#3

DYNAMIC ANALOG SPEECH SYNTHESIZER

GEORGE ROSEN

Loan Copy Only

TECHNICAL REPORT 353

FEBRUARY 10, 1960

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
RESEARCH LABORATORY OF ELECTRONICS
CAMBRIDGE, MASSACHUSETTS

The Research Laboratory of Electronics is an interdepartmental laboratory of the Department of Electrical Engineering and the Department of Physics.

The research reported in this document was made possible in part by support extended the Massachusetts Institute of Technology, Research Laboratory of Electronics, jointly by the U.S. Army (Signal Corps), the U.S. Navy (Office of Naval Research), and the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), under Signal Corps Contract DA36-039-sc-78108, Department of the Army Task 3-99-20-001 and Project 3-99-00-000, and was performed under Air Force Contract AF19(604)-6102.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

RESEARCH LABORATORY OF ELECTRONICS

Technical Report 353 February 10, 1960

DYNAMIC ANALOG SPEECH SYNTHESIZER

George Rosen

Submitted to the Department of Electrical Engineering, M. I. T., January 25, 1960, in partial fulfillment of the requirements for the degree of Doctor of Science.

Abstract

A dynamically controllable electrical analog of the vocal tract that is capable of synthesizing sequences of speech sounds is described. The acoustic transmission line between glottis and lips in the human vocal tract is realized electrically by electronically controlled variable inductance-capacitance sections. Excitation is provided for simulating the glottal tone and the noise of turbulence. The synthesizer was conceived as an instrument for research on speech production and perception. Its control system is designed to permit the synthesis of any sequence of two phonetic elements, and provides precise, flexible control over many geometric and temporal variables. Exploratory formal listening tests were conducted, with idealized geometries for the vocal tract, and piecewise-linear functions were used for describing the timing relations. An articulation score of 93 per cent was obtained for vowels made with a parabolic approximation to the tongue hump. Listeners were within 5 per cent of unanimity in their identification of some fricatives in consonant-vowel syllables. At the phoneme boundary in these syllables, the tolerance for relative timing error between control signals is approximately 10-30 msec. Experimental evidence that strongly supports the viewpoint that the articulatory level provides a natural and economical description of speech is given. Thus, the dynamic analog is potentially capable of synthesizing highly natural connected speech from signals having a low information rate.

TABLE OF CONTENTS

I. INTRODUCTION

1.1 HISTORICAL NOTE ON SYNTHESIZING SPEECH

Attempts to synthesize speech were made as early as nine hundred years ago (48). The first analog of the speech mechanism, constructed in 1791 by von Kempelen (59), consisted of various pneumatic devices imitating specific parts of the vocal tract. In 1829, Willis (60) reported on vowel-like sounds produced by reed organ pipes. More realistic vowels were produced in 1872, when Helmholtz (30) excited several of his famous acoustic resonators by tuning forks. However, the first electrical synthesizer was not reported until 1922 when Stewart (56) described a circuit that produced vowels, but not continuous speech. The first devices to accomplish the synthesis of continuous speech were the voder and vocoder of Dudley (15, 16). Two methods of synthesis were introduced in 1950: Dunn's static analog of the vocal tract (18) and the pattern playback of the Haskins group (12, 13). The resonance, or formant-coded, synthesizer, whose earliest version is due to Lawrence (38), came soon afterwards, in 1953. Two electrical analogs were also reported in that year: one by Fant (20) and one by Stevens, Kasowski, and Fant (54).

We shall compare several of these synthesizers and pay particular attention to the degree to which constraints in the human vocal tract are incorporated into the structure of each device.

Let us consider the synthesizer in the vocoder system, which produces speech whose short-time spectral envelope is specified by signals at the synthesizer input. The area under this envelope is a line spectrum during voiced sounds; at other times, except during silence, it is a continuous spectrum characteristic of random noise. Thus this device explicitly observes distinctions in excitation that are physiologically and linguistically important. However, it can produce spectrum envelopes of arbitrary shape, and must rely on information in the input signals to select only those that are characteristic of speech. The need for such information implies a cost in required channel capacity at the input; the lack of such information in the signals results in a lack of intelligibility or naturalness at the output.

The pattern playback, like the vocoder, also deals with short-time spectra. Operating on spectral patterns painted on a moving belt, it is an important research tool but is not suited for information transmission in real time.

The resonance synthesizer (22,24,38,51) is a filter network that simulates the behavior of the vocal tract as seen from its output terminals. The control input for this class describes the pole positions as functions of time and also describes the excitation. This device reflects in its structure our knowledge that the transfer function specifying transmission from glottis to lips has several poles but no zeros. But transfer functions having zeros are needed for the synthesis of most consonants. In all circuits reported thus far, transitions between these two situations have been treated in an ad hoc manner.

Synthesizers of this kind have demonstrated facility with vowels and vowel-like sounds, but have handled consonants poorly.

The electrical analog is concerned explicitly with parameters that describe the configuration and excitation of the vocal tract. The speech at the output of this device is a mapping into the acoustic domain of these articulatory parameters, in much the same manner as human speech is the acoustic correlate of physiological events. Static analogs have synthesized some highly natural isolated speech sounds, such as vowels, but they are incapable of producing connected speech. The limitations of static analogs are particularly evident in endeavors to produce consonants, since certain classes of consonants such as stops, semivowels and affricates exist only by virtue of articulatory changes. Analog production of these consonants, and of connected speech in general, requires a dynamically variable speech synthesizer. Such a synthesizer imitates the human vocal tract with respect to both geometrical and temporal variables, and it is expected that the device will be capable of producing most of the sounds that a human being can produce. Because it offers a natural and direct way of building in constraints that are present in the human speech organism the dynamic analog offers the greatest opportunity for storing at the synthesizer information about how speech is produced and how transitions from one sound to another are made.

1.2 APPLICATIONS OF SYNTHETIC SPEECH

The first speech synthesizers were mere objects of curiosity; nevertheless they provided some clues about speech production and perception. During the twentieth century, the development of synthesizers has been motivated with several applications in mind: to help reduce the channel capacity required for speech communication, to enable computers to speak, and to serve as instruments for speech research.

By using the concepts of information theory, quantitative estimates of the information transmission at various levels of the speech communication process can be made (45). These estimates show that most speech channels are used at a small fraction of their capacity, and that it is theoretically possible to recode or compress speech to reduce channel capacity requirements by three orders of magnitude. A system for compression consists of three parts: an analyzer, a channel, and a synthesizer. As a potential component of a compression system, the dynamic analog synthesizer offers advantages over other devices in its obtainable degree of compression and in the naturalness of its output.

Synthesizers are needed as output members of automata to make speech communication possible between man and machine. Such facility is necessary, for example, in language-translation systems for accepting and delivering the spoken word. Almost all computing systems, at the present time, rely on sight or touch for communication with human beings. Speech input and output devices would be advantageous in many existing systems and could extend the range of new applications of computers.

A speech synthesizer can serve as a research instrument in fields such as psychology, biophysics, speech pathology, and linguistics. A possible function of the synthesizer is to serve as a standard signal source; its settings provide a comprehensive, detailed description of each speech sound, and its repetitive utterances at the same settings are free from unknown and uncontrolled changes in important variables. Correlation techniques (5, 11) have been developed to extract neural signals from background activity. When investigators are ready to study neural responses to stimuli as complex as speech, sythesizers will be needed to provide the repetitive, phase-coherent stimuli that are often demanded by correlation techniques. Many studies in pathology and linguistics are hindered because certain manipulations and observations on the living human vocal tract are not possible. Often such studies can be pursued if a model of human speech production is available; the dynamic analog synthesizer can function in this role.

1.3 PLAN OF THE PRESENT RESEARCH

To investigate the feasibility of the dynamic analog method of speech synthesis and to explore the capabilities of this method, the following program of research was initiated:

(a) Development of a dynamic analog of the vocal tract;

(b) Development of a control system capable of maneuvering the analog through any sequence of two phonetic elements; and

(c) Evaluation of the machine through formal listening tests.

The formal listening tests were experiments for testing the usefulness of this device as a research instrument and demonstrating its capabilities as a synthesizer. The experimental work had as its objectives the production of certain classes of speech sounds, and the determination of the precision with which several of the most important control variables must be specified. Another goal was to contribute to the understanding of speech production and perception by quantifying variables whose values, at present, can only be rank-ordered.

The synthesizer, as it is now conceived, is a flexible instrument capable of generating a very large class of speech waveforms. Many studies that have been reported can be duplicated with this apparatus; but the instrument also extends considerably the area within which quantitative investigations are possible. Because of the broad scope of this area, the experiments reported here cannot be exhaustive; they should be regarded as a preliminary probing of the area.

II. THE HUMAN VOCAL TRACT

Figure 1 is a schematic drawing of the human vocal mechanism. Speech is a secondary function superimposed on apparatus whose primary functions are respiration and ingestion. In speaking, muscles surrounding the chest cavity contract it to squeeze the lungs, expelling the air in them and forcing it through the trachea, past the vocal folds (the opening between the vocal folds is known as the glottis), through the pharynx and out – either through the nasal cavities and the nose or through the oral cavity and the mouth. Some of the energy in the airstream can be transferred to the audible frequency range by three kinds of instability which thus become the following acoustic-energy

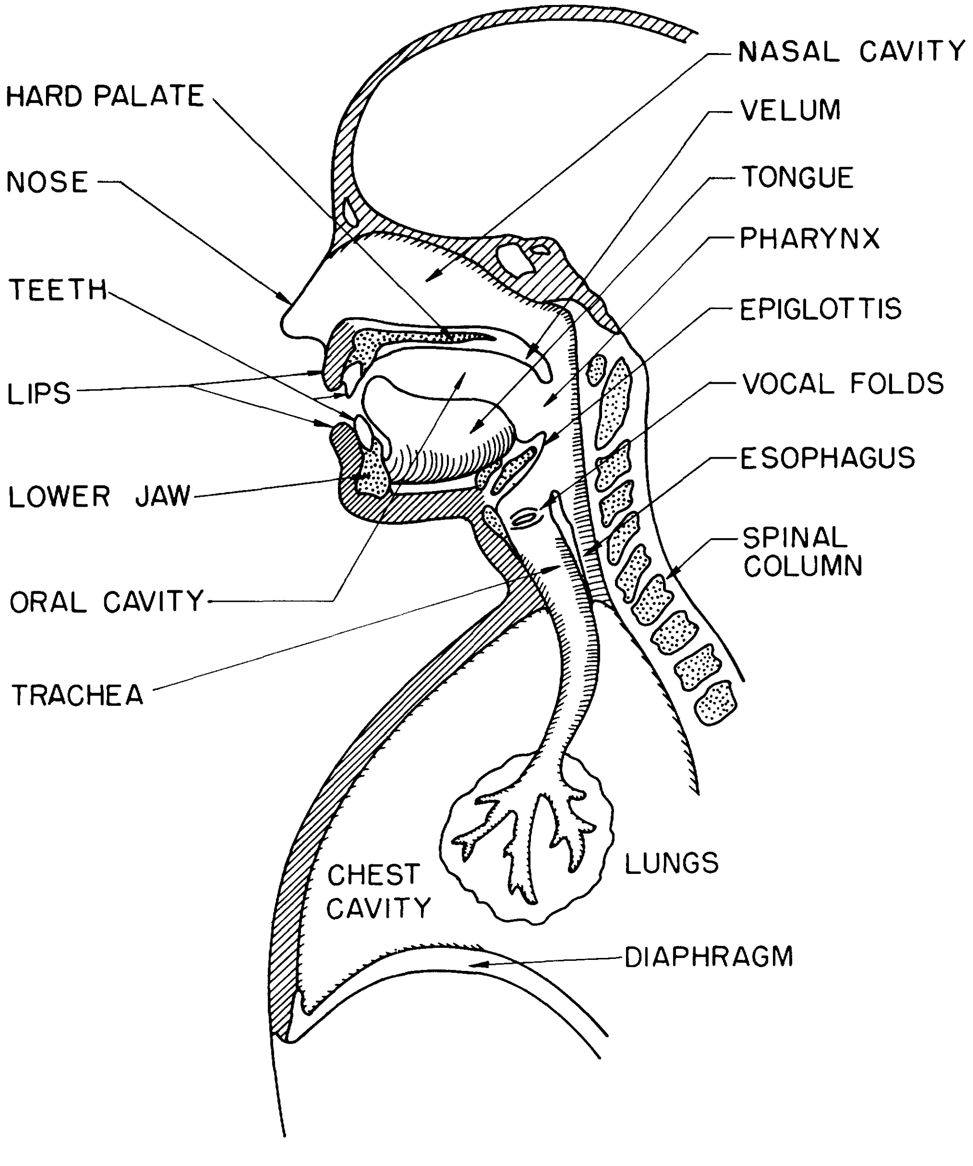


Fig. 1. Diagram of the human vocal mechanism based on a midsagittal section.

sources: (a) vibration of the vocal folds; (b) turbulence in the airstream passing through narrow constrictions or over sharp edges; and (c) sudden release after complete closure. The amount and character of the energy that enters the system at the sources and leaves the system as radiated sound in the outside air depends on the coupling provided by the vocal tract.

The function of speech (43) is to transmit reports, ideas, commands, and emotions from the speaker to the listener. Information generation and transmission are defined, respectively, as: (a) choosing, at the source, a message from among a set of possible messages; and (b) using a transmitter to send a signal through a channel to make the receiver aware of the choice made by the source. The messages can be expressed as sequences of elements chosen from a set of linguistic entities. To serve as an information channel, the acoustic output of the vocal tract must be amenable to modulation or change by the speaker (17). There are two things under the speaker's control, the source excitations, and the configuration of the coupling to the outside air. The coupling is controlled by moving the tongue and lips, directly or by displacement of the lower jaw, and by moving the velum to couple the nasal cavity to the oral cavity. The sources are controlled directly through the musculature of the vocal folds, or indirectly by producing constrictions and closures in the vocal tract and through pressure on the lungs.

The speaker's role can be examined by a study of his modulation process, that is, his manner of speech production, or by a study of the signal that he transmits to the channel, that is, his acoustical output. In studying the manner of production, one can deal with the physics of speech production, making explicit reference to the size and shape of the vocal cavities, or one can treat the vocal tract as a variable multiterminal filter and operate with its transfer and source functions.

The vocal cavity is shown in Fig. 2. It is a physical system that is too complicated for exact analysis. Its walls, consisting of the pharynx, hard and soft palates, tongue, cheeks, and lips are made of tissue that has both absorptive and reflective properties. It is a region whose shape changes with time. Within this region some waves propagate longitudinally, from glottis to mouth; others, in cross modes, reverberate from wall to wall. The opening and closing of the vocal folds, which excites the system, introduces time-variant coupling to the lungs in synchronism with that excitation. During sounds

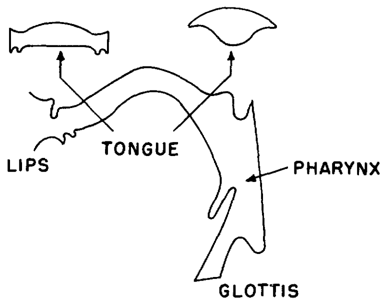


Fig. 2. Boundaries of the vocal tract during the production of the vowel [e]. A midsagittal section and two transverse sections are shown (from T. Chiba and M. Kajiyama (9)).

that require turbulent excitation there exist regions of nonlaminar compressible fluid flow, with the result that there is a spacially distributed source in the region of turbulence. Sounds produced by closure release exhibit all of the complications of turbulent sound and, furthermore, they exist under conditions that are nonstationary.

Simplifying assumptions are needed to make the system amenable to analysis or simulation. The system exhibits pressure variations that are small compared with atmospheric pressure, so that linear small-signal theory applies to the air medium. The dimensions of the human vocal mechanism are such that cross modes are negligible at frequencies below approximately 5000 cps. The air and walls appear to have but slight dissipation for longitudinal waves, so that, for many purposes, the lossless case can be assumed (21). When only longitudinal waves exist in a tube, each wave front is a cross section of the tube. Thus pressure, and also particle velocity, are constant over a cross section. Its shape need not be specified, and one parameter, area, is sufficient to describe it. The vocal tract may be regarded as a straight tube whose area at a given time is a function of distance along its axis. The tube has a variable port leading to a side branch, the nasal cavity, whose dimensions do not vary with time. For voiced sounds, the vocal tract is driven by the glottis, whose maximum opening is quite small. It may be represented by a moving piston that gives the same volume velocity as the puffs of air passing through the opening. Turbulent regions may be characterized by a few lumped sources and dissipative effects may be summarized by a single dissipative element per source added to the acoustical circuit. (Good fricatives have been synthesized by using one source. (See refs. 29 and 54 and Sec. VI.) Stop consonants, particularly those with distinct intervals of frication and aspiration may require two sources.) Some sounds made with a closure use glottal excitation, and previous remarks apply to them. Little is known about the physics of closure-release excitation. The vocal tract radiates into the outside air through the mouth opening which presents a lumped impedance load to the tube. The pressure field at a distance from the speaker resembles the field of a point source whose strength equals the volume velocity through the mouth opening.

Fant (19), Dunn (18), and Van den Berg (58) have calculated the first few resonant frequencies of the vocal tract by approximating its shape as two cylinders connected by a constriction. Dunn verified his calculations on an analog based on the same geometry. Stevens, Kasowski, and Fant (54) constructed an analog based on 35 cylinders to approximate more closely the shape of the vocal tract, and Fant (23) has built a 45-section analog.

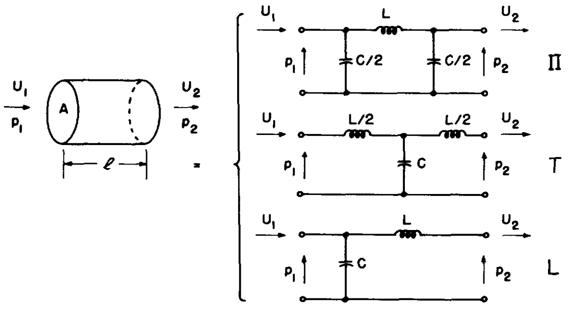


Fig. 3. A uniform acoustic tube and three electric network analogs of the tube.

Each analog that has been built approximates the shape of the tube representing the vocal tract by a cascade of cylinders, each cylinder having uniform cross-section area. Each lossless cylinder is represented in the analog by a Π, T or L section, as shown in Fig. 3. Pressure is identified with voltage and volume velocity is identified with current, so that the relations between acoustical and electrical parameters are:

$$L = \frac{\rho \ell}{kA} \tag{1a}$$

$$C = \frac{kA\ell}{\rho c^2} \tag{1b}$$

where A is the area of the cylinder; ℓ, the length of the cylinder; ρ, the density of air; c, the velocity of sound in air; and k, a constant depending on impedance level and choice of units. The analog is, therefore, a nonuniform electrical transmission line. The glottal source at the input is a high-impedance (current) generator and the terminating impedance at the mouth is predominantly inductive for frequencies below 4000 cps.

The circuit of the transmission line may be hidden from view, and thus the behavior of the line as a filter may be studied. Its properties help to define the class of waveforms and the class of spectra to which the acoustic output signal must belong. Fant (21) has shown that the transfer function relating volume velocity at the lips, U_2, to volume velocity at the glottis, U_1, is a rational fraction having no zeros, but only conjugate complex poles in the finite complex-frequency plane. The function equals unity at zero frequency, as required by continuity of air flow. In Laplace-transform notation:

$$\frac{U_2(s)}{U_1(s)} = H(s) = \prod_k \frac{s_k s_k^*}{(s-s_k)\left(s-s_k^*\right)} \tag{2}$$

where $s = \sigma + j\omega$ is the complex frequency; U(s) is the volume velocity transform; and s_k, s_k^* are conjugate complex poles. It is seen that the numerator is a constant.

A modified transfer function must be used when the vocal tract is modified by nasal coupling. The modified function has zeros at frequencies determined by nasal resonances. The spectrum of the radiation at a distance from the speaker has zeros caused by cancellation effects between nose and mouth radiation. Excitation within the vocal tract by transient or fricative sources implies excitation of the filter at terminals other than the glottal terminals. The pertinent transfer functions possess zeros at frequencies determined by resonances of the cavity behind the source. The poles exhibited by any circuit are the same for all terminals (27), provided that the generator impedances are not changed from case to case. Thus, with a given configuration, the poles for glottal excitation are the same as those for other excitations.

III. VARIABLE CIRCUIT ELEMENTS FOR THE VOCAL-TRACT ANALOG

3.1 STRUCTURE OF THE SYNTHESIZER

The dynamic analog synthesizer consists of three parts:

(a) The lumped electrical transmission line which is the analog of an acoustical system, the vocal tract.

(b) Sources of excitation for the line.

(c) A control system for both the line and its excitation.

The transmission line is shown in Fig. 4. It is a cascade of one or more of the section configurations shown in Fig. 3. For voiced sounds, the transmission line is excited by glottal current pulses of variable frequency and amplitude (on the left, in Fig. 4). For fricatives and other turbulent sounds, a noise voltage generator of controllable amplitude is inserted in series with the transmission-line inductor to correspond to the place where the turbulence is produced. For nasals, a nasal analog, which has not yet been constructed, must be coupled to the vocal-tract analog. (See the appendix for the nasal analog circuit of House and Stevens (1, 31, 33).) The line is terminated by an inductor that approximates the radiation impedance seen from the mouth. The output is the voltage across a small inductor in series with the radiation impedance. The current through the radiation impedance is treated as the strength (volume velocity) of a simple source, and hence the voltage across the small inductor behaves as the sound pressure at some distance from that source.

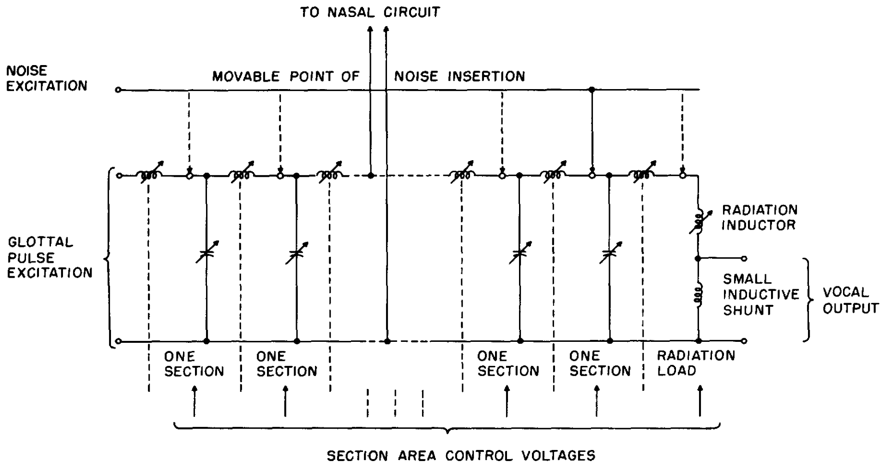


Fig. 4. Circuit of the dynamic analog of the vocal tract showing the variable elements of the transmission line. Voltages for control and excitation are derived from devices shown in Fig. 22. Noise is inserted in the line through a transformer that connects the low-impedance noise generator in series with the inductor of the appropriate section. Voltages representing radiation from the nose and mouth are summed externally.

The relations between electrical and acoustic quantities, Eqs. 1, may be restated:

$$A = \frac{\rho c}{k}\left(\frac{C}{L}\right)^{1/2} \tag{3a}$$

$$\ell = c\,(LC)^{1/2} \tag{3b}$$

To vary the area of a section, its control circuit must vary C/L. If the section is to represent a cylinder of constant length, the control must also constrain LC to remain constant.

All of the section areas of the dynamic analog can be varied together, by means of externally generated electric control voltages, to simulate the changing articulatory configurations of the human vocal tract. When buzz and noise excitation are synchronized with the articulatory changes, the synthesizer can produce connected speech.

3.2 REQUIREMENTS FOR THE TRANSMISSION-LINE ELEMENTS

The choice of the physical phenomena for realizing the transmission-line elements and of their numerical values resulted from the interplay of four factors: (a) known minimum requirements, (b) previous experience with the static analog at M.I.T. (36, 54), (c) the availability of circuits and components, and (d) the feasibility of developing new circuits and components should the existing ones be unsuitable. The variable inductor utilizes changes in the incremental permeability of a ferromagnetic material and the variable capacitor utilizes changes in the magnitude of the Miller effect. The variable elements and their realization (47) will be described.

The first requirement to be considered is that the lumped approximation to a continuous line be valid for frequencies somewhat higher than that of the third formant region. This requirement implies a maximum section length of 1.5 cm, or approximately 1/8 wavelength at 3000 cps. The switches of the static analog (0.5-cm sections) were set in groups of 3, in order to test, by informal listening, the adequacy of the coarse approximation to the shape of the vocal tract provided by the 1.5-cm sections. A combination of experience with the static analog and considerations of cost and complexity led to the choice of the number of sections to be used in the dynamic analog. It is a compromise between the simplicity of Dunn's (18) analog and the versatility of the 35-section static analog.

The static analog had an area range of 100:1 for each section (0.17 cm^2-17 cm^2). Satisfactory vowels had been produced by using only settings between 0.59 cm^2 and 17 cm^2, a ratio of 29:1, but it was considered desirable to have a full continuous range of 100:1 in the dynamic analog. The low end of the area range of a given section could be extended by switching a large fixed inductor in series with the variable inductor of that section. Informal listening with the static analog showed that switching transients were imperceptible when this was done. Thus narrow constrictions and closures for consonant production could be realized without putting impossible requirements on the variable inductor.

For each section, the length requirement fixes $(LC)^{1/2}$, and the area-range requirement fixes the range of $A = \frac{1}{k}(C/L)^{1/2}$. If a section is to have a constant length and an area range of 100:1, then both L and C must have a range of 100:1. The impedance level k, which is the same for all sections, is an arbitrary parameter of the transmission line. For reasons pertaining to the design of the variable elements, the impedance level was chosen to be 0.77×10^{-4} (electrical ohm)$^{-1}$/cm^{2} of area.

The Q (quality factor) required of the inductors and capacitors can be estimated from formant bandwidth data (34, 39, 57). The Q values for the first three formants of the vocal mechanism lie in the range 3-15. The losses responsible for these values arise, predominantly, from glottal resistance and radiation from the mouth, although losses in the vocal tract itself are relatively small. In an electrical analog of the vocal mechanism, however, most of the losses must be in the transmission line itself because most of the physical elements, with their irreducible losses, are located within the line. The resonances of the static analog, which has produced highly recognizable stimuli, are characterized by Q values similar to those mentioned above. The quality factor of its inductors, which depends on area setting, varies from, at least, 9 at 200 cps to, at most, 250 at 5000 cps. Best results were obtained when dissipation was inserted within the line to provide lower effective Q values.

Preliminary designs of variable inductors indicated that Q values above 50 at 1000 cps were attainable at the upper end of the inductance range. However, Q values much greater than 1 at 200 cps at minimum inductance are not attainable with these inductors. This is not a serious limitation, as an analysis based on energy functions shows. The definition of quality factor (27) at any frequency — not only at resonance — is

$$Q = 2\pi \frac{\text{peak energy stored}}{\text{energy dissipated per cycle}} \tag{4}$$

This definition is applied to a single isolated section, which is treated as a simple resonant circuit. For frequencies far below the resonance that occurs above 3 kcps, the expression is

$$Q = (a/Q_L + 1/Q_C)^{-1} \tag{5}$$

where Q_L is the inductor quality factor; Q_C is the capacitor quality factor; and $a = 2\pi f\,(LC)^{1/2}$ is the normalized frequency. Physically, this means that most of the energy at low frequencies is stored in the capacitor and that the Q of the inductor is, in effect, multiplied by 1/a.

Electronically variable capacitors having as low dissipation as the paper dielectric capacitors used in the static analog are clearly unattainable. In view of the limitations of the variable inductors, we decided to reduce the dissipation of the capacitors just to the point where it would not be the factor limiting the over-all performance of the

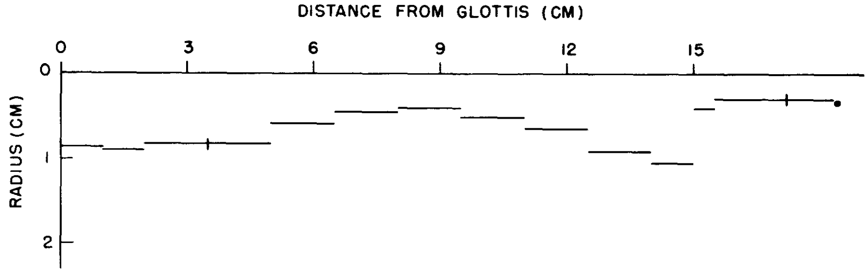


Fig. 5. Configuration for the production of vowel [u]. The vocal tract is treated as a cascade of circular cylinders and the radius of each cylinder is given. Short vertical markers show boundaries between sections set to the same radius; dot symbol indicates setting of radiation impedance.

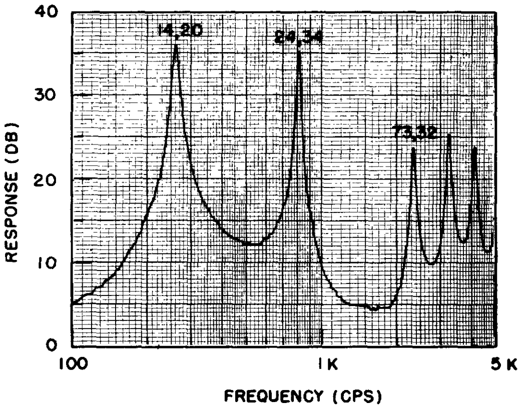


Fig. 6. Transfer function of the analog set to the configuration of Fig. 5. The square of the ratio of output voltage to input current is given in decibels relative to an arbitrary level. The first number above each of the first three formants gives the bandwidth in cps; the second gives the corresponding Q.

transmission line. Accordingly, a quality factor of 30, or more, was set as a requirement for the capacitors.

The performance of the completed transmission line is the test of the previous arguments. A vowel such as [u], which has a low first formant, is well suited to serve as a test item. A configuration for this vowel appears in Fig. 5, and the corresponding transfer function of the analog in Fig. 6. The first number above each peak at frequency f in Fig. 6 gives the bandwidth, Δf, between half-power points, and the second gives $Q = f/\Delta f$. The bandwidths, especially that of the first formant, are similar to those characteristic of natural speech. (The values of Q are slightly high. An error in this direction is easily remedied by inserting dissipative elements.) Thus, the validity of the reasoning based on Eq. 4 is substantiated, and the sufficiency of the element dissipation requirements verified.

3.3 CONTROL PATHS WITHIN THE TRANSMISSION LINE

A typical section in the transmission line consists of three parts, a variable inductance circuit, a variable capacitance circuit, and a control amplifier. The design of these circuits, both of which have a 100:1 range, is discussed in sections 3.4 and 3.5. Figure 7 is a block diagram of a typical section. The variable capacitance appears between ground and the input of the capacitor amplifier. Its value is determined by V_{CC}, the capacitor control voltage. The variable inductance, whose value is determined by V_{CL}, the inductor control voltage, appears across the signal winding of the saturable inductor. The control characteristics of both the variable capacitance circuit and the variable inductance circuit are exponential:

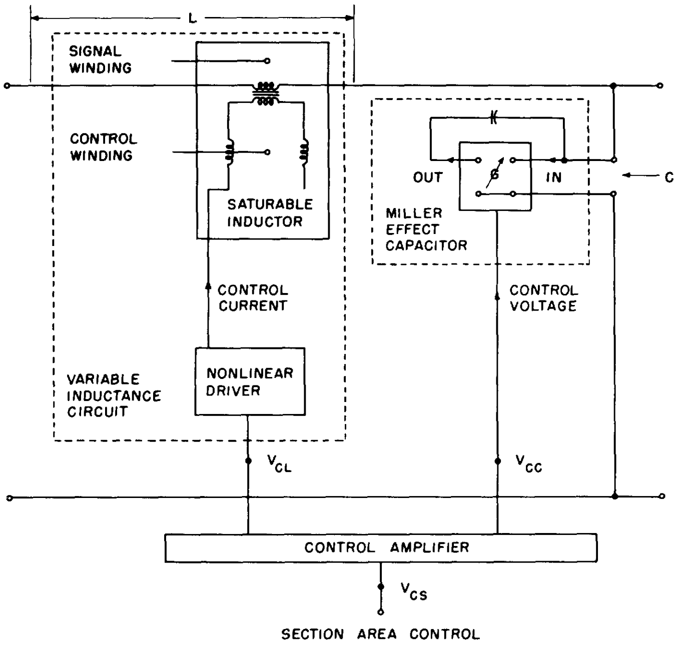


Fig. 7. Simplified block diagram of a typical section. The section-control input voltage determines the magnitudes of both variable capacitance and variable inductance. The electrical length of the section is constrained to remain constant as its area is varied. (See Fig. 10 for details.)

$$C = e^{a_C - b_C V_{CC}} \tag{6}$$

$$L = e^{a_L - b_L V_{CL}} \tag{7}$$

$$V_{CL} = d - (b_C/b_L)\, V_{CC} \tag{8}$$

where a_C, b_C, a_L, b_L, and d are constants. The control amplifier maintains a linear relation (Eq. 8) between V_{CL} and V_{CC}, so that the LC product remains constant. Both

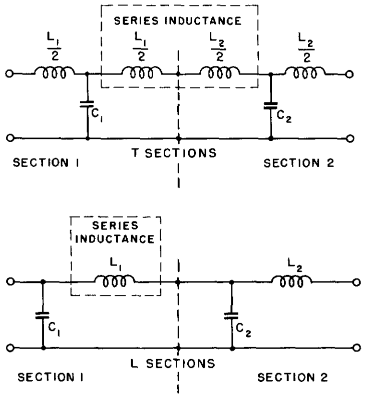


Fig. 8. Two alternative methods of approximating a distributed system with a lumped system. The T sections yield element values that approximate the distributed system better than those of L sections do.

V_{CC} and V_{CL} are determined by V_{CS}, the control input voltage that determines the section area. Thus, the length of the section remains constant as its area changes exponentially over a 100:1 range.

Each section is constructed as an L section. A nonuniform line, however, can be approximated more closely by a given number of T sections than by the same number of L sections. Figure 8 shows L- and T-section approximations to the acoustical system that consists of two short adjacent cylinders of equal length and of areas A_1 and A_2. The inductances are given by

$$L_1 = k/A_1 \tag{9a}$$

$$L_2 = k/A_2 \tag{9b}$$

where k is a constant. An L-section approximation leads to a total inductance L_1 connected between C_1 and C_2. A T-section approximation leads to a total inductance:

$$L_{\text{series}} = \frac{L_1}{2} + \frac{L_2}{2} = \frac{k(1/A_1+1/A_2)}{2} \tag{10}$$

connected between C_1 and C_2. In the general case of a nonuniform line, $A_1 \neq A_2$; and hence Eq. 10 leads to a better approximation for the series inductance than Eqs. 9 do.

Cross-coupling of the control voltages of adjacent sections allows us to approximate the desired L_{series} of Eq. 10. The input information to each of two adjacent sections constructed as in Fig. 7 is V_{CS}, which corresponds to A_1 (or A_2). The input voltage V_{CS} determines V_{CC}, the capacitor control voltage. The coupling connection (see Fig. 10 for details) sums inductance control outputs of adjacent control amplifiers to derive a new control voltage, V_{CL}, for the variable inductance circuit. The new V_{CL} lies between the values intended by either control amplifier. If the sections are of equal length, their summing coefficients are set equal. The inductance of the series inductor is an exponential function of V_{CL}; therefore its inductance is the geometric mean of the values given by Eqs. 9. For quantities differing by small ratios, the geometric mean and the arithmetic mean are almost equal, so that the series inductance approaches that given by Eq. 10. The line consisting of cross-coupled L sections may be operated as though it were constructed of T sections. The summation of control voltages for the inductor leads to an improved approximation to the nonuniform line with negligible increase in complexity, in fact, halving the number of inductors otherwise required.

The transmission line, whose nominal length is 17.5 cm, consists of 14 sections. The two sections adjacent to the glottis are T sections, each 1.0 cm in length, and have fixed areas approximately equal to 1.0 cm^2. They are followed by 8 variable T sections, each 1.5 cm in length, and the control inputs of the first two variable sections are tied together. (The first of these two sections covers a smaller area range to provide a "fairing" between the fixed and variable sections.) The T sections are followed by a half-section, 0.5 cm in length. The line is completed by two Π sections, nominally 1.0 cm in length, and the radiation inductor. The final capacitor and the radiation inductor are controlled by one control amplifier. The inductors and capacitors of the Π sections are controlled separately so that these sections may be shortened to simulate the shortening of the human vocal tract which occurs at wide mouth openings. Thirteen voltages are required to specify an articulatory configuration. Another voltage will be needed when the nasal analog is added to specify the degree of nasal coupling.

3.4 VARIABLE-CAPACITANCE CIRCUIT

The variable capacitance is realized by means of the Miller effect, which is illustrated schematically in Fig. 9. In general, Z_i is given by

$$Z_i = \frac{E_i}{I_i} = \frac{R_o + \frac{1}{j\omega C_1}}{1 + K\angle\theta} \qquad (11)$$

In the useful band of the amplifier, R_o and θ are made small so that

$$Z_i = \frac{1}{j\omega C_i} = \frac{1}{j\omega(1+K)C_1} \qquad (12)$$

and therefore

$$C_i = (1+K)\, C_1 \qquad (13)$$

In practice, $\theta \neq 0$ and $R_o \neq 0$ and the electronic capacitance exhibits dissipation. A quality factor, defined to measure dissipation, is given by

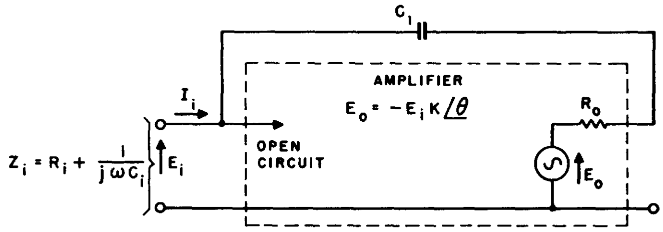
$$Q = \frac{1}{\omega R_i C_i} \qquad (14)$$

When $R_o = 0$ and $K \gg 1$, it is necessary for $0 \geqslant \theta \geqslant -1.2$, in order that $Q \geqslant 50$. When $\theta = 0$,

$$Q = \frac{1}{2\pi f R_o C_1} \qquad (15)$$

and thus when f = 3000 cps and C_1 = 500 μμfd, it is necessary for R_o < 2000 ohms, in order that $Q \geqslant 50$.

Phase shift and finite output resistance exist simultaneously; thus it is desirable to keep phase lag less than one degree, and the output resistance less than 1000 ohms. The amplifier input must be, effectively, an open circuit, as assumed in the derivation. This requirement is met by omitting a grid resistor for the amplifier input tube and providing bias and dc return through the transmission-line inductors. Parasitic capacitance C_p increases the effective capacitance of the section beyond C_i. Therefore,



C_1
AMPLIFIER
$E_o = -E_i K \angle\theta$
I_i
R_o
OPEN CIRCUIT
$Z_i = R_i + \frac{1}{j\omega C_i}$
E_i
E_o

Fig. 9. Scheme of a Miller-effect amplifier. The magnitude of C_i at the input is varied by varying the amplifier gain. C_1 is a small fixed capacitor.

$$C_{eff} = C_p + C_i = C_p + C_1 + KC_1 \qquad (16)$$

When K is small, changing K by a given factor changes C_{eff} by a much smaller factor. The gain, K, must vary from 1 to 350 in order to realize a capacitance range of 100:1. The voltage across the apparent capacitance is 1/K times the voltage at the output of the amplifier. Large values of K and moderate signal voltage across C_i imply a large signal voltage at the output of the amplifier. The maximum peak-to-peak signal at that point is comparable to the plate supply voltage of the amplifier.

The Miller-effect capacitor is shown in Fig. 10, which is a detailed block diagram of a section. The Miller amplifier consists of three parts: a high-gain input stage, an electronic attenuator, and a low-impedance output stage. Of these, the electronic attenuator departs most from conventional circuitry, and thus merits some discussion. The requirements for the attenuator are:

(a) Gain over a 350:1 range, controlled by a voltage with frequency components ranging from zero to 25 cps.

(b) Minimum disturbance of amplifier operating points and dc levels by the control voltage (to maintain suitable conditions within the amplifier and to prevent the control voltage from appearing as an unwanted component of the voltage across the completed capacitor).

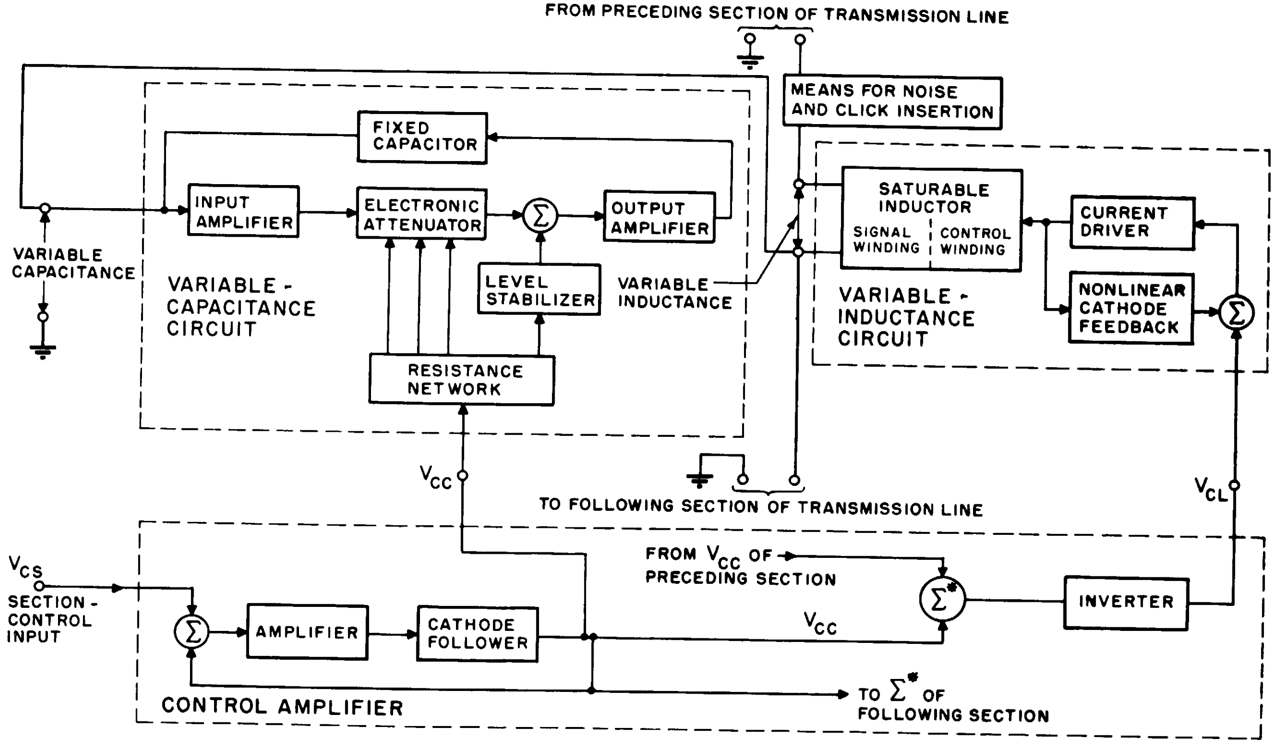


Fig. 10. Detailed block diagram of a typical section. The control voltage V_{CS} at the section input controls both variable elements by first controlling V_{CC}. Either voltage may be used when describing section performance.

(c) Ability to handle large signal voltages (compared with variations in dc levels) without excessive distortion.

(d) Less than one degree of phase shift in the useful (100 cps-3 kcps) band.

(e) Frequency range wide enough to permit inclusion in a loop having 55-db gain.

(f) Gain versus control voltage to be exponential.

(g) High input impedance and a low output impedance.

(h) Stability of calibration and absence of critical adjustments.

The strategy employed in the development of this attenuator was to:

(a) Utilize the known effect of control-voltage changes on dc levels to generate a cancellation voltage or current. (There is insufficient separation between the highest control frequency and the lowest signal frequency to permit the use of filtering for removing control-voltage components from the signal path.)

(b) Maintain the ratio of signal voltage to level change as high as possible.

(c) Use several attenuation stages to reduce requirements on individual tubes.

(d) Connect successive stages in a manner compatible with large changes in dc level.

(e) Keep the asymptotic phase shift at low and high frequencies sufficiently low to be suitable for use in a high-gain feedback loop (7).

The method for realizing this attenuator, as shown in Fig. 11, is by a ladder attenuator with vacuum tubes used as the shunt elements and fixed resistors as the series elements.

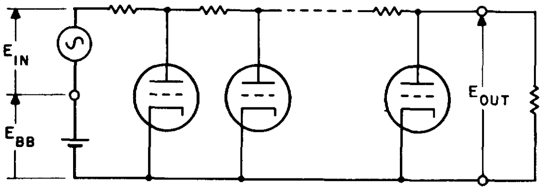


Fig. 11. A ladder attenuator network with tubes used as the shunt elements, and fixed resistors as the series elements.

The use of variable r_p, instead of variable μ, provides high signal-handling ability. Confining the connection between attenuator stages to the plate circuit permits drastic uncompensated operating point changes to occur in an early stage of the attenuator without disabling a later stage. Many stability problems are circumvented by grounded-cathode operation, in regard both to drift and high-frequency phase response. The latter is important because the unity loop-gain point is above 2 mc/sec in the completed capacitor. The attenuator is useful down to zero frequency because reactive elements or transformers are not needed for coupling between attenuator tubes. By putting small capacitors across each series resistor, the high-frequency asymptotic behavior can be made like that of a capacitive voltage divider for which the slope of the attenuation curve is zero in the neighborhood of infinite frequency. It is clearly impossible to put a capacitor across the (hypothetical) plate resistance, r_p, of a tube used as an amplifier, in

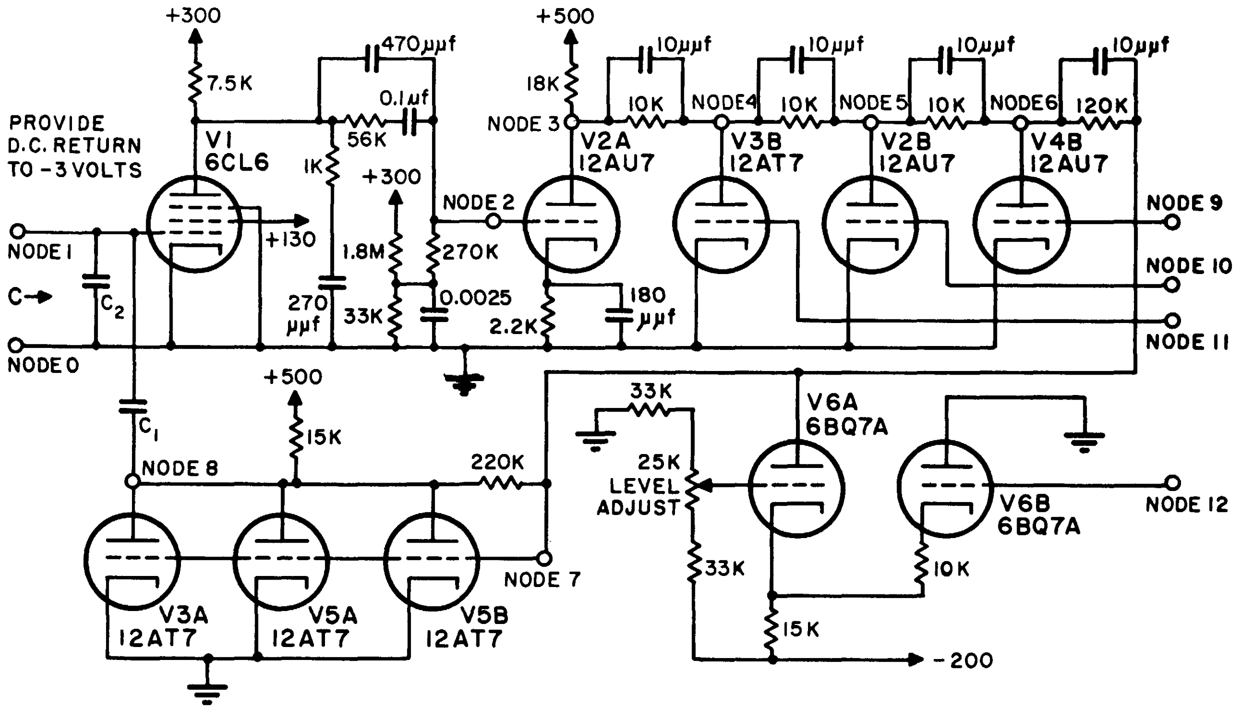


Fig. 12. Signal amplifier and level stabilizer of the variable-capacitance circuit. The resistance network (shown in Fig. 61) is connected to nodes 9 through 12 and applies control voltages which, in turn, are derived from V_{CC}, the capacitor control voltage.

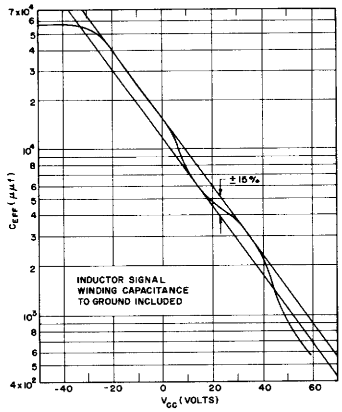


Fig. 13. Control characteristic of the variable capacitor.

order to achieve the same asymptotic behavior. A nonlinear circuit, designated "level stabilizer" in Fig. 10, is connected between the attenuator control source, V_{CC}, and the output amplifer to stabilize the dc level of the signal applied to the output amplifier against variation caused by changes of operating points in the attenuator.

The desired calibration curve (Eq. 6) was achieved by choosing the proper functional relations between the attenuator control voltage V_{CC} and the grid voltages applied to the shunt tubes. These relations are realized by use of a resistance network in conjunction with grid conduction in the attenuator tubes. This network is shown in Fig. 10 and is described in detail in the appendix. The circuit of the attenuator, together with the input and output amplifiers, is shown in Fig. 12. The control characteristic of the capacitor is shown in Fig. 13. The capacitance is an exponential function of control voltage within 15 per cent of the mean over a 40:1 range, in the presence of the expected stray capacitance to ground. Dissipation in the capacitor, described by the contour map in Fig. 14, is within the limits specified above.

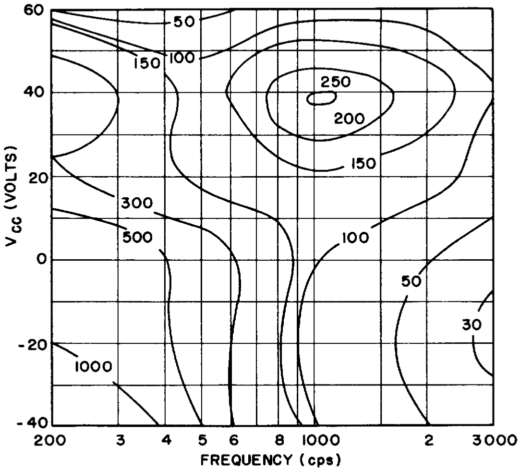


Fig. 14. Contours of Q (quality factor) of the variable capacitor. The required minimum Q is achieved at 3 kc. The Q is well above minimum at other frequencies.

3.5 VARIABLE-INDUCTANCE CIRCUIT

The variable inductance is realized as a saturable inductor. A saturable inductor has a magnetic circuit consisting of three parts, or legs, arranged in a "figure eight." A control winding is wound on the inner leg, which produces a control flux whose return paths are the outer legs. The signal winding is split into two halves, each of which is placed on an outer leg and connected in series aiding. As a result, the mutual inductances between each half of the signal winding and the control winding have opposite polarities, and hence the net coupling between the signal and control winding is very small. Control flux through the magnetic material of the outer legs varies the incremental permeability of that material through values lying between those of the demagnetized and fully saturated states. The signal-winding inductance is proportional to the permeability of the core. This inductance as a function of control current is given in Fig. 15.

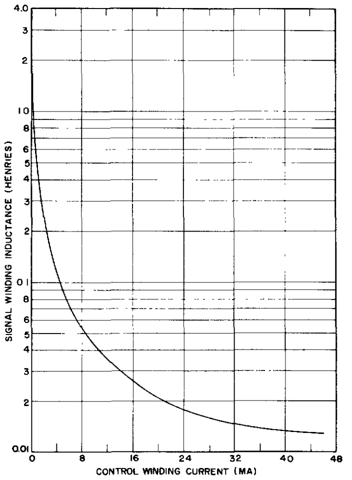


Fig. 15. Control characteristic of a typical saturable inductor.

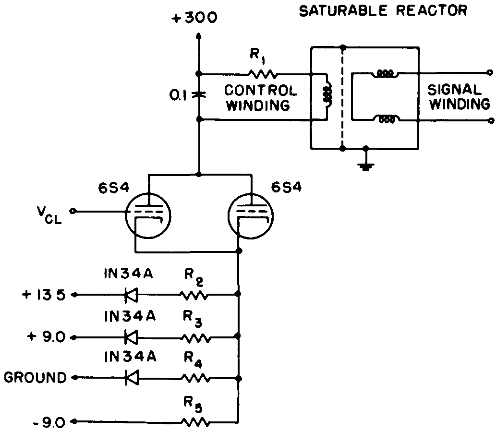


Fig. 16. Driver circuit for inductor control winding.

The inductance of the signal winding is not an exponential function of the current through the control winding. Control current must, therefore, be obtained from a nonlinear driver circuit that is so constructed that the inductance of the inductor-signal winding is an exponential function of V_{CL}, the control input voltage to the driver. Figure 15 shows that signal-winding inductance is very sensitive to changes in control current in the region of low current and high inductance. Thus the gain of the driver must be small at low currents, but its stability of calibration must be very high, and vice versa. Also, the driver output impedance must be high if rapid response is to be obtained. A triode with a nonlinear cathode resistor meets all of the stated requirements. The backward gain in the driver circuit, Fig. 16, is a monotonic function of output

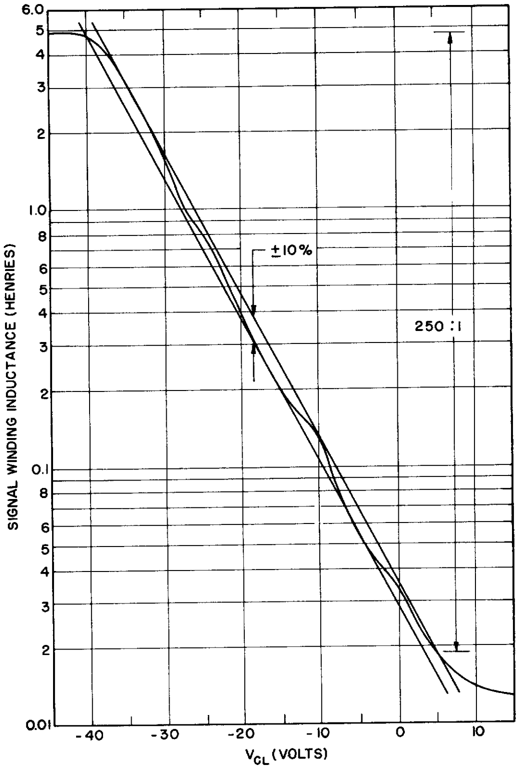


Fig. 17. Control characteristic of inductor with driver circuit. The characteristic is exponential over an inductance range of 250:1 with a tolerance of ±10 per cent.

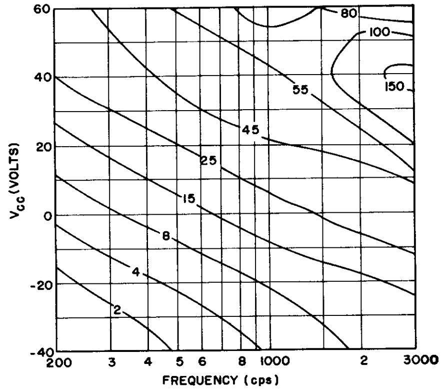


Fig. 18. Contours of Q of the inductor. (The ordinate is given in V_{CC} to make possible direct comparison with Figs. 14 and 20.) The inductor control voltage V_{CL} can be obtained from V_{CC} by use of Eq. 8.

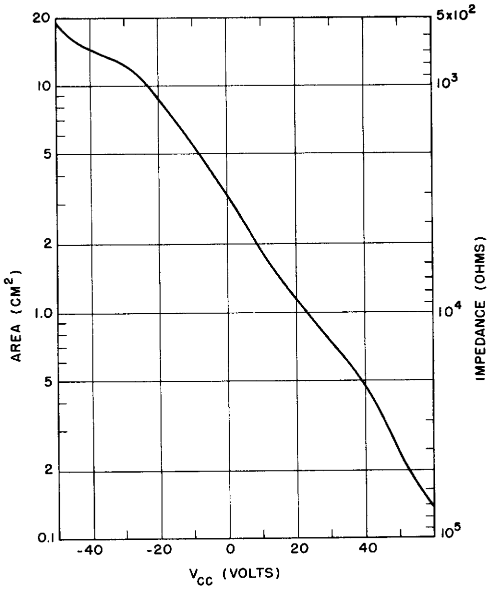


Fig. 19. Calibration curve of a typical section.

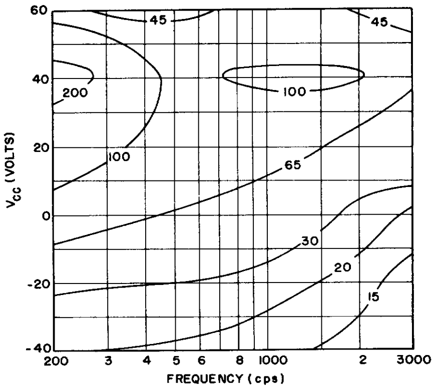


Fig. 20. Contours of Q of a single section.

current; it is greatest at low output current and least at high output current. Thus the driver circuit exhibits the greatest calibration stability in the region where the least incremental gain is required, and vice versa.

A piecewise-linear cathode resistor with three breakpoints is sufficient to obtain the over-all control characteristic shown in Fig. 17. The control characteristic of the inductor-driver combination is exponential over a range of 250:1 with a tolerance of 10 per cent. The contours of constant quality factor, Q, of the inductor are given in Fig. 18.

3.6 A COMPLETE SECTION OF THE TRANSMISSION LINE

A single section is constructed by connecting a variable capacitance circuit and a variable inductance circuit to a control amplifier (see appendix) so that V_{CC} and V_{CL} are related by Eq. 8, and both are controlled by V_{CS}. The calibration curve relating section area to V_{CS} is given in Fig. 19. The effect of summing control voltages for the inductor is eliminated from consideration by setting V_{CC} (see Fig. 10) of the adjacent section equal to V_{CC} of this section. This is tantamount to making the line uniform during calibration. A single measure of the dissipation in a completed section can be obtained by treating it as a simple resonance circuit. The definition of Q given in Eq. 13 is used without simplifying assumptions to derive an expression more general than Eq. 5. If we substitute the data of Figs. 14 and 18 in Eq. 13, the contour map of Fig. 20 is obtained.

If the section is treated again as a simple resonance circuit, its ability to maintain constant length as its area changes can be measured statically and dynamically. The resonant frequency is inversely proportional to section length. By measuring the former statically as a function of V_{CS}, the latter was found to remain within 14 per cent of the mean as the section area varied over a 100:1 range. In a dynamic test, a large step of V_{CS} was applied to the section, and the time required for the resonant frequency to reach a new steady-state value was measured as follows: The output of a white-noise generator was filtered by the section and fed to a sound spectrograph. The band of energy concentration in the resulting spectrograms moved from one frequency to another within a few milliseconds between values given by the static curve. The control signals actually used vary much more slowly than the test input. Thus, the section-length constraint is realized just as well with time-variant control inputs as it is in the steady state.

IV. APPARATUS FOR EXCITING THE VOCAL-TRACT ANALOG AND CONTROLLING ITS CONFIGURATION

4.1 RATIONALE FOR THE CONTROL SYSTEM

The plan for the control system and the size of its storage were predicated on a study of sequences of two phonetic elements. Such a control system permits the study of the essential features of connected speech with a minimum number of experimental variables.

Let us trace the events when a human talker generates a typical sequence, a fricative-vowel syllable. The timing of the events is indicated in Fig. 21. At the beginning, the speaker has his vocal tract in a constricted configuration that is appropriate for the fricative, and no sound is generated. He then produces the fricative. The amplitude of this noiselike sound increases from zero to some sustained value. Soon, he begins to move his articulators to the vowel configuration, and, as the fricative constriction disappears, the region of turbulent air flow ceases to exist and the noise amplitude falls to zero. Meanwhile, he has signaled the vocal folds to start vibration. The amplitude of vocal excitation increases from zero to its sustained value. This excitation is applied to the vocal tract, now at its vowel configuration, and thus a vowel is produced. As the vibration of the vocal folds ceases, the buzz amplitude decays to zero and the speaker is again silent. Since a human being will rarely speak in a monotone, some variation of the vocal frequency is indicated.

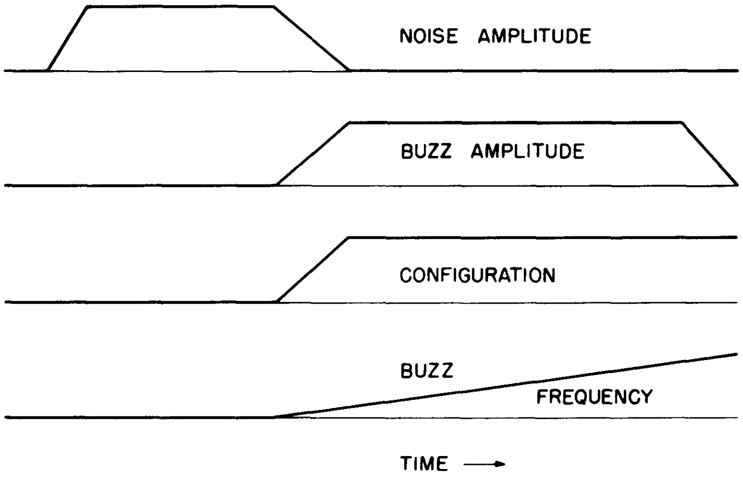


Fig. 21. Sequence of events occurring during the production of a fricative-vowel syllable. Piecewise-linear approximations are used to specify the values of the parameters as functions of time. Base lines for amplitude waveforms represent zero excitation. For the graph labeled "configuration," the first quiescent value corresponds to the fricative configuration, and the second corresponds to the vowel configuration. The ramp connecting the two values merely shows the time and duration of the transition. The buzz-frequency graph indicates an increment of frequency. Longer utterances can be described by graphs having more segments than the example shown.

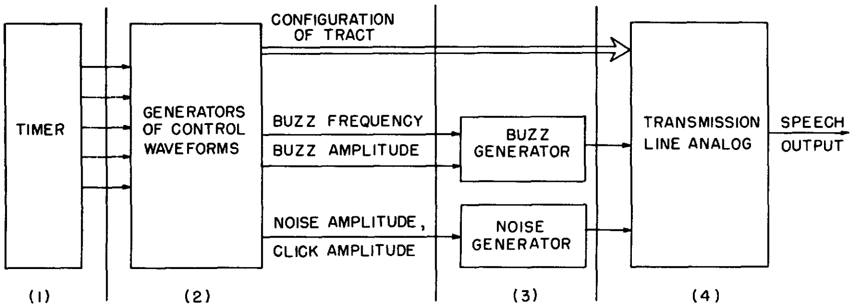


Fig. 22. Block diagram of the control system. The timer emits pulses to trigger the generators of continuous control waveforms in column 2. A pulse occurs on each timer output line at a time preset on a matrix within the timer. The starting time of each ramp is determined by a trigger pulse, and the duration of that ramp is determined by settings of the appropriate generator. The block in column 2 also contains a matrix for storing initial and final vocal-tract configurations. This matrix is driven by a waveform generator, and both devices act together as a 14-channel generator to furnish the 14-section area control signals applied to the transmission-line analog. This set of signals specifying the configuration of the vocal tract is indicated by the large arrow.

Description of the fricative-vowel syllable requires a model with the following features: a single articulatory transition, a single occurrence of inflected glottal excitation, and a single occurrence of noise excitation. This model will serve for any two-element sequence if the following generalizations are made:

(a) Events are allowed to occur in any order, in any spacing.

(b) Time parameters, such as durations and rates, have wide ranges.

(c) Specification of articulatory configuration is extended to include degree of nasal coupling — when the proposed nasal circuit is added.

(d) Single shock-pulse excitation, as well as noise excitation, is available, if necessary, to simulate closure release.

The control system is an implementation of this model. By duplicating some existing units, sounds that seem to have several noise sources can also be investigated. In its present realization, the system behaves as though all masses involved in articulation completed their movements in the same interval. This is probably not quite true of human speech organs, but the existing arrangement permits the study of first-order effects in speech production before proceeding to second-order effects.

The completed control system is not strictly limited to two-element sequences. It can produce longer sequences, provided that their structures are not too complex. The system can also control POVO or terminal analog synthesizers (52). When used with ancillary circuits, it can generate stimuli that are useful in psychoacoustics (3).

The general arrangement of the synthesizer, including the control system, is shown in Fig. 22. The output member is the dynamic analog (column 4). It must be provided with control signals and with excitation. The buzz and noise generators (column 3) furnish the excitation and require control signals themselves. These signals are furnished by waveform generators (column 2). Their outputs closely resemble the time functions shown in Fig. 21. The generators are kept in synchronism by means of the timing pulse source (column 1). The timing pulses determine the positions and durations of the trapezoidal functions. Other circuits (not shown) exist for monitoring, for synchronizing a tape recorder, and for auxiliary switching.

4.2 CONTROL OF CONFIGURATION

The purpose of the apparatus to be described here is to generate the configuration signals (Fig. 22) that are applied to the section area control inputs shown in Fig. 4. During the production of a typical two-element sequence each section passes through three states: (a) an initial quiescent state in which the section has an area appropriate for the initial vowel or consonant; (b) a transition state; and (c) a final quiescent state in which the section has an area appropriate for the final vowel or consonant. As we have mentioned, the transitions for all sections occur in the same time interval.

Let v_j be a typical waveform that resembles those signals applied to the area-control input of a typical section. Each v_j has an initial quiescent value $b_{2,j}$, a linear transition lasting from $t = t_1$ to $t = t_1 + T$, and a final quiescent value $b_{1,j}$. Each v_j can be expressed as a linear combination of two time functions, $f_1(t, t_1, T)$ and $f_2(t, t_1, T)$, having the same transition interval as v_j, but with fixed quiescent values, 0 and 1. Let

$$f_1(t, t_1, T) = \begin{cases} 0 & t \leqslant t_1 \\ \dfrac{t - t_1}{T} & t_1 \leqslant t \leqslant t_1 + T \\ 1 & t \geqslant t_1 + T \end{cases} \tag{17}$$

and

$$f_2(t, t_1, T) = \begin{cases} 1 & t \leqslant t_1 \\ \dfrac{T - t + t_1}{T} & t_1 \leqslant t \leqslant t_1 + T \\ 0 & t \geqslant t_1 + T \end{cases} \tag{18}$$

Then

$$\begin{aligned} v_j(t, t_1, T) &= b_{1,j} f_1 + b_{2,j} f_2 = b_{1,j} f_1 + b_{2,j}(1-f_1) \\ &= b_{2,j} + (b_{1,j} - b_{2,j})\, f_1 \end{aligned} \tag{19}$$

where $0 \leqslant b_{1,j} \leqslant 1$ and $0 \leqslant b_{2,j} \leqslant 1$. That is

$$v_j(t, t_1, T) = \begin{cases} b_{2,j} & t \leqslant t_1 \\ b_{2,j} + (b_{1,j} - b_{2,j}) \dfrac{t - t_1}{T} & t_1 \leqslant t \leqslant t_1 + T \\ b_{1,j} & t \geqslant t_1 + T \end{cases} \tag{20}$$

The summations necessary for generating a set of 14 control voltages, v_j, $j = 1, \ldots, 14$, are conveniently written in matrix notation:

$$[f_1, f_2] \begin{bmatrix} b_{1,1}, \ldots, b_{1,14} \\ \\ b_{2,1}, \ldots, b_{2,14} \end{bmatrix} = [v_1, \ldots, v_{14}] \tag{21}$$

All functions in Eq. 21 are to be considered normalized, so that 0 represents the value of v_j which sets the area of a given section to its minimum value and so that 1 represents the v_j required for maximum section area. Thus, b_2 is the normalized V_{CS} appropriate for the initial vowel or consonant and b_1 is the normalized V_{CS} appropriate for the final vowel or consonant.

In practice, section calibrations cannot be made to coincide precisely. The normalization should be reinterpreted as relating to the set of sections, rather than to an individual section. Then, let 0 stand for a least value of V_{CS} among all sections, and let 1 stand for a greatest value. Therefore, v_j, the normalized V_{CS} to a typical section, will cover most, but not quite all, of the interval (0, 1), in order to bring that section through its full area range.

The coefficients $b_{i,j}$ are physically realized as potentiometer settings. Changing these settings is tedious and time consuming and restricts the number of comparisons that can be made during informal listening because of the listener's limited memory span. It is desirable to store several configurations in a large matrix and select the required pair by switching. Let $[s]_{2,6}$ be a 2 × 6 switching matrix, all of whose elements are equal to zero except for two, one from each row, which are equal to one. (The symbol $[x]_{m,n}$ denotes an m × n matrix with typical element x.) Let $[a]_{6,14}$ be a 6 × 14 matrix for storing six configurations. Then

$$\begin{bmatrix} b_{1,1}, \ldots, b_{1,14} \\ \\ b_{2,1}, \ldots, b_{2,14} \end{bmatrix} = \begin{bmatrix} s_{1,1}, \ldots, s_{1,6} \\ \\ s_{2,1}, \ldots, s_{2,6} \end{bmatrix} \begin{bmatrix} a_{1,1}, \ldots, a_{1,14} \\ \\ \\ \\ a_{6,1}, \ldots, a_{6,14} \end{bmatrix} \tag{22}$$

For example, if the fifth row of $[a]_{6,14}$ is selected for the initial configuration and the third row is selected for the final one, then $s_{1,5} = 1$, $s_{2,3} = 1$, and all other $s_{i,k} = 0$. Then

$$b_{1,j} = a_{5,j} \qquad j = 1, \ldots, 14$$

and

$$b_{2,j} = a_{3,j} \qquad j = 1, \ldots, 14$$

The matrix $[f]_{1,2}$ can be expressed as

$$[f_1, f_2] = [f_1[1,-1]+[0,1]] \tag{23}$$

and Eq. 21 can be rewritten as

$$[f]_{1,2}[s]_{2,6}[a]_{6,14} = [v]_{1,14} \tag{24}$$

Note that $[v]_{1,14}$ is determined entirely by f_1 and matrix coefficients:

$$[v_1, \ldots, v_{14}] = [f_1[1,-1]+[0,1]] \begin{bmatrix} s_{1,1}, \ldots, s_{1,6} \\ \\ s_{2,1}, \ldots, s_{2,6} \end{bmatrix} \begin{bmatrix} a_{1,1}, \ldots, a_{1,14} \\ \\ \\ \\ a_{6,1}, \ldots, a_{6,14} \end{bmatrix} \tag{25}$$

Equation 25 is implemented directly by the arrangement shown in Fig. 23. The input, f_1, is obtained from a function generator that will be described in section 4.5. The input

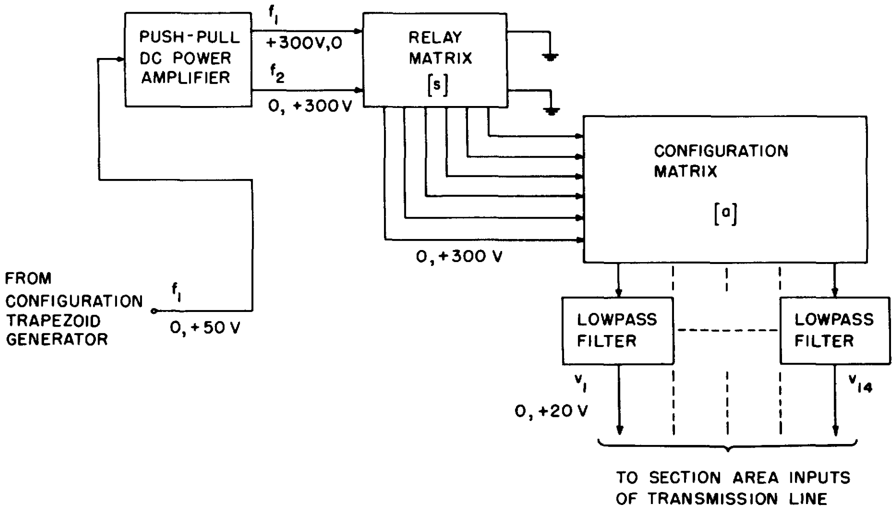


Fig. 23. Apparatus for generating area-control voltages for transmission-line sections. Steady-state configurations are stored in matrix [a]. The filters provide smooth transitions between steady-state voltages, and are intended to approximate the dynamics of vocal-tract motion.

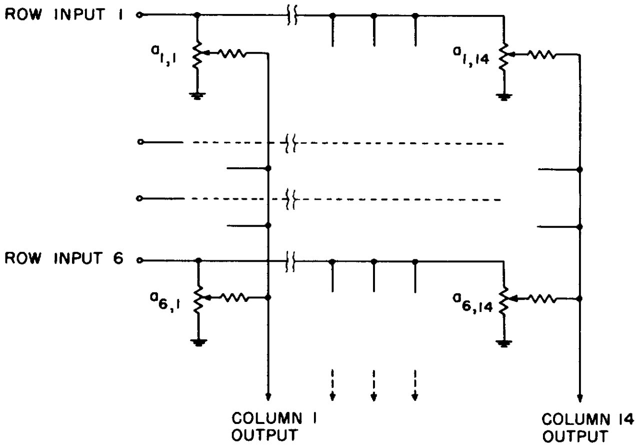


Fig. 24. Matrix for the storage of articulatory configurations. Each row of the matrix is associated with one configuration. Each column is associated with one control input to the transmission line.

is applied to the push-pull dc amplifier and yields an output $[f_1, f_2]$. The voltage level at the output of the amplifier is such that a change from one configuration to another involves a change of 300 volts at each output terminal. The relay matrix [s] connects the amplifier outputs to two row inputs of the configuration matrix [a], and grounds all other row inputs. The column outputs of [a] are the v signals of Eq. 25. Each v signal, since it is piecewise linear, has a discontinuity of derivative at the beginning and end of the transition interval. This is unsatisfactory, theoretically, because it implies infinite acceleration of the articulators and, practically, because it causes loud clicks in the output of the analog. The lowpass filters smooth the configuration signals. The filter should have dc gain that is reasonably close to unity, no overshoot, and zero initial response to a step. The simplest filter meeting these specifications, a single RC section, has been found to be satisfactory.

The circuit of the push-pull dc power amplifier is given in the appendix. The amplifier is rated at 0 to +300 volts, 0 to 70 ma output, and 5 ohms internal impedance. The circuit of matrix [a] is shown in Fig. 24. The maximum gain from any given row input to any given column output is 1/6 because there is loss in the summing network. The high-gain requirement is imposed on the filter to obviate the need for amplification between matrix and section-control input. Thirteen outputs are now in use; the fourteenth is reserved for specifying degree of nasal coupling. Unused row inputs are grounded to reduce interactions among potentiometer settings. The two active rows are effectively grounded, in this respect, because they are connected to low-impedance generators. From an over-all viewpoint, the apparatus may be regarded as 14 trapezoidal generators with individually adjustable quiescent levels. Each such generator provides a piecewise-linear control voltage for a section of the transmission line. Because the 14 units are not really independent, a ramp at the output of one generator must begin or end at the same time as the corresponding ramp at the output of another.

The [a] and [s] matrices can be used in two ways: (a) to store the configurations for a number of alternative two-element sequences, and (b) to store configurations needed for producing longer sequences. The study of the syllable [sɑ] illustrates the first mode of operation. Five alternative [s] configurations may be stored, one per row, with the sixth row used to store the [ɑ] configuration. The alternatives may be rapidly compared in informal listening or quickly recorded for formal listening.

To illustrate mode 2, consider the production of a three-element sequence, such as [bɔℓ]. The [b] configuration is stored in row 1, the [ɔ] configuration in row 2, and the [ℓ] configuration in row 3. Let the switching matrix and power amplifier apply a normalized input of 1 to row 1, a normalized input of 0 to row 2, and a ground to row 3. Assume that the input to the amplifier is a full trapezoid, that is, a function having both upward and downward ramps with excursions from 0 to 1. Initially, the amplifier input is 0; after the first transition (from [b] to [ɔ]) it is 1. Then the conditions at the input to matrix [a] are:

row 1	0 applied
row 2	1 applied
row 3	grounded.

If the switching now operates to ground row 1 and to apply a zero input to row 3, the voltages at the matrix output will be undisturbed. However, when the downward ramp at the amplifier input occurs, the transition will be from [ɔ] to [ℓ]. The switching matrix is operated by an auxiliary switching device that is triggered by the pulse source shown in column 1 of Fig. 22.

This switching scheme can be used for sequences of arbitrary length. The variety of utterances possible with any given device depends on the size of its [a] matrix and on the number of ramps available in the input waveform. The duration of each transition is independent of the duration of other transitions, and each can be of length that is appropriate to the production of the associated pair of phonetic elements.

4.3 EXCITATION

The elements belonging to column 3 of Fig. 22 will now be described. Together they provide the buzz and noise excitation for the analog.

a. Buzz Excitation

During the production of voiced sounds by the human vocal tract the vocal folds vibrate; this causes periodic changes in the size of the glottis, the opening between the vocal folds. Puffs of air giving rise to a volume velocity waveform, such as that shown in Fig. 25, are thus released into the vocal tract. The vocal fold vibration can start with various attack characteristics and can end with various decay rates. The vocal frequency is also varied by the speaker, partly to indicate emotion and partly to observe the rules of his language community (8). A more complete description of human speech must

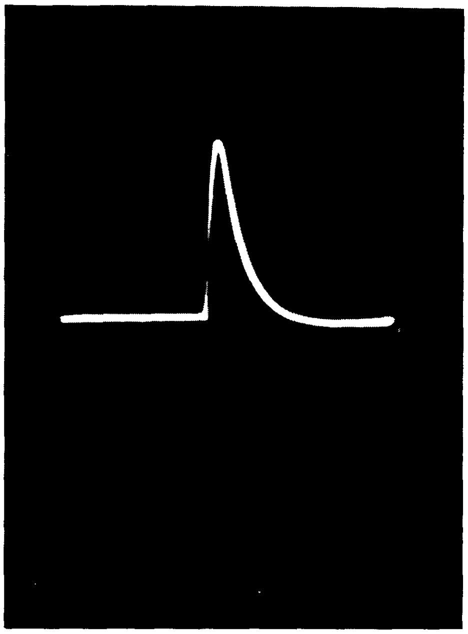


Fig. 25. Typical glottal pulse obtained from the buzz generator.

consider other phenomena; the vocal frequency exhibits slight irregularities (37). The glottal output changes with changes in vocal effort (25, 40), and is different for different speakers (44). Singers are able to vary their vocal outputs to get different qualities. The value of the synthesizer as a research tool would be increased if these effects could be included to help make the buzz generator an artificial larynx.

In the synthesizer, the buzz amplitude envelope and the buzz frequency pattern are approximated by trapezoidal functions as in Fig. 21. The linear rise and decay of the amplitude envelope trapezoid have an important advantage over other shapes such as the exponential, in that their beginnings and ends are well defined. Thus, perceptual data relating to stimuli with trapezoidal envelopes may be free of artifacts owing to "tails" of excitation. A trapezoidal signal is used to vary the buzz frequency. Variable buzz frequency makes the study of inflection patterns possible and, in phonetic identification experiments, adds to naturalness. Often the inflection pattern in a single syllable is very simple, requiring only one or two of the five linear segments available in a trapezoid.

The desired output for a buzz generator is a train of pulses whose spacing and amplitude envelope are specified by control signals. Also, each pulse must have the form of a typical glottal pulse, such as is shown in Fig. 25. In the conventional approach, a train of pulses with the required form and spacing is fed to a device, such as a balanced modulator, which imposes an amplitude envelope on that train in response to a control signal. But the classical balanced modulator has a highly nonlinear control voltage-versus-output characteristic with tails in the region of the tube cutoff. It also has distortion whose character varies with gain. The theory of the balanced modulator assumes perfectly balanced tubes and transformers, in order that the control voltage (common mode) cancel out. In practice, this degree of balance is unattainable and an appreciable portion of the control voltage appears as an unwanted component (thump) in the output. The signal-to-thump ratio of the classical balanced modulator is often less than one. Spectra of control signals and pulse train overlap, and hence thump cannot be removed by filtering as, for example, is done in AM transmitters. Thump is troublesome because, when applied to the transmission line, it can drive the Miller capacitors out of their operating regions. Thump also gives rise to clicks that impart an undesired consonantal quality to the synthesizer output.

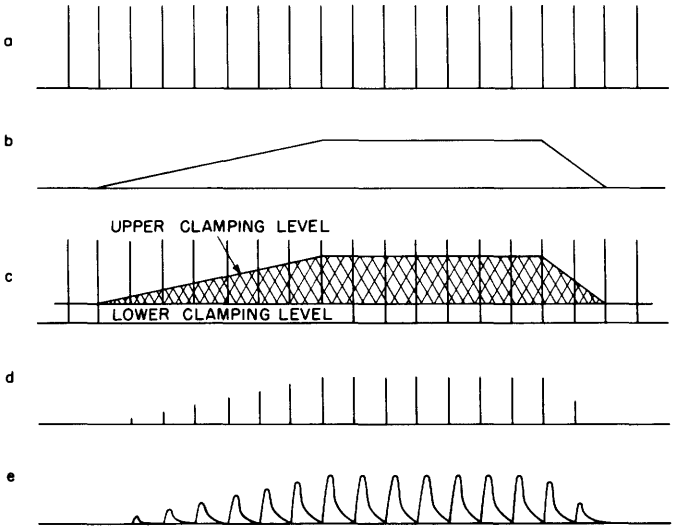


Fig. 26. Waveforms within the buzz modulator illustrating its operation. Each part of this figure shows the waveform at a node in Fig. 27 and is labeled with a corresponding letter. Crosshatching in (c) shows the window between upper and lower clamping levels. Pulses must be adjusted to extend above and below the window.

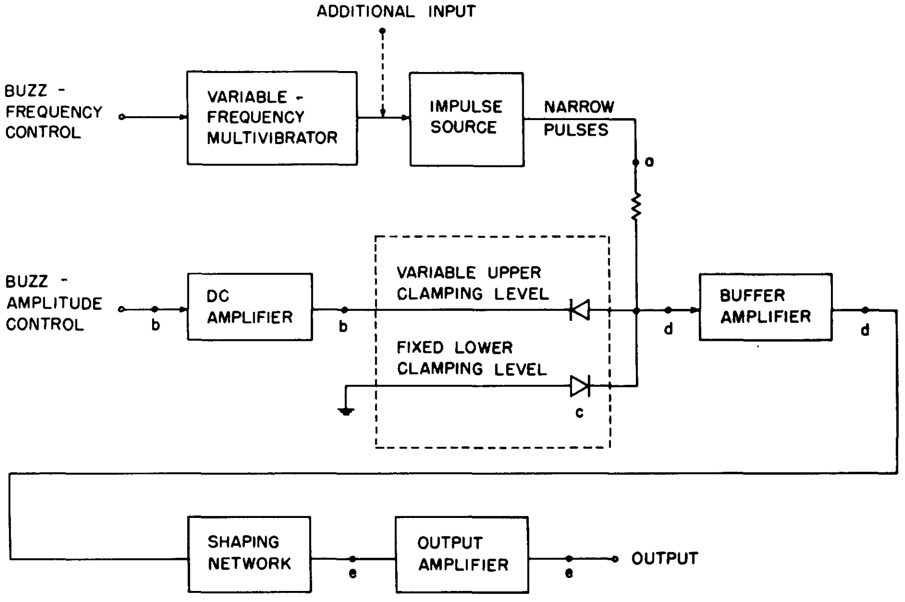


Fig. 27. Block diagram of the buzz generator. Voltages at lettered nodes correspond to similarly labeled waveforms in Fig. 26. The window in Fig. 26 exists through operation of diode clamps shown here in block c. For phase-coherent stimuli, the multivibrator may be disconnected so that glottal pulses are triggered by the timer through the additional input.

To circumvent the shortcomings of conventional modulators, a linear modulator was developed. It operates on the principle that clipping a rectangular pulse does not distort it but merely changes its amplitude. The modulator consists of a clipper connected to a dc amplifier in such a manner that the clipping level changes in response to a control voltage. The operation of the modulator can be traced with the help of the waveforms shown in Fig. 26 and in the block diagram in Fig. 27. The multivibrator generates a modified square wave whose frequency changes in response to a control voltage. The square wave drives a monostable trigger circuit that generates a train of very narrow pulses (50-μsec duration). If we assume that buzz frequency is constant, the train will appear as it does in Fig. 26a. A control voltage, b, is applied to the dc amplifier and a magnified version appears at its output. This output defines an upper clamping level that, together with a lower clamp at ground, produces a "window," c, through which the pulses, a, are seen. In the classical balanced modulator, the audio signal applied to the tube grids is small compared with the control voltage; in our modulator, the "audio" is larger than the control voltage. The waveform in d, the resultant, is applied to the shaping network through a buffer amplifier. The shaping network has an impulse response like the wave in Fig. 25, and its output, e, is sent to the output amplifier. Figure 28 shows two oscillograms of the output of the buzz generator. Absence of thump is readily apparent. The glottal frequency-control signal is constant in part b and is rising in part c. The same amplitude control signal, a, is applied in both cases. The output in b

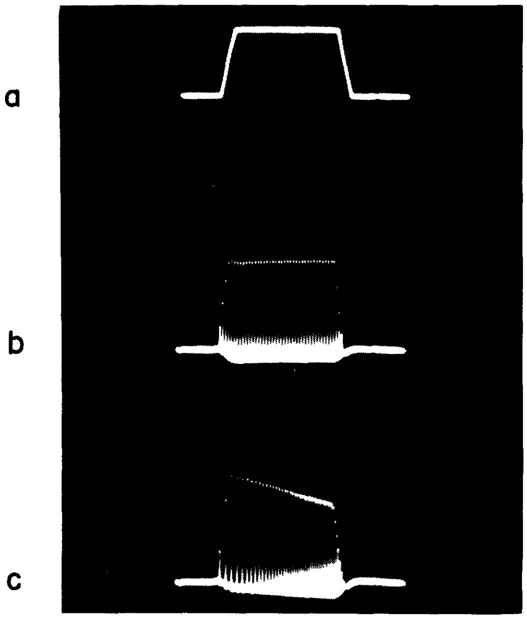


Fig. 28. (a) Oscillogram of amplitude-control input to buzz generator. Over-all duration of trapezoid, 450 msec; onset and decay times each, 50 msec. (b) Output of generator with buzz frequency constant at 100 cps. Note stability of base line and linear dependence of output on control input. (c) Output of generator with buzz frequency swept linearly from 50 cps to 200 cps in 450 msec.

is seen to be quite linearly related to the control voltage. The decreasing output amplitude in c is traceable to the lowpass action of the shaping network, and the excursion of the output around both sides of the zero line is traceable to ac coupling in the amplifiers. A rejection ratio of 60 db between "on" and "off" states is easily obtainable with this device.

b. Noise Excitation

The desired output for the noise generator is Gaussian noise, which is white to 10 kc and whose standard deviation (rms voltage) is proportional to a control signal. Amplitude control with the use of conventional modulators is unsatisfactory, for reasons previously stated. For example, the character of the distortion in a balanced modulator depends on its gain setting, and this is undesirable because it implies that the shape of the noise amplitude distribution at the output changes as the gain changes.

The clipping technique just described is applicable to noise modulation. The validity of the technique rests on the central limit theorem (14). The theorem states that if

$$y = \sum_{1}^{n} x_i \tag{26}$$

and if the x_i are statistically independent and are distributed with finite mean and variance, then the distribution of y approaches Gaussian distribution as n approaches infinity, regardless of the exact form of the x_i distributions. Figure 29 illustrates the principle of the noise modulator. The noise diode feeds an overdriven amplifier whose output is, ideally, a rectangular wave with random zero-crossing times and whose two values are equiprobable and symmetrical about zero. Such a wave has zero mean. The actual output distribution clusters about two values in a manner that approximates the ideal distribution, and the output waveform has significant frequency components up to 200 kc. Symmetrical clipping of the ideal rectangular wave does not distort it, but merely changes its amplitude scale. The push-pull dc amplifier acts as a threshold voltage source for the clipper, so that the clipping limits can change in response to the control voltage. Hence this voltage controls the amplitude scale of the clipper output.

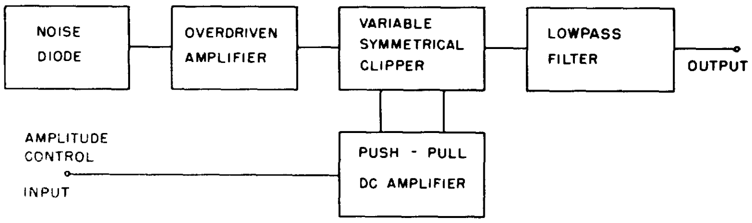


Fig. 29. Block diagram of symmetrical-clipping noise modulator. The lowpass filter acts as a summing device for noise samples.

The effect of the lowpass filter can be seen with the help of Shannon's sampling theorem (50). The input and output of the filter are bandwidth-limited to 200 kc and to 10 kc, respectively. Each sample of the filter output can be regarded as a weighted sum of effectively 20 samples of the filter input, although no lumped-parameter filter exhibits an ideal square-cornered impulse response. The conditions of the central limit theorem hold sufficiently well for the intended purposes, and the distribution of the filter output is approximately Gaussian, with the variance determined by the amplitude control voltage. Symmetrical circuitry operating on symmetrical waveforms yields an output with no dc component, so that the output voltage shown in Fig. 30 is free from thump. A modulator with many desirable characteristics has thus been achieved by taking advantage of the properties of the signal.

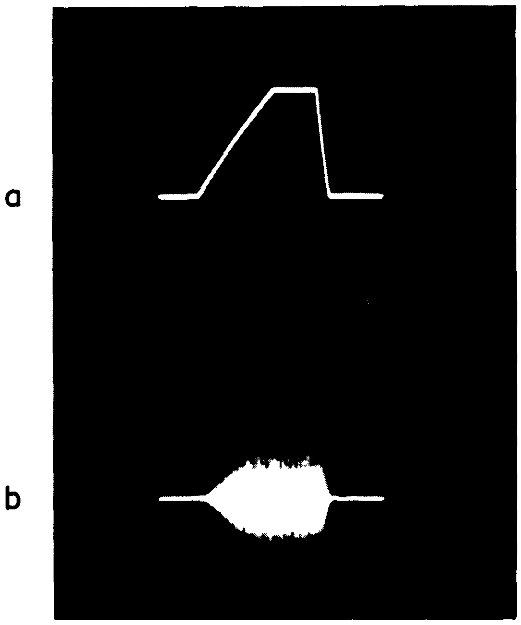


Fig. 30. (a) Oscillogram of amplitude-control input to noise generator. Over-all duration of the trapezoid, 450 msec. Onset and decay times, 250 msec and 50 msec, respectively. (b) Output of noise generator. Note stability of base line and linearity of modulation judged by visual estimation.

There is interaction between buzz and noise sources whenever they are used together. The vibrating vocal folds function as a kind of valve and cause the flow of air through the vocal tract to vary at the glottal rate. The noise power output of a turbulent region depends on the flow velocity through it, and so the noise power varies at the glottal rate. Simulation of this effect requires the addition of a glottal component to the noise control voltage. The circuit to be described uses the buzz amplitude-control voltage to control the amount of glottal component added, and introduces no glottal component at all during unvoiced intervals. The circuit can be permanently connected to the noise generator, and the synthesizer becomes thereby a more refined model of human speech production.

Figure 31 is a block diagram of the circuit that, like the modulators, uses the clipping window technique. All waveforms appearing in Fig. 32 pertain to partially overlapping buzz and noise, the buzz occurring first. Parts a and b show, respectively, the

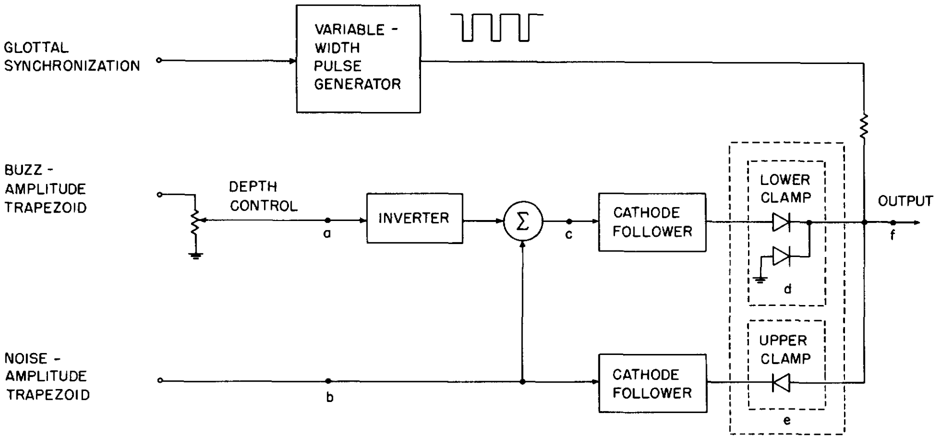


Fig. 31. Block diagram of circuit for simulating glottal modulation of noise. Waveforms existing at lettered nodes are shown in Fig. 32. The noise modulator is controlled by the output of this circuit rather than directly by a trapezoid generator.

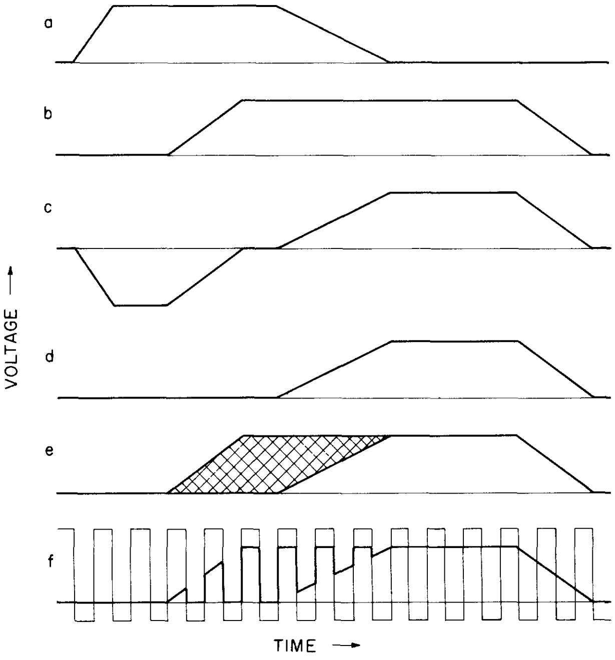


Fig. 32. Waveforms within the glottal-modulation circuit illustrating its operation. The time functions shown describe voltages or clamping levels at similarly labeled nodes or blocks of Fig. 31.

buzz-trapezoid and noise-trapezoid inputs; their difference, b – a, is given in part c. In the absence of the grounded diode, c would be the lower clamping level but, in combination, the two lower clamps operate to choose the greater of their respective levels, d. The upper clamping level is simply the noise amplitude trapezoid, b. The upper clamp, b, together with the lower clamp, d, form a window, e, through which a rectangular wave of glottal frequency is seen. The rectangular wave is synchronous with the glottal pulses and is of variable width. The view through the window, the heavy line in part f, is the final product of this circuit. Oscillograms of this control voltage and the

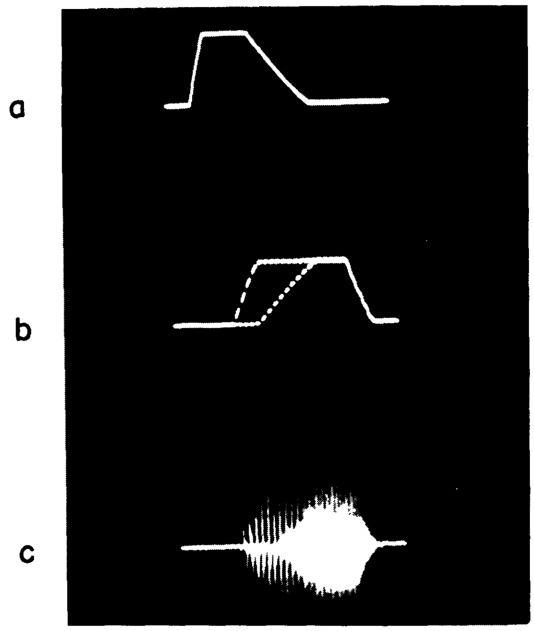


Fig. 33. (a) Oscillogram of buzz-amplitude trapezoid at input to the circuit of Fig. 31. (b) Output of the same circuit. Upper envelope of output is the same as the noise-amplitude trapezoid at the input to this circuit. (c) Output of noise generator modulated at glottal rate, showing decrease of noise modulation with decrease of buzz amplitude.

corresponding noise output appear in Fig. 33b and c. The circuit, in its present form, modulates the noise rectangularly. Notches of other shapes could be realized by shaping networks following node f. The network should have time constants comparable with the buzz period, and hence small compared with medium onset and decay times. Some experiments are needed to refine the output waveform. Informal listening demonstrates that rectangularly modulated noise yields a marked improvement over unmodulated noise.

Several other features have been incorporated to increase the usefulness of the buzz and noise generators. An external signal input on the buzz generator is provided for inserting noise at the glottis during the production of [h], and is useful for obtaining formant frequencies and transmission curves. An external input on the noise generator is useful for similar measurements, and serves for the insertion of clicks associated with some stop release sounds. An additional input on the buzz source (Fig. 27) allows for synchronization from other circuits, such as the timing trigger unit of Fig. 22. Voiced stimuli can be made phase-coherent by deriving glottal timing from the clock-pulse source that is part of that unit and will be described later. Both phase coherence and inflection may be obtained if the roles of master and satellite are interchanged by feeding pulses from the variable-frequency multivibrator (Fig. 27) to the timing unit. The buzz generator (operating at fixed frequency), together with a narrow filter, may be used as a source of gated sine waves that are not marred by switching transients. Such a wave train is useful in psychophysical experiments.

4.4 TIMING-PULSE SOURCE

The digital part of the synthesizer, represented by column 1 of Fig. 22, generates pulses that trigger all events and keep them locked to a common time base. The timing cycle is defined by 100 points equally separated in time. The time-base generator has 20 output channels, and only one pulse occurs in each channel during each cycle. The times of occurrence may be set at any one of the 100 points and the channels may be assigned to various functions as needed. (The first and last points of the interval, 00 and 99, are used for cycling and are not available for triggering. The option of having no pulse occur in any specified channel is available.) The first 8 channels are normally assigned to the trapezoid generators. Four generators, one for each trapezoid appearing in Fig. 21, are used, and each requires a trigger to initiate an upward ramp and another trigger to initiate a downward ramp. Two channels are normally connected to an oscilloscope for monitoring and time measurement and others are associated with auxiliary switching functions.

Figure 34 is a block diagram of the timing-pulse generator. The flow of control signals starts at the clock-pulse source which is usually set to give one pulse every 5 msec. The length of the timing cycle is determined by the pulse repetition rate, and this setting allows syllables up to 0.5 sec in duration to be produced. In effect, the clock pulses divide each cycle into 100 equal segments. The pulses are counted by two tubes, each

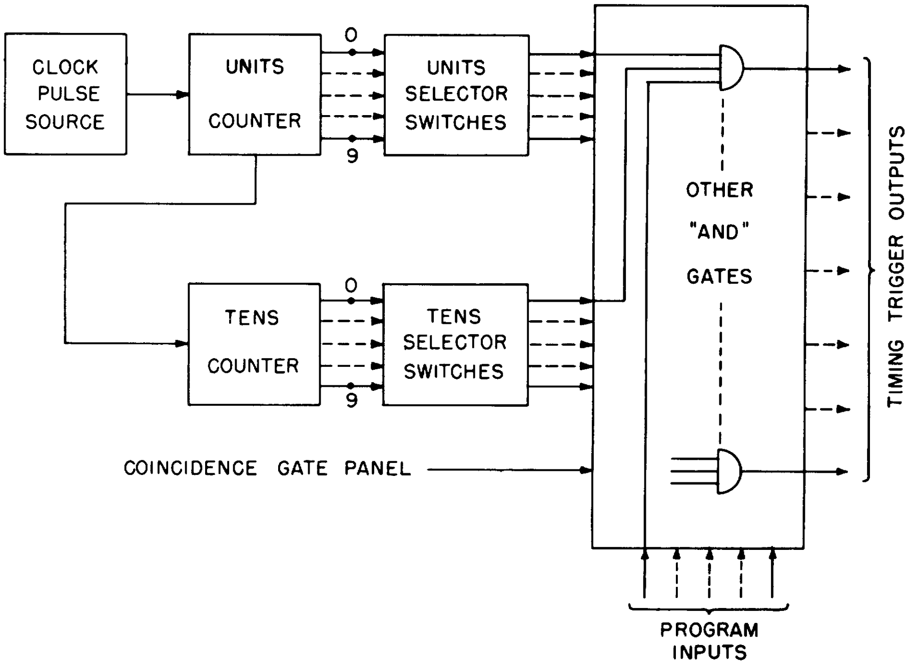


Fig. 34. Block diagram of timing-pulse generator. The clock may be driven from an internal oscillator, an external oscillator, or the buzz generator (for phase-coherent stimuli requiring variation of buzz frequency).

of which functions as a scale-of-10 counter. The counters can be stepped as quickly as 2 msec per pulse, and as slowly as desired.

The counter tubes are cold-cathode gas-discharge devices (4). Each tube is filled with neon and has 31 electrodes: a single disc-shaped anode, and 30 identical negative electrodes uniformly spaced in a ring about the anode. Ten of these are cathodes, and each cathode has its own terminal which is "high" when that cathode is conducting. The other 20 electrodes, called guides, are connected in rings of 10 guides each so that two guides, one from each ring, appear between each pair of cathodes. The counting action is obtained by stepping the glow from one cathode to the next one. The stepping depends on the ionizing property of the discharge that permits an electrode adjacent to a conducting electrode to start conduction more readily than others. In operation, the glow is transferred from the n^{th} cathode to adjacent guide 1, to guide 2, and then to the $n + 1^{th}$ cathode. One and only one output of the counter tube is high at any time, except during the transfer intervals when no cathode is conducting. Stepping the glow from the 9^{th} cathode returns it to the zero cathode. The transition from a 9 count to a 0 count in the "units" tube is used to pulse the "tens" tube whose count is thus increased by 1, and so the tens counter steps once for every 10 of the "units" tube counts. Each tens pulse is 10 times as long as a units pulse, and there are 10 times as many units pulses as there are tens pulses.

Each output channel in Fig. 34 uses a single logical element, the coincidence ("and") gate. When all three inputs to a given gate are high, the output is high, otherwise it is low. The program input is used to inhibit a gate, the use of which will be discussed in section 4.6. For the moment, assume that the program input does not exist and that each coincidence gate has only two inputs. Each units selector switch serves to connect one input of its associated gate to any desired output terminal of the units counter. The second input to each gate is similarly fed by the tens counter through its associated selector switch. A coincidence between a units pulse going into a gate and a tens pulse will cause an output at that gate. Thus, the 100-point time base has this significance: There is one point for each possible coincidence of a units pulse and a tens pulse. Any event can be triggered at any one of those 100 points and the choice is made by turning two selector switches.

4.5 TRAPEZOIDAL-FUNCTION GENERATOR

The basic control waveform in the synthesizer is the trapezoid, which is used for controlling the buzz and noise excitation, as well as the changes in the configuration of the analog. The main element of the trapezoidal function generator is a capacitor, C, so that control of the generator implies control of the current flow into C. Figure 35 illustrates the function-generating scheme. Its operation can best be understood by tracing the sequence of events occurring during the production of a trapezoid. At the start, the switch S is down, the capacitor is uncharged, V_C equals zero, and the circuit is in its initial quiescent state. The capacitor is connected to a negative voltage, E_N,

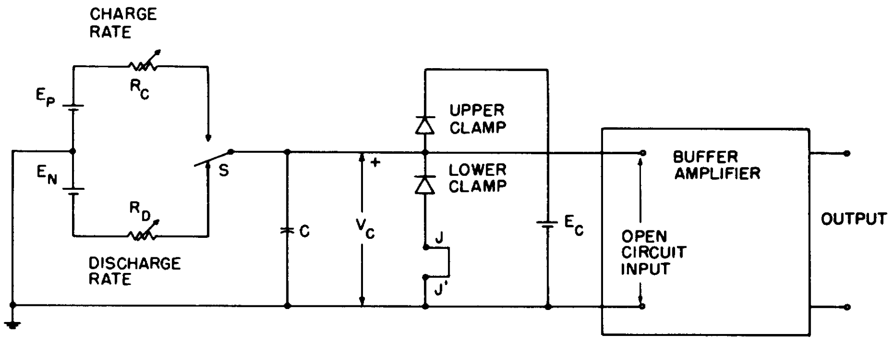


Fig. 35. Simplified circuit of the trapezoidal-function generator. The complete circuit (given in Fig. 64) uses tubes to perform the functions ascribed to switch, S.

through R_D, but it cannot charge negatively because of the lower clamp at zero. When the switch is thrown to its upper position, the capacitor is connected to a positive voltage, E_P, through the charging resistor, R_C, and C begins to charge, and generates an upgoing ramp. In the actual generator, switching is initiated by receipt of a rise-trigger pulse. Charging continues until V_C equals E_C, the upper clamping voltage. After that, the clamp operates to maintain V_C at E_C and the upper quiescent state exists. The circuit remains in this state until the switch is thrown to the lower position and the capacitor discharges through the discharge resistor, R_D, and generates a downgoing ramp. When V_C equals zero, the lower clamp operates to prevent negative charging, the capacitor voltage is held at its lower quiescent value and the circuit is in its final state. The buffer amplifier delivers the capacitor voltage waveform to the output load but does not itself load the capacitor.

If we let

$$E_P = kE_C \qquad (27a)$$

and

$$E_N = (1-k)\,E_C \qquad (27b)$$

then the voltage across R_C at the start of charging is the same as the voltage across R_D at the start of discharging, and both are equal to kE_C. Thus equal values of R_C and R_D lead to equal values of upward and downward ramp durations. The value of E_C is 50 volts, and consequently is the maximum amplitude of the output. Resistors R_C and R_D are each constructed of three decade resistors to make ramp durations variable from 1 msec to 1 sec in 1-msec steps. Charge and discharge follow exponential functions, but the initial portions of those functions are approximately linear. The approximation to linearity improves as that portion of the full curve lying between the clamping limits becomes smaller, that is, as k becomes larger. However, large values of k are impractical because they imply large values of E_P and E_N, but when k is set equal to 4, E_P is 200 volts and E_N is -150 volts and the ramps are within a few per cent of linearity.

In the completed generator (see appendix), there are no mechanical parts, the switch being replaced by a flip-flop and gating circuits. The upward ramp is initiated by receipt of a rise trigger from the time-pulse source and the downward ramp is initiated by a fall trigger. Five parameters are associated with each trapezoid: two trigger times set by selector switches on the pulse source, two ramp durations set by decade resistors on the function generator, and the amplitude that is varied by changing the gain of the buffer amplifier.

The basic circuit can be modified or used in a number of ways to produce a large variety of piecewise-linear time functions. The flip-flop can be switched many times during a counting cycle to make the generator produce a succession of trapezoids, so that the synthesis of a syllable, such as /sæʃ/, which requires two distinct intervals of noise excitation, is readily accomplished. Polysyllabic utterances are similarly handled. The nonlinear, time-variant nature of the circuit can be exploited to render unobtrusive certain switching operations. For instance, a function with three quiescent levels can be generated with the aid of auxiliary switching inserted between terminals JJ' in Fig. 35. Initially, the switching short circuits J and J' and the first 2 quiescent levels, together with the intervening upward ramp, are produced in the usual manner. An emf is then switched into the JJ' terminals during the upper quiescent interval to change the lower clamping level. Subsequently, the downward ramp is initiated, but it will terminate at the new clamping level. Auxiliary switching can also change the values of R_C and R_D. (Consider, for example, the upper quiescent interval: R_D is effectively disconnected from C so that its resistance may be changed without immediately affecting the output. The value of R_C may likewise be changed because $E_P > E_C$ and any value of R_C will, theoretically, keep the upper clamp operating.) To be unnoticed, switching of resistors must occur when no transition is in progress, otherwise a transition that has several linear segments will be generated. Thus one can conceive of a control system that uses no additional function generators but enables the synthesizer to produce connected speech of unlimited duration. Additional function generators, however, provide additional flexibility and operating convenience. Outputs of several generators can be summed to produce time functions having many segments and quiescent levels, or to produce a sequence of trapezoids. The parameters of each trapezoid are determined by its associated generator. The second mode of operation has been used to produce stimuli for a psychophysical study — a study of the perception of juxtaposed noise bursts (49).

4.6 THE CONTROL SYSTEM AND ITS OPERATION

The connections between the timing-pulse source, the function generators, and the auxiliary switching will now be described. The coincidence gate panel seen on the left in Fig. 36 has already been shown in Fig. 34. (The inverter tubes shown in Fig. 36 are circuit details rather than functional entities, and so they were omitted in Fig. 34.) Each output channel of the coincidence panel provides one pulse per cycle to the triggered units shown on the right. Each function generator and auxiliary switching cell uses a flip-flop

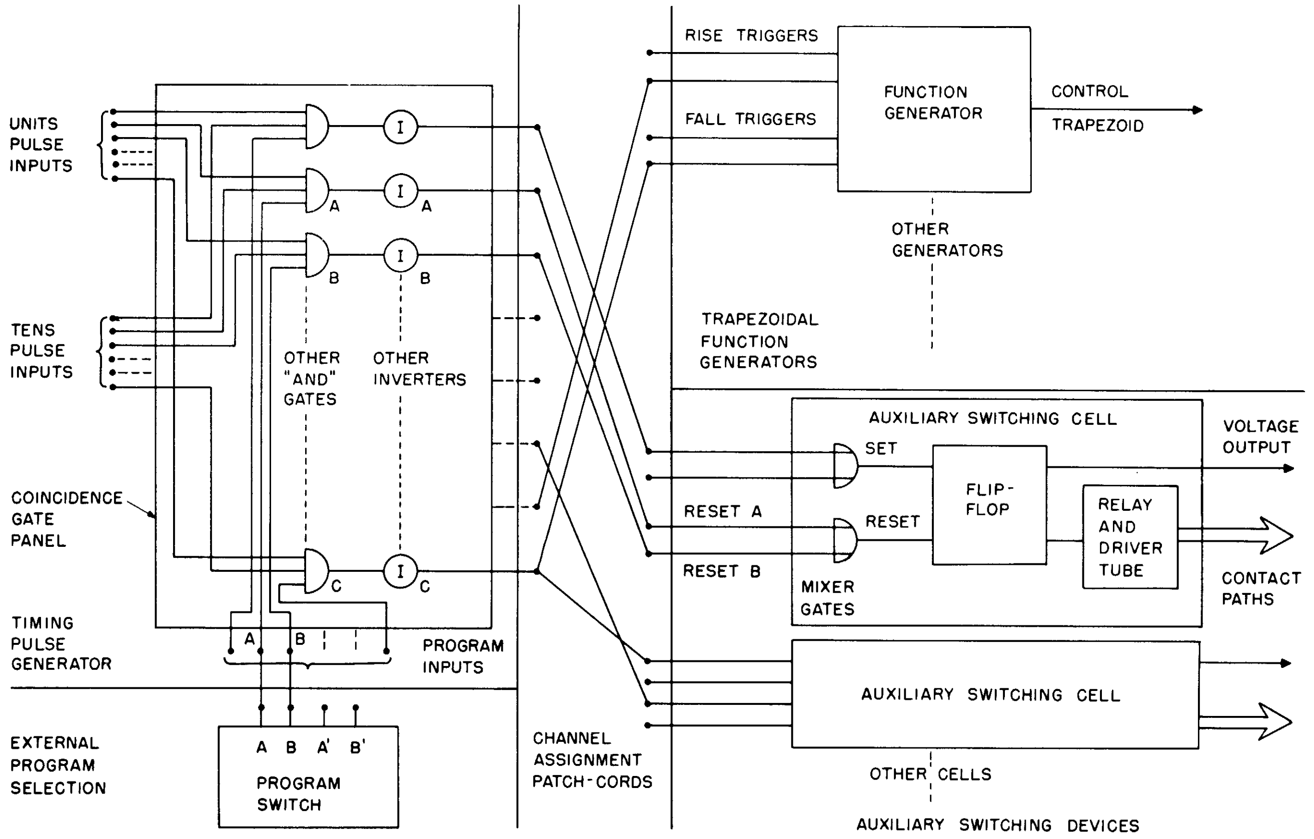


Fig. 36. Functional diagram showing relationship between timing-pulse source, function generators, and other units concerned with timing triggers. Pulse inputs at the left are connected to counter tubes through time-selection matrix.

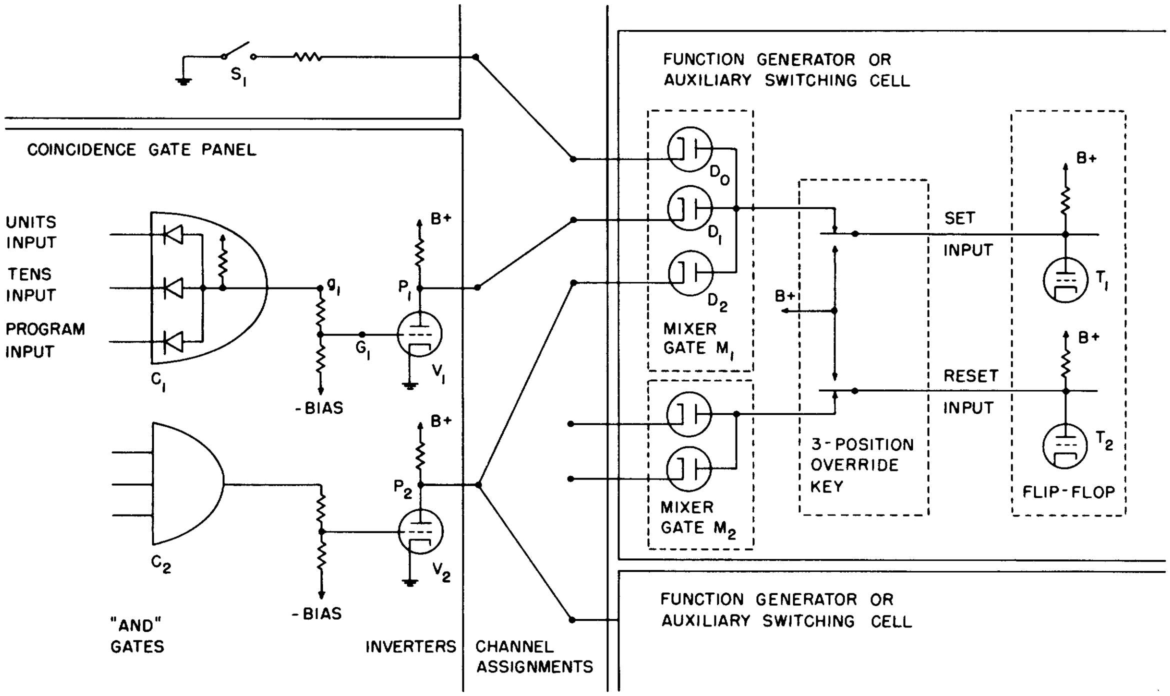


Fig. 37. Circuit components encountered in typical triggering paths.

input stage so that both trigger in the same way. Let us consider the process by which a coincidence of pulses from the counter tubes initiates a ramp or switches an auxiliary cell.

Many triggering paths are shown in Fig. 36 and one of them may be chosen as typical. Along such a path, the signal will travel from a coincidence gate, through an inverter, a patch cord, and a mixer gate, and then into a flip-flop. The circuitry typically encountered is shown in Fig. 37. Let us assume that the flip-flop is initially in its "zero" state (T_2 conducting and T_1 cutoff), that V_1, V_2, D_0, D_1 and D_2 are nonconducting, and that the program input to gate C_1 is not connected. Then let positive pulses be applied simultaneously to the tens and units input of gate C_1. (The coincidence gate uses three crystal diodes and works with pulses having a "high" value of approximately +25 volts and a "low" value of +5 volts.) This will cause a positive pulse to appear at g_1, the output of the gate, and will cause the voltage at grid G_1 of the inverter V_1 to rise to zero from a value more negative than the cutoff value for V_1. Thus, V_1 and D_1 will conduct heavily to apply a negative pulse to the "set" input of the flip-flop. The flip-flop will switch to its "one" state with T_1 conducting and T_2 cut off. Subsequent pulses applied to the set input will have no effect unless the flip-flop is first reset. Note that the flip-flop could have been set by an output of gate C_2, which would imply conduction through V_2 and D_2, or by a closure of switch S, which would imply conduction through D_0. This gating arrangement allows external inputs from relays or switches to be compatible with pulse inputs. One inverter, such as V_2, can trigger several flip-flops; the required connection at the plate of V_2 is shown. There is an option of using a direct connection and eliminating the mixer diodes by paralleling several inverter plates and connecting them to the flip-flop input, but in that case each inverter must be associated with only one flip-flop. By symmetry, triggering at the reset input is similarly effected. Multiple inputs to the function generators are needed for utterances having several distinct intervals of buzz and noise excitation or more than two articulatory transitions.

A three-position manual override key is associated with each flip-flop. The flip-flop is held in its set or reset state when the key is in its upper or lower position. The center position permits automatic operation under control of the triggering devices in the manner described. By use of such keys, the machine may be put in any desired state without disturbing the main control settings. The manual intervention capability is useful for informal listening, and for tracing malfunctioning. Other circuits also have override keys for the same reasons.

The auxiliary switching flip-flops can perform their functions either directly or through their associated relays. In typical situations, their purpose is to:

(a) Alter the characteristics of the excitation circuits.

(b) Change parameters of the trapezoidal function generators.

(c) Sequence-switch matrix S for multiple articulatory transitions.

(d) Change the point at which noise is inserted into the transmission line.

(e) Assist in program selection.

(f) Alter characteristics of appurtenances used to generate stimuli for psychological and psychophysical experiments.

The operation of the program gating will be illustrated by taking a simple example and by referring to Fig. 36. The figure shows an auxiliary switching cell with 3 pulse channels assigned to its inputs. The relay is to be operated at the same time in every cycle but is to be released at two different times, A or B, the choice depending on the external program. The switching to implement this choice is connected to the program inputs A and B so that coincidence gates A and B can be selectively inhibited, but other coincidence gates with program inputs left free will behave as two-input gates. In one state of the selection switch, a pulse will occur in channel A at a time determined by its associated tens and units switches; and in the other state, a pulse will occur in channel B. The channel assignment connections are chosen to connect channels A and B to separate inputs of the mixer gate that is used to reset the flip-flop in the cell. During each cycle, a pulse appears at one or the other of the mixer inputs and the choice between the two stored reset times is determined by the external program selection. Selection outputs A' and B' are switched in unison with A and B and could be connected to program other channels. The external programming devices may operate with switching paths or voltage levels — the gates will accept either.

Thus several similar syllables may be preset and the program input used to enable the gates needed for the desired syllable. Rapid A-B comparisons can be made during informal listening sessions, and formal listening tests can be recorded more easily.

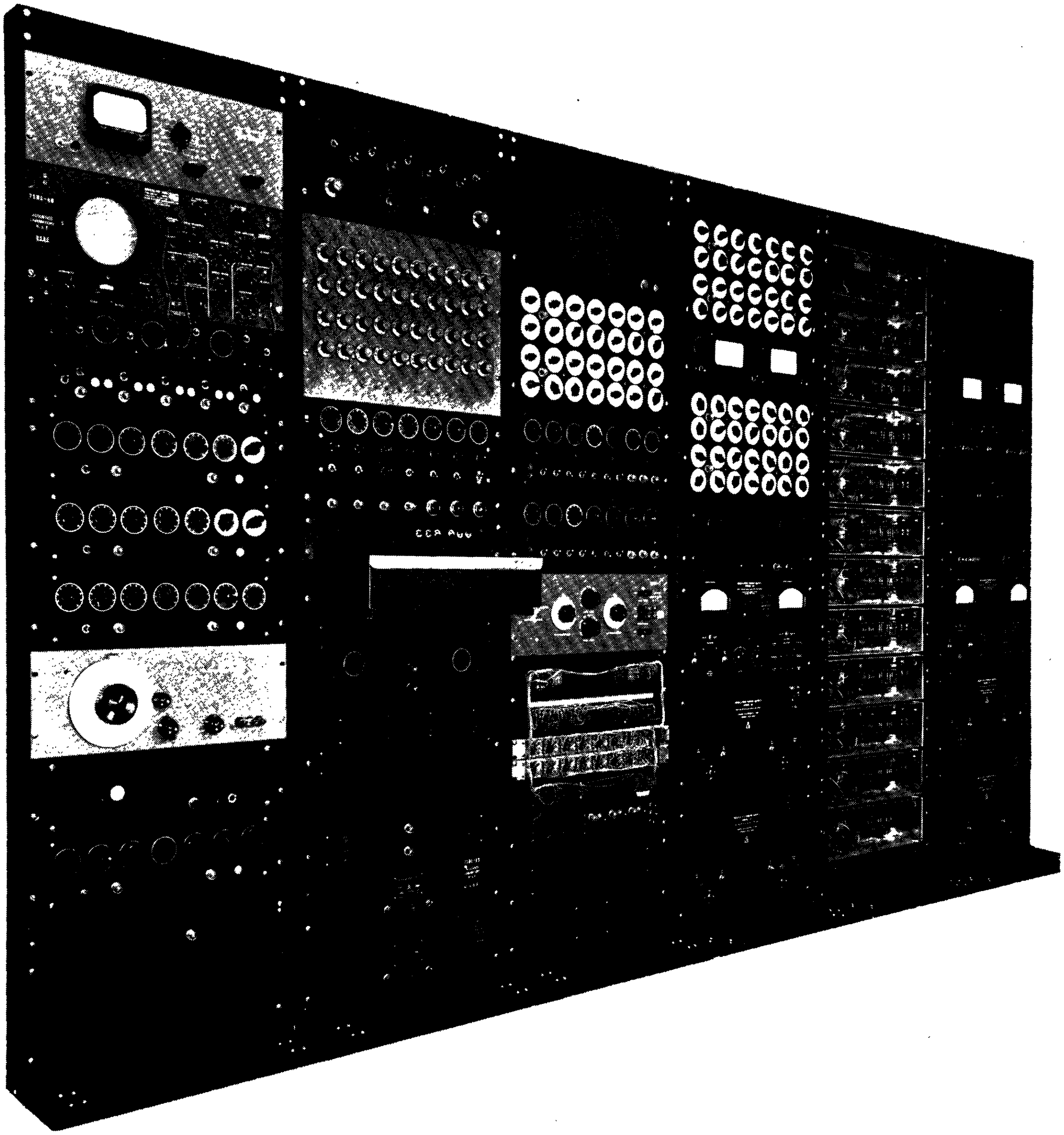


Fig. 38. Photograph of the synthesizer. Principal units, from left to right: rack 1, function generators; rack 2, timer with time-selection matrix; rack 3, buzz and noise generators and part of configuration matrix; rack 4, configuration matrix; rack 5, transmission line; rack 6, power supplies for transmission line.

Before this facility was available, informal listening was hindered because an appreciable delay occurred after one stimulus while the second was being set. The listener's memory span was exceeded, and the desired comparisons could not be made.

Research with the dynamic analog involves the judgment of large numbers of synthetic stimuli. In order to facilitate the recording of both formal and informal listening tests, remote switching for the tape transport mechanism was incorporated in the synthesizer control system. The circuits perform the following functions, in the order given: starting the recorder, allowing the tape to come up to speed, recording and identifying marker, initiating the stimulus that is to be recorded, providing for preset spacings between utterances, and stopping the recorder. Uniform spacing between stimuli is easily maintained. Accurate spacing within stimulus pairs or triplets is especially important in paired comparison or ABX tests. In the absence of proper recorder synchronization, a tape can be made by splicing all spaces to their required lengths. In practice, there is a tendency to avoid such tedious, time-consuming tasks at the expense of required test recordings. Tape-motion controls, therefore, make the synthesizer a more useful research instrument.

The control system has been connected to ancillary circuits by the author's colleagues and used for studies of inflection patterns (41), of learning in multidimensional auditory displays (3), of the detectability of irregularities in noise spectra (42), of the effect of duration on vowel identification (52), and for other investigations. This device, shown in Fig. 38, gives the investigator precise control over each stimulus that he generates, and allows him to specify the stimulus in great detail. The synthesizer and its appurtenances function as a facility for recording a great many stimuli rapidly and accurately.

V. STUDY OF VOWELS

The experimental program for the completed synthesizer can be planned with one or more of these goals:

(a) use of the synthesizer for research in speech production and perception;

(b) acquisition of data that are needed to couple the synthesizer to other elements of communication or computer systems; and

(c) building a vocabulary for the machine.

These objectives call for the presentation of the output of the machine to human listeners and for evaluation of the output in terms of their responses.

In a formal listening program, the human subjects play two roles. First, they can help us to evaluate the quality and intelligibility of the output of the synthesizer. Second, their linguistic behavior can be used to determine the articulatory and acoustic representation of the cues that function in the identification of the various classes of speech sounds. For formal listening, synthetic stimuli are recorded in a randomized sequence and presented to a panel of listeners under standardized conditions. The usefulness of the resulting data depends on the choice of stimuli, the instructions given to the listener, and the linguistic background of the listener.

The first sounds studied formally were the vowels. Satisfactory rendition of the vowels is a necessary condition for satisfactory production of connected speech. Vowels are important because they have a high probability of occurrence and because they provide the connective tissue in which the consonants are embedded. A syllable consisting of a vowel and a consonant may have poor naturalness or intelligibility because of the presence of a poor vowel or a poor consonant, or both. Many experiments with consonants will use stimuli in the form of CV or VC syllables. Therefore it seemed advisable to give attention, first, to the vowels, to reduce the likelihood that poor vowels would becloud data for consonants. Other factors favoring this decision were that vowels can be studied in isolation, and that they are relatively well understood and easily synthesized. Furthermore, if the vowels resulting from these studies are highly natural and intelligible, the correctness of certain assumptions is demonstrated. These assumptions relate to the number of elements needed in the transmission line and the permissible dissipation in those elements.

The resulting effort was concerned mainly with 9 vocal-tract configurations that are appropriate for the production of the following vowels of American English: /i I ε æ ɑ ɔ U u ʌ /. (Many of the symbols used in this report can be taken in either a phonetic or phonemic sense but when a distinction must be made, the meaning is clear from the context. Therefore, all phonetic and phonemic symbols, hereafter, will be written between slants.) This set typifies the American English vowel articulations, and provides the materials for approximating rarely used vowels and for fabricating sounds that are not strictly vowels. The section area settings were chosen to represent the various portions of the vowel "triangle," and they test the analog with a variety of articulatory configurations.

The first set of vowel configurations used on the dynamic analog was based on a previous study made on a static analog by House and Stevens (32). They used three parameters to characterize vowel configurations that were expressed graphically by giving the relation between the radius of each circular cross section of the vocal tract and the distance of that cross section from the glottis. Each graph has three parts: a throat section, a parabolic representation of the tongue, and a cylindrical representation of the lips. The three parameters are: r_o, the minimum radius of the tongue constriction; d_o, the distance of the constriction (point of minimum radius) from the glottis: and A/ℓ, the ratio of area of the lip cylinder to its length (this ratio is inversely proportional to the acoustic mass of the lip cylinder). A formal listening test served to establish a correspondence between 279 combinations of parameter values and 9 vowel categories plus a null category.

The configurations set of the static analog cannot be used directly on the dynamic analog because of certain differences between the two machines. First, the detailed shape of the throat defined in the parabolic model is ignored by the dynamic analog, whose throat consists of two fixed sections near the glottis plus two variable sections constrained to use a common area-control signal. The gross specifications, however, are similar in both cases. Second, the ranges of section areas for the two machines do not completely coincide. Third, sections of the static analog are removed by switching, which shortens the lip cylinder to help provide an A/ℓ range of 170:1. Switching is not permitted in the dynamic analog, since continuous control is necessary for generation of connected speech. Separate control of the inductors and capacitors in the lip sections is employed in lieu of switching, but the realizable range in the A/ℓ parameter is somewhat smaller than the desired range.

Starting with the House and Stevens data, we found a new set of configurations empirically and evaluated it by informal listening and by measuring formant frequencies. The curves given by Stevens and House (53) were used to estimate the effect of changing each parameter of the model. In each case, iteration was terminated either when the output was informally judged to be satisfactory or when formant frequencies could no longer be changed in the desired manner. The new set of vowel configurations was then ready for proof by formal listening. The configurations finally chosen are shown in Fig. 39, and the corresponding set of formant frequencies is given in Table I. The vocal-tract transmission curve and formant bandwidths for /u/, shown in Fig. 6, are typical of all of the vowels in the final set.

There were two kinds of variables in the design of the formal listening test: configuration and duration. Each configuration was used with two over-all durations, 180 msec and 330 msec. The buzz rise and decay times were 40 msec and buzz frequency changed linearly from 100 cps to 140 cps in each stimulus, regardless of its duration. Two tests, one for back vowels and one for front vowels, were recorded. Each test was designed to require no more than 6 configurations to obviate the need for changing the settings of the six-row matrix used for configuration storage. Subsequently, it was decided to

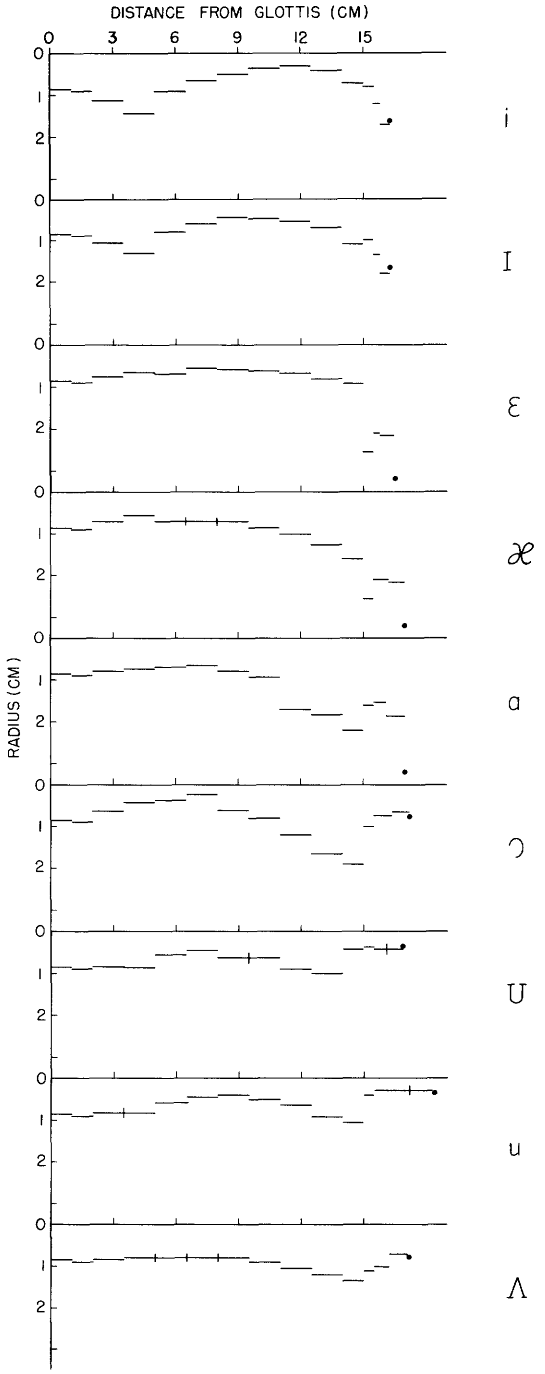


Fig. 39. Configurations set on the analog for a study of 9 synthetic vowels. The variable-length feature of the lip sections is used. Short vertical markers show boundaries between sections set to the same area; dot symbols at the right denote settings of radiation impedance.

Table I. Frequencies of the first three formants measured on analog set at configurations shown in Fig. 39.

Configuration	F_1	F_2	F_3
i	270	2250	2950
I	350	2020	2900
ɛ	555	1730	2250
æ	700	1550	2250
ɑ	725	1230	2650
ɔ	510	850	2650
U	400	1100	2750
u	270	840	2300
ʌ	560	1300	2420

Table II. Confusion matrix for vowels.

Stimuli		Responses								
Configuration	Duration	i	I	ɛ	æ	ɑ	ɔ	U	u	ʌ
i	180	27								
i	330	27								
I	180		27							
I	330	1	26							
ɛ	180			27						
ɛ	330			25	2					
æ	180			2	50	2				
æ	330				54					
ɑ	180					12				15
ɑ	330					27				
ɔ	180						26			1
ɔ	330						27			
U	180							25	1	1
U	330							23	4	
u	180								27	
u	330								27	
ʌ	180							1	1	52
ʌ	330					1		6	1	46

replace the smaller tests by one large test and the one long tape was prepared by randomized interleaving of the two short tapes. Both /æ/ and /ʌ/ appeared on each of the earlier tests, and so both appear with double frequency on the final tape. All other vowels are represented by three occurrences each of the short version and the long version. The gain of the buzz-generator output amplifier is kept constant so that stimulus intensities correspond roughly to constant vocal effort (if we assume that constant vocal effort corresponds to constant volume velocity at the glottis). The time between stimuli was 5 seconds. The subjects were given oral instructions, and those who were not familiar with the vowel symbols noted key words on their answer sheets before the test started. Five subjects were experienced; five were naive.

The correspondence between the stimuli (configurations) and the responses (phonetic judgments) is given by the confusion matrix (see Table II). This table pertains to a panel of 9 listeners and implies an articulation score of 93 per cent for the synthetic vowels. This value is quite close to the 94 per cent score obtained by Peterson and Barney (46) for natural vowels, but several distinctions must be considered when comparing these figures. Each natural vowel was embedded in a syllable beginning with /h/ and ending in /d/, while each synthetic vowel was produced in isolation with a time-invariant configuration and time-invariant formants. (Ten categories were used with equal frequency in the study of the natural vowels. The category used in that study but not used in the study of synthetic vowels is /ɝ/.) Each phoneme for the natural vowels exhibits considerable scatter in formant frequencies because the stimuli were uttered by 76 speakers, including men, women, and children. In contrast, only one configuration of the synthesizer, with a corresponding set of formants, was chosen as a candidate for each phoneme. Both experiments were carried out with available subjects who had had little screening for linguistic homogeneity.

The vowels /i/ and /u/, which are the extremities of the vowel triangle, received the highest articulation scores in both the Peterson and Barney experiment and in this experiment. In our experiment, the judgment of these vowels was unanimously correct in both long and short versions. These results, together with a /ʌ/ response to a short /ɑ/ stimulus, are consistent with Stevens' findings on the effect of duration on vowel identification (52). In Table II, a long /U/ is confused with a /u/, perhaps, because the formants for /U/, at 400 cps and 1100 cps are not unambiguous (cf. the area of the F_1F_2 plane assigned to /u/ by the Peterson and Barney data (46)). The /U/ responses to the long /ʌ/ stimulus may have a similar explanation because the first formant of the /ʌ/ stimulus is quite low, and the formants are brought close to the /U/ region. The data indicate a need for enlarging the set of formant frequencies beyond the set attainable on the present machine with plausible configurations. In the articulatory domain, this implies greater range in the lip sections and more flexible control of the throat sections.

On the whole, the synthesizer has demonstrated its capability of generating a set of vowels whose intelligibility is comparable to that of natural vowels. For this class of sounds, the assumptions made in the design of the analog are valid.

Natural vowels can be ordered along dimensions relating to their production by human speakers. Some such dimensions are place of production, size of constriction, and "rounding." On the basis of the formal tests, the vowels produced by the analog can be similarly ordered.

In generating the 18 items for the listening tests, each configuration was used with each of 2 durations. By examination of Table II, we can choose 9 items to meet the condition that duration cues reinforce configuration cues, and thus each configuration will be associated with only one duration. Instead of performing another listening test, let us estimate a 9 × 9 confusion matrix for this set by transferring half of the entries from Table II to the new matrix. The predicted articulation score obtained in this manner is approximately 98 per cent.

VI. STUDY OF FRICATIVES

6.1 INTRODUCTION

For the next series of experiments, we chose the study of fricative-vowel syllables. Synthesis of such sounds makes use of the machine's ability to make transitions, and typifies the work for which the control system was designed. This task seemed suitable for an initial problem because it is more challenging than the synthesis of dipthongs, but not quite as difficult as the production of stops. Procedures developed in these experiments can serve as a model for future experiments.

The first step, before formal listening begins, is the study of fricatives in isolation. By static operation of the analog, certain parameters can be studied independently. Temporal variables are introduced when we study fricatives embedded in syllables. The formal listening program that will be discussed here started with a study of configuration variables and was followed by a study of temporal variables.

6.2 STUDY OF CONFIGURATION VARIABLES

The vocal-tract configuration for fricative production can be idealized as a uniform tube, tapered near the glottis, having a single constriction, and excited by a single white-noise source. With this model, it was possible to investigate the manner in which fricative identification depends on position of constriction, degree of constriction, and placement of noise insertion relative to the constriction.

The details of the model are strongly influenced by the construction of the transmission line. The line is composed of sections of different length and configuration, so that it is not possible to realize the analog of a plug of fixed size and shape placed in various positions along the tube. Therefore, the term "constant degree of constriction" must be defined, preferably in a manner that is physically or linguistically meaningful. A physical definition based on acoustic mass was adopted. An acoustic mass is the single lumped element that best represents a constriction, and hence the magnitude of this element was held constant as the constriction was moved along the axis of the tube. The acoustic mass of a cylindrical constriction varies as the ratio of length to cross-section area, and is represented electrically as series inductance in the transmission line. The uniform tube must have a tapered throat because the line uses fixed glottal sections, and because the two variable sections nearest the glottis are controlled by a common control voltage. The area of the main portion of the tube was 4.5 cm^2 for all of the configurations used in this study.

The first variable, position of constriction, was given 9 different values. This variable serves to identify the sections (and one half-section) at which a constriction may be formed. (This half-section serves as a transition between the two Π-sections at the front of the tract and the T-sections used in the rest of the tract. A constriction formed here is said to be in position 3.) The values establish an ordering in which

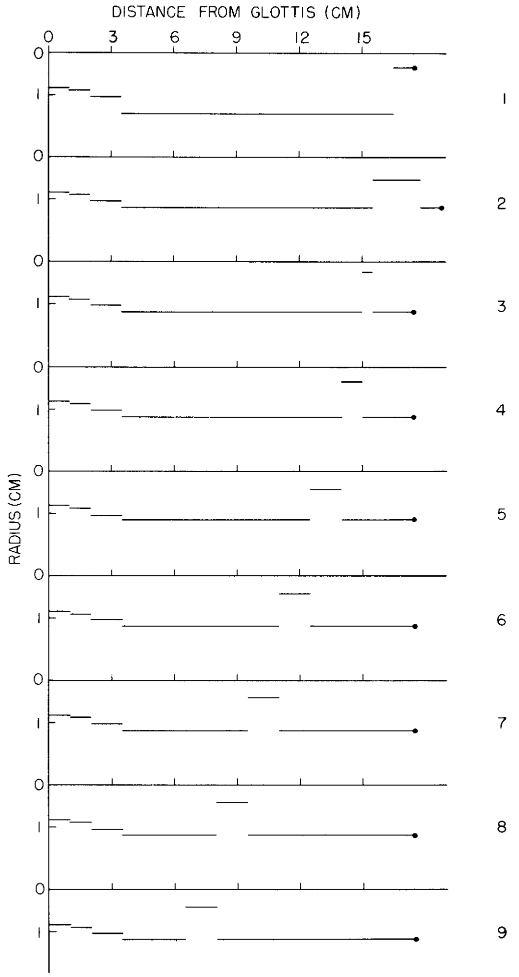


Fig. 40. Configurations derived from an idealized model of fricative production with open constrictions. Numbers at the right refer to the position of constriction.

position "one" is farthest forward and in which the numbers increase as the constriction is moved backward. Sections differ in length so that position numbers are not proportional to distance. The number of positions to be investigated was determined partly by informal listening and partly by physiological considerations. Constrictions beyond the 9^{th} position were found to be unrealistic, and hence of no interest. The fricatives occurring in English are produced near the front of the mouth by using the lips, teeth, and the front of the tongue. Postpalatal and velar fricatives, produced with the middle and back parts of the tongue, occur in many languages (German and Russian included). The velar fricatives were included partly for completeness in the study of the position variable, and partly as an illustration of the capabilities of the analog and its potential use in producing sounds in many different languages.

The constriction was made in two degrees, open and close. The corresponding acoustic masses can be represented by cylindrical constrictions, 1 cm in length. The

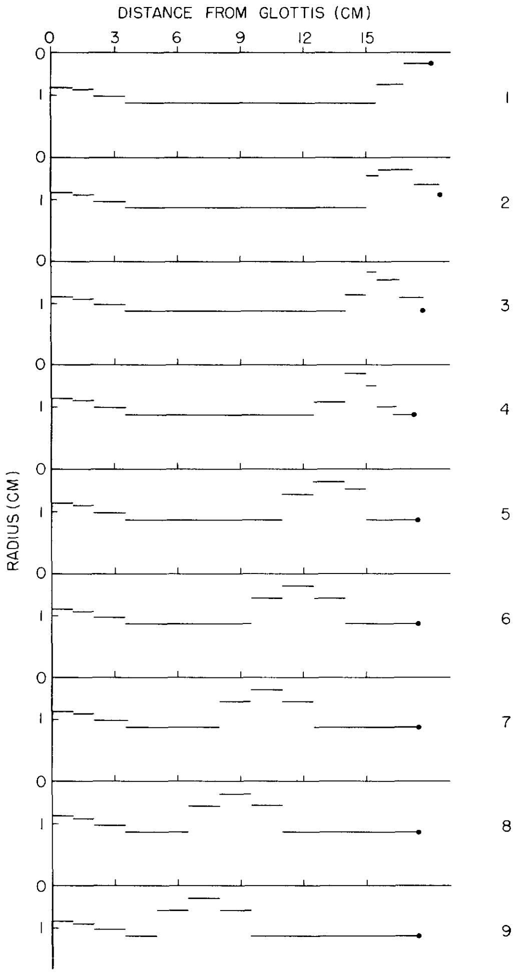


Fig. 41. Configurations used for .studying fricatives produced with close constrictions.

areas of the constrictions in these two cases are 0.4 cm^2 and 0.13 cm^2, respectively. The open constriction was set at the greatest degree that could, in all cases, be represented as a single plug realized with only one section, or, equivalently, by a single inductor set at its greatest value. Because inductance is proportional to section length, the maximum inductance of the half-section is less than the maximum inductance of any other section. Thus the inductance of any other section must be set at the maximum value for the short section. All of the close constrictions were made by reducing the area of the section at the constriction to its minimum value and then adding whatever additional acoustic mass was required by reducing the areas of adjacent sections. Figures 40 and 41 show the resulting configurations for the open and close cases.

The third variable in this experiment is the location at which noise is to be inserted. A region of turbulence that produces noise in the vocal tract is treated as a pressure source, which is represented electrically as a voltage source. The voltage is inserted in series with the inductor of the section at the given location by the secondary of a

Table III. Combinations of variables used for making listening tests.

Location of				Position of Constriction					
Noise	9	8	7	6	5	4	3	2	1
1					x	x	x	x	x
2					x	x	x	x	x
3				x	x	x	x	x	x
4				x	x	x	x	x	x
5			x	x	x	x	x	x	
6		x	x	x	x	x	x		
7	x	x	x	x	x	x			
8	x	x	x	x	x				
9	x	x	x	x					
10	x	x							
11	x								

transformer. The transformer primary is connected to a white-noise generator having low output impedance. By informal listening, all physically possible combinations of noise location and place of constriction were examined; the set specified in Table III was chosen for formal testing.

The stimuli for the listening test were recorded in randomized order so that each item in Table III appeared three times, giving a total of 156 judgments per listener per test. All of the stimuli were consonant-vowel syllables, always containing the vowel /ɑ/. The relative levels of buzz and noise excitation and the temporal relations between buzz amplitude, buzz frequency, noise amplitude, and configuration change were determined by informal listening. [The levels were set so that the full scale of the buzz-amplitude trapezoid corresponds to 25 volts peak-to-peak at the output of the buzz generator. When buzz frequency is set at 100 cps, the present shaping results in a reading of 3.5 volts on an average-reading meter calibrated in rms of sine wave. The same meter reads 0.007 volt at the secondary of the noise-insertion transformer. There is a 470 K resistor between the buzz generator and the input to the transmission line.]

The temporal relations shown in Fig. 42 were used for all stimuli. Successive stimuli were spaced 6 seconds apart. Marker tones were recorded after every 10th item, and the end of each column of 25 items on the answer sheet was marked by a 15-second rest. The stimuli were presented through earphones to a panel of 8 subjects, who were instructed to make the responses /s, ʃ, f, θ, ç, x /. Two tests were given on different days, for each degree of constriction; one test was forced choice, the other was not.

The stimuli included many sounds resembling the velar fricatives /ç/and /x/, which do not occur in English and were not in the phonetic vocabulary of most of the subjects. All subjects were given training and examination in these sounds. A table of random numbers was used to prepare a list of 20 items, each of which was either the sequence

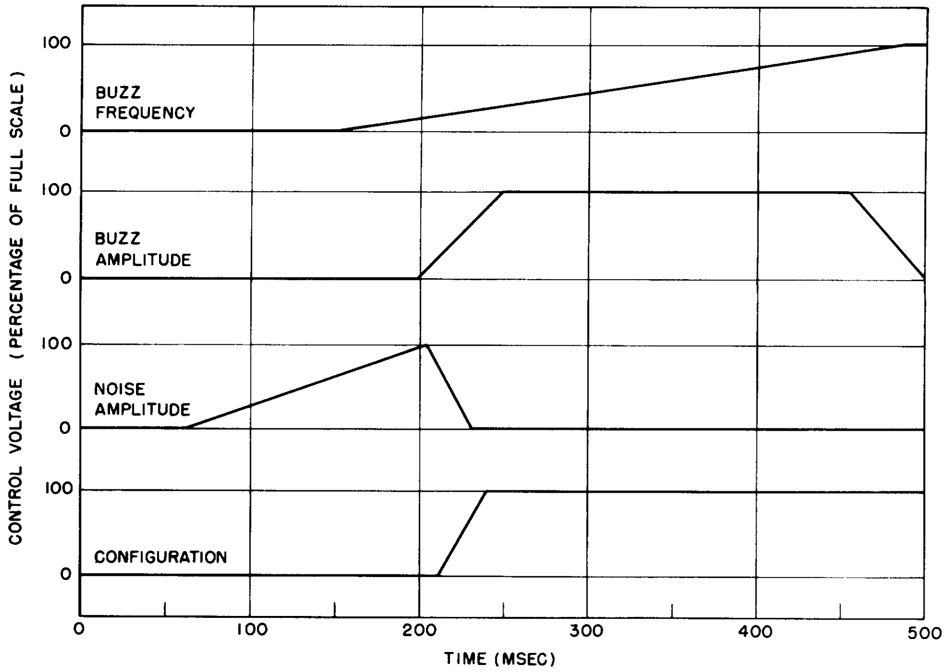
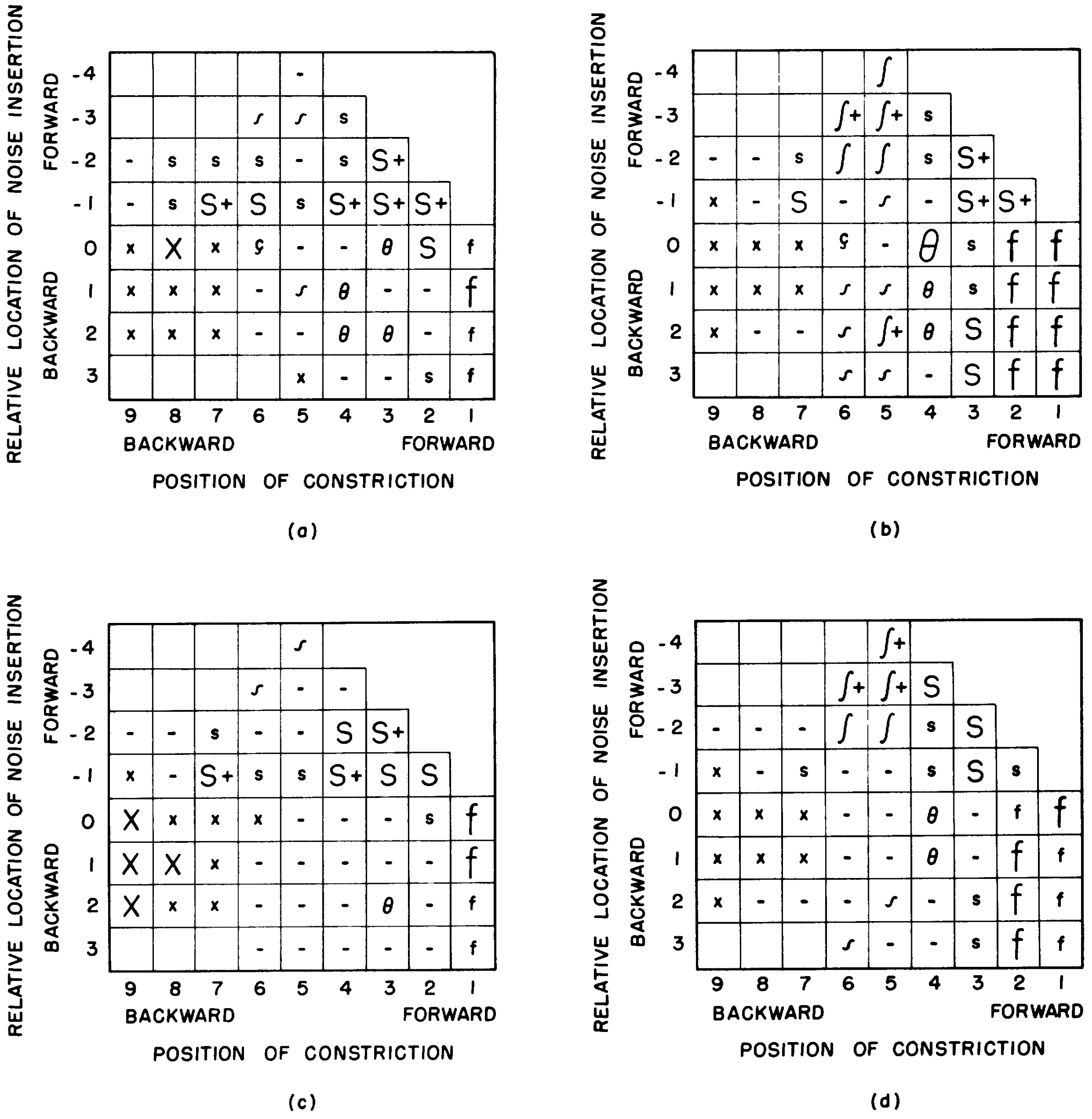


Fig. 42. Timing patterns describing all stimuli generated for the study of configuration variables. Each of the syllables used begins with a fricative and ends with the vowel /ɑ/.

/ɪç/, /ɑx/ or that sequence reversed. The list was recorded by a fluent speaker of German and played back to the subjects, who looked at the list as they listened to the items that were being pronounced. They were then given a preliminary test, in which they were asked to identify 20 randomized recorded occurrences of the two velar fricatives spoken in isolation by the same speaker. Then the main test with synthetic stimuli was administered. This was followed by a supplementary test of the same length as that of the preliminary test. The over-all preliminary-test and supplementary scores for the series of four main tests were 100 per cent for each of seven subjects and 92 per cent for the eighth.

The listeners' responses are listed in Tables IV-IX, one table being provided for each phonetic category. Each table has four parts, one for each listening test. In all of the tables we made use of a parameter giving the location of noise insertion relative to the position of constriction. That parameter is equal to the absolute noise location number minus the constriction position number. The data are summarized in the response maps of Fig. 43, showing responses at the 50 per cent, 75 per cent, and 90 per cent levels. It is seen that the various fricatives occur in the expected order as the constriction moves from front to back. For example, most /f/ responses were obtained with a frontal constriction, whereas /x/ responses were obtained with constrictions at, or posterior to, velar positions. A good /ʃ/ requires a close constriction, whereas a good /x/ requires an open constriction. A highly intelligible /ç/ has



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

(a)

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

(b)

	9	8	7	6	5	4	3	2	1
-4					ʃ				
-3				ʃ	-	-			
-2	-	-	s	-	-	S	S+		
-1	x	-	S+	s	s	S+	S	S	
0	X	x	x	x	-	-	-	s	f
1	X	X	x	-	-	-	-	-	f
2	X	x	x	-	-	-	θ	-	f
3				-	-	-	-	-	f

(c)

	9	8	7	6	5	4	3	2	1
-4					ʃ+				
-3				ʃ+	ʃ+	S			
-2	-	-	-	ʃ	ʃ	s	S		
-1	x	-	s	-	-	s	S	s	
0	x	x	x	-	-	θ	-	f	f
1	x	x	x	-	-	θ	-	f	f
2	x	-	-	-	ʃ	-	s	f	f
3				ʃ	-	-	s	f	f

(d)

Fig. 43. Response maps for fricatives: (a) open constriction, nonforced instructions; (b) close constriction, nonforced instructions; (c) open constriction, forced instructions; (d) close constriction, forced instructions. In all maps: unruled area in upper right corner indicates physically impossible combinations. Symbol conventions: blank, unused combination of no interest; dash, response level below 50 per cent; small phonetic symbol, response level 50-75 per cent; large phonetic symbol, response level above 75 per cent; plus sign, response level for indicated category above 90 per cent.

Table IV. Number of /f/ responses. The four parts of this and the next five tables give data for the following conditions: (a) open constriction, forced instructions; (b) close constriction, forced instructions; (c) open constriction, non-forced instructions; (d) close constriction, non-forced instructions. (A dash indicates one response or no response and the largest number possible is 24.)

	Relative Location				Position of Constriction					
	of Noise	9	8	7	6	5	4	3	2	1
(a)	-4					-				
	-3				-	-	-			
	-2	6	2	-	-	-	-	-		
	-1	-	-	-	-	-	-	-	-	
	0	3	-	-	-	-	2	5	-	17
	1	4	-	-	-	-	-	-	7	19
	2	3	2	2	-	2	-	-	6	17
	3				-	-	4	-	4	15
(b)	-4					-				
	-3				-	-	-			
	-2	3	-	-	-	-	-	-		
	-1	-	-	-	-	-	-	-	-	
	0	3	2	-	-	-	-	-	21	21
	1	3	-	-	-	-	-	-	21	21
	2	-	-	-	-	-	-	-	21	21
	3				-	-	-	-	21	21
(c)	-4					-				
	-3				-	-	-			
	-2	5	-	-	-	-	-	-		
	-1	4	-	-	-	-	-	-	-	
	0	3	3	-	-	-	2	4	-	19
	1	3	2	2	-	-	-	3	10	19
	2	3	3	-	-	-	-	3	7	16
	3				-	-	-	-	5	16
(d)	-4					-				
	-3				-	-	-			
	-2	5	-	-	-	-	-	-		
	-1	4	-	-	-	-	-	-	-	
	0	4	3	2	-	-	-	-	18	23
	1	2	3	-	-	-	-	-	20	17
	2	3	-	-	-	-	-	-	19	18
	3				-	-	-	-	20	18

Table V. Number of /s/ responses.

Relative Location of Noise		Position of Constriction								
		9	8	7	6	5	4	3	2	1
(a)	-4					7				
	-3				2	8	12			
	-2	5	15	15	16	4	17	23		
	-1	6	13	23	19	13	24	23	23	
	0	-	-	-	-	-	2	-	20	-
	1	-	-	-	3	2	6	4	6	-
	2	-	-	3	-	-	2	5	6	-
	3				-	-	3	6	14	-
(b)	-4					-				
	-3				-	-	16			
	-2	-	8	12	-	-	18	23		
	-1	3	9	21	4	3	19	22	23	
	0	-	-	2	-	-	3	17	-	-
	1	-	-	2	-	-	3	15	-	-
	2	2	7	-	-	-	5	30	-	-
	3				-	-	2	19	-	-
(c)	-4					-				
	-3				-	6	10			
	-2	-	10	13	10	3	18	22		
	-1	2	10	23	17	12	23	21	19	
	0	-	-	-	-	-	2	-	18	-
	1	-	-	-	-	-	3	2	4	-
	2	-	-	-	-	-	5	3	6	-
	3				-	-	-	6	9	-
(d)	-4					-				
	-3				-	-	20			
	-2	-	8	8	-	-	15	20		
	-1	-	9	16	4	3	16	20	17	
	0	-	-	-	-	-	3	9	-	-
	1	-	-	-	-	-	3	9	-	-
	2	-	5	-	-	-	5	16	-	-
	3				-	-	2	18	-	-

Table VI. Number of /θ/ responses.

Relative Location		Position of Constriction								
of Noise		9	8	7	6	5	4	3	2	1
(a)	-4					–				
	-3				–	–	–			
	-2	–	–	–	–	–	–	–		
	-1	–	–	–	–	–	–	–	–	
	0	–	3	3	2	4	10	12	–	6
	1	–	3	3	2	–	13	11	8	5
	2	–	5	2	–	–	12	12	10	7
	3				2	3	2	8	6	9
(b)	-4					–				
	-3				–	–	–			
	-2	–	–	–	–	–	–	–		
	-1	–	–	–	–	4	2	–	–	
	0	–	–	–	–	7	18	6	3	3
	1	–	–	–	–	3	15	8	3	3
	2	–	–	–	–	2	15	–	3	3
	3				–	–	8	2	3	2
(c)	-4					–				
	-3				–	–	–			
	-2	2	–	–	–	–	–	–		
	-1	–	–	–	–	–	–	–	–	
	0	–	–	–	–	4	11	5	–	5
	1	–	–	–	–	–	11	10	7	4
	2	–	–	–	–	3	8	13	7	6
	3				–	–	4	8	5	7
(d)	-4					–				
	-3				–	–	–			
	-2	–	–	–	–	–	–	–		
	-1	–	–	–	–	3	2	–	–	
	0	–	–	–	–	5	18	7	3	–
	1	–	–	–	–	3	18	9	2	3
	2	2	–	–	–	–	11	2	3	3
	3				–	–	7	–	3	4

Table VII. Number of /ʃ/ responses.

	Relative Location of Noise	Position of Constriction								
		9	8	7	6	5	4	3	2	1
(a)	-4					11				
	-3				14	13	6			
	-2	-	-	-	3	8	2	-		
	-1	-	-	-	2	3	-	-	-	
	0	-	-	-	2	7	-	-	-	-
	1	-	-	-	2	12	2	-	-	-
	2	-	-	-	2	8	3	-	-	-
	3				4	2	-	2	-	-
(b)	-4					21				
	-3				22	23	3			
	-2	6	3	8	19	21	2	-		
	-1	3	-	-	10	13	-	-	-	
	0	-	-	-	4	5	-	-	-	-
	1	-	-	-	12	16	-	-	-	-
	2	2	-	4	16	22	-	-	-	-
	3				15	14	2	-	-	-
(c)	-4					14				
	-3				12	7	5			
	-2	2	-	-	8	10	-	-		
	-1	-	-	-	-	2	-	2	-	
	0	-	-	-	-	5	-	2	-	-
	1	-	-	-	4	9	-	-	-	-
	2	-	-	-	-	5	5	-	-	-
	3				2	3	2	-	-	-
(d)	-4					22				
	-3				23	24	-			
	-2	4	-	7	21	20	3	-		
	-1	2	-	-	3	11	-	-	-	
	0	-	-	-	-	2	-	-	-	-
	1	-	-	-	9	11	-	-	-	-
	2	-	-	4	11	18	-	-	-	-
	3				12	11	-	-	-	-

Table VIII. Number of /ç/ responses.

Relative Location					Point of Constriction					
of Noise		9	8	7	6	5	4	3	2	1
(a)	-4					6				
	-3				3	-	5			
	-2	7	-	3	3	9	3	-		
	-1	7	4	-	2	4	-	-	-	
	0	-	-	3	12	7	7	2	2	-
	1	2	4	5	8	8	3	2	2	-
	2	3	-	5	11	9	5	-	2	-
	3				5	6	5	5	-	-
(b)	-4					2				
	-3				-	-	2			
	-2	3	3	4	4	-	3	-		
	-1	2	3	2	8	4	2	-	-	
	0	4	5	4	15	10	-	-	-	-
	1	3	7	8	7	4	6	-	-	-
	2	3	6	9	8	-	2	3	-	-
	3				6	5	9	2	-	-
(c)	-4					4				
	-3				4	2	4			
	-2	4	3	2	4	7	5	2		
	-1	3	2	-	4	5	-	-	-	
	0	-	4	5	6	7	-	-	-	-
	1	-	2	7	6	4	2	-	-	-
	2	-	3	3	11	5	2	2	-	-
	3				9	9	4	-	-	-
(d)	-4					-				
	-3				-	-	-			
	-2	2	6	2	-	-	-	-		
	-1	3	4	-	5	-	-	-	-	
	0	4	6	7	9	7	-	-	-	-
	1	4	3	8	6	2	-	-	-	-
	2	-	6	10	4	-	-	-	-	-
	3				5	5	4	-	-	-

Table IX. Number of /x/ responses.

Relative Location of Noise		Position of Constriction								
		9	8	7	6	5	4	3	2	1
(a)	-4					–				
	-3				3	3	–			
	-2	5	5	4	–	–	–	–		
	-1	9	6	–	–	3	–	–	–	
	0	17	19	18	8	–	–	5	–	–
	1	17	16	16	8	–	–	5	–	–
	2	18	16	12	9	4	2	4	–	–
	3				12	12	10	2	–	–
(b)	-4					–				
	-3				–	–	2			
	-2	11	10	–	–	–	–	–		
	-1	14	11	–	–	–	–	–	–	
	0	17	16	12	4	–	2	–	–	–
	1	16	16	13	5	–	–	–	–	–
	2	15	10	18	–	–	–	–	–	–
	3				2	3	3	–	–	–
(c)	-4					2				
	-3				2	2	–			
	-2	4	5	6	–	–	–	–		
	-1	12	7	–	–	–	–	–	–	
	0	19	17	18	17	5	3	7	–	–
	1	20	20	15	10	3	3	3	–	–
	2	20	15	17	11	8	–	–	–	–
	3				11	10	5	2	–	–
(d)	-4					–				
	-3				–	–	–			
	-2	8	5	–	–	–	–	–		
	-1	12	6	–	–	2	–	–	–	
	0	13	15	13	6	–	–	–	–	–
	1	13	17	14	5	–	–	–	–	–
	2	16	10	7	–	–	–	–	–	–
	3				–	–	2	–	–	–

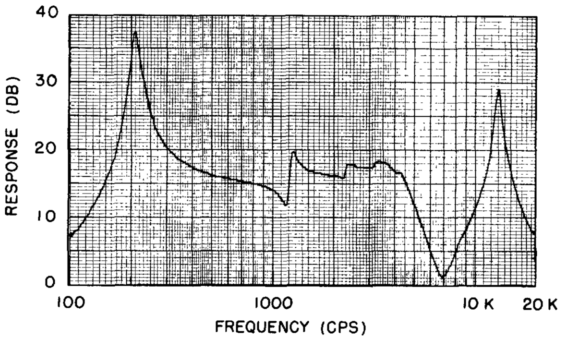


Fig. 44. Vocal-tract transmission for a fricative judged to be /f/. Curve drawn by automatic recorder shows ratio of voltage across small output inductor of vocal tract to voltage across noise-insertion transformer. The excitation is inserted at location 1; the constriction is close and at position 1.

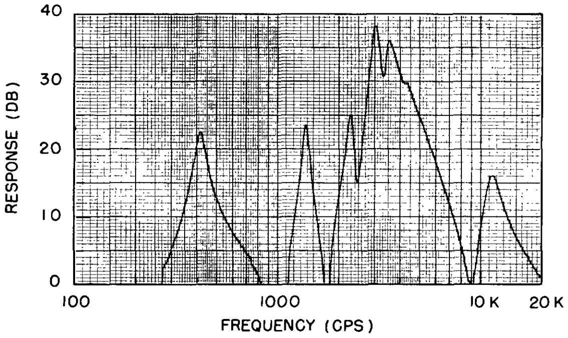


Fig. 45. Transmission for fricative made with open constriction at position 3 and judged to be /s/. Excitation is inserted at location 1.

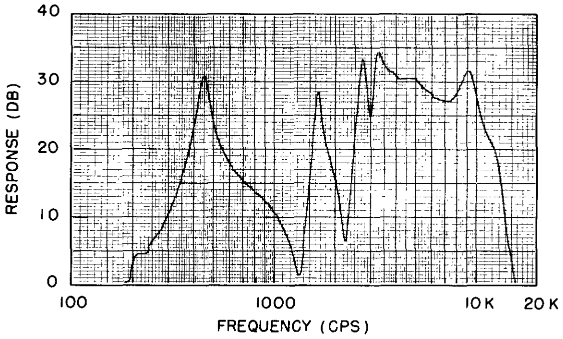


Fig. 46. Transmission for fricative made with open constriction at position 4 and judged to be /s/. Excitation is inserted at location 3.

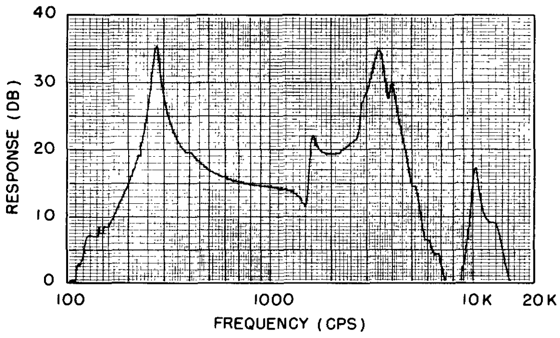


Fig. 47. Transmission for fricative made with close constriction at position 4 and judged to be /θ/. Excitation is inserted at location 4.

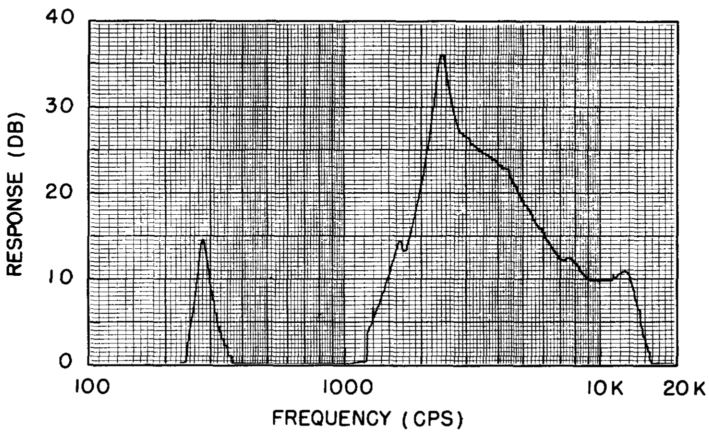


Fig. 48. Transmission for fricative made with close constriction at position 5 and judged to be /ʃ/. Excitation is inserted at location 2.

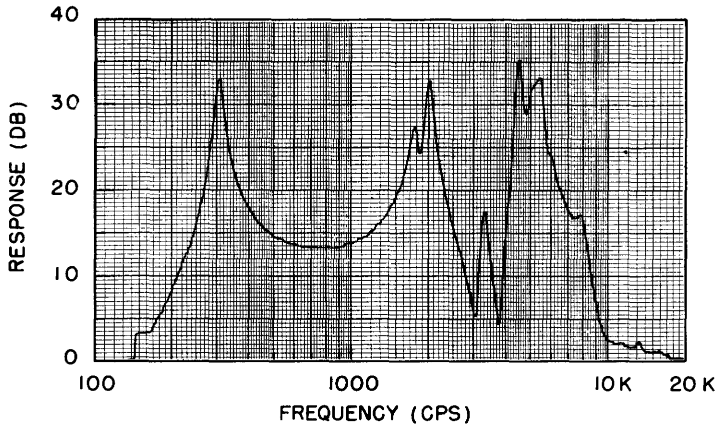


Fig. 49. Transmission for fricative made with close constriction at position 6 and judged to be /ç/. Excitation is inserted at location 6.

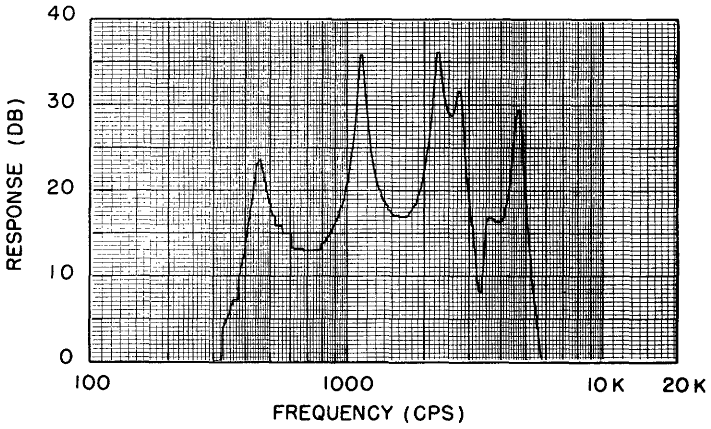


Fig. 50. Transmission for fricative made with open constriction at position 9 and judged to be /x/. Excitation is inserted at location 11.

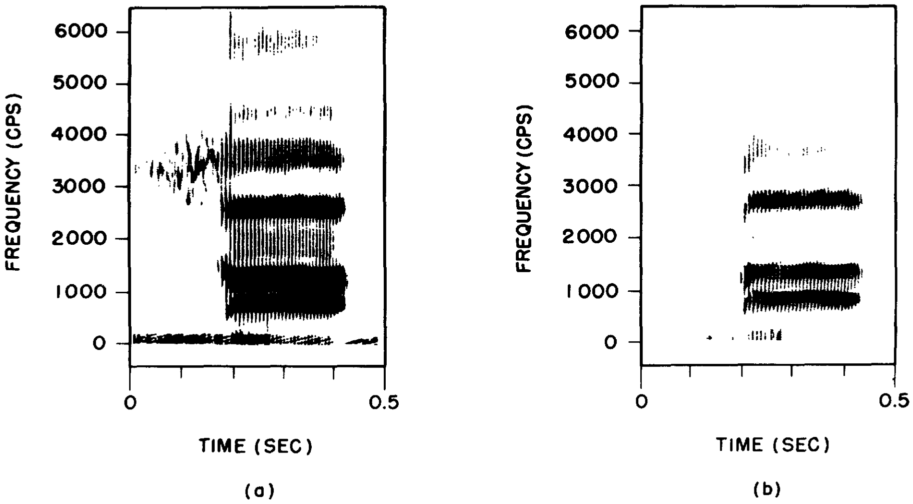


Fig. 51. (a) Spectrogram of /θɑ/ showing downward second-formant transition and noise energy between 2500 and 4000 cps. Area indicating noise is triangular with apex at the left. This shape is an artifact of the slow onset of noise which gradually brings more of the spectral peak above the marking threshold of the spectrograph. (b) Spectrogram of /fɑ/ showing upward second-formant transition. Noise spectrum for /f/ has no pronounced peak corresponding to the one seen in (a). With the pre-emphasis network used in making the spectrogram, no portion of the fricative spectrum rises above the marking threshold.

not yet been obtained.

The acoustic properties of the fricatives can be studied in terms of their spectra. A specimen for each fricative was selected by examining Tables IV-IX and choosing the item showing the greatest number of responses. The vocal tract was set to the corresponding configuration and point of excitation and then its transmission as a function of frequency was measured. Because noise with a flat spectrum was used to generate the stimuli, the fricative spectrum is the same as the transmission function. The curves shown in Figs. 44-50 were obtained with an automatic-response recorder linked to an oscillator. For each configuration, levels were set to keep the graphs within the range of the recorder.

Gated samples of American English fricatives produced by human subjects have been studied by Hughes and Halle (35). They found that the spectra of voiceless fricatives can be characterized in terms of gross regions of energy concentration. In both the samples of natural speech and in the fricatives produced by the analog, the series /ʃ, s, f/ corresponds to energy concentrated at successively higher frequencies. Details of natural fricative spectra vary considerably from one utterance to the next. Similarly, different versions of the same synthetic fricative differ in detail, as may be seen by comparing Figs. 45 and 46. Both curves describe stimuli yielding /s/ responses above the 95 per cent level. By using a tape-splicing technique, Harris (28) was able to join the noise burst of one natural fricative-vowel syllable to the vocalic portion of another. She found that $/\theta/$ and /f/ are distinguished from each other by transitional cues and that those sounds, as a class, are distinguished from /ʃ/ and /s/ by means of the noise spectrum. The spectra of $/\theta/$ and /f/ in Figs. 44 and 47 are quite similar, but in Fig. 51 the spectrograms of the corresponding syllables show contrasting second-formant transitions. Stevens and Nakata (55) worked with CV syllables made with shaped noise having a single spectral peak. Their perceptual studies led to conclusions corroborating all of those mentioned above.

The high-frequency energy concentration in the /f/ spectrum is caused by resonance of the very small front cavity of the /f/ configuration. This resonance is not a reliable cue for those whose hearing is impaired at high frequencies, and for those listening to most speech-transmission apparatus, even with high-quality earphones. If this peak is ignored, the spectrum (Fig. 44) becomes relatively flat and resembles the natural and synthetic /f/ spectra studied by Fant (21). Flat portions of fricative spectra are, in many cases, attributable to secondary noise sources. The spectrum for /x/ (Fig. 50) is comparable to those obtained by Fant, using a Russian subject.

Spectra produced by the dynamic analog differ in one respect from those previously published. Each spectrum shown for the dynamic analog has a very pronounced first resonance whose frequency increases as the back cavity is shortened. This effect can be observed by arranging the spectra for the open constriction in order of increasing position number and noting corresponding peaks at 420, 450, and 560 cps. Similar ordering for close constrictions yields frequencies of 190, 280, 290, and 350 cps. The influence of the back cavity is suggested in another way by Fig. 48. In this case, the noise is inserted greatly anterior to the constriction, the source is relatively decoupled from the back cavity, and the peak is not too prominent. The pressure pattern for the back-cavity resonance has a maximum at the glottis. In the existing machine, the glottal termination of the transmission line closely resembles a current source, and so the need for damping at that point is apparent. This conclusion is strengthened by an examination of Fig. 6, in which a vowel-transmission curve with a quite narrow first formant is shown.

6.3 STUDY OF TEMPORAL VARIABLES

Stimuli can easily be modified by changing parameters of the control system and, for temporal variables, these changes are readily described in terms of trapezoidal functions. However, measuring the effect of those changes on a listener requires some care, and the method must be chosen with full cognizance of the experimental goals. The general purposes of these experiments are: to find those values of temporal variables that are typical of fricative production, to determine how critical these values are, and to study the effect of timing on voiced-voiceless differentiation.

The changing of time parameters can have one of three effects on a voiceless fricative: improvement, degradation, or transformation into another sound. To measure the first two effects some definition of "goodness" must be used to rank stimuli that are

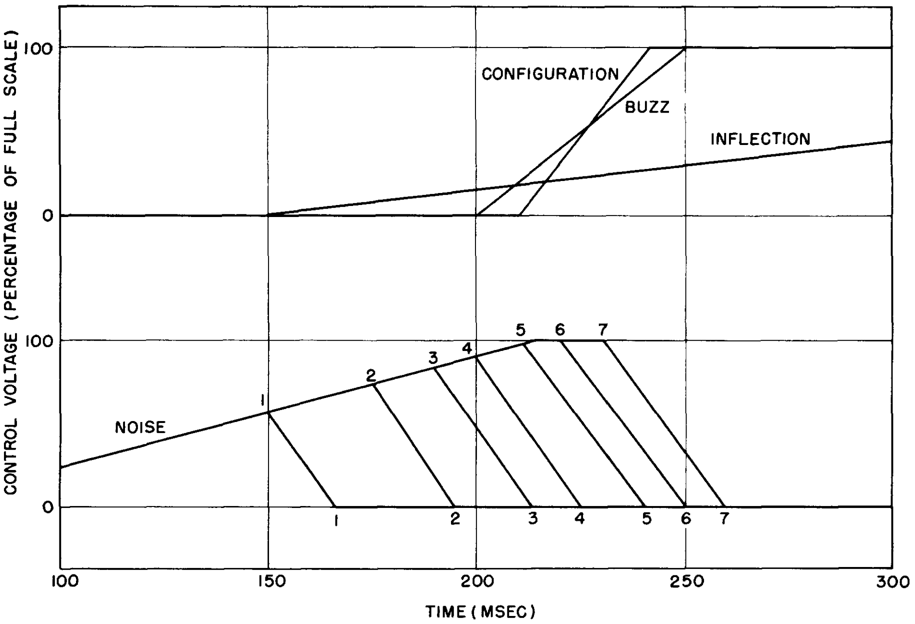


Fig. 52. Timing patterns describing stimuli used for studying the effect of changing the noise-cessation time. Control waveforms, unaltered from case to case, appear at the top.

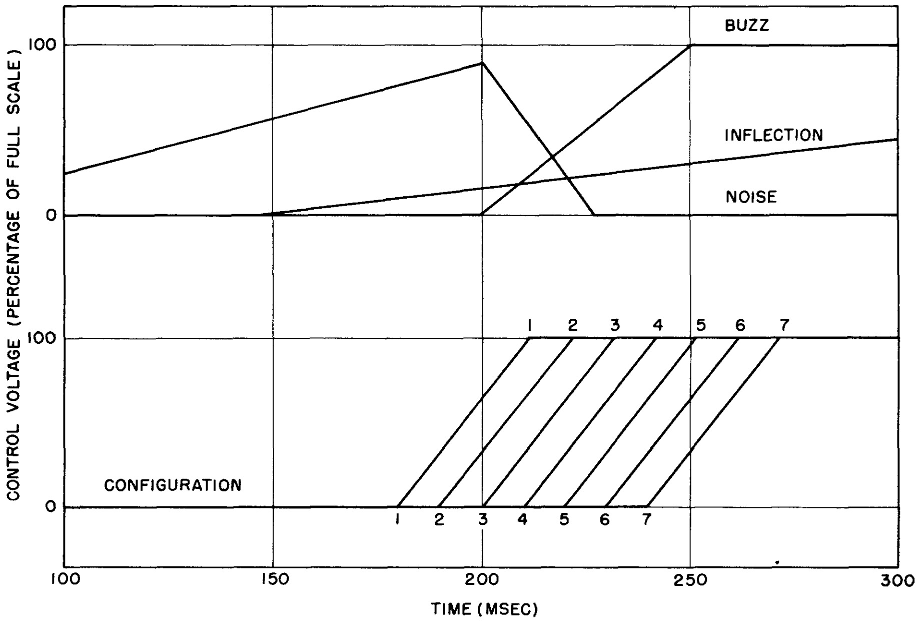


Fig. 53. Timing patterns describing a stimulus set with differing starting times for configuration change.

known to have the same phonetic value. This can be done by means of the method of paired comparisons (26), in which each stimulus is compared, in turn, with every other stimulus. Pairs of stimuli are recorded in randomized order and presented to listeners who are asked to vote for one member of each pair according to some stated criterion. A series of paired comparison tests was recorded to study the effects of changing these three variables: time at which noise begins to decay; time at which the configuration change begins; and time of buzz onset.

Data from the configuration study was examined and an item was found which received unanimous /ʃ/ responses in a non-forced test. Configuration variables for the present study were then fixed at values used in the production of that item: constriction close, constriction at position 5, and noise at location 2 (relative location -3). The context was the same as before, CV syllables always with the vowel /ɑ/. Both initial and final configurations were identical for all syllables. The device for modulating noise with buzz was always connected into the noise-generator circuit. This modulator, described in section 4.3, is important in connection with stimuli in which noise and buzz excitations overlap in time; otherwise it is of little consequence.

The timing patterns shown in Fig. 42 and used in the configuration study also provide the neutral values for this study. The patterns for the noise cessation and configuration change tests are shown in Figs. 52 and 53. For each item on the answer sheet, there is an item on the tape, of the following form: the pair of syllables to be compared, then a pause of 1.25 seconds, a repetition of the pair, and then a pause of 5 seconds for writing the answer. Three sample items precede the test, which consists of 42 items in randomized order. That is, each of 7 syllables representing 7 timing patterns are compared with each of the remaining 6 patterns so that each intrapair ordering, AB and BA, appears once somewhere in the test. Each group of 5 items is punctuated with a marker tone and a hiatus. Each subject received written instructions (reproduced in the appendix) requesting that he vote for the more "natural" member of each pair. The score is the total number of times a given stimulus is judged to be more natural than the one paired with it in an item. Five subjects participated in each of these tests; therefore the maximum score that any stimulus could receive is 60, the minimum is zero, and the average is 30.

The results of the noise-cessation study are shown in Fig. 54. There is a rough plateau, approximately 30 msec wide, approximately between items 3 and 6. Under the conditions of the experiment, the tolerance for this parameter seems to be approximately 3 pitch periods of male speech if a 100-cps glottal rate is assumed. In items 1 and 2, judged to be the least natural, the decay of the noise is completed before either the articulatory change or the buzz excitation begins. These perceptual data are consistent with articulatory phenomena. For, in natural speech, movement of the articulators from the fricative configuration destroys the constriction, together with its associated region of turbulence, and ends the noise. Items 1 and 2 violate the time constraints thus implied and, consequently, are poor.

The responses for the articulation test, Fig. 55, corroborate the need for overlap between the noise ramp and the articulation ramp. Items 6 and 7, with no such overlap, were judged to be poor. In item 1, which is also poor, more than half of the change of configuration is completed before either of the new ramps for control of excitation begins. In the items judged best, items 3 and 4, the entire change of configuration occurs during the interval of increasing buzz excitation, the change being completed when the buzz is within 5 db of its upper quiescent value. Furthermore, there is a greater than 50 per cent overlap of the articulation and noise-decay ramps, which is measured relative to either one. The width of the plateau in the figure is 10 msec. Examination of all of the data shows that configuration-change time is the most critical parameter studied, although the tolerance for it is still close to one pitch period.

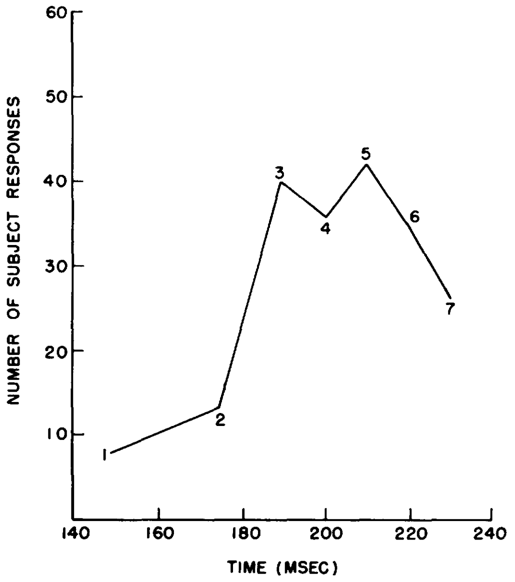


Fig. 54. Subject responses obtained for study of noise-cessation time. Item numbers correspond to those in Fig. 52. Ordinate shows total number of times a given stimulus was voted "more natural" than the stimulus with which it was paired.

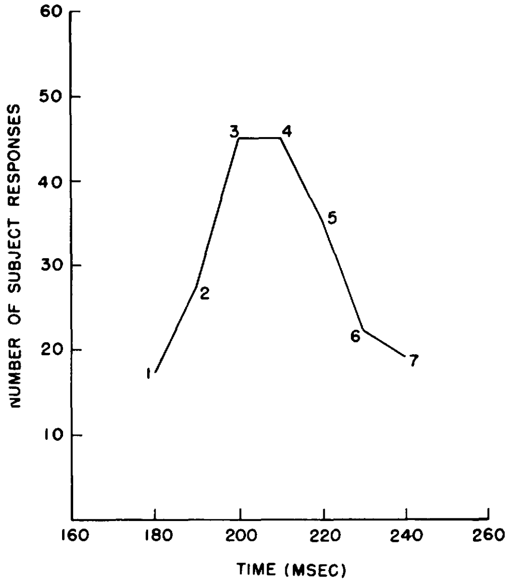


Fig. 55. Subject responses to stimuli with differing times of configuration change.

The study of buzz-onset time required recording 3 tapes, two for paired comparison tests, and one, which will be discussed later, for an absolute identification test. Timing patterns for all stimuli are shown in Fig. 56. Our plan was to find the most natural /ʃɑ/ and the most natural /ʒɑ/ in two separate tests that were intended to function independently, although they had some stimuli in common. The first set of instructions (see appendix), in which we asked for the most natural /ʃɑ/ in one test and for the most natural /ʒɑ/ in the other, proved to be troublesome. Consider the test intended to find

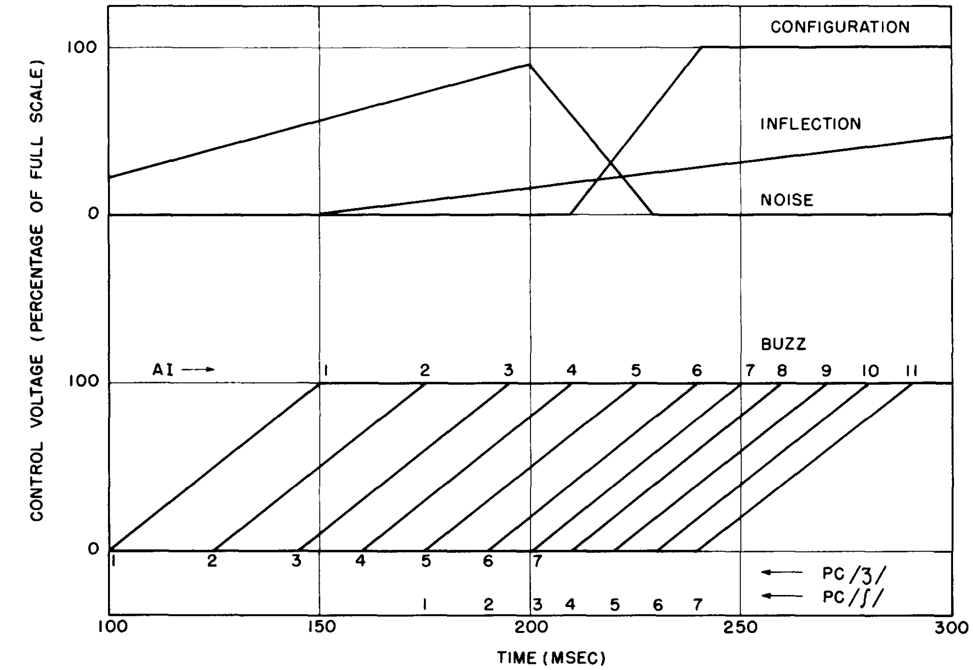


Fig. 56. Patterns describing stimuli used in 3 tests for investigating buzz-onset time. Control waveforms, unchanged during the series, are shown at the top. Item numbers for the absolute identification test are above the buzz ramps, and item numbers for the 2 paired comparison tests are below.

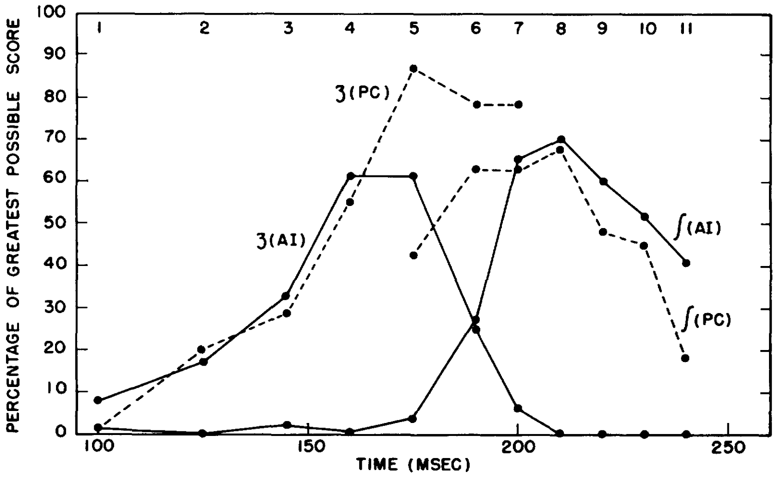


Fig. 57. Test scores, for two fricatives, as a function of buzz-onset time. Scores derived from paired comparison tests are shown by dotted lines; those derived from the absolute identification test are shown by solid lines. Curves are normalized to facilitate comparison, and item numbers correspond to those in Fig. 56.

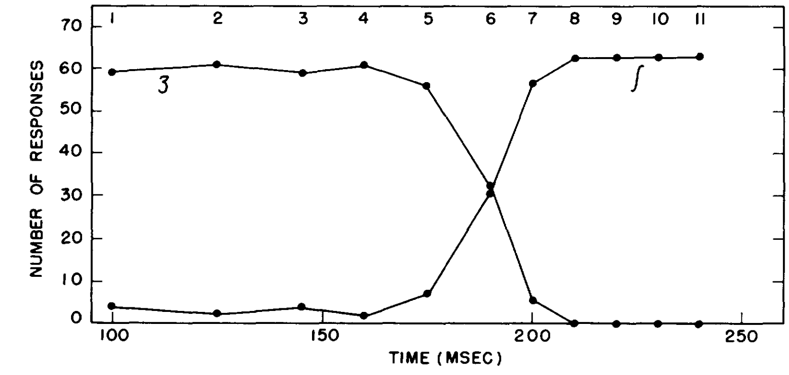


Fig. 58. Subject responses in absolute identification test as a function of buzz-onset time. The number of times each item was judged to be /ʒ/ or /ʃ/ is shown totaled for all subjects.

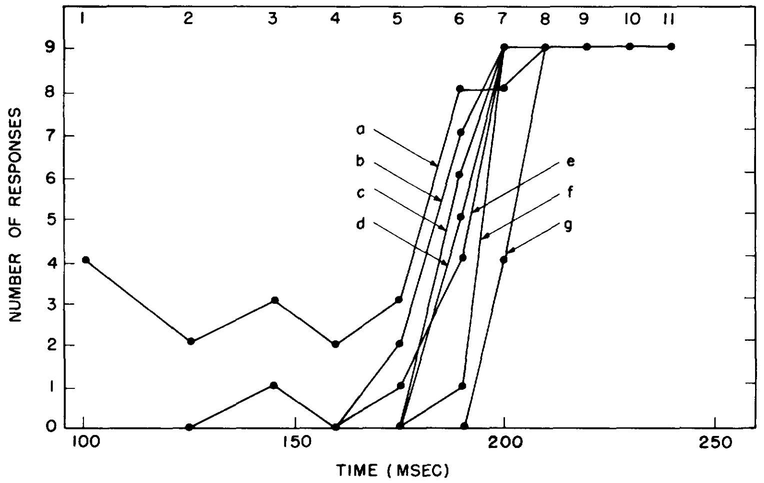


Fig. 59. Individual responses for 7 subjects in absolute identification test. Number of times each item was perceived as /ʃɑ/ is given. Number of /ʒɑ/ responses is the nines complement of the number of /ʃɑ/ responses.

the most natural voiceless stimulus and the subject's dilemma when he hears item one. With buzz onset at 175 msec, it sounds natural but voiced. Does he prefer it because it is natural or does he vote against it because it is voiced? Some subjects responded in one way, and some in the other.

To circumvent this difficulty, the role of the two paired comparison tests was revised. The new role is that of defining a region in which either /ʃa/ or /ʒa/ may be produced, and of exploring its extremities. The division of the region into voiced and voiceless compartments was made later with a supplementary absolute identification test. In this subsequent listening session, the subjects were instructed (see appendix for new instructions) to vote for the more natural stimulus, regardless of its phonetic identity. The responses for the two paired comparison tests are indicated by the two dotted curves in Fig. 57. These curves show how rapidly the sounds deteriorate as the extremities of the combined region are approached, but their meaning within the region is not clear. Consider items 5, 6, and 7, which are definitely within the combined region and are common to both tests. In the new role of the paired comparison tests, these items merely serve to provide a background against which the other items may be judged.

The absolute identification test for buzz-onset time was recorded to find the boundary between the voiced and voiceless regions, and also to obtain a second measure of naturalness. The test used all of the 11 stimuli described in Fig. 56 and each stimulus appeared 9 times in randomized order. Samples, marker tones, and pauses were provided as in the previous tests. The instructions to the subjects (see appendix) required a /ʒa/, /ʃa/ judgment for each stimulus, and a naturalness rating on a three-point scale. Seven subjects served on the listening panel.

The averaged phonetic responses are shown in Fig. 58. The decision of the majority changes from /ʒa/ to /ʃa/ when the onset time is changed by 20 msec. Stimulus 6 is the only one not definitely voiced or voiceless. In Fig. 54, the corresponding buzz ramp begins 20 msec before the configuration ramp and ends when the configuration ramp ends. The buzz ramps for stimuli 7 and 8 also straddle the configuration ramp, but these stimuli are perceived as voiceless. The individual subjects whose responses are shown in Fig. 59 were more sensitive to changes in onset time than the averaged curve indicates. One subject switched his answers in response to a 10-msec change in onset time. The smaller slope of the graph of averaged responses is a consequence of a 20-msec spread in thresholds for the individual subjects. This spread is comparable to the 25-msec difference of onset time between items 5 and 7 (the closest items in Fig. 58 that are clearly distinguishable).

The secondary purpose of the absolute identification test was to provide an alternative measure of naturalness. This measure is calculated by weighted summation of responses for each onset time, one tally being kept for voiced responses and another for unvoiced. The stimuli rated "good" by the subjects were given weight 2; those rated "fair" were given weight 1; and those rated "poor" were not counted. Thus, a given stimulus may have a low score as /ʃa/ because it yields few /ʃa/ judgments, or because it results in sounds that are definitely /ʃa/ but of poor quality. For example, item 11 is clearly not voiced, and still it is a poor /ʃa/ because of a wide gap between buzz and noise. This hybrid nature is exemplified by the curve marked "ʒ(AI)" in Fig. 57, which follows the curve marked "ʒ(PC)" as far as item 4 and then behaves like the /ʒa/ ogive in Fig. 58. There is close agreement between the two quality measures at the extremities of the combined region so that each measure gives some weight to the validity of the other. Assuming that the other timer settings are at the neutral values, we may conclude from Fig. 58 that the production of /ʒa/ requires that buzz onset occur between 160 msec and 175 msec after the start of the clock; the corresponding interval for /ʃa/

lies between 200 msec and 215 msec. The tolerance is approximately 15 msec in each case, which is clearly greater than one pitch period of male speech.

VII. CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

7.1 KNOWLEDGE GAINED FROM DEVELOPMENT AND OPERATION OF THE SYNTHESIZER

In this research our objectives have been to develop a dynamic analog of the vocal tract, to develop a control system for the analog, and to carry out experiments with the finished synthesizer. We shall now discuss these points.

a. Development of the Analog

The success of the transmission line depended on the validity of certain simplifying assumptions regarding the acoustics of speech production (stated in Section II), on assumptions (stated in Section III) regarding the size and number of sections needed, and on the development of variable inductive and capacitive elements meeting certain requirements. Relying on the reasoning outlined in Section II, we have simulated a distributed system, partly nonlinear, by a linear lumped-parameter system. This reasoning has been tested by earlier analogs, all of which were static, but the dynamic analog has added to the variety of conditions under which the assumptions have been tested and has corroborated their validity.

Construction of the analog was preceded by the development of dynamic capacitive and inductive elements, each of which was variable over a 100:1 range. A question about the adequacy of the inductor quality factor has received an affirmative answer which is supported by a margin of safety. The number of fixed and variable sections in the transmission line was found to be adequate, since the line demonstrated suitable behavior at higher frequencies, at which a lumped approximation is most severely tested. Production of fricatives with their important high-frequency cues served as a test for this frequency region. Although the geometry of the lips and teeth is only grossly approximated during fricative production, listeners responding to sounds produced with the machine have consistently identified certain stimuli with appropriate fricative categories.

The calibration of the line is reasonably stable; the first 3 formants may be expected to remain within 5 per cent of appropriate nominal values throughout an entire day after a 15-minute warm-up. Although the signal-to-noise ratio of the analog is lower than that of comparable resonance synthesizers, a figure of 40 db can be obtained for vowel outputs. For this measurement, the buzz excitation is set so that no overload occurs for any vowel configuration in Fig. 39, and noise is measured for the most noisy configuration.

b. Development of the Control System

The control system, which is a physical implementation of a simplified description of the events in speech production, was designed to accomodate all sequences of two phonetic elements. The settings of the present control system, together, function as a comprehensive, detailed description of a given sound that is complete in the sense that no information need be added by external control signals. Experience has shown that the design of the control system permits parameter values to be easily changed and the stimuli to be expeditiously produced. Within the scope of our experiments, performance of the control system has been quite satisfactory, the number of control voltages has proved sufficient, and the trapezoidal functions adequately describe the timing patterns. Results obtained, thus far, have indicated that piecewise-linear functions should be retained until they prove unsatisfactory for the synthesis of certain sounds, or until physical or physiological studies lead to more plausible functions for describing the time course of excitation and configuration. The six-row configuration matrix, designed with factors of cost and complexity in mind, is suitable for isolated syllables and short words, but not for connected speech.

A new control system for generating synthetic-speech outputs of arbitrary length would closely resemble the present system but would use a different pulse source. The present timer would be replaced by a unit accepting a sequence of phonetic symbols and emitting trigger pulses for microprogramming of transitions from one sound to the next. This unit would also choose the appropriate ramp durations for each transition, and execute these choices through switching within the existing trapezoidal generators. The existing configuration matrix, which is relatively small and composed of variable elements, would be replaced by another matrix storing between 20 and 40 fixed configurations. In section 7.3 there is a discussion of such a control system.

c. Experiments with the Synthesizer

The synthesizer has demonstrated its capability of generating a set of vowels whose intelligibility is comparable to that of natural vowels. The formant bandwidths and the ranges of formant frequencies exhibited by this set of vowels are also characteristic of human male speech. Natural vowels can be ordered along dimensions that relate to their production by human speakers. Examples of such dimensions are: place of production, size of constriction, and "rounding." On the basis of the formal listening tests, the vowels produced by the analog can be similarly ordered.

The synthesizer has also generated a set of 6 voiceless fricatives, and several of these, embedded in CV syllables, have yielded almost unanimous listener identifications in formal listening tests. The configurations and points of insertion of the noise source that are appropriate for each fricative response corresponded closely to configurations and excitations that were expected on the basis of knowledge of human-speech production. In the cases studied, the use of a single noise-source spectrum and a single movable point of insertion of the source seems justified, and the degree to which configurations can be approximated on the present transmission line seems adequate. Study of several temporal control parameters showed tolerances of from 10 msec to 30 msec for those parameters.

The voiced-voiceless distinction in fricatives has been investigated formally and semivowels and stops have been studied informally, with results that indicate that the repertoire of sounds that can be generated by the existing dynamic analog can be considerably extended.

7.2 SUGGESTIONS FOR FURTHER RESEARCH

Future research on components and circuits for a dynamic analog speech synthesizer should place emphasis on alternative variable capacitors that require less power and are less complex than the existing ones. A variable capacitor might be realized with a solid-state device constructed as a dual of the existing variable inductor. A nasal analog should be added to the vocal-tract analog, and the variable coupling to the nasal analog may consist, in the beginning, of a single saturable inductor.

Since there is a dearth of data on the dynamics of the human articulators, the design of the configuration-control circuits is based on the simplest of assumptions. Future work should concentrate on studies of the dynamics of the human articulators; even kinematic data would help to fill a large gap in our knowledge of speech production. Production of the syllables in our experiments required a transition from a fricative configuration based on a straight tube model to a vowel configuration based on a parabolic model. It would be desirable to develop a more general model for configurations which would subsume, as special cases, those configurations that are suitable for the various classes of speech sounds. Such a model could lead to a synthesizer controlled by articulatory parameters.

Other goals for future research are the extension of the repertoire of the synthesizer and the development of a set of rules that can operate on a sequence of phonetic symbols to derive a description of the control signals to be applied to the analog. In principle, these rules can be reduced to a table that would deal with short overlapping segments of speech and would be entered by specifying a phonetic symbol and its context. The listening tests, thus far, have been concerned mainly with unvoiced fricatives in CV syllables, and most of the spaces in the desired table, which are now blank, must be filled in through work with other speech sounds in a variety of environments. (Such research with the dynamic analog would deal with many of the phenomena investigated by the Haskins group with the pattern playback.) The stops, voiced and unvoiced, offer a particularly challenging opportunity for exploiting the ability of the synthesizer to simulate articulatory movements.

Many of the future studies would be expected to follow the paradigm of the temporal studies reported in section 6.3. In such experiments, a panel of listeners establishes an ordering of stimuli according to some criterion for "goodness" so that some items are found suitable for inclusion in the machine vocabulary and others are rejected. New psychophysical techniques for evaluating "quality" or "naturalness" should be investigated with a view to developing a measure suitable for an extensive listening program involving a large number of stimuli. Such a measure should be economical of listening time and should facilitate comparison of data from different experiments.

Another goal of future research is an increased understanding of speech production and perception. In many experiments, the analog, together with its control system, might act as a standard speech source or "signal generator." Perhaps, phonetic entities may be defined in terms of this instrument. The possiblities and limitations of the synthesizer in this role should be explored.

A great deal of speech research may be regarded as an exploration of the mappings between the articulatory, acoustic, and linguistic domains. Of particular interest is the relation between continuous physical parameters and distinctive features that are defined in terms of contrasting pairs. Certain physical axes may be divided into two regions, each of which corresponds to a member of a pair. The locations of boundaries between such regions should be studied. An example is the /s/, /t/ opposition (continuant-interrupted), which is related to parameters such as noise duration and noise-onset rate.

7.3 INTEGRATION OF THE SYNTHESIZER WITH OTHER DEVICES

Most applications for speech synthesizers call for the production of connected speech in response to information that may ultimately be traced to a human or to a computer.

The input to a speech-compression system is a speech wave from a human speaker. Because the input is acoustic, the analyzer must operate in the acoustic domain. If, however, a dynamic analog is to be used as the synthesizer in the speech-compression link, it must be controlled with articulatory information. The acoustic signals must be converted to articulatory signals either at the entrance or exit of the channel. Thus, the most straightforward system entails the development of a device for converting one set of continuous signals (at the acoustic level) into another set of continuous signals (at the articulatory level). Design of such a system is a formidable engineering task, partly because the signals specifying configuration and excitation of the analog must be closely synchronized at certain critical times. After passing through the analyzer and the converter, the signals must exhibit a relative timing error that, in some cases, must not exceed 10 msec. Timing requirements are most severe at phoneme boundaries and are relatively lax at other times. One can propose a device to effect temporal sharpening with the aid of decision elements that perform operations upon linguistic entities. This device would examine the continuous signals to recognize phoneme boundaries and classify successive phonemes with respect to their manner of production. The device would then emit precisely synchronized signals that are appropriate to the given transition. In such a scheme, the analyzer need preserve only gross timing patterns. The signals in the channel may be slowly varying, without sharp wave fronts. The high-frequency response needed in other schemes to preserve occasional sharp wave fronts is not needed in this one.

Discrete codes must be handled by the synthesizer serving in a computing system or in a speech transmission system that is operating at high compression ratios. Coding at the phonetic level is of particular interest because speech cannot be segmented into shorter linguistic elements, and because phonetic elements are building blocks for phonemes, words, sentences, and so on. Information coded at higher linguistic levels can be resolved into phonetic symbols by a device distinct from the synthesizer. A phonetic control system for the analog is essentially an instrumentation of the table discussed in section 7.2. The control system memory must store the symbol that is being uttered plus the subsequent symbol, in order to choose the appropriate transitions in excitation and configuration and to execute their microtiming. The start of the transition can be signaled by receipt of the subsequent symbol. Such a system allows for variable phone duration, and can accommodate itself to different speaking rates and to different patterns of emphasis. Information about prosodic features is needed to complete specification of an utterance. In compression systems such information can be carried in parallel channels. In computing systems, signals can be constructed from known rules (see, for example, Chomsky, Halle, and Lukoff (10)) when the syntactic structure underlying the utterance is available in the computer memory.

7.4 THE DYNAMIC ANALOG AND THE TESTING OF A PHILOSOPHY

Phoneticians have for many years suggested that the articulatory level plays a central role in the speech-communication process. It acts as a bridge between the linguistic level and the acoustic level when the production of speech is considered, and it has also been suggested that it mediates between the acoustic and linguistic levels during the perception of speech. Much of the speech research at Massachusetts Institute of Technology has followed this philosophy, since it has been based on the conviction that a physical description of speech is most naturally and economically achieved at the articulatory level. This viewpoint provided the original motivation for the construction of the dynamic analog of the vocal tract.

Experiments with the completed machine have shown that highly intelligible stimuli can be generated with simple timing patterns and with rather rough approximations to human articulatory configurations. Several configurations that are physiologically similar have yielded the same high intelligibility, even though the spectra of the stimuli are quite different. These experiments are to be contrasted with acoustically oriented studies in which difficulty has been experienced in obtaining high response levels to certain stimuli, and in which responses are quite sensitive to small changes in the acoustic signal.

Thus research with the dynamic analog has provided strong support for the point of view that the articulatory level provides a natural and economical description of speech. The machine provides, furthermore, a facility that can be used to extend our understanding of the articulatory level, and thus to broaden our knowledge of the entire speech-communication process.

APPENDIX

1. CIRCUIT DIAGRAMS OF CRITICAL OR NOVEL PARTS FOR THE ANALOG AND ITS CONTROL SYSTEM

Figure 60 shows the nasal analog (2) added to a static analog of the vocal tract (54) for studies of nasality (31, 33). A nasal circuit for the dynamic analog can be expected to be similar to this circuit.

Figure 61 shows the resistance network, a detail of the variable-capacitance circuit. The calibration curve that is desired from the capacitor is obtained through proper choice of element values in the resistance network.

Figure 62 shows the control amplifier that controls both the capacitance and inductance of its section, to keep the length of the section constant.

Figure 63 shows the circuit diagram of the dc amplifier for deriving the configuration matrix. The amplifier is rated at 0 to +300 volts, 0 to 70 ma output, and 5 ohms internal impedance.

Figure 64 is a complete circuit diagram of the trapezoidal-function generator showing details omitted in the simplified circuit of Fig. 35. The 6AL5 tube at the right of Fig. 64

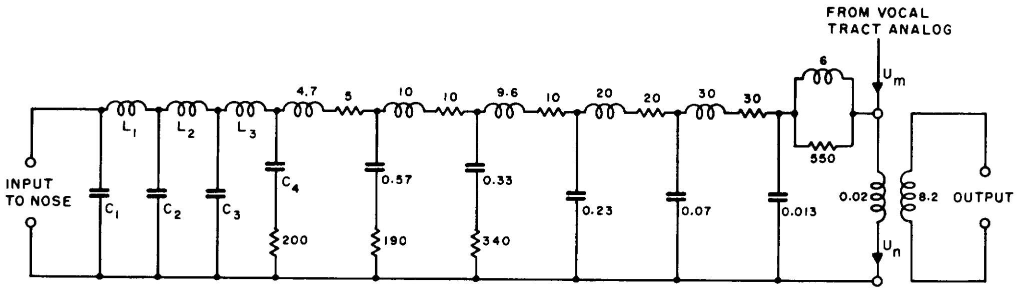


Fig. 60. Nasal analog designed for use with a static analog of the vocal tract. (Resistance in ohms, inductance in mh, capacitance in μfd.)

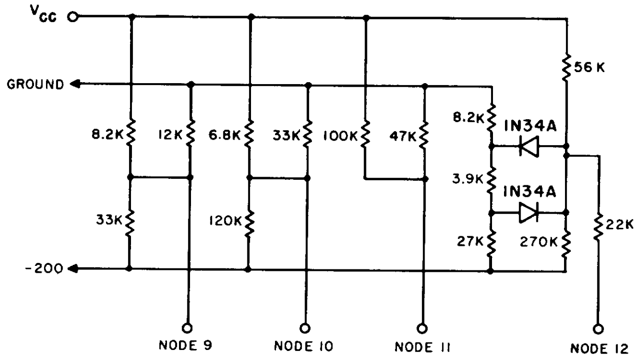


Fig. 61. Resistance network in grid circuit of electronic attenuator. Grids connected to nodes 9, 10, and 11 conduct for certain ranges of V_{CC}.

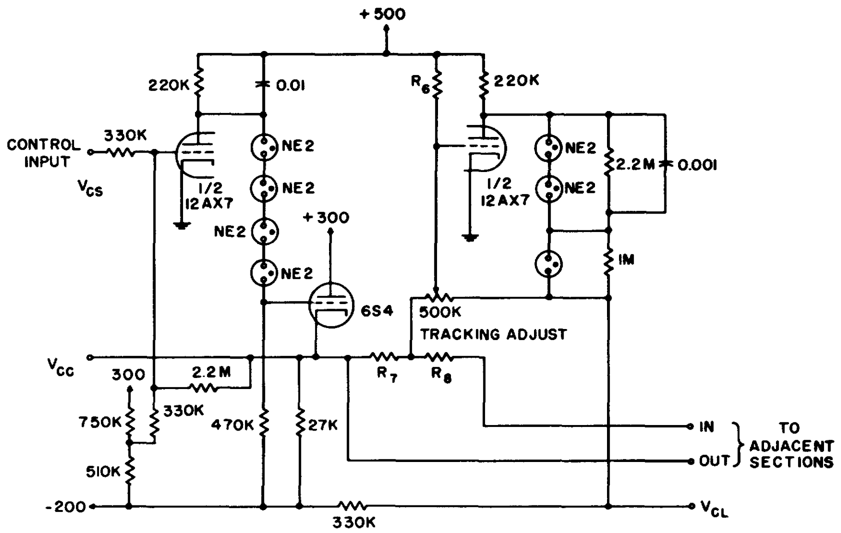


Fig. 62. Circuit diagram of control amplifier for a variable section.

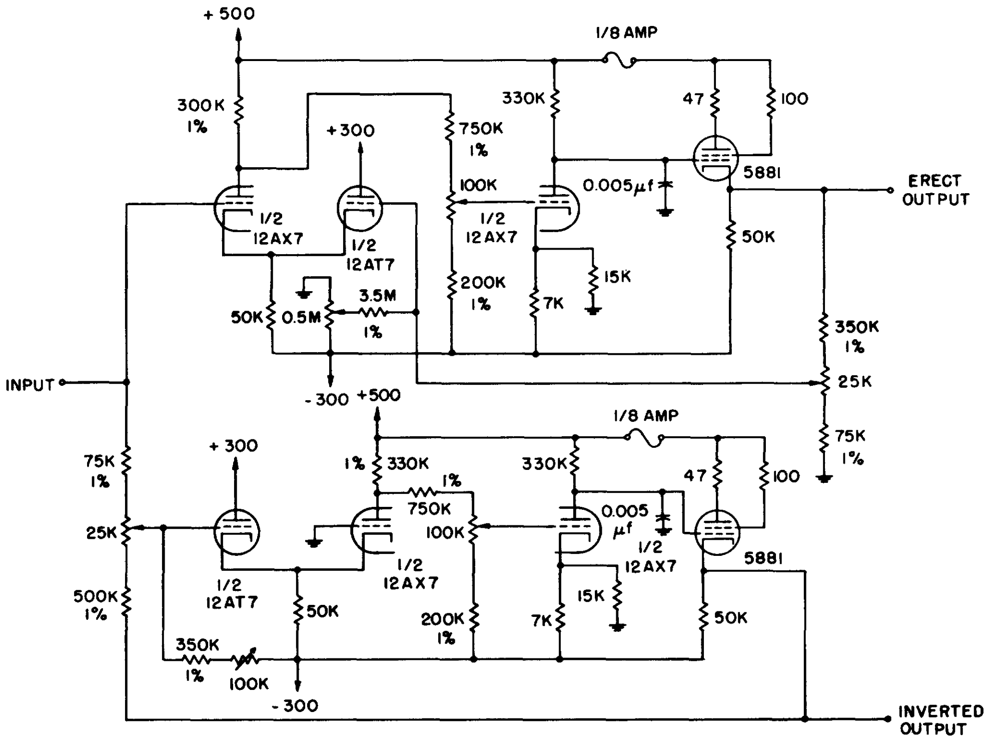


Fig. 63. Circuit diagram of dc amplifier that yields both erect and inverted versions of the input waveform.

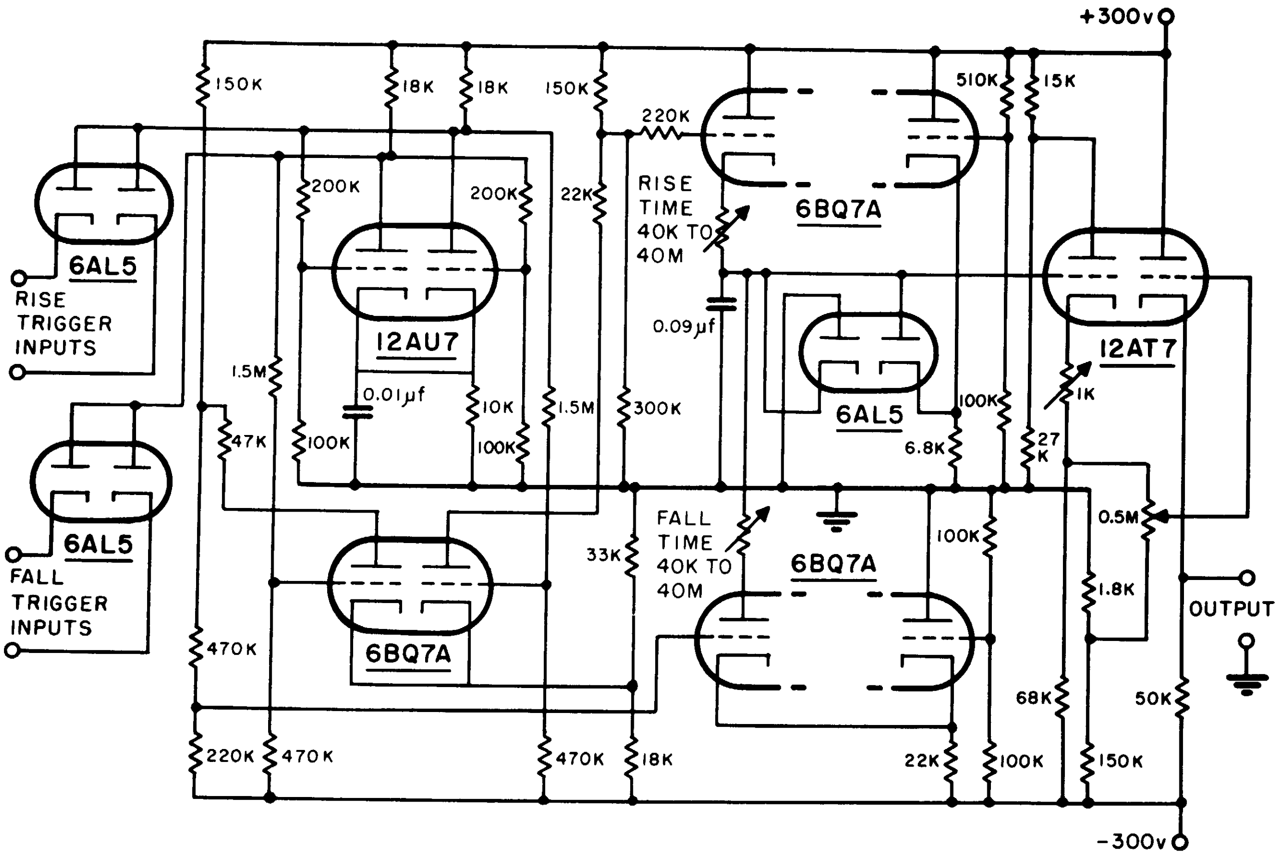


Fig. 64. Schematic diagram of the trapezoidal-function generator.

provides both the upper and lower clamps. The 12AT7 tube is the buffer amplifier. All of the other tubes, including the 12AU7 flip-flop, serve to replace the switch, S, of Fig. 35. The variable resistors to control rise time and fall time are constructed as decade resistors.

2. TEST INSTRUCTIONS TO SUBJECTS

The instructions that follow were given to subjects for the listening tests in which time of noise cessation and time of configuration change were varied. These instructions were also used in a study of buzz-onset time made with values appropriate for the production of /ʃ/.

> In this test you will hear pairs of sounds and you are to pick the one which is the more "natural," in that it is more like human speech. Each sound is a syllable consisting of a consonant followed by a vowel. The vowel is identical in all cases. The consonant has differing degrees of similarity to the sh sound in ship, shore, shoe or shine. If the first sound of a pair is more natural write 1; if the second is more natural write 2. Each pair will be repeated, the second occurrence followed by a pause for writing your answer. The test will start with three sample groups. Do not write answers for them. Start answering after you hear the high pitch tone three times. The tone also appears after every fifth item. There are 42 test items.

The following instructions were used in a study of buzz-onset time made with values appropriate for the production of /ʒ/.

> In this test you will hear pairs of sounds and you are to pick the one which is the more "natural," in that it is more like human speech. Each sound is a syllable consisting of a consonant followed by a vowel. The vowel is identical in all cases. The consonant has differing degrees of similarity to the zh sound in Asia, beige or azure. If the first sound of a pair is more natural write 1, if the second is more natural write 2. Each pair will be repeated, the second occurrence followed by a pause for writing your answer. The test will start with three sample groups. Do not write answers for them. Start answering after you hear the high pitch tone three times. The tone also appears after every fifth item. There are 42 test items.

The following instructions reflect a revision of purpose for the study of buzz-onset time by the method of paired comparisions.

> In this test you will hear pairs of sounds and you are to pick the one which is the more "natural," in that it is more like human speech. Each sound is a syllable consisting of a consonant followed by a vowel. The vowel is identical in all cases. The consonant may resemble the sh sound in ship, shore, shoe or shine or it may resemble the zh sound in Asia, beige or azure. The consonants in a pair may be the same or they may be different. In any case, if the first sound of a pair is more natural write 1; if the second is more natural write 2. Each pair will be repeated, the second occurrence followed by a pause for writing your answer. The test will start with three sample groups. Do not write answers for them. Start answering after you hear the high-pitch tone three times. The tone also appears after every

fifth item. There are 42 test items.

These instructions, asking for absolute identifications, were given to subjects listening to stimuli with differing buzz-onset times.

> You will hear a string of syllables, each consisting of a consonant followed by a vowel. The vowel is identical in all cases. The consonant resembles either the sh sound (as in ship, shore, shoe, shine) or the zh sound (as in Asia, beige, azure). Some specimens are better approximations to sh or zh than are others. Right after you hear each sound on the test, write the symbol sh if you think the sound is more like sh than zh and write the symbol zh if the sound is more like zh than sh.
>
> Rate the quality of the sound by writing 1, 2 or 3 after the symbol. Use 1 for good and 3 for poor. A sound may be "good" because it "sounds natural" or because it sounds very much like a sh or zh sound or because you think it is good. Use any criterion or combination of criteria you wish.
>
> To form your opinion of how good or bad the sounds can be you will hear 33 samples. Do not write anything for the samples. Start writing for the items following the series of 3 beeps. You will then hear 99 items with a beep after every 10 items.

Acknowledgment

The author wishes to express his gratitude to Professor Kenneth N. Stevens for the inspiration and continued encouragement that he has given throughout this research program. The helpful comments and suggestions of Professor Morris Halle and Professor Walter A. Rosenblith are also very much appreciated. For their stimulation and cooperation, the author wishes to thank the members of the Speech Communication Group, of Research Laboratory of Electronics, past and present, particularly Dr. Arthur S. House, Dr. James L. Flanagan, John M. Heinz, and Charles I. Malme.

This research would not have been possible without the facilities and support of the Acoustics Laboratory, M. I. T., the Research Laboratory of Electronics, M. I. T., and the Cambridge Research Center of the U. S. Air Force.

Bibliography

1. Speech Compression Research: Final Report. AFCRC-TN-57-166, Acoustics Laboratory, M.I.T., 1957.

2. Research on Speech Synthesis, Scientific Report No. 17. AFCRC-TN-58-140, Acoustics Laboratory, M.I.T., 1958.

3. J. B. Arnold, M. Halle, T. T. Sandel, and K. N. Stevens, Perception of speech-like sounds, Quarterly Progress Report No. 54, Research Laboratory of Electronics, M.I.T., July 15, 1959, pp. 167-171.

4. R. C. Bacon and J. R. Pollard, The Dekatron, Electronic Engineering 22, 173-177 (1950).

5. J. S. Barlow, M. A. B. Brazier, and W. A. Rosenblith, The application of autocorrelation analysis to electroencephalography, Proc. First National Biophysics Conference (Yale University Press, New Haven, 1959), pp. 622-626.

6. L. L. Beranek, Acoustics (McGraw-Hill Book Company, Inc., New York, 1954).

7. H. Bode, Network Analysis and Feedback Amplifier Design (D. Van Nostrand Company, Inc., New York, 1945).

8. T. Chiba, A Study of Accent: Research into its Nature and Scope in the Light of Experimental Phonetics (Fuzanbo Publishing Company, Tokyo, 1935).

9. T. Chiba and M. Kajiyama, The Vowel, Its Nature and Structure (Tokyo-Kaiseikan Publishing Company, Tokyo, 1941).

10. N. Chomsky, M. Halle, and F. Lukoff, On accent and juncture in English, For Roman Jakobson (Mouton and Company, N.V., The Hague, 1956).

11. Communications Biophysics Group of Research Laboratory of Electronics and W. M. Siebert, Processing Neuroelectric Data (The Technology Press, Massachusetts Institute of Technology, Cambridge, Mass., 1959); Technical Report 351, Research Laboratory of Electronics, M.I.T., July 7, 1959.

12. F. S. Cooper, A. M. Liberman, and J. M. Borst, Preliminary studies of speech produced by a pattern playback, J. Acoust. Soc. Am. 22, 678 (1950).

13. F. S. Cooper, A. M. Liberman, and J. M. Borst, The interconversion of audible and visible patterns as a basis for research in the perception of speech, Proc. Nat. Acad. Sci. 37, 318-325 (1951).

14. W. B. Davenport, Jr. and W. L. Root, An Introduction to the Theory of Random Signals and Noise (McGraw-Hill Book Company, Inc., New York, 1958).

15. H. W. Dudley, Synthesizing speech, Bell Labs. Record 15, 98-102 (1936).

16. H. W. Dudley, Remaking speech, J. Acoust. Soc. Am. 11, 169-177 (1939).

17. H. W. Dudley, The carrier nature of speech, Bell System Tech. J. 29, 147-160 (1940).

18. H. K. Dunn, The calculation of vowel resonances, and an electrical vocal tract, J. Acoust. Soc. Am. 22, 740-753 (1950).

19. C. G. M. Fant, Transmission Properties of the Vocal Tract with Application to the Acoustic Specification of Phonemes, Technical Report No. 12, Acoustics Laboratory, M.I.T., 1952.

20. C. G. M. Fant, Ingen.-Vetensk. Akad. (Stockholm) 24, 331-337 (1953).

21. C. G. M. Fant, Acoustic Theory of Speech Production, Report No. 10, Royal Institute of Technology, Stockholm, Sweden, 1958.

22. C. G. M. Fant, Modern instruments and methods for acoustic studies of speech, Acta Polytechnica Scandinavica, Ph 1, No. 2 (246/1958).

23. C. G. M. Fant, Acoustic analysis and synthesis of speech with applications to Swedish, Ericsson Technics 15, 3-108 (1959).

24. J. L. Flanagan, A Speech Analyzer for a Formant-Coding Compression System, Sc.D. Thesis, Department of Electrical Engineering, M.I.T., 1955; Scientific Report No. 4, Acoustics Laboratory, M.I.T., May 1955.

25. J. L. Flanagan, Some properties of the glottal sound source, J. Speech Hearing Res. 1, 99-116 (1958).

26. J. P. Guilford, Psychometric Methods (McGraw-Hill Book Company, Inc., New York, 2d ed., 1954).

27. E. A. Guillemin, Introductory Circuit Theory (John Wiley and Sons, Inc., New York, 1953).

28. K. S. Harris, Cues for the identification of the fricatives of American English, J. Acoust. Soc. Am. 26, 952 (1954).

29. J. M. Heinz, A terminal analog of fricative consonant articulation, Quarterly Report, Acoustics Laboratory, M.I.T., September 1957.

30. H. L. F. Helmholtz, Die lehre von den Tonempfindungen als Physiologische Grundlage für die Theorie der Musik (Vieweg, Braunschweig, 1863).

31. A. S. House, Analog studies of nasal consonants, J. Speech and Hearing Disord. 22, 190-204 (1957).

32. A. S. House and K. N. Stevens, Auditory testing of a simplified description of vowel articulation, J. Acoust. Soc. Am. 27, 882-887 (1955).

33. A. S. House and K. N. Stevens, Analog studies of the nasalization of vowels, J. Speech and Hearing Disord. 21, 218-232 (1956).

34. A. S. House and K. N. Stevens, Estimation of formant band widths from measurements of transient response of the vocal tract, J. Speech Hearing Res. 1, 309-315 (1958).

35. G. H. Hughes and M. Halle, Spectral properties of fricative consonants, J. Acoust. Soc. Am. 28, 303-310 (1956).

36. S. E. Kasowski, A Speech Sound Synthesizer, S.M. Thesis, Department of Electrical Engineering, M.I.T., 1952.

37. K. Kleinschmidt, The Effect of Quasi-Periodicity on the Naturalness of Synthetic Speech, S.B. Thesis, Department of Electrical Engineering, M.I.T., 1957.

38. W. Lawrence, The synthesis of speech from signals which have a low information rate, Communication Theory, edited by W. Jackson (Butterworths Scientific Publications, London, 1953); see Chapter 34.

39. D. Lewis, Vocal resonance, J. Acoust. Soc. Am. 8, 91-99 (1936).

40. J. C. R. Licklider, M. E. Hawley, and R. A. Walking, Influences of variations in speech intensity and other factors upon the speech spectrum, J. Acoust. Soc. Am. 27, 207 (1955).

41. P. Lieberman, Vowel intonation contours, Quarterly Progress Report, Research Laboratory of Electronics, M.I.T., July 15, 1958, pp. 156-161.

42. C. I. Malme, Detectability of small irregularities in a broadband noise spectrum, Quarterly Progress Report No. 56, Research Laboratory of Electronics, M.I.T., Jan. 15, 1959, pp. 139-142.

43. G. A. Miller, Language and Communication (McGraw-Hill Book Company, Inc., New York, 1951).

44. R. L. Miller, Nature of the vocal cord wave, J. Acoust. Soc. Am. 31, 667-677 (1959).

45. G. E. Peterson, Information Theory: 2. Application of information theory to research in experimental phonetics, J. Speech and Hearing Disord. 17, 175-188 (1952).

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

47. G. Rosen, A Prototype Section for a Dynamic Speech Synthesizer, S.M. Thesis, Department of Electrical Engineering, M.I.T., 1955; AFCRC-TN-55-196, Acoustics Laboratory, M.I.T., 1955.

48. G. O. Russell, The Vowel (Ohio State University Press, Columbus, Ohio, 1928).

49. T. T. Sandel and K. N. Stevens, Studies of the perception of speech-like sounds, Quarterly Progress Report, Research Laboratory of Electronics, M.I.T., July 15, 1958, pp. 155-156.

50. C. E. Shannon, Communication in the presence of noise, Proc. IRE 37, 10-21 (1949).

51. K. N. Stevens, Synthesis of speech by electrical analog devices, J. Audio Eng. Soc. 4, 2-3 (1956).

52. K. N. Stevens, The role of duration in vowel identification, Quarterly Progress Report No. 52, Research Laboratory of Electronics, M.I.T., Jan. 15, 1959, pp. 136-139.

53. K. N. Stevens and A. S. House, Development of a quantitative description of vowel articulation, J. Acoust. Soc. Am. 27, 484-493 (1955).

54. K. N. Stevens, S. Kasowski, and C. G. M. Fant, An electrical analog of the vocal tract, J. Acoust. Soc. Am. 22, 734-742 (1953).

55. K. N. Stevens and K. Nakata, Synthesis and perception of fricative consonants, Quarterly Progress Report, Research Laboratory of Electronics, M.I.T., July 15, 1958, pp. 150-152.

56. J. Q. Stewart, An electrical analog of the vocal organs, Nature 110, 311-312 (1922).

57. T. von Tarnóczy, Resonanzdaten der Vokalresonatoren, Akust. Z. 8, 22-31 (1943).

58. J. Van den Berg, Calculations on a model of the vocal tract for vowel /i/ (meat) and on the larynx, J. Acoust. Soc. Am. 27, 332-338 (1955).

59. W. von Kempelen, Le Mechanisme de la Parole Suivi d'une Description d'une Machine Parlante (Vienna, 1791).

60. R. Willis, On vowel sounds and on reed organ pipes, Trans. Cambridge Phil. Soc. 3, 231-268 (1829).