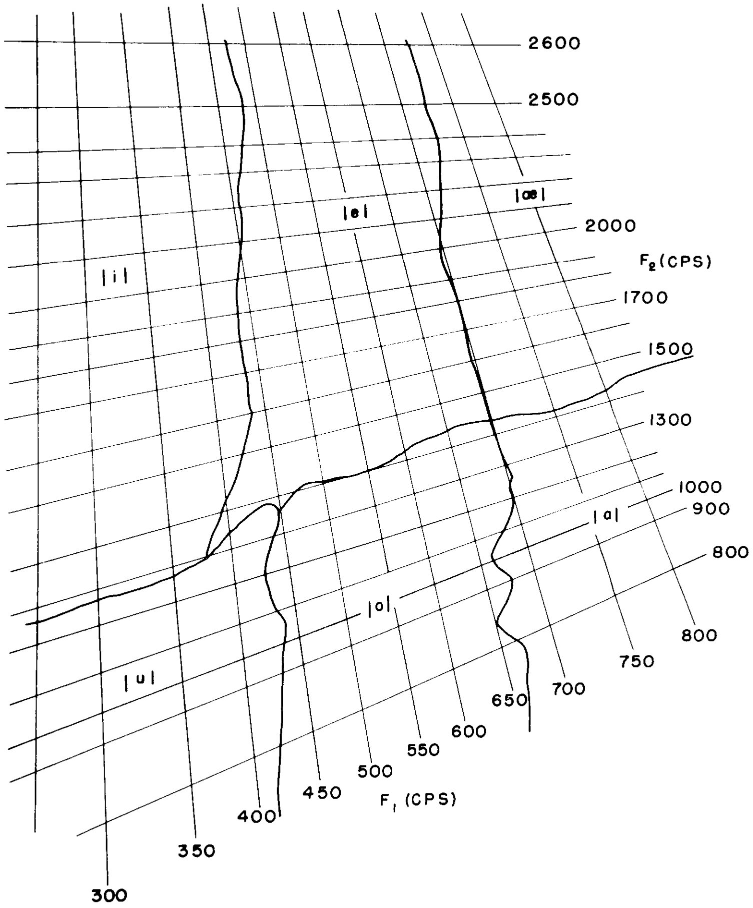

Fig. X-1. A plot on the F1-F2 plane of the boundaries between phoneme categories with the vowel identifier excited by OVE; fundamental pitch frequency, 120 cps.

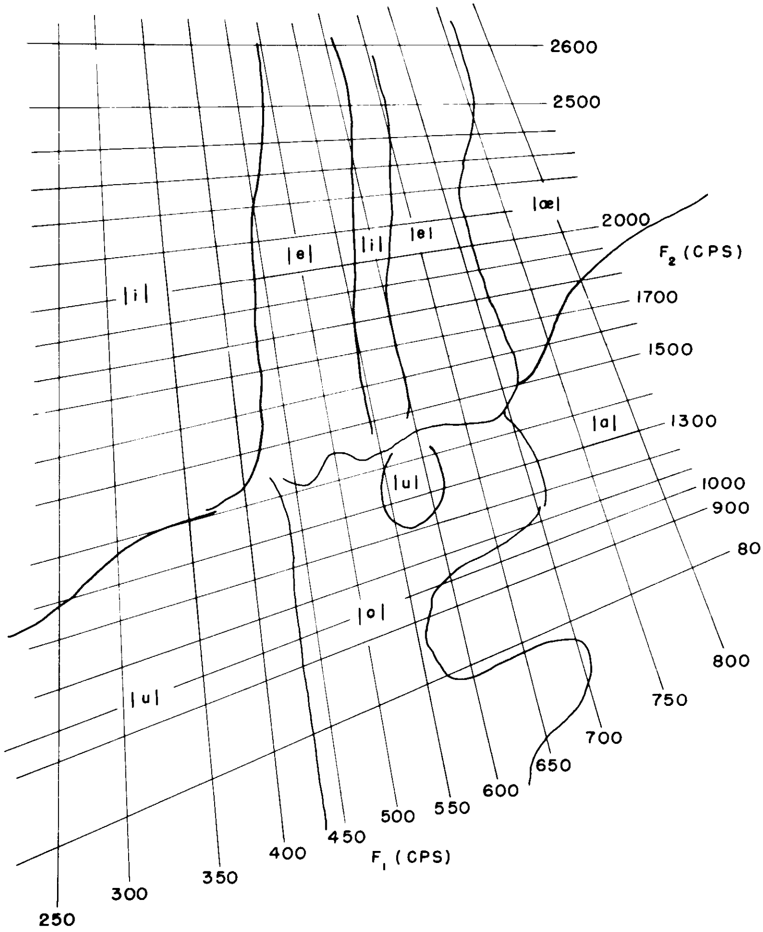

Fig. X-2. A plot on the F1-F2 plane of the boundaries between phoneme categories with the vowel identifier excited by OVE; fundamental pitch frequency, 200 cps.

same conditions except that the pitch was raised to approximately 200 cps. The wider spacing of the harmonics causes some distortion of the boundary lines and the appearance of spurious regions, since the energy in a given frequency band is more sensitive to variations in harmonic frequencies when the fundamental frequency is high (widely spaced harmonics) than when many harmonics are present in the band (low fundamental).

APPENDIX. DESCRIPTION OF THE PPM CIRCUIT

This part of the identifier (see Fig. X-3) consists of six separate, identical circuits that transform the output of each of the filters into a single pulse. The time between a given pulse output and a zero reference time is proportional to the rectified and smoothed voltage of the corresponding filter output. This voltage-to-time transformation is accomplished by detecting a coincidence between a downward linear sweep and the rectified filter output voltage. The three essential parts of the circuit are: (1) the amplifier, rectifier, and smoothing RC integrator; (2) a phantastron sweep generator; and (3) a multiar coincidence detector with associated multivibrator and pulse-shaping circuits.

Briefly, the operating cycle of one of the six identical channels is as follows: Initially, the sweep voltage at one input to the coincidence circuit is stationary at +10 volts, and the rectified-integrated voltage from the filter at the other input to the coincidence circuit is zero. As a vowel is spoken into the microphone the rectified-integrated voltage rises until the channel whose bandpass filter is passing maximum energy reaches +10 volts. A coincidence pulse which starts the sweep downward is generated. The sweep reaches bottom at slightly below zero volt in about 30 msec, while the RC time constant of the integrator is 300 msec. As the sweep passes the integrated voltages on the other channel coincidence circuits, pulses whose time order represents the filters in order from highest to lowest output are produced. These pulses, as well as the initial one which defines time zero, are fed to the relay selector chassis which carries out the logical procedure indicated in Tables X-2 and X-3 for vowel identification. When the sweep reaches bottom, a pulse is generated, triggering a relay circuit that resets all integrator outputs to zero and extinguishes the thyratrons in the selector circuit. Thus the initial conditions are restored and the cycle is repeated until the vowel input to the microphone ceases.

This discussion is based, in large part, on two theses that were submitted in partial fulfillment of the requirements for the S. B. degree in the Department of Electrical Engineering, M.I.T., 1956: "A device for locating peaks and intensities of the vowel spectra," by Harry James Jacobsen, and "Automatic vowel recognizer operating on distinctive feature principles," by Alan I. Engel.

G. W. Hughes, H. J. Jacobsen, A. I. Engel, M. Halle

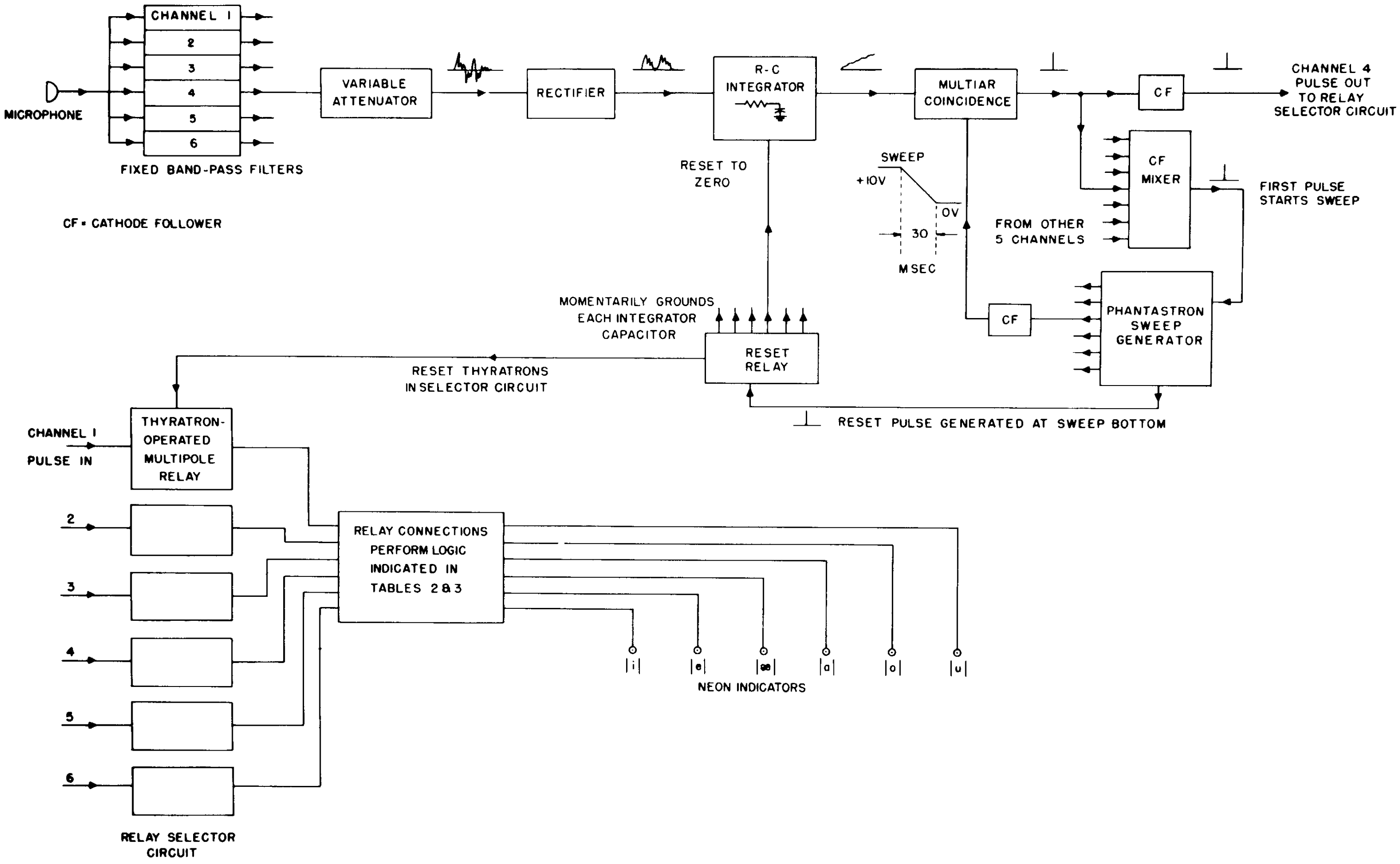

Fig. X-3. Block diagram showing circuit connections in the vowel identifier. The top row represents only one of the six identical channels.

B. STOP STUDIES

The two major cues for stop consonants are the stop burst and the transitions of the formants in the adjacent vowels. Detailed energy density spectra of the isolated stop bursts were prepared, and criteria for identification of the spectra were developed and tested. The reliability of these criteria was compared with the reliability of human listeners in identifying the isolated bursts. It was found that human listeners, especially if trained, can identify stop bursts correctly in a large number of cases. The objective criteria mentioned above, although not quite as reliable as the best listeners, provided correct identification in the large majority of instances.

Transitions were studied by means of sonagrams. Great difficulties were experienced in stating simple criteria, like those proposed for synthetic speech. Transitions in the formants of adjacent vowels were studied by means of sonagrams of isolated words recorded by several speakers. Unlike synthetic speech, natural speech did not permit the formulation of simple criteria. The transitions, although not uncorrelated with the nature of the following stop, failed to provide sufficient information for a definitive identification. This conclusion was apparently supported by the evidence obtained in perceptual tests with vowel and stop sequences in which the stop bursts had been gated out so that the stop cue was entirely in the vowel transitions. The identifications of these stimuli were about as reliable as would have been predicted from an examination of the sonagrams.

Details of this investigation are included in a paper by Halle, Hughes, and Radley, "Acoustic properties of stop consonants," accepted for publication by the Journal of the Acoustical Society of America, January 1957.

M. Halle, G. W. Hughes