33
insert
copy

DYNAMIC ANALOG SPEECH SYNTHESIZER

by

GEORGE ROSEN

B.S., University of Pennsylvania
(1952)

S.M., Massachusetts Institute of Technology
(1955)

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF SCIENCE
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
January, 1960

Signature redacted

Signature of Author.....
Department of Electrical Engineering, January 25, 1960

Signature redacted

Certified by.....
Thesis Supervisor

Signature redacted

Accepted by.....
Chairman, Departmental Committee on Graduate Students

NR

33
Thesis

DYNAMIC ANALOG SPEECH SYNTHESIZER
1960

by

GEORGE ROSEN

B.S., University of Pennsylvania
(1952)

S.M., Massachusetts Institute of Technology
(1955)

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
January, 1960

.....
Signature of Author.....
Department of Electrical Engineering, January 25, 1960

.....
Certified by.....
Thesis Supervisor

.....
Accepted by.....
Chairman, Departmental Committee on Graduate Students

DYNAMIC ANALOG SPEECH SYNTHESIZER

by

GEORGE ROSEN

Submitted to the Department of Electrical Engineering on 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 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 using idealized geometries for the vocal tract and piecewise-linear functions to describe 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 about 10 to 30 milliseconds. Experimental evidence is given strongly supporting the viewpoint that the articulatory level provides a natural and economical description of speech. Thus, the dynamic analog is potentially capable of synthesizing highly natural connected speech from signals having a low information rate.

Thesis Supervisor: Kenneth N. Stevens

Title: Associate Professor of Electrical Engineering

ACKNOWLEDGEMENT

The author wishes to express his gratitude to Professor Kenneth Stevens for the inspiration and continued encouragement given throughout this research program. Also, the helpful comments and suggestions of Professors Morris Halle and Walter Rosenblith are very much appreciated. For their stimulation and cooperation, the author wishes to thank the members at the Speech Group, past and present, particularly Dr. Arthur House and Dr. James Flanagan, John Heinz and Charles Malme.

This research would not have been possible without the facilities and support of the MIT Acoustics Laboratory, the MIT Research Laboratory of Electronics and the Cambridge Research Center of the United States Air Force.

TABLE OF CONTENTS

ABSTRACT	Page 2
ACKNOWLEDGMENT	3
TABLE OF CONTENTS	4
LIST OF FIGURES	6
LIST OF TABLES	11
CHAPTER 1 - Introduction	12
History and survey of synthetic speech	12
Applications of synthetic speech	15
Plan of present research	17
CHAPTER 2 - The human vocal tract	19
CHAPTER 3 - Variable circuit elements for the vocal tract analog	29
Over-all arrangement	29
Requirements for transmission line elements	32
Control paths within the transmission line	39
The variable capacitance	44
The variable inductance	53
A complete section	60
CHAPTER 4 - Apparatus for exciting the vocal tract analog and for controlling its configuration	66
Rationale and over-all arrangement	66
Control of configuration	71
Excitation	81
Timing pulse source	99
Trapezoidal function generator	103
The control system and its operation	109
CHAPTER 5 - Study of vowels	118
CHAPTER 6 - Study of fricatives	130
Introduction	130
Study of configuration variables	130
Study of temporal variables	159

CHAPTER 7 - Conclusions and suggestions for further research	176
Knowledge gained by development and operation of the synthesizer	176
Suggestions for further research	180
Integration of the synthesizer with other devices	183
The dynamic analog and the testing of a philosophy	186
APPENDIX	188
Circuit diagrams	188
Instructions to subjects	194
BIOGRAPHICAL NOTE	198
BIBLIOGRAPHY	199

LIST OF FIGURES

FIGURE		PAGE
2.1	Diagram of the human vocal mechanism based on a mid-sagittal section.	20
2.2	Boundaries of the vocal tract during the production of the vowel [e] .	23
2.3	Showing a uniform acoustic tube and three electric network analogs of the tube.	26
3.1	Circuit of the dynamic analog of the vocal tract showing the variable elements of the transmission line.	30
3.2	Configuration for the production of the vowel [u] .	37
3.3	Transfer function of the analog set to the configuration of Fig. 3.2.	38
3.4	Simplified block diagram of a typical section.	40
3.5	Two alternative methods of approximating a distributed system with a lumped system.	42
3.6	Scheme of a Miller effect amplifier.	45
3.7	Detailed block diagram of a typical section.	49
3.8	A ladder attenuator network using tubes as the shunt elements and fixed resistors as the series elements.	51
3.9	Signal amplifier and level stabilizer of the variable capacitance circuit.	54
3.10	Control characteristic of the variable capacitor.	55
3.11	Contours of quality factor (Q) of the variable capacitor.	56
3.12	Control characteristic of a typical saturable inductor.	58

FIGURE	PAGE
3.13 Driver circuit for inductor control winding.	59
3.14 Control characteristic of inductor with driver circuit.	61
3.15 Contours of quality factor (Q) of inductor.	62
3.16 Calibration curve of a typical section.	63
3.17 Contours of quality factor (Q) of a single section.	65
4.1 Sequence of events occurring during the production of a fricative-vowel syllable.	67
4.2 Block diagram of the control system.	70
4.3 Apparatus for generating area control voltages for transmission line sections.	77
4.4 Matrix for the storage of articulatory configurations.	79
4.5 A typical glottal pulse obtained from the buzz generator.	83
4.6 Waveforms within the buzz modulator illustrating its operation.	86
4.7 Block diagram of the buzz generator.	87
4.8 Oscillogram of amplitude control input to buzz generator and of buzz generator output.	89
4.9 Block diagram of symmetrical-clipping noise modulator.	91
4.10 Oscillogram of amplitude control input to noise generator and of noise generator output.	93
4.11 Block diagram of circuit for simulating glottal modulation of noise.	95
4.12 Waveforms within the glottal modulation circuit illustrating its operation.	96

FIGURE	PAGE
4.13 Oscillogram of buzz amplitude trapezoid at input to glottal modulation circuit, output of glottal modulation circuit, and output of noise generator.	97
4.14 Block diagram of timing pulse generator.	101
4.15 Simplified circuit of the trapezoidal function generator.	104
4.16 Functional diagram showing relationship between timing pulse source, function generators, and other units concerned with timing triggers.	110
4.17 Circuit components encountered in typical triggering paths.	111
4.18 Photograph of the synthesizer.	117
5.1 Configurations set on the analog for a study of nine synthetic vowels.	123
6.1 Configurations derived from an idealized model of fricative production for the case of an open constriction.	134
6.2 Configurations used for studying fricatives produced with close constrictions.	135
6.3 Timing patterns describing all stimuli generated for the study of configuration variables.	139
6.4 Response maps for fricatives.	147
6.5 Vocal tract transmission for a fricative judged to be /f/.	149
6.6 Transmission for fricative made with open constriction at position 3 and judged to be /s/.	150
6.7 Transmission for fricative made with open constriction at position 4 and judged to be /s/.	151
6.8 Transmission for fricative made with close constriction and judged to be /θ/.	152

FIGURE		PAGE
6.9	Transmission for fricative made with close constriction at position 5 and judged to be //.	153
6.10	Transmission for fricative made with close constriction at position 6 and judged to be / ζ /.	154
6.11	Transmission for fricative made with open constriction at position 9 and judged to be / χ /.	155
6.12	Spectrogram of / θa / showing downward second-formant transition and spectrogram of / $f a$ / showing upward second-formant transition.	157
6.13	Timing patterns describing stimuli used for studying the effects of changing noise cessation time.	162
6.14	Timing patterns describing stimulus set with differing starting times for configuration change.	163
6.15	Subject responses obtained for study of noise cessation time.	165
6.16	Subject responses to stimuli with differing times of configuration change.	166
6.17	Patterns describing stimuli used in three tests for investigating buzz onset time.	168
6.18	Test scores for two fricatives as a function of buzz onset time.	170
6.19	Subject responses in absolute identification test as a function of buzz onset time.	172
6.20	Individual responses for 7 subjects in absolute identification test.	173
A.1	The nasal analog designed for use with a static analog of the vocal tract.	189
A.2	Resistance network in grid circuit of electronic attenuator.	190

FIGURE	PAGE
A.3 Circuit diagram of control amplifier for a variable section.	191
A.4 Circuit diagram of dc amplifier for driving the configuration storage matrix.	192
A.5 Schematic of trapezoidal function generator.	193

LIST OF TABLES

TABLE		PAGE
5.1	Frequencies of first three formants measured on analog set to configurations shown in Fig. 5.1.	124
5.2	Confusion matrix for vowels.	126
6.1	Combinations of variables used for making listening tests.	137
6.2	Number of /f/ responses.	141
6.3	Number of /s/ responses.	142
6.4	Number of /θ/ responses.	143
6.5	Number of /ʃ/ responses.	144
6.6	Number of /ʒ/ responses.	145
6.7	Number of /x/ responses.	146

CHAPTER 1 INTRODUCTION

1.1 HISTORY AND SURVEY OF SYNTHETIC SPEECH

Known attempts to synthesize speech date as far back as nine centuries.⁴⁸ The first analog of the speech mechanism, constructed as early as 1791 by Von Kempelen,⁵⁹ consisted of various pneumatic devices imitating specific parts of the vocal tract. In 1829, Willis⁶⁰ reported on vowel-like sounds produced by reed organ pipes. More realistic vowels were produced in 1872, when Helmholtz³⁰ excited several of his famous acoustic resonators by tuning forks. However, the first electrical synthesizer was not reported until 1922 when Stewart⁵⁶ 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 tract¹⁸ and the pattern playback of the Haskins group.^{12,13} The resonance or formant coded synthesizer, whose earliest version is due to Lawrence, came soon afterwards, in 1953.³⁸ Two electrical analogs were also reported that year, one by Fant²⁰ and one by Stevens, et al.⁵⁴

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

Let us first 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 and at other times, except silence, it is a continuous spectrum characteristic of random noise. Thus this device explicitly observes distinctions in excitation which 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. However, transfer functions having zeros are needed for the synthesis of most consonants. In all circuits reported to date, 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 handle 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 much 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 the case of consonants, since certain classes of consonants such as stops, semi-vowels 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 can produce.

Because it offers a natural and direct way of building in constraints present in the human speech organism the dynamic analog offers the greatest opportunity for storing information at the synthesizer about how speech is produced and how transitions from one sound to another are made.

1.2 APPLICATIONS OF SYNTHETIC SPEECH

Historically, the first speech synthesizers were mere objects of curiosity but nevertheless they provided some clues about speech production and perception. In this century, the development of synthesizers has been motivated by several applications: to help reduce the channel capacity required for speech communication, to enable computers to speak, and to serve as instruments for speech research.

Using the concepts of information theory, one can make quantitative estimates of the information transmission at various levels of the speech communication process*. 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 some 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

*See, for example, reference 45.

compression system, the dynamic analog synthesizer offers advantages over other devices, both in obtainable degree of compression and in naturalness of its output.

Synthesizers are needed as output members of automata to make possible speech communication between man and machine. Such facility is necessary, for example, in language translation systems accepting and delivering the spoken word. Almost all computing systems presently rely on sight or touch for communication with humans. 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. One function of the synthesizer is 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, synthesizers will be needed to provide the repetitive, phase coherent stimuli often demanded by the 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 and the dynamic analog synthesizer can function in this role.

1.3 PLAN OF 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:

- 1) Development of a dynamic analog of the vocal tract,
- 2) Development of a control system capable of maneuvering the analog through any sequence of two phonetic elements, and
- 3) Evaluation of the machine through formal listening tests.

The formal listening tests were experiments which tested the usefulness of this device as a research instrument and demonstrated its capabilities as a synthesizer. The experimental work had as its objectives the production of certain classes of speech sounds and determination of the precision with which several of the most important control variables must be specified. An additional 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 conceived, is a flexible instrument capable of generating a very large class of speech waveforms. Many studies reported in the literature can be duplicated with this apparatus, but the instrument also extends considerably the area within which quantitative investigations are possible. The experiments reported here cannot be exhaustive; they should be regarded as an initial probing of that area.

CHAPTER 2 THE HUMAN VOCAL TRACT

Figure 2.1 is a schematic 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*, 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 instabilities which thus become acoustic energy sources. They are: 1) vibration of the vocal folds 2) turbulence in the airstream passing through narrow constrictions or over sharp edges and 3) sudden release after complete closure. The amount and character of the energy which enters the system at the sources and which leaves the system as radiated sound in the outside air depends on the coupling provided by the vocal tract.

The function of speech⁴³ is to transmit reports, ideas, commands and emotions from the speaker to the listener. Information generation and transmission are defined respectively as 1) choosing, at the source, from among a set

*The opening between the vocal folds is known as the glottis.

Fig. 2.1 Diagram of the human vocal mechanism based on a mid-sagittal section. The velum (soft palate) can be raised to decouple the nasal cavity from the oral cavity.

of possible messages, and 2) using a transmitter to send something 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 either 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 either directly through the musculature of the vocal folds, or indirectly by producing constrictions and closures in the vocal tract and through pressures on the lungs.

The speaker's role can be examined by a study of his modulation process, i.e., his manner of speech production, or by a study of the signal he has transmitted to the channel, i.e., 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 multi-terminal filter and operate with its transfer functions and source functions.

The vocal cavity is shown in Fig. 2.2. It is a physical system too complicated to analyze exactly. Its walls, consisting of the pharynx, hard and soft palates, tongue, cheeks, and lips are made of tissue having 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-varying coupling to the lungs in synchronism with that excitation. During sounds using turbulent excitation there exist regions of non-laminar compressible fluid flow resulting in a spacially distributed source in the region of turbulence. Sounds produced by closure release exhibit all of the complications of turbulent sounds and furthermore exist under conditions which are non-stationary.

Simplifying assumptions are needed to make the system amenable to analysis or simulation. The system exhibits pressure variations small compared to 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 below about 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

Fig. 2.2 Boundaries of the vocal tract during the production of the vowel [e]. A mid-sagittal section and two transverse sections are shown (from Chiba and Kajiyama, see reference 9)

assumed²¹. When only longitudinal waves exist in a tube, each wavefront is a cross-section thereof. Thus, pressure and likewise particle velocity are identical 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 replaced by a moving piston which 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 summarized by a single dissipative element per source added to the acoustical circuit. 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

*Good fricatives have been synthesized using one source. (See references 29, 54, and Chapter 6) Stop consonants, particularly those with distinct intervals of frication and aspiration may require two sources.

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.

Fant¹⁹, Dunn¹⁸, and Van den Berg⁵⁸ have calculated the first few resonant frequencies of the vocal tract by approximating its shape with two cylinders connected by a constriction. Dunn verified his calculations on an analog based on the same geometry. Steven, et al.⁵⁴ have constructed an analog based on 35 cylinders to approximate more closely the shape of the vocal tract and Fant²³ has built a 45-section analog.

Each analog built to date approximates the shape of the tube representing the vocal tract by a cascade of cylinders, each cylinder having uniform cross-sectional area. Each lossless cylinder is represented in the analog by a Pi, T or L section as in Fig. 2.3. Pressure is identified with voltage and volume velocity is identified with current so that the relations between acoustical and electrical parameters are:

$$L = \frac{\rho \ell}{kA} \quad (2.1)$$

$$C = \frac{kA \ell}{\rho c^2} \quad (2.2)$$

Fig. 2.3 Showing a uniform acoustic tube and three electric network analogs of the tube.

where:

A = area of cylinder

ℓ = length of cylinder

ρ = density of air

c = velocity of sound in air

k = constant depending on impedance level and
choice of units

The analog is seen to be a non-uniform 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 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. Fant²¹ 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:

$$\frac{U_2(s)}{U_1(s)} = H(s) = \prod_K \frac{s_K s_K^*}{(s-s_K)(s-s_K^*)} \quad (2,3)$$

where:

$s = \sigma + j\omega$ (complex frequency)

$U(s)$ = volume velocity transform

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 possesses zeros at frequencies determined by resonances of the cavity behind the source. The poles exhibited by any circuit are the same for all terminals²⁷ provided the generator impedances are not changed from case to case. Thus, given a configuration, the poles for glottal excitation are the same as those for other excitations.

CHAPTER 3 VARIABLE CIRCUIT ELEMENTS FOR THE VOCAL TRACT ANALOG

3.1 OVER-ALL ARRANGEMENT

The dynamic analog synthesizer consists of three parts:

- 1) The lumped electrical transmission line which is the analog of an acoustical system, the vocal tract;
- 2) Sources of excitation for the line;
- 3) A control system for both the line and its excitation.

The transmission line is shown in Fig. 3.1. It is a cascade of one or more of the section configurations shown in Fig. 2.3. For voiced sounds, the transmission line is excited at the left by glottal current pulses of variable frequency and amplitude. For fricatives and other turbulent sounds, a noise voltage generator of controllable amplitude is inserted in series with the transmission line inductor corresponding to the place where the turbulence is produced. For nasals, a nasal analog*, not yet constructed, must be coupled to the vocal tract analog. The line is terminated by an inductor which approximates the radiation impedance seen

*See appendix for the nasal analog circuit of House and Stevens.^{1,31,33}

Fig. 3.1 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. 4.2 Noise is inserted in the line through a transformer which connects the low impedance noise generator in series with the inductor of the appropriate section. The nasal circuit has not yet been built but is expected to be another transmission line less elaborate than the one shown in the figure. Voltages representing radiation from the nose and mouth are summed externally.

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.

The relations between electrical and acoustic quantities, Eq. 2.1 and Eq. 2.2 are restated in Eq. 3.1 and 3.2.

$$A = \frac{\rho c}{k} \sqrt{\frac{C}{L}} \quad (3.1)$$

$$l \neq c \sqrt{LC} \quad (3.2)$$

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 synamic analog can be varied together by means of externally generated electrical 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

Both the physical phenomena for realizing the transmission line elements and their numerical values were chosen through the interplay of three factors: 1) known minimum requirements, 2) previous experience with the static analog at M.I.T.,^{36,54} and the state of the electronic art. 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. Both variable elements and their realization⁴⁷ will be described in later sections.

The first requirement to be considered is that the lumped approximation to a continuous line be valid for frequencies somewhat above the third formant region. This requirement implies a maximum section length of $1\frac{1}{2}$ cm, or about $1/8$ wave length at 3000 cps. The switches of the static analog (0.5 cm sections) were set in groups of 3 to test, by informal listening, the adequacy of the coarse approximation to the shape of the vocal tract provided by the $1\frac{1}{2}$ 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 used in the dynamic analog. It is a compromise between the simplicity of Dunn's¹⁸ 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 to 17 cm^2). Satisfactory vowels had been produced 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) $^{-1}/\text{cm}^2$ of area.

The quality factor (Q) 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 between 3 and 15.

The losses responsible for these values are due predominantly to glottal resistance and to radiation from the mouth, while 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 since 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 varies from at least 9 at 200 cps to at most 250 at 5000 cps depending on area setting. 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 one at 200 cps at minimum inductance were not attainable in the present state of the art. This is not a serious limitation as an analysis using energy functions shows. The definition of quality factor²⁷ at any frequency, not only at resonance, is

$$Q = 2\pi \frac{\text{peak energy stored}}{\text{energy dissipated per cycle}} \quad (3.3)$$

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

$$Q = (a/Q_L + 1/Q_C)^{-1} \quad (3.4)$$

where

Q_L = inductor quality factor

Q_C = capacitor quality factor

$a = 2\pi f (LC)^{1/2} = \text{normalized frequency.}$

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, it was decided to reduce the dissipation of the capacitors just to the point where it would not be the factor which limits the over-all performance of the 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 as a test item. A configuration for this vowel appears in Fig. 3.2 and the corresponding transfer function of the analog in Fig. 3.3. The first number above each peak at frequency f in Fig. 3.3 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*. Thus, the validity of the reasoning based on Eq. 3.3 is substantiated and the sufficiency of the element dissipation requirements verified.

*The values of Q are a bit high. An error in this direction is easily remedied by inserting dissipative elements.

Fig. 3.2 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. The short vertical markers show boundaries between sections set to the same radius. Dot symbol indicates setting of radiation impedance.

Fig. 3.3 Transfer function of the analog set to the configuration of Fig. 3.2. 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 .

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, each having a 100:1 range, is discussed in a later section. Figure 3.4 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:

$$C = e^{a_C - b_C V_{CC}} \quad (3.5)$$

$$L = e^{a_L - b_L V_{CL}} \quad (3.6)$$

$$V_{CL} = d - (b_C / b_L) V_{CC} \quad (3.7)$$

where a_C, b_C, a_L, b_L , and d are constants. The control amplifier maintains a linear relation (Eq. 3.7) between V_{CL} and V_{CC} so that the LC product remains constant. Both V_{CC} and V_{CL} are determined by V_{CS} , the control input voltage that determines

Fig. 3.4 Simplified block diagram of a typical section. The section control input voltage determines the magnitudes of both the variable capacitance and the variable inductance. The electrical length of the section is constrained to remain constant as its area is varied. See Fig. 3.7 for details.

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 non-uniform line, however, can be approximated more closely by a given number of T-sections than by the same number of L-sections. Figure 3.5 shows L- and T-section-approximations to the acoustical system consisting of two short adjacent cylinders of equal length and of areas A_1 and A_2 . The inductances are given by

$$L_1 = k/A_1 \quad (3.8)$$

$$L_2 = k/A_2 \quad (3.9)$$

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

$$L_{\text{series}} = L_1/2 + L_2/2 = k(1/A_1 + 1/A_2)/2 \quad (3.10)$$

connected between C_1 and C_2 . In the general case of a non-uniform line $A_1 \neq A_2$ and hence Eq. 3.10 leads to a better approximation for the series inductance than do Eqs. 3.8 or 3.9.

Fig. 3.5 Two alternative methods of approximating a distributed system with a lumped system. The T-sections yield element values which approximate the distributed system better than do those of L-sections.

A cross coupling of the control voltages of adjacent sections allows us to approximate the desired L_{series} of Eq. 3.10. The input information to each of two adjacent sections constructed as in Fig. 3.4 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* 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. 3.8 and 3.9. 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. 3.10. 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 at negligible increase in complexity, halving the number of inductors otherwise required.

*See Fig. 3.7 for details.

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 eight variable T-sections, each 1.5 cm in length, and the control inputs of the first two variable sections are tied together*. The T-sections are followed by a half-section 0.5 cm in length. The line is completed by two Pi-sections nominally 1.0 cm in length and the radiation inductor. The final capacitor and the radiation inductor are controlled by a single control amplifier. The inductors and capacitors of the Pi-sections are controlled separately so that those sections may be shortened to simulate the shortening of the human vocal tract which occurs at wide mouth openings. In all, 13 voltages are required to specify an articulatory configuration. Another will be needed when the nasal analog is added to specify degree of nasal coupling.

3.4 THE VARIABLE CAPACITANCE

The variable capacitance is realized by means of the Miller effect, shown schematically in Fig. 3.6. In general, Z_1 is given by

*The first of these sections covers a smaller area range to provide a "faring" between the fixed and variable sections.

Fig. 3.6 Scheme of a Miller effect amplifier. The magnitude of C_1 at the input is varied by varying the amplifier gain. C_1 is a small fixed capacitor.

$$Z_1 = \frac{E_1}{I_1} = \frac{R_o + \frac{1}{j\omega C_1}}{1 + K \angle \theta} \quad (3.11)$$

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

$$Z_1 = \frac{1}{j\omega C_1} = \frac{1}{j\omega(1+K)C_1} \quad (3.12)$$

and therefore

$$C_1 = (1 + K) C_1 \quad (3.13)$$

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

$$Q = \frac{1}{\omega R_1 C_1} \quad (3.14)$$

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

$$Q = \frac{1}{2\pi f R_o C_1} ; \quad (3.15)$$

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

Phase shift and finite output resistance exist simultaneously, thus making it desirable to keep phase lag less than 1° and to keep 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 via the transmission line inductors. Parasitic capacitance C_p increases the effective capacitance of the section beyond C_1 , therefore,

$$C_{eff} = C_p + C_1 = C_p + C_1 + KC_1 \quad . \quad (3.16)$$

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_1 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. 3.7 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 input stage. Of these, the electronic attenuator departs the most from conventional circuitry and thus merits some discussion. The requirements for the attenuator are:

- 1) Gain over a 350 to 1 range controlled by a voltage with frequency components from zero to 25 cps;
- 2) Minimum disturbance of amplifier operating points and dc levels by the control voltage (both 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)
- 3) Ability to handle large signal voltages (compared to variations in dc levels) without excessive distortion;
- 4) Less than one degree of phase shift in the useful (100 cps to 3 kcps) band;
- 5) Frequency range wide enough to permit inclusion in a loop having 55 db of gain;
- 6) Gain vs. control voltage to be exponential;

Fig. 3.7 Detailed block diagram of a typical section. The variable capacitance uses a level stabilizer in lieu of filtering to remove control voltage components from the signal path. 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.

- 7) A high input impedance and a low output impedance;
- 8) Stability of calibration and absence of critical adjustments.

The strategy in the development of this attenuator was:

- 1) Utilizing the known effect of control voltage changes on dc levels to generate a cancellation voltage or current;*
- 2) Maintaining the ratio of signal voltage to level change as high as possible;
- 3) Using several attenuation stages to reduce requirements on individual tubes;
- 4) Connecting successive stages in a manner compatible with large changes in dc level;
- 5) Keeping the asymptotic phase shift at low and high frequencies low enough to be suitable for use in a high-gain feedback loop.⁷

The method for realizing this attenuator, shown in Fig. 3.8, is a ladder attenuator using vacuum tubes as the shunt elements and fixed resistors as the series elements. The use of variable r_p instead of variable μ provides high

*There is insufficient separation between the highest control frequency and the lowest signal frequency to permit use of filtering to remove control voltage components from the signal path.

Fig. 3.8 A ladder attenuator network using tubes as the shunt elements and fixed resistors as the series elements.

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, both in regard to drift and in regard to high-frequency phase response. The latter is important because the unity loop gain point is above two megacycles per second in the completed capacitor. The attenuator is useful 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 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 to achieve the same effect. A nonlinear circuit, designated in Fig. 3.7 as "level stabilizer", is connected between the attenuator control source V_{CC} and the output amplifier to stabilize the dc level of the signal applied to the output amplifier against variation due to changes of operating points in the attenuator.

The desired calibration curve (Eq. 3.5) 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 appears in Fig. 3.7 and is described in detail in the appendix. The circuit of the attenuator together with the input and output amplifiers is given in Fig. 3.9. The control characteristic of the capacitor is shown in Fig. 3.10. The capacitance is an exponential function of control voltage within 15 per cent of the mean over a 40 to 1 range in the presence of the expected stray capacitance to ground. Dissipation in the capacitor, described by the contour map in Fig. 3.11, is within the limits specified above.

3.5 THE VARIABLE INDUCTANCE

The variable inductance is realized as a saturable inductor. A saturable inductor has a magnetic circuit consisting of three portions or legs arranged in a "figure eight". A control winding is wound on the inner leg producing a control flux whose return paths are the outer legs. The signal winding is split into two halves each placed on an outer leg and connected in series aiding. As a result the mutual inductances between each half of the signal winding

Fig. 3.9 Signal amplifier and level stabilizer of the variable capacitance circuit. The resistance network, shown in the appendix, is connected to nodes 9 through 12 and applies control voltages which, in turn, are derived from V_{CC} , the capacitor control voltage.

Fig. 3.10 Control characteristic of the variable capacitor.

Fig. 3.11 Contours of quality factor (Q) of the variable capacitor. The minimum required Q is achieved at 3 kc. The Q is well above minimum at other frequencies.

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 which lie 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. 3.12.

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 such that the inductance of the inductor signal winding is an exponential function of V_{CL} , the control input voltage to the driver. Figure 3.12 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. In addition the driver output impedance must be high if rapid response is to be obtained. A triode with a nonlinear cathode resistor meets all the stated requirements. The backward gain in the driver circuit, Fig. 3.1.3, is a monotonic function of output current; it is greatest at

Fig. 3.12 Control characteristic of a typical saturable inductor.

Fig. 3.13 Driver circuit for inductor control winding.

low output current and least at high output current. Thus the driver circuit exhibits the greatest calibration stability 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. 3.14. The control characteristic of the inductor-driver combination is exponential over a range of 250 to 1 with a tolerance of 10 per cent. The contours of constant quality factor (Q) of the inductor are given in Fig. 3.15.

3.6 A COMPLETE SECTION

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. 3.7 and both are controlled by V_{CS} . The calibration curve relating section area to V_{CS} is given in Fig. 3.16. The effect of summing control voltages for the inductor is eliminated from consideration by setting V_{CC} (see Fig. 3.7) 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 in as a simple resonance circuit. The definition of Q , Eq. 3.13, is used without simplifying assumptions to derive an expression more

Fig. 3.14 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. 3.15 Contours of quality factor (Q) of inductor. The ordinate is given in V_{CC} to make possible direct comparison with Figs. 3.11 and 3.17. The inductor control voltage V_{CL} can be obtained from V_{CC} by use of Eq. 3.7.

Fig. 3.16 Calibration curve of a typical section. The effective cross-sectional area is plotted as a function of capacitor control voltage.

general than Eq. 3.4. Upon substituting the data of Figs. 3.11 and 3.15 therein, the contour map of Fig. 3.17 is obtained.

If the section is treated again as above, 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 to 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 as well with time-varying control inputs as it is realized in the steady state.

Fig. 3.17 Contours of quality factor (Q) of a single section.

CHAPTER 4 APPARATUS FOR EXCITING THE VOCAL TRACT ANALOG AND FOR CONTROLLING ITS CONFIGURATION

4.1 RATIONALE AND OVER-ALL ARRANGEMENT

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. 4.1. At the beginning, the speaker has his vocal tract in a constricted configuration appropriate for the fricative and no sound is generated. He then produces the fricative. The amplitude of this noise-like 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

Fig. 4.1 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. Baselines for amplitude waveforms represent zero excitation. For the graph labeled "configuration", the 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.

is once again silent. Since a human will rarely speak in a monotone, some variation of the vocal frequency is indicated.

Description of the fricative-vowel syllable requires a model having the following features: 1) a single articulatory transition, 2) a single occurrence of inflected glottal excitation and 3) a single occurrence of noise excitation. This model will serve for any two-element sequence if one makes the following generalizations:

- 1) Events are allowed to occur in any order, in any spacing;
- 2) Time parameters, such as durations and rates, have wide ranges;
- 3) Specification of articulatory configuration include degree of nasal coupling, if any;
- 4) Single shock pulse excitation, in addition to 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 which seem to have several noise sources can also be investigated. As presently realized, 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 structure is not too complex. The system can also control POVO or terminal analog synthesizers.⁵² When used with ancillary circuits, it can generate stimuli useful in psychoacoustics.³

The general arrangement of the synthesizer, including the control system, is shown in Fig. 4.2. 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 themselves require control signals. These signals are furnished by waveform generators, column 2. Their outputs closely resemble the time functions portrayed in Fig. 4.1. 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.

Fig. 4.2 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 pre-set 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 together act 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 signified by the heavy arrow.

4.2 CONTROL OF CONFIGURATION

The purpose of the apparatus to be described here is to generate the configuration signals (Fig. 4.2) which are applied to the section area control inputs shown in Fig. 3.1. During the production of a typical two-element sequence each section passes through three states:

- 1) An initial quiescent state in which the section has an area appropriate for the initial vowel or consonant,
- 2) A transition state, and
- 3) A final quiescent state in which the section has an area appropriate for the final vowel or consonant.

As discussed previously, the transitions for all sections occur in the same time interval.

Let v_j be a typical waveform which 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 which lasts 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 having fixed quiescent values, 0 and 1. Let

$$f_1(t, t_1, T) = \begin{cases} 0 & ; t \leq t_1 \\ \frac{t-t_1}{T} & ; t_1 \leq t \leq t_1+T \\ 1 & ; t \geq t_1+T \end{cases} \quad (4.1)$$

and

$$f_2(t, t_1, T) = \begin{cases} 1 & ; t \leq t_1 \\ \frac{T-t+t_1}{T} & ; t_1 \leq t \leq t_1+T \\ 0 & ; t \geq t_1+T \end{cases} \quad (4.2)$$

Then

$$\begin{aligned} v_j(t, t_1, T) &= b_{1,j} f_1 + b_{2,j} f_2 = b_{1,j} f_1 + b_{2,j} (1 - f_1) \\ &= b_{2,j} + (b_{1,j} - b_{2,j}) f_1 \end{aligned} \quad (4.3)$$

where

$$\left. \begin{aligned} 0 \leq b_{1,j} &\leq 1 \\ 0 \leq b_{2,j} &\leq 1 \end{aligned} \right\} \quad (4.4)$$

That is

$$v_j(t, t_1, T) = \begin{cases} b_{2,j} & ; t \leq t_1 \\ b_{2,j} + (b_{1,j} - b_{2,j}) \frac{t - t_1}{T} & ; t_1 \leq t \leq t_1 + T \\ b_{1,j} & ; t \geq t_1 + T \end{cases} \quad (4.5)$$

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

$$\begin{bmatrix} f_1, f_2 \end{bmatrix} \begin{bmatrix} b_{1,1} & \dots & b_{1,14} \\ b_{2,1} & \dots & b_{2,14} \end{bmatrix} = \begin{bmatrix} v_1, \dots, v_{14} \end{bmatrix} \quad (4.6)$$

All functions in Eq. 4.6 are to be considered normalized so that 0 represents the value of v_j which sets the area* of a

*Because section calibrations do not precisely coincide, 0 stands for a least value of V_{CS} among all sections and 1 stands for a greatest value. The normalization is for the set, not an individual section. 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.

given section to its minimum value and so that l 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.

The coefficients $b_{1,j}$ are physically realized as potentiometer settings. Changing these settings is tedious and time consuming and restricts the comparisons one can make during informal listening because of the listener's limited span of memory. It is desirable to store several configurations in a large matrix and to select the required pair by switching. Let $[s]_{2,6}$ be a 2 by 6 switching matrix*, all of whose elements are equal to zero except for two, one from each row, which are equal to one. Let $[a]_{6,14}$ be a 6 by 14 matrix for storing 6 configurations. Then

$$\begin{bmatrix} b_{1,1}, \dots, b_{1,14} \\ b_{2,1}, \dots, b_{2,14} \end{bmatrix} = \begin{bmatrix} s_{1,1}, \dots, s_{1,6} \\ s_{2,1}, \dots, s_{2,6} \end{bmatrix} \begin{bmatrix} a_{1,1}, \dots, a_{1,14} \\ a_{6,1}, \dots, a_{6,14} \end{bmatrix} \quad (4.7)$$

*The symbol $[x]_{m,n}$ denotes an m by n matrix with typical element x .

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,

$$s_{1,5} = 1, \quad (4.8)$$

$$s_{2,3} = 1, \quad (4.9)$$

and all other $s_{i,k} = 0$. Then

$$b_{1,j} = a_{5,j} \quad j = 1, \dots, 14 \quad (4.10)$$

and

$$b_{2,j} = a_{3,j} \quad j = 1, \dots, 14. \quad (4.11)$$

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

$$[f_1, f_2] = [f_1[1, -1] + [0, 1]] \quad (4.12)$$

and Eq. 4.6 can be rewritten as

$$[f]_{1,2}[s]_{2,6}[a]_{6,14} = [v]_{1,14}. \quad (4.13)$$

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

matrix coefficients:

$$\begin{aligned}
 & [v_1, \dots, v_{14}] \\
 &= [f_1[1, -1] + [0, 1]] \begin{bmatrix} s_{1,1}, \dots, s_{1,6} \\ s_{2,1}, \dots, s_{2,6} \end{bmatrix} \begin{bmatrix} a_{1,1}, \dots, a_{1,14} \\ a_{6,1}, \dots, a_{6,14} \end{bmatrix} \quad (4.14)
 \end{aligned}$$

Equation 4.14 is implemented directly by the arrangement shown in Fig. 4.3. The input, f_1 , is obtained from a function generator to be described in a later section. The input is applied to the push-pull dc amplifier yielding $[f_1, f_2]$ such that $f_1=0$ corresponds to +300 volts. 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. 4.14. Each v signal, being 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 low-pass filters smooth the configuration signals. The filter should have dc gain which is reasonably

Fig. 4.3 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.

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 and 0 to 70 ma output and 5 ohms internal impedance. The circuit of the matrix [a] appears in Fig. 4.4. The maximum gain from any given row input to any given column output is one sixth 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 in use at present, 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 regard 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.

Fig. 4.4 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.

The [a] and [s] matrices can be used in two ways, 1) to store the configurations for a number of alternative two-element sequences and 2) to store configurations needed to produce longer sequences. The study of the syllable [sa] illustrates the first mode of operation. Five alternative [s] configurations may be stored, one per row, with the sixth row used to store the [a] 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 [bʔl]. The [b] configuration is stored in row 1, the [ʔ] configuration in row 2, and that of [l] in row 3. Let the switching matrix and power amplifier apply a normalized input of 1 to row 1, anormalized input of 0 to row 2 and a ground to row 3. Assume that the input to the amplifier is a full trapezoid i.e., 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 [ʔ]) it is 1. Then the conditions at the input to matrix [a] are these:

row 1..... 0 applied,
 row 2..... 1 applied, and
 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 [ɛ]. The switching matrix is operated by an auxiliary switching device which is triggered by the pulse source shown in column 1 of Fig. 4.2.

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 appropriate to the production of the associated pair of phonetic elements.

4.3 EXCITATION

The elements belonging to column 3 of Fig. 4.2 are described in this section. Together they provide the buzz and noise excitation for the analog.

4.31 BUZZ EXCITATION

During the production of voiced sounds by the human vocal tract, the vocal folds vibrate, causing periodic changes in the size of the glottis, the opening between the vocal folds. Puffs of air are thus released into the vocal tract giving

rise to a volume velocity waveform* such as depicted in Fig. 4.5. 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

In the synthesizer, the buzz amplitude envelope and the buzz frequency pattern are approximated by trapezoidal functions as in Fig. 4.1. 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 possible study of inflection patterns and, in phonetic

*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 enhanced if these effects could be included to help make the buzz generator an artificial larynx. Another possible refinement is the addition of slight randomness to the glottal frequency³⁷ thereby rendering the buzz less precise and "mechanical".

Fig. 4.5 A typical glottal pulse obtained from the buzz generator.

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. In addition, each pulse must have the form of a typical glottal pulse, such as is shown in Fig. 4.5. 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 tube cut-off. 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 is done, for example, in AM transmitters. Thump

is troublesome because, when applied to the transmission line, it can drive the Miller capacitors out of their operating regions. In addition, thump gives rise to clicks which impart an undesired consonantal quality to the synthesizer output.

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 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 depicted in Fig. 4.6 and the block diagram in Fig. 4.7. The multivibrator generates a modified square wave whose frequency changes in response to a control voltage. The square wave drives a monostable trigger circuit which generates a train of very narrow pulses (50 μ sec. duration). Assuming that buzz frequency is constant, the train appears as in Fig. 4.6a. 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 which, 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 this modulator, the "audio" is larger

Fig. 4.6 Waveforms within the buzz modulator illustrating its operation. Each part of this figure shows the waveform at a node in the block diagram, Fig. 4.7 and is labeled with a corresponding letter. Cross-hatching in part c shows the window between upper and lower clamping levels. Pulses must be adjusted to extend above and below the window.

Fig. 4.7 Block diagram of the buzz generator. Voltages at lettered nodes correspond to similarly labeled waveforms in Fig. 4.6. The window in that figure 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 via the additional input.

than the control voltage. The waveform in d, the resultant, is applied to the shaping network via a buffer amplifier. The shaping network has an impulse response like the wave in Fig. 4.5 and its output, e, is sent to the output amplifier. Figure 4.8 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 is seen to be quite linearly related to the control voltage. The decreasing output amplitude in c is traceable to the low-pass 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.

4.32 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 using 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

Fig. 4.8 a) Oscillogram of amplitude control input to buzz generator. Over-all duration of trapezoid is 450 msec. Onset and decay times each are 50 msec. b) Output of generator with buzz frequency constant at 100 cps. Note stability of baseline 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.

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

$$y = \sum_{1}^n x_1 \quad (4.15)$$

and if the x_1 are statistically independent and are distributed with finite mean and variance, then the distribution of y approaches the Gaussian as n approaches infinity, regardless of the exact form of the x_1 distributions. Figure 4.9 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 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.

Fig. 4.9 Block diagram of symmetrical-clipping noise modulator. The lowpass filter acts as a summing device for noise samples.

Hence this voltage controls the amplitude scale of the clipper output.

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 twenty 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. 4.10 is free from thump. A modulator with many desirable characteristics has thus been achieved by taking advantage of the properties of the signal.

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

Fig. 4.10 a) Oscillogram of amplitude control input to noise generator. The over-all duration of the trapezoid is 450 msec. Onset and decay times are 250 msec and 50 msec respectively. b) Output of noise generator. Note stability of baseline and linearity of modulation judged by visual estimation.

described uses the buzz amplitude control voltage to control the amount of glottal component added, introducing no glottal component at all during unvoiced intervals. The circuit can be permanently connected to the noise generator, and the synthesizer thereby becomes a more refined model of human speech production.

Figure 4.11 is a block diagram of the circuit which, like the modulators, uses the clipping window technique. All waveforms appearing in Fig. 4.12 pertain to the case of partially overlapping buzz and noise, the buzz occurring first. Parts a and b respectively show the buzz and noise trapeziod inputs while 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 trapeziod, b. The upper clamp, b, together with the lower clamp, d, form a window, e, through which is seen a rectangular wave of glottal frequency. 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 corresponding noise output appear in Fig. 4.13 b and c. The circuit, in its present form,

Fig. 4.11 Block diagram of circuit for simulating glottal modulation of noise. Waveforms existing at lettered nodes are shown in Fig. 4.12. The noise modulator is controlled by the output of this circuit rather than directly by a trapezoid generator.

Fig. 4.12 Waveforms within the glottal modulation circuit illustrating its operation. The time functions shown in this figure describe voltages or clamping levels at similarly labeled nodes or blocks of Fig. 4.11.

Fig. 4.13 a) Oscillogram of buzz amplitude trapezoid at input to glottal modulation circuit. b) Output of glottal modulation circuit. Upper envelope of output is same as noise amplitude trapezoid at input to this circuit. c) Output of noise at glottal rate and decrease of modulation of noise with decrease of buzz amplitude.

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 additional 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. 4.7) allows for synchronization from other circuits, such as the timing trigger unit of Fig. 4.2. Voiced stimuli can be made phase coherent by deriving glottal timing from the clock pulse source which is part of that unit and which 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. 4.7)

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 which are not marred by switching transients. Such a wave train is useful in psychophysical experiments.

4.4 TIMING PULSE SOURCE

The digital portion of the synthesizer, represented by column 1 of Fig. 4.2, generates pulses which 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 eight channels are normally assigned to the trapezoid generators. Four generators, one for each trapezoid appearing in Fig. 4.1, are used and each requires a trigger to initiate an upward ramp and another trigger to initiate a downward ramp. Two channels are normally

*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.

connected to an oscilloscope for monitoring and time measurement and others are associated with auxiliary switching functions.

Figure 4.14 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 five milliseconds. 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 of which functions as a "scale of 10" counter. The counters can be stepped as quickly as two milliseconds per pulse and as slowly as desired.

The counter tubes are cold cathode gas discharge devices.⁴ Each tube is filled with neon and each 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 placed so that two guides appear between each pair of cathodes. All 10 guides adjacent clockwise to a cathode are connected together to form ring 1; the other 10 forming ring 2. The counting action is

Fig. 4.14 Block diagram of timing pulse generator. The clock may be driven from an internal oscillator, from an external oscillator, or from the buzz generator. The latter mode is used for phase coherent stimuli requiring buzz frequency variation.

obtained by stepping the glow from one cathode to the next higher numbered 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 $n+1^{\text{st}}$ cathode. One and only one output of the counter tube is high at any time except for 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 one and so the tens counter steps once for every 10 units counts. Each tens pulse is ten 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. 4.14 uses a single logical element, the coincidence gate ("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 and its use 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 via 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 4.15 illustrates the function-generating scheme; the operation can best be understood by tracing the sequence of events occurring during the production of a trapezoid. Initially, 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 , through R_D but it cannot charge negatively because of the lower clamp

Fig. 4.15 Simplified circuit of the trapezoidal function generator. The complete circuit, given in the appendix, uses tubes to perform the functions ascribed to switch S.

at zero. When the switch is thrown to its upper position, the capacitor is connected to a positive voltage, E_P , via the charging resistor, R_C , and C begins to charge, generating an upgoing ramp. In the actual generator, switching is initiated by receipt of an "up trigger" pulse. Charging continues until V_C equals E_C , the upper clamping voltage. After that instant 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 , generating 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

$$E_P = kE_C \quad (4.16)$$

and

$$E_N = (1-k)E_C \quad (4.17)$$

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 one millisecond to one second in one millisecond 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 which lies between the clamping limits becomes smaller, i.e., 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 percent 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 trapeziod: 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 which 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 causing the generator to produce a succession of trapezoids so that the synthesis of a syllable, such as /sæ/, which requires two distinct intervals of noise excitation is readily accomplished. Polysyllabic utterances are similarly handled. The nonlinear, time-varying 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. 4.15. Initially, the switching short circuits J and J' and the first two 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 . To be unnoticed, switching of resistors must occur when no transition is in progress, otherwise a

transition having several linear segments is generated*.

Thus one can conceive of a control system which uses no additional function generators but which 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 being determined by its associated generator. The latter mode of operation has been used to produce stimuli for a psychophysical study, a study of the perception of juxtaposed noise bursts.⁴⁹

*Consider, for example, the upper quiescent interval. In that case, 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 (in theory) keep the upper clamp operating.

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 appearing at the left in Fig. 4.16 was shown previously in Fig. 4.14.* Each output channel of the coincidence panel provides one pulse per cycle to the triggered units shown at the right. Each function generator and auxiliary switching cell uses a flip-flop input stage so that both trigger in the same way. Let us consider first 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. 4.16 and one of those may be chosen as typical. Along such a path, the signal will travel from a coincidence gate, through an inverter, through a patch cord, through a mixer gate, and into a flip-flop. The circuitry typically encountered is shown in Fig. 4.17. Referring to this figure, assume that the flip-flop is initially in its "zero" state (T_2 conducting and T_1 cut off), that V_1 , V_2 , D_0 , D_1 and D_2 are non-conducting and that the program input to gate C_1 is not connected. Then let positive pulses be simultaneously

*The inverter tubes shown in Fig. 4.16 are circuit details rather than functional entities and so were omitted from Fig. 4.14.

Fig. 4.16 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 via time selection matrix.

Fig. 4.17 Circuit components encountered in typical triggering paths.

applied to the tens and units input of gate* C_1 . 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 cut-off. 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 causing conduction through V_2 and D_2 or by a closure of switch S causing 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 and 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 then each inverter must be associated with only one flip-flop. By symmetry, triggering at the reset input is similarly effected. Multiple inputs

*The coincidence gate uses three crystal diodes and works with pulses having a "high" value of about +25 volts and a "low" value of +5 volts.

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 states when the key is respectively in its upper or lower positions. 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 malfunctions. 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 they would:

- 1) Alter the characteristics of the excitation circuits;
- 2) Change parameters of the trapezoidal function generators;
- 3) Sequence switching matrix S for multiple articulatory transitions;
- 4) Change the point at which noise is inserted into the transmission line;

- 5) Assist in program selection; and
- 6) Alter characteristics of appurtenances used to generate stimuli for psychological and psychophysical experiments.

The operation of the program gating can be illustrated with the aid of a simple example and Fig. 4.16. The figure shows an auxiliary switching cell with three 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, depending on an external program choice. External program selection switching 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 made to connect channels A and B to separate inputs of the mixer gate used to reset the flip-flop in the cell. Then, 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 program input selection devices may operate with switching paths or voltage levels and 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. Before this facility was available, informal listening was hindered because appreciable delay occurred after one stimulus while the second was being set. The listener's span of memory was exceeded, making the desired comparisons impossible.

Research with the dynamic analog involves the judgement 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: start the recorder, allow the tape to come up to speed, record an identifying marker, initiate the stimulus to be recorded, provide for pre-set spacings between utterances, and stop 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 has been used for studies of inflection patterns,⁴¹ for studies of learning in multidimensional auditory displays,³ for studies of the detectability of irregularities in noise spectra,⁴² for the study of the effect of duration on vowel identification,⁵² and for other investigations. This device, shown in Fig. 4.18, 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.

Fig. 4.18 Photograph of the synthesizer. Principal occupants of the racks, from left to right, are: 1) function generators, 2) timer with time selection matrix, 3) buzz and noise generators, part of configuration matrix, 4) configuration matrix, 5) transmission line, and 6) power supplies for transmission line.

CHAPTER 5 STUDY OF VOWELS

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

- 1) Use of the synthesizer for research in speech production and perception,
- 2) Acquisition of data needed to couple the synthesizer to other elements of communication or computer systems, and
- 3) Building of a vocabulary for the machine.

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

In the formal listening program, the human subjects function in two roles. First, they can help us to evaluate the quality and intelligibility of the output of the synthesizer. Secondly, their linguistic behavior can be used to determine the articulatory and acoustic representations 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 listeners, and the linguistic backgrounds of the listeners.

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 owing to 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 early attention 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 resulting vowels 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 nine vocal tract configurations appropriate for the production of the

following vowels of American English: /ɪ I ɛ æ ʌ ʊ U u ʌ/.*

This set typifies the American English vowel articulations, and provides the materials for approximating rarely used vowels and for fabricating sounds 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 were based on a previous study made on a static analog by House and Stevens.³² They used three parameters to characterize vowel configurations 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 portions; a

*The configurations set on the machine are merely physical descriptions relating to sounds. The choice of configurations, however, reflects the phonemic structure of the experimenter. Responses of subjects in the listening tests described in this report are phonemic. Many of the symbols in the text can be taken in either a phonetic or phonemic sense, but, in each case where a distinction must be made, the meaning is clear from the context. Therefore, all subsequent phonetic and phonemic symbols will be written between slashes.

throat section, a parabolic representation of the tongue, and a cylindrical representation of the lips. The three parameters are; r_0 , the minimum radius of the tongue constriction, d_0 , 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 on the one hand and nine vowel categories plus a null category on the other hand.

The configurations set of the static analog can not 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. Secondly, the ranges of section areas for the two machines do not completely coincide. Finally, sections of the static analog are removed by switching to shorten the lip cylinder to help provide an A/l range of 170 to 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, a new set of configurations was found empirically and evaluated by informal listening and by measuring formant frequencies. The curves given by Stevens and House⁵³ 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. 5.1 and the corresponding set of formant frequencies is given in Table 5.1. The vocal tract transmission curve and formant bandwidths for /u/, shown in Fig. 3.3, 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 milliseconds and 330 milliseconds. The buzz rise and decay times were 40 milliseconds and buzz frequency changed

Fig. 5.1 Configurations set on the analog for a study of nine synthetic vowels. The variable-length feature of the lip sections has been used. Short vertical markers appear between sections set to the same area. Dot symbols at right denote settings of the radiation impedance.

Table 5.1 Frequencies of first three formants measured on analog set to configurations shown in Fig. 5.1.

CONFIG.	F ₁	F ₂	F ₃
i	270	2250	2950
I	350	2020	2900
ε	555	1730	2250
æ	700	1550	2250
α	725	1230	2650
ɔ	510	850	2650
U	400	1100	2750
u	270	840	2300
Λ	560	1300	2420

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 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 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 was kept constant so that stimulus intensities correspond roughly to the case of constant vocal effort (assuming constant vocal effort corresponds to constant volume velocity at the glottis). The time between stimuli was 5 seconds. The subjects were given spoken instructions and those unfamiliar with the vowel symbols wrote key words on their answer sheets before the test started. Five subjects were experienced; five were naive.

The correspondence between the stimuli, which are configurations, and the responses, which are phonetic judgments, is given by the confusion matrix, Table 5.2.

Table 5.2 Confusion matrix for vowels.

STIMULI		RESPONSES								
CONFIG.	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

Pertaining to a panel of 9 listeners it 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 Barney⁴⁶ for natural vowels, but several distinctions must be considered when comparing these figures. Each natural vowel was imbedded in a syllable that began with /h/ and ended in /d/ while each synthetic vowel was produced in isolation with a time-invariant configuration and time-invariant formants. 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 done with available subjects with little screening for linguistic homogeneity.

The vowels /i/ and /u/, which are the extremities of the vowel triangle, received the highest articulation scores both in the Peterson and Barney experiment and in this

*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 /ɜ/.

experiment. Here they were judged correctly unanimously in both long and short versions. These results, together with a /ʌ/ response to a short /a/ stimulus, are consistent with Stevens' findings on the effect of duration on vowel identification.⁵² In Table 5.2, a long /U/ is confused with a /u/ perhaps because the formants for /U/, at 400 cps and 1100 cps are not unambiguous; c.f., the area of the F_1F_2 plane assigned to /u/ by the Peterson and Barney data. The /U/ responses to the long /ʌ/ stimulus may be similarly explained because the first formant of the /ʌ/ stimulus is quite low, bringing the formants near the /U/ region. The data indicate a need for enlarging the set of formant frequencies beyond that 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 shown to be 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 two durations. By examination of Table 5.2, one can choose 9 items such that duration cues reinforce configuration cues, and thus each configuration is associated with only one duration. Instead of performing another listening test, let us estimate a 9 by 9 confusion matrix for this set by transferring half of the entries from Table 5.2 to the new matrix. The predicted articulation score found in this manner is approximately 98 per cent.

CHAPTER 6 STUDY OF FRICATIVES

6.1 INTRODUCTION

For the next series of experiments, we chose the study of fricative-vowel syllables. Synthesis of such sounds uses the machine's ability to make transitions and typifies the work for which the control system was designed. This task seems well suited as an initial problem because it is more challenging than the synthesis of diphthongs 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 to be discussed here consisted of a study of configuration variables 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 how 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 which is physically or linguistically meaningful. A physical definition based on acoustic mass was adopted. An acoustic mass is the single lumped element which 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-sectional 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 the configurations used in this study.

The first variable, position of constriction, was given nine different values. This variable serves to identify the sections (and one half-section*) at which a constriction may be formed. The values establish an ordering in which position one is furthest 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 ninth position were found to be unrealistic and of no interest. The fricatives occurring in English are produced near the front of the mouth using the lips, teeth, and the front of the tongue. Post-palatal and velar fricatives, produced with the mid and rear portions of the tongue, occur in many languages, including German and Russian. The latter fricatives were included partly for completeness in the study of the position variable and partly as an illustration of the capabilities of the analog which make it

*This section serves as a transition between the two Pi-sections at the front of the tract and the T-sections used in the remainder of the tract. A constriction formed here is said to be in position "three".

possible to produce sounds in a great variety of 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 with areas respectively equal to 0.4 cm^2 and 0.13 cm^2 . 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 then 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 6.1 and 6.2 show the resulting configurations for the open and close cases respectively.

The third variable in this experiment is the location at which noise was 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

Fig. 6.1 Configurations derived from an idealized model of fricative production for the case of an open constriction. Numbers at right refer to position of constriction.

Fig. 6.2 Configurations used for studying fricatives produced with close constrictions.

of the section at the given location by the secondary of a 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 and the set specified in Table 6.1 was chosen for formal testing.

The stimuli for the listening test were recorded in randomized order so that each item on Table 6.1 appeared three times, giving a total of 156 judgements per listener per test. All of the stimuli were consonant-vowel syllables, always with the vowel / α /. The relative levels of buzz and noise excitation* and the temporal relations between buzz

*The levels were set so that 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 volts at the secondary of the noise insertion transformer. There is 470 K between the buzz generator and the input to the transmission line.

Table 6.1 Crosses specify combinations of variables used for making listening tests.

LOCATION OF NOISE	POSITION OF CONSTRICTION								
	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								

amplitude, buzz frequency, noise amplitude and configuration change were determined by informal listening. The temporal relations shown in Fig. 6.3 were used for all stimuli.

Successive stimuli were spaced 6 seconds apart. Marker tones were recorded after every tenth 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, /, f, θ, ɸ, χ/. Two tests, each on different days, were given for each degree of constriction; one test was forced choice, the other non-forced.

The stimuli included many sounds resembling the velar fricatives /ɸ/ and /χ/ which do not occur in English and which were not in the vocabulary of most of the subjects. All subjects were given training and examination in those sounds. A table of random numbers was used to prepare a list of twenty items, each of which was either the sequence /Iɸ/, /αχ/ 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 being pronounced. They were then given a pre-test in which they were asked to identify twenty randomized recorded occurrences of the two velar fricatives spoken in isolation by the same speaker. Then the main test with synthetic stimuli was

Fig. 6.3 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 / α /.

administered and followed by a post-test of length equal to that of the pre-test. The over-all pre-test and post-test 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 listener responses are shown in Table 6.2 through Table 6.7, one table being provided for each phonetic category. Each table has four parts, one for each listening test. All tables use 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. 6.4, showing responses at 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 /ɣ/ has not yet been obtained.

Table 6.2 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 OF NOISE		POSITION OF CONSTRICTION								
		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 6.3 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 6.4 Number of / Θ / responses.

RELATIVE LOCATION OF NOISE		POSITION OF CONSTRICTION								
		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	-	-	-	-	58	18	7	3	-
	1	-	-	-	-	3	18	9	2	3
	2	2	-	-	-	-	11	2	3	3
	3				-	-	7	-	3	4

Table 6.5 Number of /s/ 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 6.6 Number of / ζ / responses.

RELATIVE LOCATION OF NOISE		POSITION OF CONSTRICTION								
		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 6.7 Number of / χ / 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. 6.4 Response maps for fricatives. Maps corresponding to open constrictions are at the left and those for non-forced instructions are at the top. Unruled area at upper right corner of each map corresponds to physically impossible combinations. Symbol conventions are: blank, unused combination of no interest; dash, response level below 50 per cent; small phonetic symbol, response level 50 to 75 per cent; large phonetic symbol, response level above 75 per cent; plus sign, response level for indicated category above 90 per cent.

The acoustic properties of the fricatives can be studied in terms of their spectra. A specimen for each fricative was selected by examining Tables 6.2 through 6.7 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. 6.5 through 6.11 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. The series $/\int, s, f/$ corresponds to energy concentrated at successively higher frequencies, both in the samples of natural speech and in the fricatives produced by the analog. Details of natural fricative spectra vary considerably from one utterance to the next. Different versions of the same synthetic fricative similarly differ in detail as may be seen by comparing Fig. 6.6 and Fig. 6.7. Both curves describe

Fig. 6.5 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. 6.6 Transmission for fricative made with open constriction at position 3 and judged to be /s/. Excitation is inserted at location 1.

Fig. 6.7 Transmission for fricative made with open constriction at position 4 and judged to be /s/. Excitation is inserted at location 3.

Fig. 6.8 Transmission for fricative made with close constriction at position 4 and judged to be / θ /. Excitation is inserted at location 4.

Fig. 6.9 Transmission for fricative made with close constriction at position 5 and judged to be / $\int$ /. Excitation is inserted at location 2.

Fig. 6.10 Transmission for fricative made with close constriction at position 6 and judged to be / ζ /. Excitation is inserted at location 6.

Fig. 6.11 Transmission for fricative made with open constriction at position 9 and judged to be / χ /. Excitation is inserted at location 11.

stimuli yielding /s/ responses above the 95 per cent level. Using a tape splicing technique, Harris²⁸ was able to join the noise burst of one natural fricative-vowel syllable to the vocalic portion of another. She found that /θ/ 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 /θ/ and /f/ in Fig. 6.5 and Fig. 6.8 are quite similar but, in Fig. 6.12, the spectrograms of the corresponding syllables show contrasting second formant transitions. Stevens and Nakata⁵⁵ 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 having high-frequency hearing loss and for those listening to most speech transmission apparatus, even high-quality earphones. If this peak is ignored, the spectrum, Fig. 6.5, 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 /χ/, Fig. 6.11, is comparable to those obtained by Fant using a Russian subject.

(a)

(b)

Fig. 6.12 a) Spectrogram of / θa / 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 gradually bringing more of the spectral peak above marking threshold of the spectrograph. b) Spectrogram of / $f a$ / showing upward second-formant transition. Noise spectrum for / f / has no pronounced peak corresponding to the one seen in part (a). With the pre-emphasis used in making the spectrogram, no portion of the fricative spectrum rises above marking threshold.

Spectra produced by the dynamic analog contrast in one respect with those seen in the literature. Each of the former 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. 6.9. In this case, the noise is inserted much 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. 3.3 which shows a vowel transmission curve with a quite narrow first formant.

6.3 STUDY OF TEMPORAL VARIABLES

Stimuli can easily be modified by changing parameters of the control system and, in the case of 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 in cognizance of one's goals. The general purposes of these experiments are to find those values of temporal variables typical of fricative production, to determine how critical those 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 known to have the same phonetic value. This can be done by means of the method of paired comparisons²⁶ 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:

- 1) Time at which noise begins to decay,
- 2) Time at which configuration change begins, and
- 3) Time of buzz onset.

Data from the configuration study was examined and an item was found which received unanimous / $\int$ / 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 the fifth position, and noise at location 2 (relative location -3). The context was the same as before, CV syllables always

*An absolute identification test could also be devised for the same purpose. Such a test could use stimuli of several fricative categories with the subjects making non-forced responses. Poorer stimuli would receive more null responses and would probably receive a set of responses spread over several phonetic categories. The need for populating several phonetic categories and for many replications of each stimulus renders such a test much more cumbersome than the method chosen. The comparison technique places stimuli in sharp contrast before the listener and therefore can be expected to make the values of the physical variables appear critical.

with the vowel / α /. Both initial and final configurations were respectively identical for all syllables. The device for modulating noise with buzz was always connected into the noise generator circuit. This modulator, described in Sect. 4.3, is important in connection with stimuli in which noise and buzz excitations overlap in time, but otherwise it is of little consequence.

The timing patterns shown in Fig. 6.3 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. 6.13 and 6.14 respectively. On the tape, corresponding to each item on the answer sheet, there is an item of the following form: the pair of syllables to be compared spaced 450 milliseconds apart, then a pause of 1.25 seconds, a repetition of the pair, and then a pause of five seconds for writing the answer. Three sample items precede the test which consists of 42 items in randomized order. That is, each of seven syllables representing seven timing patterns are compared with each of the remaining six so that each intrapair ordering, AB and BA appears once somewhere in the test. Each group of five items is punctuated with a marker tone and an hiatus. Each subject received written instructions (reproduced in the appendix) which requested him to vote

Fig. 6.13 Timing patterns describing stimuli used for studying the effect of changing noise cessation time. Control waveforms unaltered from case to case appear at the top of the figure.

Fig. 6.14 Timing patterns describing stimulus set with differing starting times for configuration change.

for the more "natural member of each pair. Each score is the total number of times a given stimulus was judged to be more natural than the one paired with it in an item. Five subjects participated in each of these tests and therefore the maximum possible score that any stimulus may receive is 60, the minimum is zero, and the average is 30.

The results of the noise cessation study are shown in Fig. 6.15. There is a rough plateau about 30 milliseconds wide extending approximately between items 3 and 6. Under the conditions of the experiment, the tolerance for this parameter seems to be about 3 pitch periods of male speech assuming a 100 cps glottal rate. 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 begin. These perceptual data are consistent with articulatory considerations. 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 should be poor.

The responses for the articulation test, Fig. 6.16, corroborate the need for overlap between the noise ramp and the articulation ramp. Items 6 and 7, with no such overlap, were

Fig. 6.15 Subject responses obtained for study of noise cessation time. Item numbers correspond to those in Fig. 6.13. Ordinate shows total number of times a given stimulus was voted "more natural" than the stimulus with which it was paired.

Fig. 6.16 Subject responses to stimuli with differing times of configuration change.

judged to be poor. In item 1, also poor, the change of configuration is more than half completed before either of the new ramps for control of excitation begin. 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 greater than 50 percent overlap of articulation ramp and noise decay ramp, measured relative to either one. The width of the plateau in the figure is 10 milliseconds. Examination of all of the data will show that configuration change time is the most critical parameter studied yet the tolerance for it is still about one pitch period.

The study of buzz onset time required the recording of three tapes, two for paired comparison tests and one, to be discussed later, for an absolute identification test. Timing patterns for all stimuli are shown in Fig. 6.17. The plan was to find the most natural $//\alpha/$ and the most natural $/3\alpha/$ in two separate tests intended to function independently although they had some stimuli in common. The first set of instructions* asking for the most natural $//\alpha/$ in one test and for the most natural $/3\alpha/$ in the other proved to be

*See appendix.

Fig. 6.17 Patterns describing stimuli used in three tests for investigating buzz onset time. Control waveforms unchanged during the series are shown at the top of the figure. Item numbers for the absolute identification test appear above the buzz ramps and item numbers for the two paired comparison tests appear below.

troublesome. Consider the test intended to find the most natural voiceless stimulus and the subject's dilemma when he hears item one. With buzz onset at 175 milliseconds, 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 the other way.

To circumvent this difficulty, the role of the two paired comparison tests was revised to that of defining a region in which either $//\alpha/$ or $/3\alpha/$ may be produced and to explore 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* to vote for the more natural stimulus regardless of its phonetic identity. The responses for both paired comparison tests are shown by the two dotted curves in Fig. 6.18. 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 over-all region and which are common to both tests. In the new role of the paired comparison tests, these items merely serve to provide

*See appendix for new instructions.

Fig. 6.18 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. 6.17.

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 11 stimuli in Fig. 6.17 and each appeared 9 times in randomized order. Samples, marker tones, and pauses were provided as in previous tests. The instructions to the subjects* required a /ʒa/, /ʃa/ judgement 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. 6.19. The decision of the majority changes from /ʒa/ to /ʃa/ when the onset time is changed by 20 milliseconds. Stimulus 6 is the only one not definitely voiced or voiceless. In Fig. 6.15, the corresponding buzz ramp is seen to begin 20 milliseconds before the configuration ramp and to end when the configuration ramp ends. The buzz ramps for stimuli 7 and 8 also straddle the configuration ramp, but those stimuli are perceived as voiceless. The individual subjects whose responses are shown in Fig. 6.20 were more sensitive to changes in onset time than indicated in the averaged curve. One subject switched his answers in response to a

*See appendix.

Fig. 6.19 Subject responses in absolute identification test as a function of buzz onset time. The number of times each item was judged to be /3/ or /ʃ/ is shown totaled for all subjects.

Fig. 6.20 Individual responses for 7 subjects in absolute identification test. Number of times each item was perceived as /sa/ is given. Number of /3a/ responses is nine's complement of number of /sa/ responses.

10 millisecond change in onset time. The smaller slope of the graph of averaged responses is a consequence of a 20 millisecond spread in thresholds for the individual subjects. This spread is comparable to the 25 millisecond difference of onset time between items 5 and 7, the closest items in Fig. 6.19 that are clearly distinguished.

The secondary purpose of the absolute identification test was to provide an alternative measure of naturalness. The 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 and those rated "fair" were given weight 1 while those rated "poor" were not counted. Thus, a given stimulus may have a low score as $//\alpha/$ because it yields few $//\alpha/$ judgements or because it results in sounds which are definitely $//\alpha/$ but are of poor quality. For example, item 11 is clearly not voiced yet it is a poor $//\alpha/$ because of a wide gap between buzz and noise. This hybrid nature is exemplified by the curve marked "3(AI)" in Fig. 6.18 which follows the curve marked "3(PC)" as far as item 4 and then behaves like the $/3\alpha/$ ogive in Fig. 6.19. There is close agreement between the two quality measures at the extremities of the over-all region so that each measure gives some weight to the validity of the other.

One may conclude from the figure that the production of /3α/ requires buzz onset to occur between 160 milliseconds and 175 milliseconds after the start of the clock when other timer settings are at the neutral values and that corresponding figures for /fα/ are 200 milliseconds and 215 milliseconds. The tolerance is about 15 milliseconds in each case, clearly greater than one pitch period of male speech.

CHAPTER 7 CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

7.1 KNOWLEDGE GAINED BY DEVELOPMENT AND OPERATION OF THE SYNTHESIZER

As stated in Chapter 1, the goals of this research were the development of a dynamic analog of the vocal tract, development of a control system for the analog, and experimentation with the finished synthesizer. We will discuss these points in that order.

7.11 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 Chapter 2), on assumptions (stated in Chapter 3) 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 in Chapter 2, we have simulated a distributed system, partly nonlinear, by a linear lumped-parameter system. This reasoning has been tested by earlier analogs, all atatic, 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 development of dynamic capacitive and inductive elements each variable

over a 100:1 range. A question about the adequacy of 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 adequate since the line demonstrated suitable behavior at higher frequencies where 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 three formants may be expected to remain within 5 per cent of appropriate nominal values throughout an entire day after a 15 minute warmup. 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. 5.1 and noise is measured for the most noisy configuration.

7.12 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 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 necessary for microprogramming of transitions from one sound to the next. This unit would also choose the ramp durations appropriate for each transition and would effect 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 one storing some 20 or 40 fixed configurations. See Section 7.3 for further discussion of such a control system.

7.13 EXPERIMENTS WITH THE SYNTHESIZER

The synthesizer has demonstrated 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 are also characteristic of human male speech. Natural vowels can be ordered along dimensions relating 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 six 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 appropriate for each fricative response corresponded closely to configurations and excitations 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 in the order of 10 to 30 milliseconds for those parameters.

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

7.2 SUGGESTIONS FOR FURTHER RESEARCH

Future research concerning components and circuits for a dynamic analog speech synthesizer should place emphasis on alternative variable capacitors that require less power and that 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 initially 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 suitable for the various classes of speech sounds. Such a model could lead to a synthesizer controlled by articulatory parameters.

Further goals of future research are the extension of the vocabulary 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 which deals with short overlapping segments of speech and which is 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, now blank, must be filled through future work* with other speech sounds in a variety of environments. 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 are 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.

*Such research with the dynamic analog would deal with many of the phenomena investigated by the Haskins group using the pattern playback.

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 possibilities and limitations of the synthesizer in this role should be explored.

Much 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 which are defined in terms of contrasting pairs. Certain physical axes may be divided into two regions, each corresponding 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 which 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, how-

ever, 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 performing operations upon linguistic entities. This device would examine the continuous signals to recognize phoneme boundaries and classify successive phonemes as to their manner of production. The device would then emit precisely synchronized signals 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 and high-frequency response otherwise necessary to preserve occasional sharp wavefronts is not needed.

Discrete codes must be handled by the synthesizer serving in a computing system or in a speech transmission system 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, etc. 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 the previous section. The control system memory must store the symbol currently being uttered plus the following symbol in order to choose the appropriate transitions in excitation and configuration and to execute their micro-timing. The start of the transition can be signaled by receipt of the following symbol. Such a system allows for variable phone duration and can accommodate 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* when the syntactic structure underlying the utterance is available in the computer memory.

*See, for example, reference 10.

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 one considers the production of speech, 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 M.I.T. 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 gross approximations to human articulatory configurations. Several physiologically similar configurations have yielded the same high intelligibility even though the spectra of the stimuli are quite different. These experiments contrast 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 to the point of view that the articulatory level provides a natural and economical description of speech. The machine provides, furthermore, a facility which can be used to extend our understanding of the articulatory level and thus to broaden our knowledge of the entire speech communication process.

APPENDIX

A.1 CIRCUIT DIAGRAMS

Fig. A.1 The nasal analog designed for use with a static analog of the vocal tract. See reference 2.

Fig. A.3 Circuit diagram of control amplifier for a variable section.

Fig. A.4 Circuit diagram of dc amplifier for driving the configuration matrix.

A.2 INSTRUCTIONS TO SUBJECTS

The instructions which follow were given to subjects listening to the 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 / $\int$ /.

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 /3/.

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 study of buzz onset time by the method of paired comparisons.

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.

BIOGRAPHICAL NOTE

George Rosen was born in Philadelphia, Pennsylvania on May 24, 1930. He attended the University of Pennsylvania and the Massachusetts Institute of Technology, receiving the degree of Bachelor of Science in Electrical Engineering from the former in 1952 and the degree of Master of Science in Electrical Engineering from the latter in 1955. His research was carried out at the Acoustics Laboratory, MIT, and at the Research Laboratory of Electronics, MIT. He is a member of Sigma Xi, Tau Beta Pi, Eta Kappa Nu, Pi Mu Epsilon, the Acoustical Society of America, the Association for Computing Machinery and the Institute of Radio Engineers.

BIBLIOGRAPHY

1. Anon. Speech Compression Research: Final Report. Acoustics Laboratory, Massachusetts Institute of Technology. AFCRC-TN-57-166. ASTIA Doc. No. 117049 (1957)
2. Anon. Research on Speech Synthesis. Scientific Report No. 17, Acoustics Laboratory, Massachusetts Institute of Technology. AFCRC-TN-58-140. ASTIA Doc. No. 117049 (1958)
3. Arnold, J.B. Halle, M. Sandel, T.T. and Stevens, K.N. Perception of Speech-Like Sounds. Quarterly Progress Report, Research Laboratory of Electronics, Massachusetts Institute of Technology. 167-171 (July, 1959)
4. Bacon, R.C. and Pollard, J.R. The Dekatron. Electronic Engineering 22, 173-177 (1950)
5. Barlow, J.S. Brazier, M.A.B. and Rosenblith, W.A. The Application of Autocorrelation Analysis to Encephalography. Proc. National Biophysics Conf. (New Haven, Yale University Press, 1959)
6. Beranek, L.L. Acoustics (New York: McGraw-Hill Book Co. Inc., 1954)
7. Bode, H. Network Analysis and Feedback Amplifier Design (New York: D. VanNostrand Co. 1945)
8. Chiba, T. A Study of Accent: Research into its Nature and Scope in the Light of Experimental Phonetics (Tokyo; Fuzanbo Publishing Co., 1935)
9. Chiba, T. and Kajiyama, M. The Vowel, Its Nature and Structure (Tokyo; Tokyo-Kaiseikan Publishing Co., 1941)
10. Chomsky, N. Halle, M. and Lukoff, F. On Accent and Juncture in English. In For Roman Jakobson (The Hague, Netherlands, Mouton and Co., 1956)

11. Communications Biophysics Group of Research Laboratory of Electronics and Siebert, W. M. Processing of Neuroelectric Data (New York, John Wiley and Sons Inc., 1959)
12. Cooper, F. S. Liberman, A. M. and Borst, J. M. Preliminary Studies of Speech Produced by a Pattern Playback. J. Acoust. Soc. Am. 22, 678 (1950)
13. Cooper, F. S. Liberman, A. M. and Borst, J. M. The Interconversion of Audible and Visible Patterns as a Basis for Research in the Perception of Speech. Proc. National Academy of Sciences 37, 318-325 (1951)
14. Davenport, W. B. Jr. and Root, W. L. An Introduction to the Theory of Random Signals and Noise (New York, McGraw-Hill Book Co., 1958)
15. Dudley, H. W. Synthesizing Speech. Bell Laboratories Record 15, 98-102 (1936)
16. Dudley, H. W. Remaking Speech. J. Acoust. Soc. Am. 11, 169-177 (1939)
17. Dudley, H. W. The Carrier Nature of Speech. Bell System Tech. J. 29, 147-160 (1940)
18. Dunn, H. K. The Calculation of Vowel Resonances, and an Electrical Vocal Tract. J. Acoust. Soc. Am. 22, 740-753 (1950)
19. Fant, C. G. M. Transmission Properties of the Vocal Tract with Application to the Acoustic Specification of Phonemes. Technical Report No. 12, Acoustics Laboratory, Massachusetts Institute of Technology (1952)
20. Fant, C. G. M. IVA 24, 331-337 (1953)
21. Fant, C. G. M. Acoustic Theory of Speech Production Report No. 10, 1958, Royal Institute of Technology, Stockholm, Sweden (To be published as a book by Mouton and Co., The Hague, Netherlands)
22. Fant, C. G. M. Modern Instruments and Methods for Acoustic Studies of Speech. Acta Polytechnica Scandinavia, Ph 1 (246/1958)

23. Fant, C. G. M. Acoustic Analysis and Synthesis of Speech with Applications to Swedish. Ericsson Technics 15, 3-108 (1959)
24. Flanagan, J. L. A Speech Analyzer for a Formant-Coding Compression System. ScD. Thesis, Department of Electrical Engineering, Massachusetts Institute of Technology (1955). The thesis is available in published form as Scientific Report No. 4, Acoustics Laboratory, Massachusetts Institute of Technology May, 1955.
25. Flanagan, J. L. Some Properties of the Glottal Sound Source. J. Speech Hearing Research 1, 99-116 (1958)
26. Guilford, J. P. Psychometric Methods (New York, McGraw-Hill Book Co. Inc., 1936)
27. Guillemin, E. A. Introductory Circuit Theory (New York, John Wiley and Sons., 1953)
28. Harris, K. S. Cues for the Identification of the Fricatives of American English. J. Acoust. Soc. Am. 26, 952 (1954)
29. Heinz, J. M. A Terminal Analog of Fricative Consonant Articulation. Quarterly Report, Acoustics Laboratory, Massachusetts Institute of Technology (Sept. 1957)
30. Helmholtz, H. L. F. Die Lehre von den Tonempfindungen als Physiologische Grundlage fur die Theorie der Musik (Braunschweig, Vieweg, 1863)
31. House, A. S. Analog Studies of Nasal Consonants. J. Speech Hearing Disorders 22, 190-204 (1957)
32. House, A. S. and Stevens, K. N. Auditory Testing of a Simplified Description of Vowel Articulation. J. Acoust. Soc. Am. 27, 882-887 (1955)
33. House, A. S. and Stevens, K. N. Analog Studies of the Nasalization of Vowels. J. Speech Hearing Disorders 21, 218-232 (1956)
34. House, A. S. and Stevens, K. N. Estimation of Formant Band Widths from Measurements of Transient Response of The Vocal Tract. J. Speech Hearing Research 1, 309-315 (1958)

35. Hughes, G. W. and Halle, M. Spectral Properties of Fricative Consonants. J. Acoust. Soc. Am. 28, 303-310 (1956)
36. Kasowski, S. E. A Speech Sound Synthesizer. Unpublished SM thesis, Department of Electrical Engineering, Massachusetts Institute of Technology (1952)
37. Kleinschmidt, K. The Effect of Quasi-Periodicity on the Naturalness of Synthetic Speech. SB thesis, Department of Electrical Engineering, Massachusetts Institute of Technology (1957)
38. Lawrence, W. The Synthesis of Speech from Signals which Have a Low Information Rate. Chapter 34 in Jackson, W. (ed.) Communication Theory (London, Butterworth Sci. Publ., 1953)
39. Lewis, D. Vocal Resonance. J. Acoust. Soc. Am. 8, 91-99 (1936)
40. Licklider, J. C. R. Hawley, M. E. and Walking, R. A. Influences of Variations in Speech Intensity and Other Factors upon the Speech Spectrum. J. Acoust. Soc. Am. 27, 207 (1955)
41. Lieberman, P. Vowel Intonation Contours, Quarterly Progress Report, Research Laboratory of Electronics, Massachusetts Institute of Technology 156-161 (July, 1958)
42. Malme, C. I. Detectability of Small Irregularities in a Broadband Noise Spectrum. Quarterly Progress Report, Research Laboratory of Electronics, Massachusetts Institute of Technology 139-142 (Jan. 1959)
43. Miller, G. A. Language and Communication (New York, McGraw-Hill Book Co. Inc., 1951)
44. Miller, R. L. Nature of the Vocal Cord Wave. J. Acoust. Soc. Am. 31, 667-677 (1959)
45. Peterson, G. E. Information Theory: 2 Application of Information Theory to Research in Experimental Phonetics. J. Speech Hearing Disorders 17, 175-188 (1952)

46. Peterson, G. E. and Barney, H. L. Control Methods Used in a Study of the Vowels. J. Acoust. Soc. Am. 24, 175-184 (1952)
47. Rosen, G. A Prototype Section for a Dynamic Speech Synthesizer. SM thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, (1955) ; also available as AFCRC-TN-55-196
48. Russel, G. O. The Vowel (Columbus, Ohio State University Press, 1928)
49. Sandel, T. T. and Stevens, K. N. Studies of the Perception of Speech-Like Sounds, Quarterly Progress Report, Research Laboratory of Electronics, Massachusetts Institute of Technology 155-156 (July 1958)
50. Shannon, C. E. Communication in the Presence of Noise. Proc. IRE 37, 10-21 (1949)
51. Stevens, K. N. Synthesis of Speech by Electrical Analog Devices. J. Audio Eng. Soc. 4, 2-8 (1956)
52. Stevens, K. N. The Role of Duration in Vowel Identification. Quarterly Progress Report, Research Laboratory of Electronics, Massachusetts Institute of Technology 136-139 (Jan. 1959)
53. Stevens, K. N. and House, A. S. Development of a Quantitative Description of Vowel Articulation. J. Acoust. Soc. Am. 27, 484-493 (1955)
54. Stevens, K. N. Kasowski, S. and Fant, C. G. M. An Electrical Analog of the Vocal Tract. J. Acoust. Soc. Am. 22, 734-742 (1953)
55. Stevens, K. N. and Nakata, K. Synthesis and Perception of Fricative Consonants. Quarterly Progress Report, Research Laboratory of Electronics, Massachusetts Institute of Technology 150-152 (July 1958)
56. Stewart, J. Q. An Electrical Analog of the Vocal Organs. Nature 110, 311-312 (1922)
57. Tarnoczy, T. Resonanzdaten der Vokalresonatoren. Akust. Z. 8, 91-99 (1936)

58. van den Berg, J. 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. von Kempelen, W. Le Mechanisme de la Parole Suivi d'une Description d'une Machine Parlante (Vienna, 1791)
60. Willis, R. On Vowel Sounds and on Reed Organ Pipes. Trans. Cambridge Philosophical Soc. 3, 231-268 (1829)