Speech Synthesis by Rule: An Acoustic Domain Approach

By LAWRENCE RABINER

(Manuscript received August 30, 1967)

A new approach to speech synthesis by rule has been formulated and evaluated. A discrete set of symbols (phonemes and stress marks) is converted to a continuous acoustic waveform by a two-step transformation. The first step involves conversion from phonemes to control signals capable of driving a terminal analog speech synthesizer. The second step is conversion from control signals to the acoustic waveform.

This paper presents a design for the terminal analog synthesizer and discusses the new features of this device. It discusses in detail the method of converting from phonemes to control signals. It places primary emphasis on determining the formant frequency control signals and the fundamental frequency contour, and presents models for determining these contours from the input data. The paper includes an experimental evaluation of the entire technique in terms of word intelligibility scores and consonant confusion matrices.

I. INTRODUCTION

Speech synthesis by rule is the method of converting from a discrete representation of speech in linguistic units, that is, phonemes and stress marks, to a continuous acoustic waveform. Fig. 1 shows the technique for carrying out this transformation. The figure shows that the discrete input is converted to continuous control signals by the synthesis strategy. The synthesis strategy contains stored information about the phonemes and stored rules about the mutual effects of adjacent phonemes. The stored rules operate on the input sequence to produce the control signals for the synthesizer. The speech synthesizer converts the control signals to continuous speech. The synthesizer may be a terminal analog, a dynamic analog of the vocal tract, or a combination.

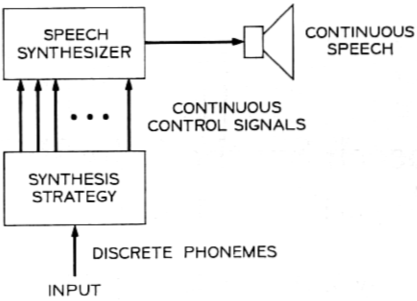

Fig. 1 — Technique of Speech Synthesis.

During the past ten years there have been many attempts at synthesis by rule. The primary goal of these attempts has been to produce natural sounding, intelligible speech. Of secondary importance has been the preservation, in some natural way, of the dynamics of speech production by embodying in the scheme the constraints imposed by the human vocal tract.

Previous methods of synthesis by rule are generally classified as either articulatory or acoustic domain approaches. An articulatory approach uses physiological parameters such as tongue-tip position, and lip opening as the control signals for the synthesizer. The stored data of the synthesis strategy are of the form of vocal tract configurations. An acoustic domain approach uses parameters such as formant values and fundamental frequency as control signals. The stored data include such information as target positions of formants and relative amplitudes of phonemes.

Articulatory domain approaches to synthesis by rule[1, 2] have been most successful in modelling the dynamics of the speech producing mechanism. Acoustic domain methods, such as the one presented here, can impose the natural constraints of the vocal tract only indirectly, that is, by rules which often lack a firm physiological basis. However, acoustic domain approaches have enjoyed the most success in producing intelligible, high quality speech,[3, 4, 5, 6] thus justifying and motivating efforts along these lines. The technique for synthesis by rule, described in this paper, is an acoustic domain approach.

The next section gives a general description of the synthesizer. Terminal analog synthesizers of this type are common[7, 8, 9] and we discuss only the new features at any length.

II. SYNTHESIZER

2.1 *General Description*

A terminal analog synthesizer models the speech-producing mechanism, which includes the vocal tract, excitation sources, and radiation impedance. The transfer function of the vocal tract can be reduced to either a cascade of complex conjugate pole and zero pair networks, or a parallel addition of complex pole pair networks. The cascade representation was used because it reduced the complexity of the synthesis strategy by reducing the number of synthesizer control parameters.

Fig. 2 is a block diagram of the synthesizer used in this work. The synthesizer was simulated on a computer at 20 kHz sampling frequency. There are two sources of excitation, a pitch impulse generator, and a frication (noise) generator. To produce voiced speech (vowels, nasals, voiced stops, and voiced fricatives) the pitch impulse generator output is gated by the switch to the upper arm of the synthesizer. The nasal network is included in the upper arm only for nasal consonants. To produce whispered or aspirated speech, the frication generator is gated by the switch to the upper arm of the synthesizer.

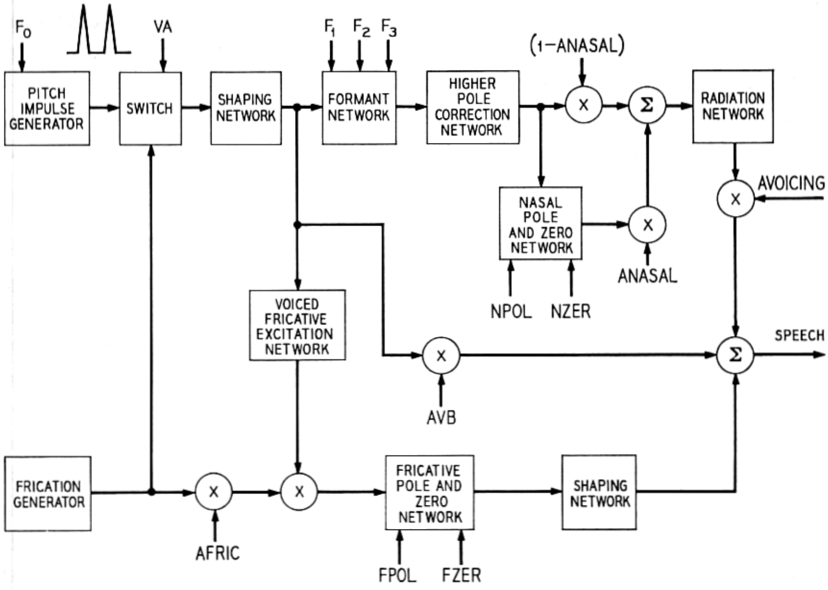

Fig. 2 — Synthesis used by author.

To produce a voiceless fricative or the unvoiced component of voiced fricatives, the frication generator excites the lower arm of the synthesizer. For a voiceless fricative, the output of the voiced fricative excitation network is constant. For a voiced fricative the output of the voiced fricative excitation network modulates the frication generator output. The details of this network are explained in Section 2.1 because this is an original design.

The higher pole correction network in the upper arm of the synthesizer compensates for the missing higher order poles.[9, 10] A new design for this network, based on the properties of sampled data systems, has been formulated and is discussed in Gold and Rabiner's paper.[11]

One last feature of the synthesizer is provision for generating a voice bar. (A voice bar is the quasi-periodic low frequency energy radiated from the region of the vocal cords during the closure interval of voiced stop consonants. During this interval the vocal cords are vibrating thus acting as the source of energy for the voice bar.) To produce a voice bar, the middle arm of the synthesizer is used with the switch gating the pitch impulse generator output to the shaping network. The voice bar has energy only at low frequency similar to voice bars of natural speech.

The outputs of the three arms of the synthesizer are added to produce the speech. The synthesizer control signals are indicated in Fig. 2 by arrows. These include four amplitude controls (avoicing, anasal, avb, afric); a derived amplitude control (1-anasal); 14 pole-zero controls (both center frequency and bandwidth of F_1, F_2, F_3, npol, nzer, fpol, fzer); a switch control (va); and a fundamental frequency control (F_0).

2.2 *Voiced Fricative Excitation Network*

The network connecting the pitch impulse generator to the lower arm of the synthesizer is used to provide the excitation for the unvoiced component of voiced fricatives. Fig. 3 shows the relevant details of this network. (For clarity, certain components of the synthesizer have been omitted from Fig. 2.)

The ouput of the pulse generator is shaped to produce a suitable pitch pulse. A complex conjugate pole pair resonator was used, but any suitably chosen network could have been used. The pitch pulses excite a resonator tuned to the first formant of the fricative sound. A single resonance is the first order approximation to the transfer function of volume velocity (the signal of interest in Fig. 3) from the

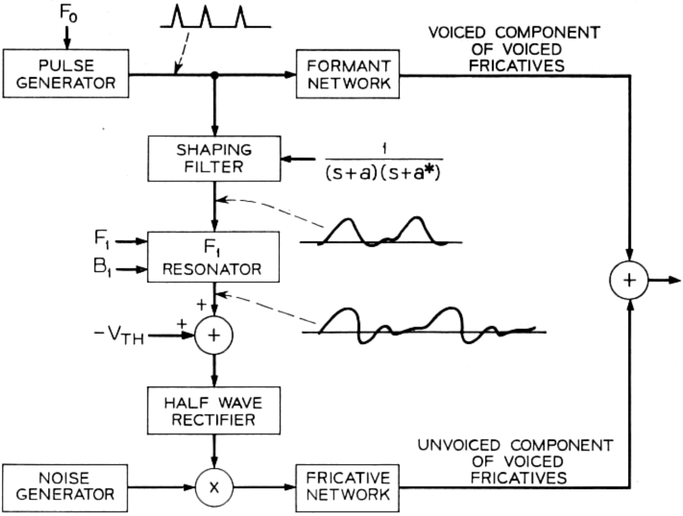

Fig. 3 — Excitation network for voiced fricatives.

glottis through the point of constriction of the vocal tract. A threshold level is subtracted from the output of the resonator and the result is half-wave rectified. These operations model the physical situation where turbulence is not produced until the volume velocity of the airflow exceeds a threshold value. The output of the half-wave rectifier modulates the output of a noise generator, prcducing a pitch synchronous excitation for the unvoiced component of the fricative. The final unvoiced component is produced by exciting the fricative network by this excitation. The voiced component is produced in the standard manner, that is, by exciting the formant network by the pitch pulses.

Spectrograms of voiced fricatives produced by the above technique are quite similar to spectrograms from natural speech. Experimental evidence presented later shows that the synthetic fricatives are highly intelligible.

III. SYNTHESIS STRATEGY

Since formant contours are crucial to speech intelligibility (see pages 220-234 of Ref. 9), the first step in transforming a discrete set of input symbols to the synthesizer control signals is to generate the formant contours. The method is explained in Sections 3.1, 3.2, and 3.3.

Once the formant contours are specified, the remaining control parameter contours are determined by delimiting certain characteristic times along the formant contours. At these times, motion of the other parameters is initiated or terminated. The techniques for generating these contours are discussed in Section 3.4.

Part IV treats the generation of a fundamental frequency contour. Table I shows international phonetic alphabet symbols and the letter equivalents used in the following sections.

3.1 *Static Phoneme Characterizations*

In order to generate the formant contours from the phoneme input sequence, certain information must be supplied. Corresponding to

TABLE I—LETTER EQUIVALENTS OF INTERNATIONAL PHONETIC ALPHABET

Letter Symbol	IPA Symbol
IY	*i*
I	*I*
E	ε
AE	*ae*
UH	Λ
A	*a*
OW	ɔ
U	*U*
OO	*u*
ER	ɜ
W	*w*
L	*l*
R	*r*
Y	*y*
B	*b*
D	*d*
G	*g*
P	*p*
T	*t*
K	*k*
M	*m*
N	*n*
NG	ŋ
F	*f*
TH	θ
S	*s*
SH	ʃ
V	*v*
THE	ð
Z	*z*
ZH	ʒ
CH	*t*ʃ
J	dʒ

each possible phoneme there must be data on formant target positions. These data, along with other data, are included in a phoneme characterization table. Certain durational data are necessary, such as for stressed vowels. Finally a technique for generating formant transitions must be supplied.

Each phoneme has a characterization independent of adjacent phonemes. The characterization includes formant information, source characteristics in the production of the phoneme, a description of whether it is nasal or fricative, and a set of frequency regions surrounding the formant positions.

The formant information is a set of target positions for both center frequency and bandwidth of formants one, two and three. The source characteristics describe the condition of the vocal cords during the production of the phoneme. If the vocal cords are vibrating the sound is voiced. The frequency regions of a phoneme represent the degree to which certain acoustic parameters must approximate the target values of these parameters in the context of connected speech. In an articulatory analog, the corresponding concept would be the extent to which a given vocal tract configuration must approximate the target configuration for the phoneme.

The frequency regions represent a compromise between choosing a single characterization for a phoneme and considering it inviolate, and the realization that there are many acceptable characterizations for a phoneme—especially in the context of connected speech.

Table II shows the phoneme characterizations we have used. The first three columns list the formant target positions of the phonemes. The second three columns show the frequency regions of the phonemes. (The figures represent both ± values.) The final three columns describe nasality, fricative, and voicing characteristics of the phonemes. A + in any column indicates the presence of the feature and a − indicates its absence. The voicing condition of the voiced fricatives z and ZH is ± indicating the two sources used to produce these sounds on the synthesizer. The bandwidths of F_2 and F_3 are held fixed at 100 Hz and 120 Hz for all phonemes. The bandwidth of F_1 is 60 Hz except for nasals where it is 150 Hz. When a + appears in the nasality or fricative columns, a table look-up procedure is used to specify pole-zero locations.*

* All data referred to but not included in this paper are available in the author's Ph.D. thesis which is available from the MIT library. (See Ref. 12.)

TABLE II—PHONEME CHARACTERIZATION

Phoneme	$F1$	$F2$	$F3$	$\Delta 1$	$\Delta 2$	$\Delta 3$	Nasal	Fricative	Voiced
IY	270	2290	3010	75	75	150	−	−	+
I	390	1990	2550	75	75	110	−	−	+
E	530	1840	2480	75	80	110	−	−	+
AE	660	1720	2410	75	75	110	−	−	+
UH	520	1190	2390	75	75	75	−	−	+
A	730	1090	2440	37	75	115	−	−	+
OW	570	840	2410	75	75	115	−	−	+
U	440	1020	2240	75	75	90	−	−	+
OO	300	870	2240	75	80	90	−	−	+
ER	490	1350	1690	75	80	100	−	−	+
W	300	610	2200	25	40	150	−	−	+
L	380	880	2575	25	80	150	−	−	+
R	420	1300	1600	30	80	100	−	−	+
Y	300	2200	3065	25	110	200	−	−	+
B	0	800	1750	50	75	120	−	−	+
D	0	1700	2600	30	50	160	−	−	+
G	0	2350	2000	15	50	100	−	−	+
M	280	900	2200	17	17	40	+	−	+
N	280	1700	2600	17	17	100	+	−	+
NG	280	2300	2750	17	17	100	+	−	+
P	0	800	1750	50	40	80	−	−	−
T	0	1700	2600	30	30	100	−	−	−
K	0	2350	2000	10	30	70	−	−	−
F	175	900	2400	30	50	120	−	+	−
TH	200	1400	2200	30	40	100	−	+	−
S	200	1300	2500	30	40	70	−	+	−
SH	175	1800	2000	30	100	150	−	+	−
V	175	1100	2400	30	50	120	−	+	+
THE	200	1600	2200	30	40	100	−	+	+
Z	200	1300	2500	30	40	70	−	+	±
ZH	175	1800	2000	30	100	150	−	+	±

(The data for $\Delta 1$, $\Delta 2$, $\Delta 3$ were determined experimentally.)

3.2 *Duration and Amplitude*

Vowel duration is specified only for stressed vowels. The durations of unstressed vowels are determined by the methods illustrated in Section 3.3. The duration of a stressed vowel is modified by its following phoneme. The longest vowels are those followed by voiced fricative consonants; the shortest are followed by voiceless stop consonants.[13]

For certain consonants maximum durations are specified. Consonant duration (as measured from human speech) is not a fixed quantity but is very dependent on context. (Four example, initial consonants are much longer than medial consonants.) The synthesis strategy generates consonants whose duration is variable within certain limits. Maximum durations are specified to prevent the consonant from being unnaturally long, hence objectionable.

Maximum stop gap durations are specified for stops; aspiration duration, as a function of the succeeding phoneme, is specified for voiceless stop consonants. Values of amplitude control signals are specified for all phonemes. Rates of change of control signals are specified for various phoneme classes (that is, vowels, nasals, fricatives, and stops).

3.3 *Formant Motion*

The technique for generating formant transitions (and hence formant contours) is a new one. We present it in detail because the entire synthesis strategy is built around it.

As we stated, the motion of formants is one of the most significant factors contributing to the intelligibility of speech. Smooth, continuous formant transitions are generally observed on spectrograms of real speech. To match these characteristics, we used the solution to a critically-damped second degree differential equation to describe the transitions of formants. We chose a second degree equation because it provided a good fit to data on formant transitions. We used a critically-damped solution because it was completely specified from a single time constant. Values of time constants were determined from examining formant transitions for real speech on spectrograms.

The input to the differential equation represents the formant target position appropriate for the current phoneme. Since the current phoneme changes its value discretely, the input to the differential equation changes in a steplike manner. The formant motion, in response to this step input, is smooth and continuous. Thus motion from steady state value Ai to target value Af, beginning at time $t = 0$ is of the form:

$$x(t) = Af + (Ai - Af)(1 + t/\tau) \exp(-t/\tau)$$

where τ is the time constant of motion and $x(t)$ the formant position at time t.

In general, motion between target positions does not proceed from a steady state condition; that is, there are initial conditions. Motion to a target whose formant value is Af from an initial formant position Ai with an initial formant velocity $Vi = dx/dt|_{0-}$ is of the form:

$$x(t) = Af + (Ai - Af) \exp(-t/\tau) + \left[Vi + \frac{(Ai - Af)}{\tau}\right] t \exp(-t/\tau); \qquad t \geqq 0.$$

At times when the input to the differential equation is changed discretely, both the output value *and* slope are continuous. Thus the concept of smooth, continuous formant transitions is realized in all cases.

The time constants of the differential equation are functions of the individual formant *and* the pair of phonemes between which the transition is being made. Hence, for each possible pair of phonemes, and each formant, a time constant is specified. Certain simplifying approximations reduce (by an order of magnitude) the number of time constants that have to be specified.

The inputs to the differential equations change discretely in time in a steplike manner. However, provision is made for delaying, to any formant, the steplike change in formant target position. Thus, in the most general case, formants move independently of each other with unequal time constants of motion. This delay feature was found necessary for only a few cases.

The phonemic goals change discretely in time. The decision of when to make the discrete changes, that is, when to initiate motion to new sets of formant targets, is based on the criterion that the formants must first be within the phoneme frequency regions of the targets, and then satisfy durational requirements of the phoneme, if there are any.

Formants, in general, are in motion towards target values appropriate for the phonemes to be generated. Their motion is characterized by the solution to a differential equation. The time constant of motion is a function of the phoneme from which motion began and the phoneme which is being generated. Each formant moves with its own time constant and there is provision for delay in time of initiation of the motion of formants. When all formants are within the frequency regions of the target, a decision is made. If a stressed vowel is being generated, then a table look-up procedure determines the correct vowel duration and motion continues for the specified time. Once a vowel of proper duration has been generated, motion towards target positions characteristic of the next phoneme is begun. If the current phoneme is not a stressed vowel, motion towards the new phoneme targets is initiated as soon as all the formants are within their specified frequency regions.

The decision to start motion to new target values results in three separate operations. First, new time constants for each formant are inserted into the respective difference equations. Second, the forcing functions (input) to the difference equations are changed in a steplike manner indicating the changes in target positions. Finally, the initial conditions of the difference equation are set to preserve continuity of formant values and formant velocities. If the motion of any formant is to be delayed, the changes in the difference equation for that formant are delayed appropriately.

Fig. 4 shows a typical cycle of events. Initially formants one and two (we shall neglect formant three in this example) are at target positions appropriate for phoneme 1. At time t_1 motion is initiated to phoneme 2. Formants one and two begin motion simultaneously (no delay is used here) with time constants τ_{12}^1 and τ_{12}^2 respectively. τ_{12}^1 is much smaller than τ_{12}^2 so formant one moves more rapidly to its target value than formant two. Periodically the formant values are tested to see whether they are within the specified frequency regions of the targets. (The frequency regions are indicated by $\Delta 1$, $\Delta 2$ in Fig. 4.) If they are not, the formants continue their motion, thus moving closer to target. For the example in Fig. 4, formant one enters its frequency region prior to formant two. Until formant two enters its frequency region at time t_2, formant one moves closer to its target position. At time t_2 both formants are within the specified frequency regions and so a check is made on whether phoneme 2 is a stressed vowel or not. In this example phoneme 2 is not a stressed vowel, so motion to phoneme 3 is initiated at t_2. However, we now have the

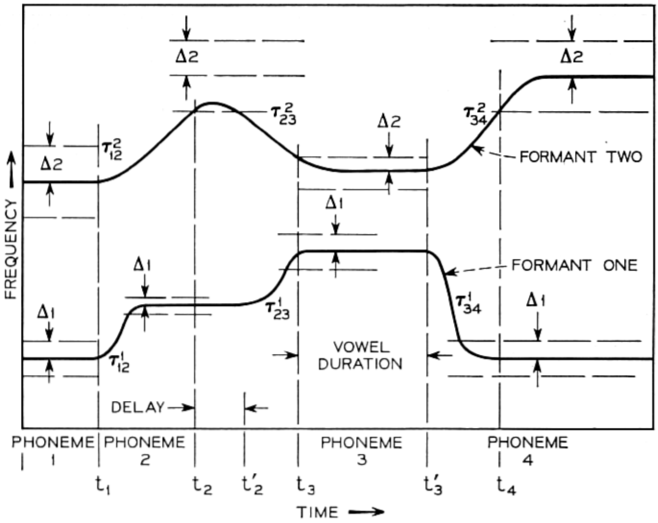

Fig. 4 — Simplified example of formant motion.

case when the time of initiation of motion of formant one is delayed. Hence at t_2 the target value and time constant for formant two is changed, but formant one's target is unchanged. At time t_2'' the delay is terminated and formant one begins its motion.

Phoneme 3 of Fig. 4 is a stressed vowel. So at time t_3, when both formants are within the frequency regions for phoneme 3, motion to targets for phoneme 3 continues for the specified vowel duration. At time t_3'', following the vowel duration, motion begins toward targets for phoneme 4. New time constants and targets are again inserted in the equations of motion. The process continues in this manner until all the input phonemes have been generated.

3.4 *Remaining Synthesizer Controls*

The motion of the remaining synthesizer control parameters (that is, nasal and fricative poles, and zeros and source amplitudes) is time-locked to the formant motion. The source amplitudes (avoicing, anasal, afric, avb) begin to switch approximately one time constant after the discrete phoneme goal is changed. The amplitudes change linearly at predetermined rates. The nasal and fricative poles and zeros initiate motion at the time the phoneme goal is changed. The motion is linear and the slopes are arranged so that the poles and zeros just reach their targets at the time the source amplitudes are switched. The target positions are specified in a table.

For nonnasal sounds, the target positions of the nasal zero and pole are set to 1400 Hz. Thus the pole and zero will cancel each other in these cases. Furthermore, for nasal sounds, the bandwidth of formant one (nominally 60 Hz) is changed linearly to 150 Hz for the duration of the nasal. The bandwidth begins to change 50 msec before the amplitudes switch and is linearly changed back to its nominal value in 50 msec after the nasal. For nonfricative sounds the fricative pole and zero target positions are set to 1500 Hz.

IV. FUNDAMENTAL FREQUENCY

Our model for generating fundamental frequency data is based on the assumption that these data can be derived from data on laryngeal tension (LT) and subglottal pressure (Ps). A description of an utterance in terms of these variables is then used to produce the desired fundamental frequency data.

The model is based on that of P. Lieberman, in which the breath-group is defined as an underlying phonetic feature of American English.[14] The unmarked breath-group is characteristic of a simple, declarative sentence, whereas the marked breath-group characterizes a simple interrogative sentence.

The feature breath-group is converted to a global description of an utterance in terms of Ps and LT. Fig. 5 shows the archetypal Ps contour, as suggested by Lieberman's data. Ps increases over the first 300 msec of the utterance, and then remains constant until the last 300 msec of the breath-group, at which point it decreases rapidly to zero. The LT contour for an unmarked breath-group is constant, whereas for a marked breath-group it is characterized by a steady increase over the last 175 msec of phonation. Fundamental frequency is linearly proportional to both Ps and LT. Since the archetypal Ps contour falls at the end of a marked breath-group, the increase in LT must compensate for the decrease in Ps to give fundamental frequency a rising terminal contour. A slope of 0.6 Hz/msec, for the last 175 msec of phonation, was assigned to the LT contour. This resulted in a terminal rise of 60 Hz in fundamental frequency for a question.

The subglottal pressure contour is modified by both consonants and vowels. Two levels of stressed vowels are adopted. One level is referred to as emphasis and only one vowel in a breath-group is emphasized. The emphasized vowel provides the highest peak in the Ps contour. All other stressed vowels are treated similarly. When a vowel is stressed, there is an increase in subglottal pressure for a pe-

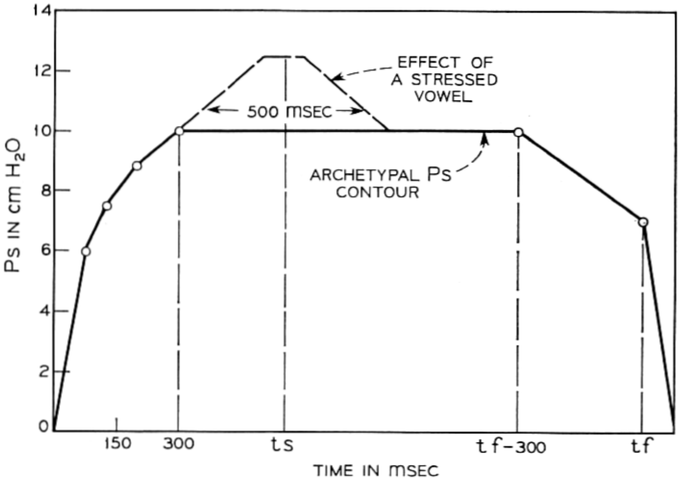

Fig. 5 — Archetypal subglottal pressure contour showing effects of vowel stress.

riod of 500 msec, centered on the vowel. Fig. 5 shows an example of this effect. The dashed curve shows the effects of placing stress on a vowel at the beginning of the breath-group. There is a rise in Ps early in the breath-group and the increase is centered at ts, the midpoint of the vowel steady state. If the stressed vowel had been the emphasized one, the only difference in Fig. 5 would be the amplitude of the increased Ps. It would have been 2.5 cm H_2O as compared to 1.0 cm H_2O.

The effects of consonants on subglottal pressure have also been included in our scheme. For a voiceless consonant, subglottal pressure automatically increases whereas subglottal pressure automatically decreased for voiced consonants. Thus the consonants introduce local perturbations to the Ps contour. The change in Ps for consonants is ±1.0 cm H_2O, and this change occurs over a period of 150 msec centered on the consonant.

V. TYPICAL INPUT SEQUENCES

All vowels are unstressed except those followed by the symbol *strss*. The symbol *strss1* signifies the emphasized vowel. Word boundaries are signified by the symbol *space* and pauses by the symbol *pause*. A question is signified by the symbol *ques*. The sentence boundary is indicated by the symbol *end*. The examples are:

(*i*) This is an olive.
THE I *strss* S *space* I Z *space* AE N *space* A *strss1* L I V *end*.

(*ii*) Why are you sad?
W A *strss1* IY *space* A R *space* Y OO *space* S AE *strss* D *end*.

(*iii*) We sang all day.
W IY *space* S AE *strss1* NG *space* OW L *space* D E IY *strss end*.

Whenever a word boundary (*space* in our code) occurs, consonants on either side of the word boundary are affected. In this strategy an initial consonant is lengthened by about 20 percent, whereas a final consonant is shortened by a similar amount. A word boundary has no effect on phonemes which do not lie on either side of the word boundary.

VI. EVALUATION TESTS

Intelligibility tests were conducted to evaluate the scheme. To test the rules in a limited environment, consonant intelligibility tests were run. One test was intended to test perception of consonants in prestressed position. The schwa vowel UH always preceded the consonant and was used as a perceptual cue for stop consonants because it provided a basis for perceiving the stop gap. The second test was intended to test perception of consonants in post-stressed position. The schwa vowel UH always followed the final consonant—again providing a basis for perceiving stop gap duration, bursts, and aspiration for stops.

Sixteen consonants were used: B, D, G, P, T, K, M, N, F, TH, S, SH, V, THE, Z, and ZH. Five vowels (besides schwa) were used: IY, AE, A, OW, and OO. For each test there were 80 possible stimuli. Twenty additional stimuli were used, ten initiating the test and ten concluding it, giving a total of 100 stimuli per test. Only the middle 80 were used for evaluation, and these were presented in random order.

Three subjects were tested. Their results are summarized in the two confusion matrices shown in Table III. In prestressed position (UH-C-V), 73 percent were correct; in post-stressed position (V-C-UH), 77 percent were correct. If F, TH and V, THE responses are pooled, as is often done, then the correct percentages increase to 79 in prestressed position and 81 in post-stressed position. Ten prestressed consonants were identified correctly more than 75 percent of the time: B, D, P, T, N, TH, S, SH, Z, and ZH. The post-stressed consonants identified correctly more than 75 percent of the time were B, P, T, K, TH, S, SH, Z, and SH. The consonants which were identified incorrectly most often were G, M, D, and K.

An examination of the errors in the confusion matrices of Table III shows:

(*i*) The voiced stop G was often confused with T and K; and D in post-stressed position was often confused with G.

(*ii*) The unvoiced stop K was often confused with T in prestressed position.

(*iii*) The nasal M was often confused with B and V.

(*iv*) The fricative pairs V, THE and F, TH were often confused.

These errors were the major confusions in the tests. The stop confusions primarily were caused by errors in frication burst positions. The fricative pair errors were anticipated because of the acute acoustic similarities between these particular fricatives. The cause for the confusions between M and other phonemes is unknown. Further work remains to be done in this area.

A second series of tests, using sentences as test material, were run.

One test contained simple declarative, interrogative and imperative sentences. A second test contained sentences chosen from a list of sentences used often in intelligibility tests.[15]

The sentences were presented to listeners who wrote down what they heard. They were told to guess whenever in doubt. The sentences were played a second time and the listeners were allowed to make changes. The tests were scored on the number of words which were correctly identified (excluding only *the* and *a* as words). The results of the tests are as follows. For the test using simple sentences, eight listeners had an average of 92 percent of the words correct after one try, and 95 percent after the second try. For the test using the longer

TABLE III—CONSONANT CONFUSION MATRICES

		CONSONANT RECEIVED															
		B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH
CONSONANT SENT	B	13												2			
	D		15														
	G		1	3	2	6	3										
	P				13	2											
	T				1	14											
	K				2	7	6										
	M	2		4				4						5			
	N								12					2	1		
	F									10	2	3					
	TH									2	12	1					
	S											15					
	SH												15				
	V	6												7	2		
	THE		1											8	6		
	Z															15	
	ZH																15

UH-C-V MATRIX
(3 SUBJECTS)

CONSONANT RECEIVED

CONSONANT SENT	B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH	?
B	12												3				
D		7	8														
G			6		1	8											
P				14		1											
T				1	14												
K				1	2	12											
M	4						4	3					4				
N			6					9									
F									11	3	1						
TH										12	3						
S											15						
SH												15					
V	3						1						10	1			
THE													4	11			
Z															15		
ZH																15	

V-C-UH MATRIX
(3 SUBJECTS)

standard sentences four listeners had an average of 83 percent of the words correct after one try; and 86 percent after the second try.

As shown above, the percent intelligibility scores for sentences were significantly higher than for isolated syllables, primarily because of the context of speech in a meaningful utterance. However, the longer the utterance, the less intelligible it became. This is because rhythm and timing are much more important for a long sentence than for a short, simple one.

VII. SPECTROGRAPHIC EXAMPLES OF SYNTHETIC SPEECH

Fig. 6a shows wideband spectrograms of the utterance "Larry and Bob are here." The spectrogram in the upper half of the figure is the

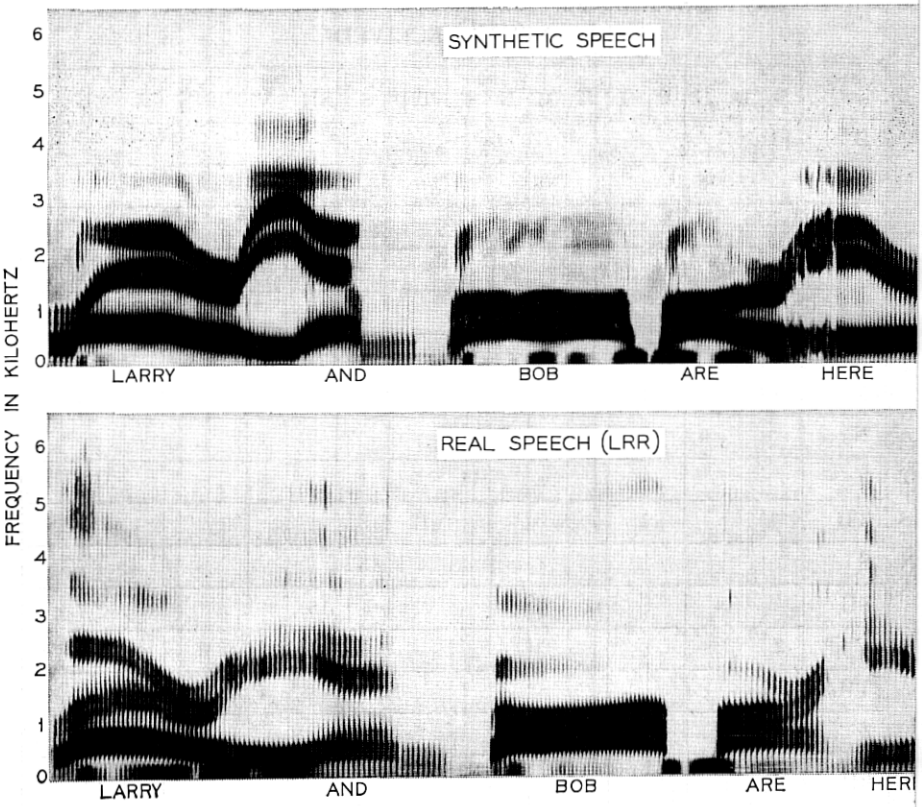

Fig. 6a — Wideband spectrograms of synthetic (top) and natural versions of "Larry and Bob are here."

synthetic version. The lower spectrogram was made from the author's speech. (The synthetic utterance was in no way modelled after or modified by the natural utterance.) Fig. 6b shows narrowband spectrograms of both the synthetic and natural versions of this utterance.

This is a high degree of similarity between the spectrograms of the real and synthetic speech. The durations of both the synthetic and natural utterances are comparable. Fig. 6a shows that the variation of the formants for both versions is quite similar. Even the fundamental frequency contours for these utterances are quite similar. As Fig. 6b shows, both contours are peaked during the stressed vowels A in BOB and I in HERE. A careful examination of the narrowband spectrograms shows the decrease of fundamental frequency, for both utterances, during the initial and final B of BOB.

The stressed vowels of this utterance can easily be identified from either the long steady state duration of Fig. 6a or the peak in the fundamental frequency contour of Fig. 6b.

VIII. FURTHER WORK

The results of the consonant intelligibility tests showed that most consonants were reproduced accurately. The sentence intelligibility tests also produced good intelligibility scores, indicating a high degree of success for the major goal of this project.

Many of the listeners made informal comments concerning the machine-like quality of the speech, but no formal tests were run to measure the naturalness or quality of the synthetic speech. Current studies

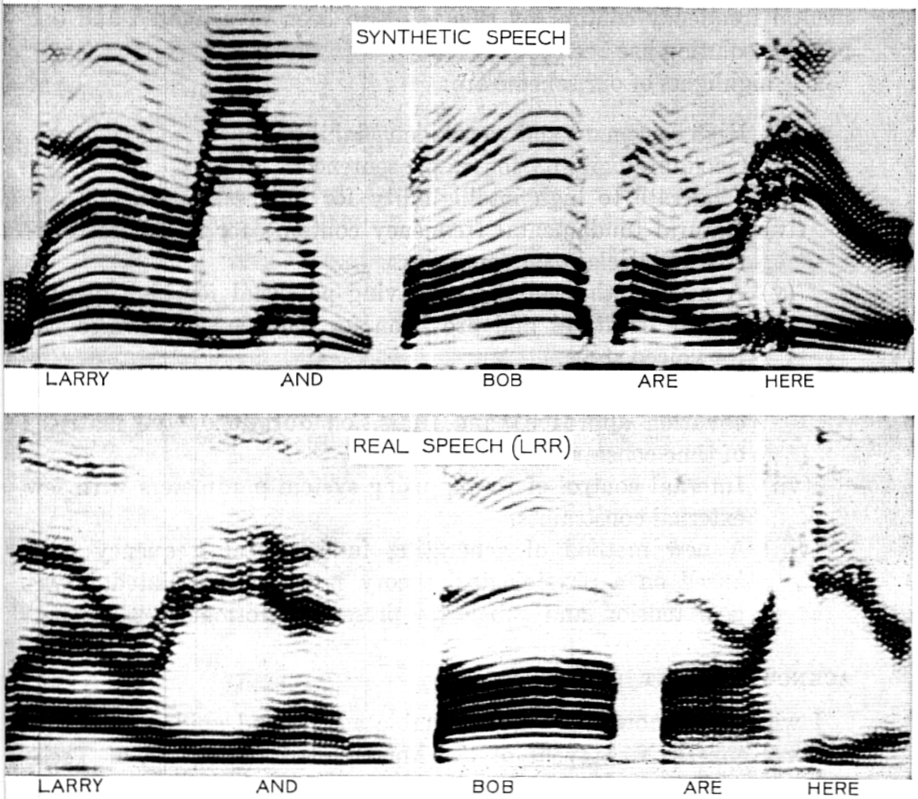

Fig. 6b — Narrowband spectograms of synthetic (top) and natural versions of "Larry and Bob are here."

about the characteristics of the source of voiced speech are expected to produce valuable information about the determinants of synthetic speech quality.

Among the topics that will be considered for future work are the effects of stress and rhythm on timing of an utterance, the inclusion of more than one breath-group in an utterance, and studies of further correlates of word boundaries.

IX. SUMMARY

An acoustic domain approach to speech synthesis by rule has been formulated and programmed on a digital computer. Samples of speech have been generated using our scheme and their intelligibility has been measured. The problem of automatically generating a fundamental frequency contour by rule has also been investigated and one possible solution has been found.

The highlights of our scheme are:

(i) High consonant and vowel identifiability.
(ii) Good intelligibility for simple sentences.
(iii) Moderate to high intelligibility for long sentences.
(iv) Natural fundamental frequency contours for both interrogative and declarative sentences.
(v) A new synthesizer design having potential for high quality voiced fricatives and provision for inclusion of a voice bar for voiced stops.
(vi) A new method of handling formant transitions (a differential equation approach) and transition durations (the matrices of time constants).
(vii) Internal control of timing using system parameters with few external constraints.
(viii) A new method of generating fundamental frequency data based on a physiological theory involving simulated laryngeal tension and subglottal pressure information.

ACKNOWLEDGMENT

I wish to acknowledge the invaluable advice and guidance of Professor Kenneth N. Stevens of the Massachusetts Institute of Technology and Doctor James L. Flanagan of the Bell Telephone Laboratories. I would also like to thank Doctors Harry Levitt and Peter Brady for valuable comments and criticisms of this manuscript.

REFERENCES

1. Henke, W., Dynamic Articulatory Model of Speech Production Using Computer Simulation, Ph.D. thesis, Massachusetts Institute of Technology, 1966.
2. Coker, C. H. and Fujimura, O., Model for Specification of the Vocal Tract Area Function, J. Acoust. Soc. Amer. *40,* 1966, p. 1271, (A).
3. Holmes, J., Mattingly, I. and Shearme, J., Speech Synthesis by Rule, Language and Speech, *7,* 1964, pp. 127–143.
4. Cooper, F., Liberman, A., Lisker, L. and Gaitenby, J., Speech Synthesis by Rules, Speech Communication Seminar, Stockholm, August 29 to September 1, 1962, Paper F2
5. Liberman, A., Ingemann, F., Lisker, L., Delattre, P., and Cooper, F., Minimal Rules for Synthesizing Speech. J. Acoust. Soc. Amer. *31,* 1959, pp. 1490–1499.
6. Peterson, G., Wang, W., and Sivertsen, E., Segmentation Techniques in Speech Synthesis. *J. Acoust. Soc. Amer. 30,* 1958, pp. 739–742.
7. Fant, G., and Martony, J., Speech Synthesis. *Speech Transmission Laboratory, Quarterly Progress Report* Univ. of Stockholm, 1962.
8. Tomlinson, R. S., SPASS—An Improved Terminal-Analog Speech Synthesizer, J. Acoust. Soc. Amer., 1965, *38,* p. 940.
9. Flanagan, J. L., *Speech Analysis Synthesis and Perception,* Academic Press, Inc., New York, 1965.
10. Fant, G., *Acoustic Theory of Speech Production,* 's-Gravenhage: 1960 Mouton and Co.
11. Gold, B. and Rabiner, L., Analysis of Digital and Analog Formant Synthesizers. Paper presented at 1967 Conference on Speech Communication and Processing, Massachusetts Institute of Technology, and accepted for publication in IEEE Trans. Audio and Elec., March 1968.
12. Rabiner, L., Speech Synthesis by Rule: An Acoustic Domain Approach, Ph.D. thesis, Massachusetts Institute of Technology, 1967.
13. House, A., On vowel duration in English, *J. Acoust. Soc. Amer., 33,* 1961, pp. 1174–1178.
14. Lieberman, P., Intonation, Perception and Language, Research Monograph 38, MIT Press, Cambridge, Mass., 1967.
15. Beranek, L., *Acoustic Measurements,* John Wiley and Sons, Inc., New York, 1949, pp. 773–777.