

Sept. 22, 1970

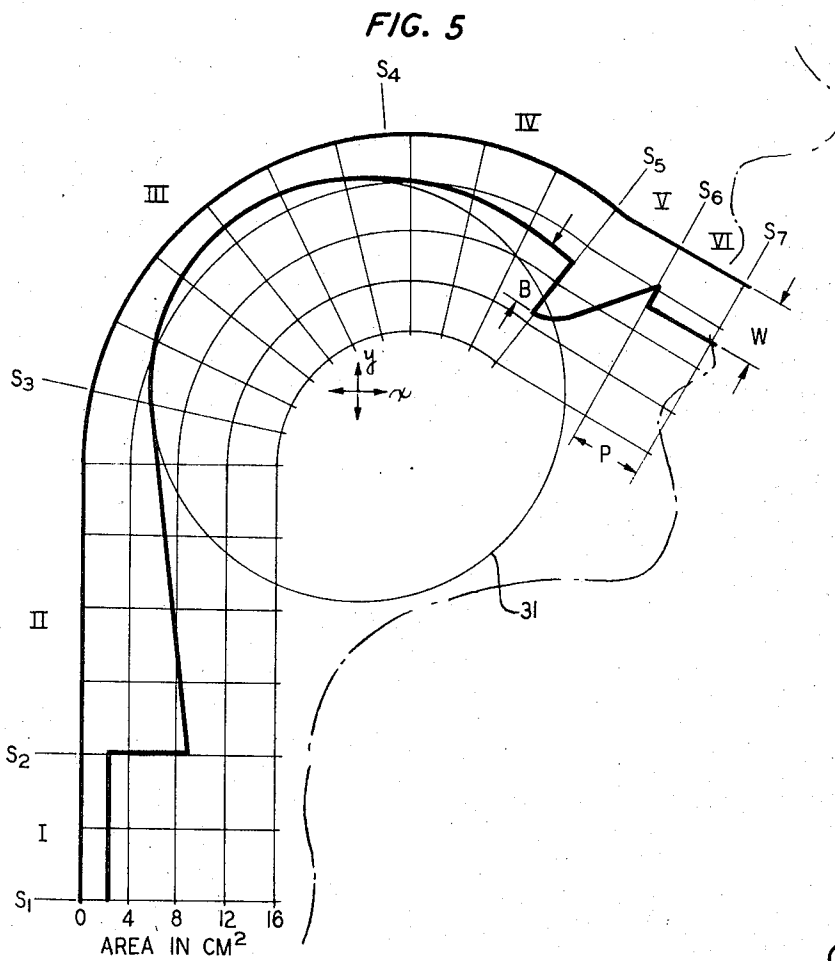
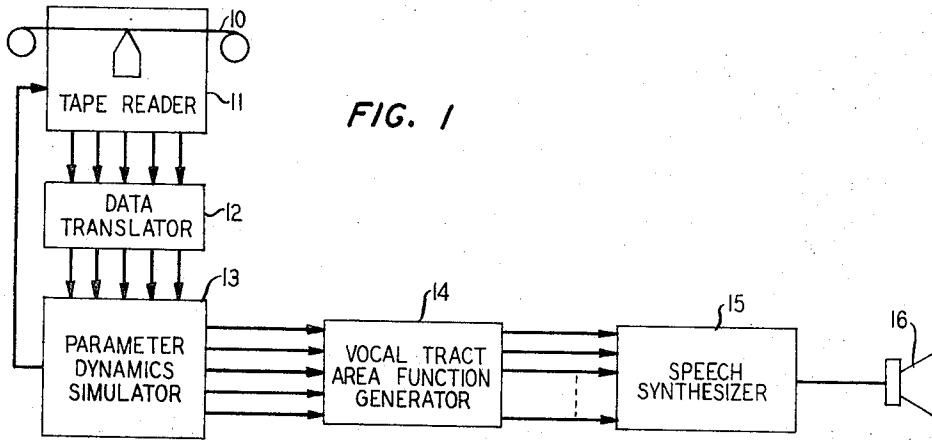
C. H. COKER

3,530,248

SYNTHESIS OF SPEECH FROM CODE SIGNALS

Filed Aug. 29, 1967

5 Sheets-Sheet 1



INVENTOR  
C. H. COKER  
BY  
G. E. Hirsch  
ATTORNEY

Sept. 22, 1970

C. H. COKER

3,530,248

SYNTHESIS OF SPEECH FROM CODE SIGNALS

Filed Aug. 29, 1967

5 Sheets-Sheet 2

FIG. 2

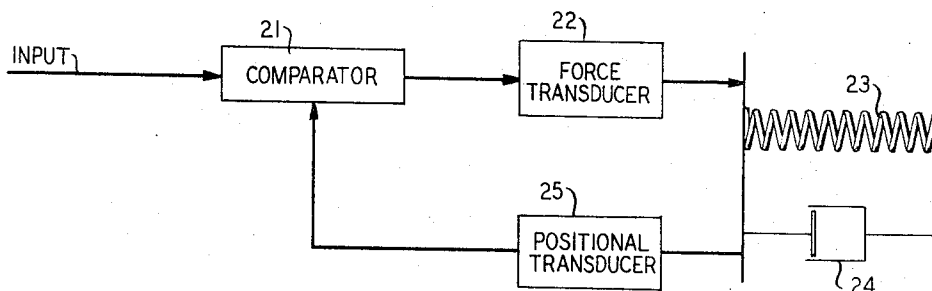


FIG. 3A

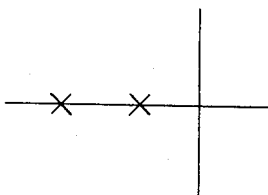


FIG. 3B

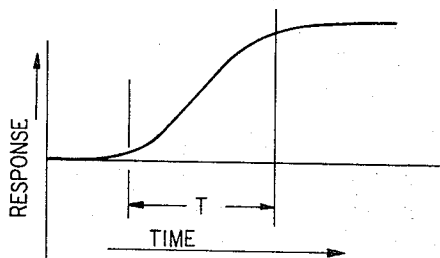


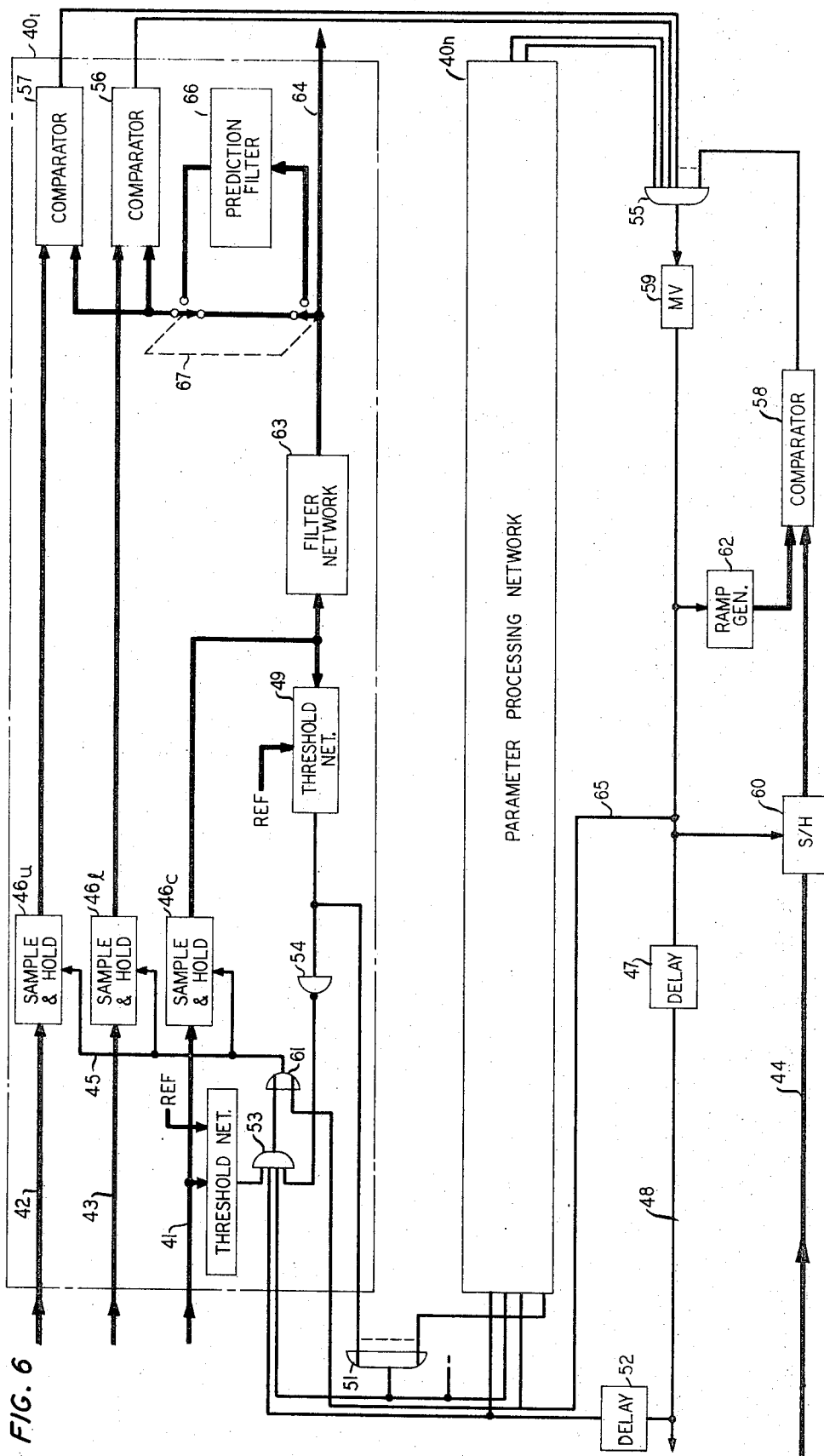
FIG. 4

| PARAMETER | RESPONSE TIME |
|-----------|---------------|
| $\alpha$  | 100 MS        |
| $\beta$   | 150 MS        |
| B         | 50 MS         |
| P         | 300 MS        |
| W         | 50 MS         |

SYNTHESIS OF SPEECH FROM CODE SIGNALS

Filed Aug. 29, 1967

5 Sheets-Sheet 3



Sept. 22, 1970

C. H. COKER

3,530,248

SYNTHESIS OF SPEECH FROM CODE SIGNALS

Filed Aug. 29, 1967

5 Sheets-Sheet 4

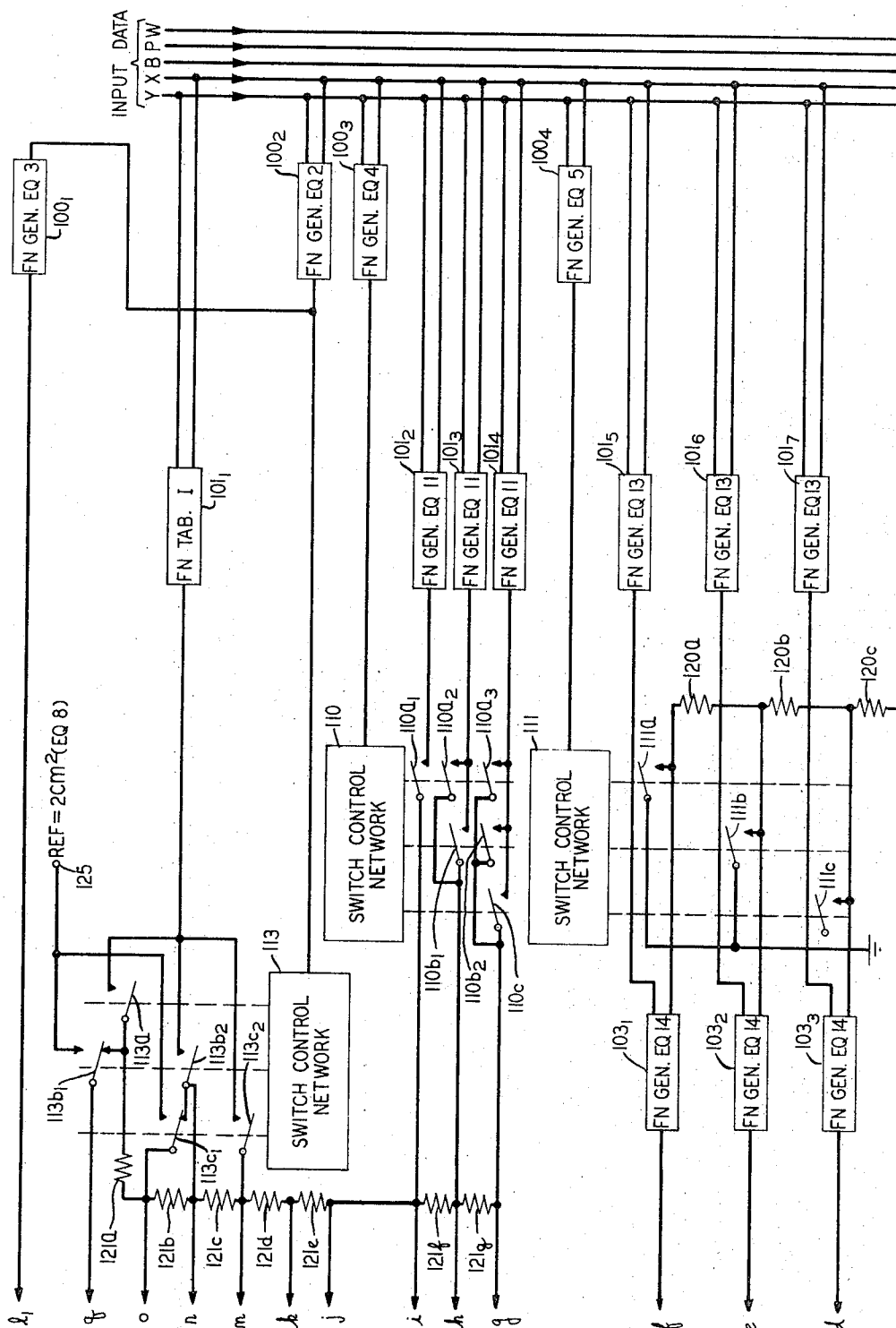


FIG. 7

**Sept. 22, 1970**

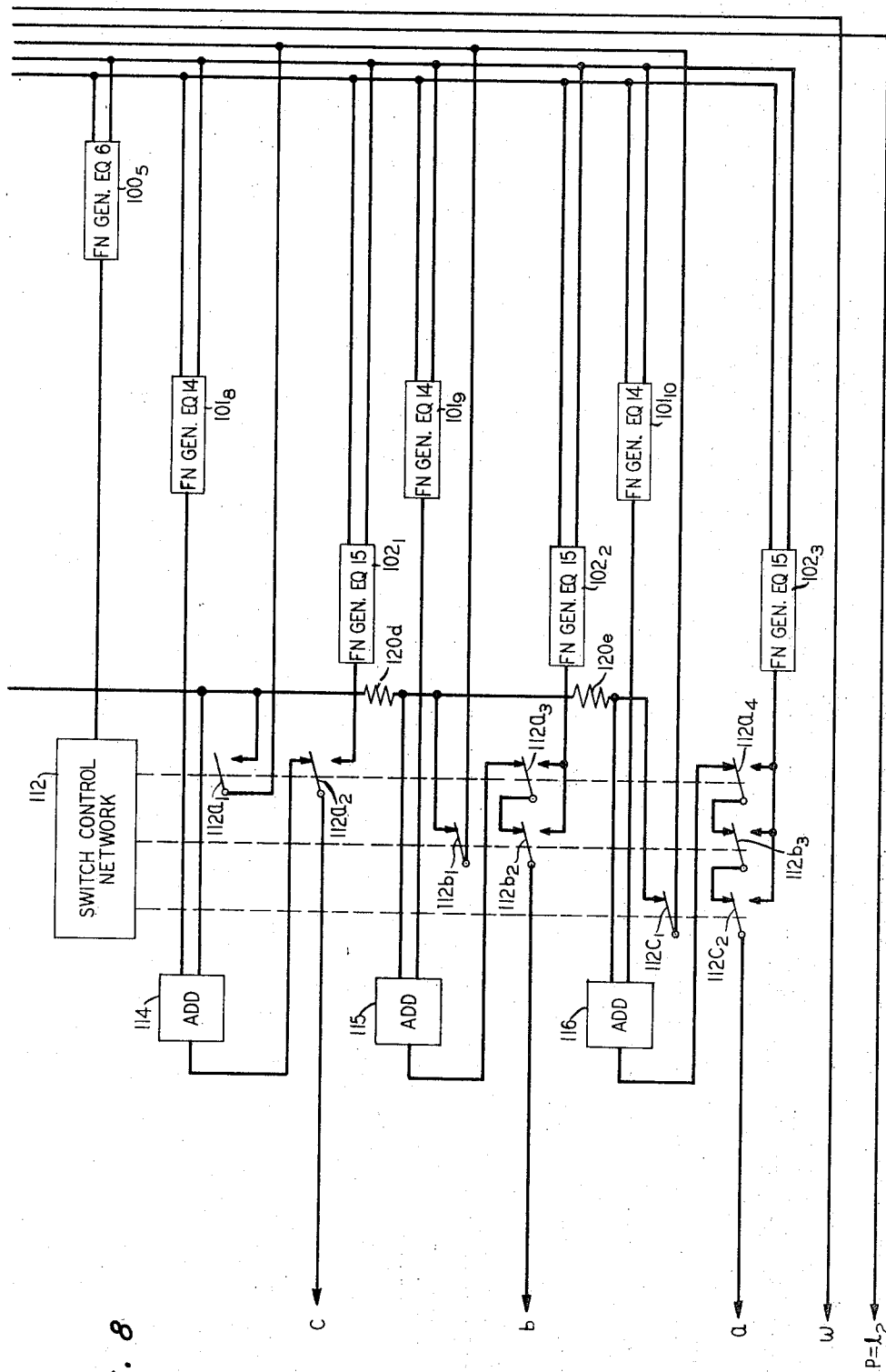
C. H. COKER

3,530,248

# SYNTHESIS OF SPEECH FROM CODE SIGNALS

Filed Aug. 29, 1967

5 Sheets-Sheet 5



1

3,530,248

## SYNTHESIS OF SPEECH FROM CODE SIGNALS

Cecil H. Coker, Chatham Township, Morris County, N.J.,  
assignor to Bell Telephone Laboratories, Incorporated,  
Murray Hill and Berkeley Heights, N.J., a corporation  
of New York

Filed Aug. 29, 1967, Ser. No. 664,129

Int. Cl. G101 1/00

U.S. Cl. 179—1

12 Claims

### ABSTRACT OF THE DISCLOSURE

In producing synthetic speech in response to recorded phonemic symbols, a number of signals must be developed for adjusting a synthesizer used for producing speech sounds and the transitional sounds between them. These control signals may be prepared by apparatus that simulates the action of the human vocal tract during phonation. An important parameter of such vocal tract apparatus is the momentary glottal area. A speech synthesis system is described which utilizes phonemic data to develop signals proportional to glottal areas.

### BACKGROUND OF THE INVENTION

This invention relates to speech synthesis, and, more particularly, to a technique for controlling an electrical analog of the human vocal tract in order to synthesize sequences of speech sounds.

It is an object of the invention to generate electrical control voltages that simulate the changing articulatory configurations of the human vocal tract and to use them in controlling a vocal tract analog speech synthesizer.

#### Field of the invention

Synthesis of human speech by artificial means is generally accomplished by way of a source of excitation signals which simulate activity of the human vocal chords; an adjustable electrical network supplied with excitation signals and designed to simulate, in one form or another, the function of the human vocal tract; and, finally, a system for controlling the frequency and intensity of the excitation and for adjusting the electrical configuration of the adjustable network in accordance with a desired sound. A variety of synthesis networks have been proposed which can be made to work well, provided that a satisfactory translation can be made between phonetic input data and synthesizer control signals.

#### Description of the prior art

One approach to the production of synthesizer control signals is to deal with variations in resonances of the vocal tract. Such resonances, called formants, constitute a rather small set of variables which, if properly generated, are sufficient to control the production of understandable speech. Synthesizing speech from a formant description is difficult, however, because there exist no rules for deriving the correct formant values for dynamic fluent speech with accuracy approaching that necessary for good quality. This is to say that, while certain very general rules exist for describing the formant transitions to produce a given set of words, the rules only crudely approximate the formant values a human speaker would produce in saying the same words. Thus, for example, while there are some similarities in the /g/'s of /igα/, /igi/ and /αgα/, there is no simple definition of /g/ in terms of formants that holds accurately in spite of changes in context.

It is not unreasonable to expect such difficulties in defining phonemes in terms of formants. There are no natural laws to force simple structures to exhibit formant behavior. In deed, the physics of human speech production quite easily justifies complex formant patterns. For

2

example, X-ray motion pictures of speakers show the vocal tract moving through a configuration of approximate symmetry with a constriction at the midpoint in the utterances /iu/ (you) and /ui/ (we). This configuration produces formants  $F_2$  and  $F_3$  that are very close in frequency. This condition, however, occurs to varying lesser degrees or not at all in other diphthongs and vowel to vowel junctures such as, for example, /au/ (as in out). Thus, /iu/ and /au/ cannot both be generated well by the same rule.

One method to circumvent this problem is to develop separate definitions for connecting each pair of speech sounds. That is, in a sense, to pre-record all possible combinations of sounds taken two at a time. This method, which substitutes storage for rules, still does not account for all effects observable in the formant behavior of human speakers. In short consonants and consonant clusters, the effect of one phoneme may be present two or more phonemes later, thus making it desirable to include combinations taken three at a time, and so on. Storage requirements are significantly expanded by this technique.

Another approach toward the production of synthetic speech involves the use of a controllable electrical analog of the vocal tract, such as a transmission line. The line is excited by a buzz source to simulate a glottal tone, by a noise source to simulate the noise of turbulence, or by both together. Control voltages are employed to adjust the line to different articulatory configurations in a desired sequence and smoothing circuits employed to provide the transitions from one configuration to the next. Such a synthesizer, given the correct area data, is capable of producing correct formant transitions. The problem is not solved, however, but merely transformed to that of supplying signals representative of vocal tract areas rather than of formants. This is, in a sense, a more difficult problem in that much more data must be provided. From 10 to 30 area values with accuracy from 6 to 10 bits must be supplied for such a form of synthesis, as compared to perhaps three formants with an accuracy of 4 to 6 bits for formant synthesis. Much of the redundancy in the input data must be properly included by correctly modeling physiological constraints to the shapes that a human vocal tract can assume. Furthermore, the movements from shapes defined for one phoneme to shapes defined for another must be realistically interpolated. Finally, this interpolation must be done with the proper timing.

### SUMMARY OF THE INVENTION

This invention deals with a source of area data for a vocal tract synthesizer that incorporates most of the natural constraints of the human vocal system. Simple interpolations are performed which approximate, closely, the way the human interpolates motions from one desired vocal tract shape to another. The dynamics of the human tract are modeled, in accordance with the invention, so that tongue tip motion, for example, is rapid compared to the motion of the central part of the tongue; lip protrusion is very slow, but lip closure is quite fast. A natural separation of coordinates of the model is thus established according to characteristic time-constants of the vocal tract. This leads to especially simple modeling of the time dependencies of vocal tract motion.

It is in accordance with this invention to so model the human vocal tract that control signals necessary for the generation of synthetic speech may be produced by subjecting applied phonemic input data to certain modifications according to specified rules. The rules define the manner by which a large number of area function control signals are developed from a smaller set of descriptive parameters. The area function control signals delineate the changing cross-section configuration of

human vocal tract at a number of points spaced between the vocal chords and the outer lips during phonation. An electrical analog synthesizer requires such area data to adjust the electrical elements used to control applied excitation in a fashion comparable to the control exerted by the vocal tract on voiced energy passing through the vocal tract.

The parameters which describe and control the development of the area control signals are defined, in accordance with the invention, by a series of equations derived from studies made of the human vocal system during the production of speech. From these studies it has been found possible to develop a model of the vocal tract which permits linear or simple interpolations to be made from one area shape to another with a high degree of accuracy. By arranging the model in a "curved" configuration, which more nearly resembles the configuration of the human vocal tract, a system of equations are formulated in terms of connected straight line segments and circular portions. Such sections are amenable to relatively simple definition and, moreover, provide the necessary approximations to permit linear interpolations as speaking takes place. Thus, it is possible to characterize the motion between two known configurations of the vocal tract, derived, for example, from numerous examinations of the human vocal tract, without the necessity of specifying all intermediate positions.

In accordance with the invention, this simplification is achieved by separating the vocal tract into a number of distinct regions and defining each region by a number of parameter values which are closely correlated with actual vocal tract configurations. The individual parameters thus define the shape and motion of the physical parts of the vocal tract, such as articulators, lips, lip protrusions, tongue, et cetera. Parametric signals are developed in response to input data specifying certain phonetic symbols, stresses, and pause information. The parameters include two coordinates locating the central part of the tongue; a parameter describing the position of the tongue tip, and two parameters describing, independently, lip opening and lip protrusion.

Apparatus constructed in accordance with this invention typically employs a system for reading phonemic symbols and associated instructions from a paper tape or the like, and a translator for converting these data into electrical control signals, preferably in digital form. The control signals are then subjected to modification according to stored rules and further modified in accordance with parameter data maintained in storage. A sequence of smoothly changing analog signals is then developed which is sufficient to define the changing areas in the vocal tract, stresses and pauses in excitation, and variations in the length of the vocal tract both between the vocal chords and the pharynx and in the length of the tract between the pharynx and the teeth or outer lips. Parameter data is thus converted into a sequence of area control signals which may be applied to a vocal tract analog speech synthesizer and used to produce audible speech signals.

The invention will be fully apprehended from the following detailed description of an illustrative embodiment thereof taken in connection with the appended drawings.

#### BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 shows, in block diagram form, the various circuit units employed in the practice of the invention;

FIG. 2 is a simplified schematic illustration of a model of a vocal tract organ;

FIGS. 3A and 3B illustrate typical pole-zero and wave form configurations useful in specifying the characteristics of filter apparatus used in the practice of the invention;

FIG. 4 is a table of typical response times for different vocal tract parameters;

FIG. 5 shows diagrammatically a model of the

human vocal tract indicating the manner in which area function data is defined;

FIG. 6 illustrates in block diagram form a signal dynamics translator suitable for use in the assembly of elements included in the apparatus of FIG. 1; and

FIGS. 7 and 8 together illustrate in block diagram form a suitable implementation of an area translator that may be used in the practice of the invention.

#### DETAILED DESCRIPTION OF THE INVENTION

FIG. 1 illustrates apparatus for translating from printed intelligence to spoken intelligence. A paper tape 10, or the like, containing sequences of phonetic symbols, together with punctuation and pause information, is supplied to paper tape reader 11 wherein data on the tape is converted into electrical signals. Reading is accomplished in response to a command signal supplied from signal dynamics translator 13. Signals supplied from the reader are delivered to translator 12 wherein a logical analysis of the supplied data is performed. In dependence on the identity of successive phonetic symbols, a sequence of analog signals is developed representing vocal tract target shapes, in parametric form, and necessary stress and pause information. Suitable read and translate apparatus is described, for example, in Gerstman et al., Pat. 3,158,685, granted Nov. 24, 1964.

The translated signals are supplied to parameter dynamics simulator 13 where they are processed to produce smoothly changing vocal tract parameters. Another function of simulator 13 is to govern the rate of reading of new phonetic symbols or other characters by tape reader 11. In general, the simulator provides filtering of the vocal tract parameters according to the characteristic speed of change of the individual parameters which, in some cases, is in turn dependent upon the particular sequence of phonemes represented by the taped symbols. In addition, simulator 13 may be used to modify the duration of a phoneme according to context, thus conditioning the reading of the next phoneme character from tape 10 upon the time required for the parameters to change from their previous positions to within an acceptable approximation of the ideal values for that phoneme. Details of the construction and operation of a suitable parameter dynamics simulator will be disclosed hereinafter with reference to FIG. 6.

Smoothly varying signals produced by generator 13 are applied to a vocal tract parameter-to-area function generator 14. Function generator 14 derives from the input parameters a number of signals representing vocal tract areas at selected points along the tract. These signals are suitable for controlling vocal tract analog speech synthesizer 15. Details of the construction and operation of function generator 14 are disclosed below with reference to FIGS. 7 and 8 and a suitable vocal tract speech synthesizer is described by G. Rosen in 30 Journal of the Acoustical Society of America 3, (March 1958), beginning at page 201. Analog speech signals developing by synthesizer 15 may be supplied to any desired utilization apparatus such as, for example, loud speaker 16.

Before describing apparatus suitable for developing the necessary parameter signals for energizing synthesizer 15, it is believed that a discussion of the various relations that lead to a specification of vocal tract area parameters will be helpful.

#### ANALYTICAL DISCUSSION

The problem essentially is one of developing a set of output signals representative of vocal tract areas from a set of continuous input parameters defined in terms of sequences of phonemes. For practical reasons, a limited number of different areas at selected points along the length of the vocal tract are chosen although, of course, the more areas selected the better the control of the synthesizer and, hence, the better the quality of the reproduced speech.

Basic vocal tract area information is available from the numerous studies of the human vocal tract that have been reported in the literature. In essence, these studies specify the area and the shape of the vocal tract at various points along its length for various standard sounds. Obviously these data are stylized in that different speakers with different accents and different vocal tract configurations employ different vocal tract configurations for producing the same or similar sounds. Moreover, these data are static; they define the vocal tract at an instant during the utterance of a classified sound. The transitions between such sounds cause changes in the vocal tract configuration that are extremely complex and, again, are different for different speakers and for different combinations of sounds made by the same speaker.

In accordance with the present invention, a model of the human vocal tract is established which closely approximates the human vocal tract. A unique feature of this model is that parts of the real vocal system, that are separately controllable and have distinct response characteristics, appear in the model as separate parts. This permits the simulation of speech dynamics for interpolation between phonemes by direct modeling of the characteristics of distinct physical elements rather than by empirical approximation of the combined effects of the overall speech system. Parameter information for the model is determined, in a large measure, from published data acquired from numerous examinations of human vocal tract systems.

A possible model for an organ of the vocal tract, or for a parameter of this model, is illustrated schematically in FIG. 2. The model is organized as a feed-back control system with force applied to a mechanical system exhibiting mass, represented by spring 23, and viscosity, represented by dash pot 24. The driving force is controlled by comparator 21 which compares the present position of the system in response to the applied force with the desired position. Force transducer 22 establishes the desired position and positional transducer 25 provides an indication of the desired position.

Several nonlinearities are manifest in the feed-back system. Force transducer 22 limits the available force to different values for positive and negative polarities. The mechanical system consisting of mass element 23 and viscosity element 24 exhibit nonlinearity in restoring springiness and viscosity, for example, increasing when some other part of the anatomy obstructs further motion. Positional transducer 25 establishes nonlinearity in the feed-back path. Thus, additional feed-back is provided when closure occurs at some point in the vocal tract.

FIGS. 3A and 3B shows, respectively, typical pole-zero configurations and wave forms of an actual filter network constructed to exhibit the characteristics of the model of FIG. 2, assuming a linear approximation.

FIG. 4 tabulates typical response times,  $T$  in FIG. 3B, for different area parameters. To a first order, nonlinearities may be ignored and linear filters, having the response characteristic of FIGS. 3A and 3B, may be employed.

FIG. 5 illustrates a simplified model of the human vocal tract developed in accordance with the invention. It is divided into six basic segments, I through VI, and is further divided by a superimposed grid. Within reasonable accuracy, the different segments correspond to distinct regions of the human vocal tract as follows:

Segment I. Portion of the tract from vocal cords to the top of the larynx.

Segment II. From the larynx to a point approximately 1 cm. below the soft palate.

Segment III. From approximately 1 cm. below the soft palate to the highest point of the tongue (with tip low).

Segment IV. High point of the tongue to tongue tip.

Segment V. Tongue tip to teeth.

Segment VI. Lips.

Mathematical specifications of vocal tract area at points along the vocal tract may be derived empirically from a geographic representation of the vocal tract created by superimposing a stylized model of the vocal tract on a graphic scale as shown in FIG. 5. In this coordinate system, which is curved to permit linear interpolations to be made from one area position to another, the throat is represented by linear Segments I and II. The tongue is represented in Segment III by a portion of the periphery of a superimposed circle 31 which is fitted, in diameter and coordinate position, to match expected curvature of the tongue. The tongue is represented as a part of a circle because the normal tongue position exhibits a substantially constant radius for most human vocal tract shapes. The high point of the tongue is represented by Segment IV, which again is characterized in the graphical representation by a curved portion. Segment V denotes the segment of the tract between the tongue tip and the teeth. It, too, is formed of curve segments. Finally, Segment VI is a linear portion which represents the lips. As noted above, the lips may be characterized as closed or protruding, depending upon the particular phonetic sound being voiced.

By virtue of the simplified representation of the vocal tract, the division of the model into distinct segments, and the use of simple geometric forms, the areas of the vocal tract at points within the several regions may be defined by five quantities as follows: two ( $X$ ,  $Y$ ) to represent the position coordinates of circle 31; one ( $B$ ) to specify raising of the tongue tip or blade; one ( $W$ ) to specify lip opening; and another ( $P$ ) to denote lip protrusion.

Locations of the segment boundaries ( $S_0$  through  $S_7$ ) are obtained with relation to the  $X$  and  $Y$  coordinates of the system. The  $S_0$  boundary is selected as the reference point:

$$S_0 = 0 \quad (1)$$

The  $S_2$  boundary is defined to indicate that the pharynx lengthens as the center tongue moves upward to the back:

$$S_2 = -.5X + .866Y + 13 \text{ cm.} \quad (2)$$

Since the distance from the vocal chords to the esophagus is a constant,

$$S_1 = S_2 + 2 \text{ cm.} \quad (3)$$

the point of transition from the linear to the curved portion of the system is

$$S_3 = 11 \text{ cm.} - .7Y \quad (4)$$

$S_4$  defines the upper portion of the tongue hump circle, as follows:

$$S_4 = 4 \text{ cm.} - .7X \quad (5)$$

Similarly, the forward portion of the tongue circle becomes:

$$S_5 = .86X + .5Y \quad (6)$$

and the terminal boundary is thus represented by

$$S_7 = P \quad (7)$$

Areas for positions within the six segments are obtained as follows:

Segment I

$$A(s) = 2 \text{ cm.}^2 \quad (8)$$

Segment II

$$A(s) = \frac{s - S_2}{S_1 - S_2} F(X \cos 60^\circ + Y \sin 60^\circ) + \frac{S_1 - s}{S_1 - S_2} CIR(X, Y, s) \quad (9)$$

These equations represent linear interpolations between boundaries  $S$ . In Equation 9, which represents an area at  $S_1$ ,  $A(S_1) = F(X \cos 60^\circ + Y \sin 60^\circ)$ , the function  $F(u)$  is defined as a nonlinear function fitting the 75 points of Table I.

7  
TABLE I

| <i>u</i> : | <i>F(u)</i> |
|------------|-------------|
| -.2        | 4           |
| 0          | 6           |
| .5         | 9           |
| 1          | 11          |
| 2          | 13          |

This term approximates the "pinching down" of the pharynx area as the tongue is moved toward the lower pharynx.

$$A(S_2) = CIR(X, Y, s) \quad (10)$$

is defined to correspond to the "circle within a circle" nature of the model for the tongue:

$$CIR(X, Y, s) = 9 + 6[X \cos \alpha + Y \sin \alpha + \sqrt{1 - (X \sin \alpha - Y \cos \alpha)^2}] \quad (11)$$

where

$$\alpha \equiv \pi(s+3)/14 \text{ radians} \quad (12)$$

Segment III

$$A(s) = CIR(X, Y, s) \quad (13)$$

Segment IV

$$A(s) = CIR(X, Y, s) + \left( \frac{S_3 - s}{S_3 - S_4} \right)^2 B \quad (14)$$

This is the circle function again with an added term for raising the tongue tip or blade.

Segment V

$$A(s) = -6[X \cos 30^\circ + Y \sin 30^\circ] + F(s) \quad (15)$$

where  $F(s)$  is a nonlinear function fitting Table II.

TABLE II

| <i>s</i> : | <i>F(s)</i> |
|------------|-------------|
| 0          | 6.5         |
| .5         | .8          |
| 1          | 9.5         |
| 1.5        | 10.5        |
| 2          | 11          |

This table is chosen to approximate the shape of the teeth and bone structure behind the teeth.

Segment VI

$$A(s) = W \quad (16)$$

The basic principle involved in generating vocal tract parameter control signals in accordance with the invention is thus to provide sequentially, according to an input string of phonetic symbols, vocal tract shapes in the form of target values of parameters of the model. It is the nature of the unique model developed in accordance with the invention that simple filtering of the particular input parameter will provide natural transitions between discretely prescribed targets.

A deficiency of simple low-pass filtering of control signals for a vocal tract synthesizer is that it does not consistently provide temporal overlap in closures of the tract at different points for consonant clusters. For example, the /db/ in "bad boy" or the /bd/ in "drab day" demand overlap in closures. Thus, it is common in saying the words "bad boy" for the lips to begin moving towards closure during the "d" sound in anticipation of the "b" sound.

A technique for correcting this deficiency, which essentially parallels human vocal tract action, is to define an "important" feature associated with certain parameters of the model and to utilize several phonemic input signals to control parameter development. "Important" (I) features may thus be associated with the vertical coordinate  $Y$  for the central tongue segment, for the tongue blade  $B$ , for lip opening  $W$ , and for vocal tract length  $S$ . If the given parameter for a particular phoneme is above a threshold in the direction to cause closure or to cause lip

protrusion, the parameter is said to be "important." In presenting parameter values to the filter networks when the current phoneme is identified as an I parameter and the next phoneme also is identified as an I parameter, the I parameter of the next coming phoneme is applied to the filter network only after other parameters of the current phoneme have been presented.

From the above analysis of vocal tract action, it is thus evident that in some cases it is advantageous to stagger the time of presentation of new values to the several filters used in interpolating between different parameter values. Staggering is specified, in accordance with the invention, at the boundaries between two phonemes which require closure at different points in the vocal tract. Staggering thus avoids an interval between the two phonemes devoid of closure, i.e., avoids the inclusion of an interposed voiced sound between two stop consonants, as between "d" and "b" in "bad boy."

Additionally, the nature of the particular set of parameters allows the inclusion of some of the durational effects which are observed in many languages. For example, the vowels /i/, /ae/, /a/, /o/, and /u/, are generally longer in duration than the vowels /I/, /e/, and /A/. These differences are justified physically since the former vowels are produced by relatively extreme positions of the vocal tract and, hence, require more time either to force articulators far from their relaxed positions, or to settle the natural physiological filters to their target positions with the necessary accuracy.

Sluggishness of the vocal tract is the most limiting factor in the speed of speech communication. Some investigators believe that the human being controls his speed of speaking by a technique of "base touching," i.e., of moving on to the next phoneme as soon as the relevant part of the vocal tract has converged within acceptable limits (but not exactly) for producing the current phoneme. This accounts for the fact that a given phoneme can be, and generally is, said more rapidly in some contexts than in others.

A relation between configuration distance to a phoneme and the time allotted for traversing the distance is readily provided in this invention in terms of the parameters of the model. Current values of the model parameters, or predictions of future values as processed through prediction filters, may be compared with acceptance limits established for the parameter in the current phoneme. A new phoneme is then called whenever all parameters are found to be acceptable. Error bounds are specified along with target values. Independent upper and lower limits are provided and maintained in storage for use during parameter processing. Upper and lower limits are employed because of nonlinearities in the effectiveness of parameters in altering the acoustic output of the system (a parameter becomes more sensitive as it approaches a tract closure). Stressing of a phoneme either narrows its acceptance bounds or enforces a minimum prescribed duration in addition to the required parameter convergence toward closure.

## APPARATUS

FIG. 6 shows the elements of a vocal tract dynamics simulator which turns these considerations to account, and which is suitable for use in the apparatus of FIG. 1. It employs a combination of analog and digital networks and serves to transform phoneme target data, obtained from translator 12, into vocal tract parameter data which is used by generator 14 in the development of synthesizer control signals.

Simulator 13 examines phoneme data supplied to it and performs the necessary filtering to interpolate between known phoneme values. In addition, simulator 13 detects I phonemes that call for overlap or stress in the generation of control signals. In order to conduct this examination, each supplied phoneme representation is sampled and held for an interval sufficient to permit a

number of logical decisions to be made. This requires, of course, that sufficient delay be provided between the delivery of input data and the transfer of output data to the control generator. If no overlap or emphasis is required, normal processing takes place. If, however, an I phoneme is detected, a special signal is developed which initiates the generation of overlap or stress data. Furthermore, a signal is produced which is delivered to tape reader 11 (FIG. 1) to request delivery of a new set of phoneme data.

Simulator 13 employs a number of identical logic networks 40<sub>1</sub> through 40<sub>n</sub>, one for each parameter required for area specification. In one application, it has been found that five parameters, viz., X, Y, B, P, and W, as discussed above, are sufficient. Hence, *n* is equal to 5 in this example. Each network includes a filter network 63 which performs the basic interpolation operation in the fashion discussed above. Each filter 63 preferably exhibits a characteristic similar to that illustrated in FIGS. 3A and 3B. In addition, the networks incorporate simple delay to assure that the outputs of all filters are synchronized.

Center value input signals, delivered to parameter networks 40 on leads 41, are thus delivered to filters 63 and processed to develop proper intermediate values which are then delivered via leads 64 as the five inputs to generator 14 of FIGS. 1, 7, and 8.

Additional apparatus is required to provide for overlap and stress called for by I phonemes. Accordingly, signals which include sufficient data to specify upper and lower acceptance limits for each parameter are delivered to networks 40 on leads 42 and 43, respectively, and a minimum duration signal is delivered via lead 44 to apparatus associated with networks 40.

Translator 12 (FIG. 1) normally operates to supply data for one phoneme in advance of the time that it is to be utilized by synthesizer 15. This procedure allows time to advance the application of an I parameter to one or more of parameter processing networks 40. When a new phoneme is desired, a pulse is developed on line 45 which causes the output signals applied by translator 12 to be sampled and held by networks 46, through 46<sub>5</sub>. After a suitable interval, provided by element 47 equipped with a delay time sufficient to permit circuits 46 to operate, translator 12 is ordered to fetch signals for a new phoneme by way of a signal delivered to the translator on line 48.

If an I parameter is present, and if certain other qualifications are met, the I parameter is sampled in advance of its normal sampling time. That all of the conditions for producing early sampling have been met is determined by AND gate 53. Center frequency signals for the current phoneme are applied, from sample and hold network 46<sub>5</sub>, to threshold comparator 49. Center frequency signals for the next incoming phoneme are those available at the input to threshold comparator 50. If the parameter for the current phoneme is above a selected threshold, as established by comparator 49, or by the counterpart of comparator 49 in any one of networks 40, a signal is produced that indicates that an I parameter is involved somewhere in the system for the current phoneme. These signals are supplied to OR gate 51 and thence to one input of AND gate 53.

Normally, all parameters are sampled simultaneously in response to signals on lead 65 by way of OR gate 61, and held until the next pulse appears on line 65. However, if some other parameter is important in the present phoneme, then the parameter important in the upcoming phoneme is sampled early by virtue of the pulse which appears at the output of AND gate 53 and which is delivered to OR gate 61.

Thus, if the parameter is important for the next incoming phoneme, threshold comparator 50 also produces the signal which is delivered to gate 53. If the parameter is not important for the current phoneme in a particular processing network (OR gate 51 delivers a signal if an I

parameter is detected in any one of networks 40), NOT circuit 54, which may be an inverter, delivers an enabling signal to AND gate 53. Finally, AND gate 53 receives an enabling signal via delay element 52 when the time is appropriate for an early transfer of new data. This signal, derived from the signal on lead 48, was utilized to call for the transfer of the current phoneme and fetch the upcoming one.

If all four conditions are met, AND gate 53 delivers a signal to denote an early transfer of data. This signal is applied to OR gate 61 and then to sample and hold networks 46.

The proper time for a normal transfer of data into the processing networks 40 is achieved by the fulfillment of a number of conditions as determined by AND gate 55. If all current parameter values being processed in networks 40 are above their required low limits (low limit is specified by the signal delivered on lead 43) and are below their required upper limits (upper limit is specified by the signal delivered on lead 42), signals are forwarded by comparator networks 56 and 57 to AND gate 55. In some applications, predictions of future values may be used instead of current values. If this is the case, signals from filter network 63 are additionally processed in prediction filter 66 before being delivered to comparators 56 and 57. Prediction filter 66 may be any one of a variety of arrangements for predicting the value of a signal on the basis of a number of past, or future, values of the signal. Such filter arrangements are known variously as prediction networks, linear predictors, or the like. One suitable arrangement is described, for example, in B. M. Oliver Pat. 2,701,274, granted Feb. 1, 1955. Typically, prediction apparatus employs a transversal filter proportioned individually to weight delayed values of a supplied signal and to develop a composite signal from selected weighted values. Switching arrangement 67 permits this additional processing to be employed as desired.

Additionally, the time elapsed since the last transfer of data must be greater than the minimum specified duration of the current phoneme with its indicated degree of stress. Thus, a timing circuit such as ramp generator 62 is triggered by a transfer pulse and, after an interval, delivers a signal to comparator 58. Comparator 58 is supplied with a signal from network 60 which denotes a minimum acceptable time. A comparison is carried out in comparator 58 and, if the indicated transfer time is greater than the minimum of acceptable time for the indicated phoneme, a signal is delivered to AND gate 55.

When enabled, AND gate 55 triggers a pulse generator, such as the one-shot multivibrator 59, which, in turn, delivers a transfer signal to OR gate 61 and, via delay element 47, to tape reader 11. This pulse is also supplied to sample and hold network 60.

The entire analysis process as described above is carried out for each set of phoneme data supplied from translator 12. As a consequence, intermediate value parameter signals (five in this case) are transferred from filters 63 in networks 40 to vocal tract area function generator 14.

One suitable implementation of a parameter-to-area translator 14 for use in the apparatus of FIG. 1 is shown in FIGS. 7 and 8, taken together. Analog computer techniques are employed for generating voltages that are functions, either linear or nonlinear, of one or more input voltages. Comparable digital techniques may, of course, be employed as desired.

Input signals W, P, B, X, and Y, received from the five parameter circuits 40 of simulator 13, are supplied to each of function generators 100<sub>1</sub> through 100<sub>5</sub>, to be each of function generators 101<sub>1</sub> through 101<sub>10</sub> and to each of function generators 102<sub>1</sub> through 102<sub>3</sub>. Function generators 100 serve to develop voltages according to Equations 2 through 6, established above, which represents the points along the vocal tract for transitions from one mode of area computation to the next. Function

generators 101 are implementations of Equations 9 through 15. Their outputs represent vocal tract areas or components of area at certain points along the vocal tract. Similarly, function generators 102 are implementations of Equation 15.

Output signals developed by function generators 100<sub>2</sub> through 100<sub>5</sub> are selectively applied to a number of gate control networks 110 through 113, which, in turn, control a number of associated switches 110<sub>a</sub>, 110<sub>b</sub>, 110<sub>c</sub>, and so on. Each driving network is structured to control one row of switch elements, i.e., *a*, *b*, or *c*, in response to different applied voltages. Thus, for example, if the voltage at the output of function generator 100<sub>2</sub> is below threshold *a*, the switches on row *a* are actuated. If the voltage is between threshold *a* and *b*, the switches on row *b* are actuated.

Controller 113 determines whether outputs *q* through *m* are established by Equation 8 or by Equation 9. Controller 110 determines whether outputs *g* through *i* are computed according to Equation 9 or Equation 13; and controller 111 determines whether outputs *d* through *f* are derived by Equation 13 or by Equation 14. Similarly, controller 112 determine whether outputs *a* through *c* are established by Equation 14 or by Equation 15. In addition, these controls establish the necessary voltages on resistors 120 and 121 to drive component terms for those equations.

The switches serve to permit function signals developed by selected ones of the function generators to pass directly to output terminals, there to be delivered to vocal tract synthesizer 15, or FIG. 1. Thus, for example, if switches 113<sub>b1</sub> and 113<sub>b2</sub> are actuated by gate control network 113, an output is delivered from terminal 125, as a fixed voltage representing the quantity 2 cm.<sup>2</sup>, to output terminal *q*. This signal represents, according to Equation 8, an area specification for Segment I. With the same switching arrangement, a signal is obtained from function generator 101<sub>1</sub>, which develops a signal in accordance with the nonlinear relation of Table I. This voltage is supplied via switch 113<sub>b2</sub> to the junction of resistors 121<sub>b</sub> and 121<sub>c</sub> to provide linear interpolation voltages at terminals *g* through *u*. Because no voltage appears across resistors 121, the voltage output of generator 101<sub>1</sub> appears at terminal *o*.

As another example, if switch 110<sub>c</sub> is actuated by gate control network 110, the other end of the voltage divider formed by resistors 121<sub>a</sub> through 121<sub>g</sub> is energized in order that the terms of Equation 11 may be supplied from output terminals *g* through *o*.

Similar methods are employed to supply voltages to resistors 120<sub>a</sub> through 120<sub>g</sub> to produce interpolation terms for Equation 14. Either of two alternatives may be employed in squaring this term as required by the equation. One is to use an approximation for the squaring by tapering the values of resistors 120. Another is to supply the square root of *B* (obtained from simulator 13) to the end of the interpolation network and to square the outputs of the divider network. For simplicity, the former method is illustrated in the figure.

To complete the generation of area functions, according to Equation 14, several terms are added together. This is accomplished by selectively interconnecting the output signals produced by function generators 101<sub>5</sub> through 101<sub>10</sub> in order to add them, as required, to the output signals produced by function generators 102<sub>1</sub> through 102<sub>3</sub>. Adder networks 114, 115, 116 serve to combine the necessary terms.

In the fashion thus described, the apparatus of FIGS. 7 and 8 utilizes the five parameter signals, *W*, *P*, *B*, *X*, and *Y* supplies from translator 13 and prepares suitable area control signals on output leads *a* through *k*, 1<sub>1</sub>, 1<sub>2</sub>, *m*, *n*, *o*, and *q*. These signals are sufficient to actuate synthesizer 15 or FIG. 1.

All function generators are constructed according to principles well known in the analog computing art. These

principles are discussed, for example, in Korn and Korn, "Electronic Analog Computers," 2nd ed., McGraw-Hill, New York, 1956; Saroka, "Analog Methods in Computation and Simulation," McGraw-Hill, New York, 1954; and Karplus, "Analog Simulation," McGraw-Hill, New York, 1958.

Gate control networks 110 through 113 are threshold comparison devices that selectively actuate one of a number of sets of gates or switches, depending upon the magnitude of the applied input voltage. A suitable gate control network is described in connection with FIG. 5 of C. H. Coker, Pat. 3,327,057, granted June 20, 1967.

The above-described arrangements are, of course, merely illustrative of the application of the principles of the invention. Numerous other arrangements may be devised by those skilled in the art without departing from the spirit and scope of the invention. For example, additional independent vocal tract parameters, such as jaw angle, and tongue positions relative to the jaw, may be included in the vocal tract model.

What is claimed is:

1. Apparatus for the synthesis of speech from code signals, which comprises, in combination,
  - means for generating signals representative of a selected phoneme sequence,
  - means for generating a plurality of area function signals representative of physical portions of the human vocal tract at each of a number of points spaced between the vocal cords and the outer lips,
  - means responsive to said selected phoneme signals for modifying said area function signals to delineate changes in vocal tract configuration during phonation,
  - a controllable vocal tract speech synthesizer, and
  - means responsive to said modified area function signals for controlling said vocal tract speech synthesizer to produce speech sounds according to said phoneme sequence.
2. Apparatus which comprises, in combination,
  - means for generating signals representative of a selected sequence of phoneme representations,
  - means for generating signals representative of selected parameters which define characteristics of the human vocal tract between the glottis and the outer lips,
  - means responsive to said phoneme signals for modifying said parameter signals in accordance with changes in the areas of the cross section configurations of the human vocal tract at selected points therein,
  - a controllable analog of the vocal tract for synthesizing speech sounds, and
  - means for utilizing said modified vocal tract area signals for controlling said vocal tract analog.
3. Apparatus as defined in claim 2, wherein said parameter signals represent individual vocal tract parameters representative of motions of selected physical parts of the human vocal tract.
4. Apparatus as defined in claim 2, wherein said parameter signals represent individual vocal tract parameters including (1) lip protrusion, (2) lip closure, (3) tongue tip location, (4) horizontal position of the tongue hump, and (5) vertical position of the tongue hump.
5. Apparatus for producing artificial speech sounds, which comprises, in combination,
  - means for translating coded representations of a selected sequence of phonemes into electrical signals,
  - means responsive to said electrical signals for generating a plurality of parameter signals each definitive of the configuration of the human vocal tract assumed in voicing a sound corresponding to a phoneme in said sequence,
  - means responsive to said parameter signals for generating a plurality of control signals proportional to said vocal tract areas at selected points,
  - a controllable analog of the vocal tract for synthesizing speech sound, and

13

means for utilizing said vocal tract area signals for controlling said speech synthesizer.

6. Apparatus as defined in claim 5, wherein said translating means includes means for developing electrical signals which denote the frequency range and duration of a vocal sound specified by each of said sequence of phonemes.

7. Apparatus as defined in claim 6, wherein said means for generating parameter signals includes means for developing from said frequency range and duration signals interpolations for each of said parameters representative of a transition from one voiced sound to another.

8. Apparatus as defined in claim 7, wherein said means for developing a plurality of interpolations includes a filter network for each of said parameters.

9. Apparatus as defined in claim 7, wherein said means for developing interpolations of said parameter signals is responsive to said selected sequences of applied phoneme signals.

10. Apparatus as defined in claim 7, wherein said means for developing interpolations of said parameter signals includes means for predicting the value of contextually related phonemes.

11. Apparatus as defined in claim 7, wherein said means for generating control signals includes a plurality of function generators individually proportioned to develop from applied parameter signals analog voltages which represent

14

transitional control functions for controlling said speech synthesizer.

12. The method of controlling an electrical analog of the human vocal tract in order to synthesize sequences of speech sounds, which comprises the steps of:

establishing a number of parameters for defining the configurations of each of a number of distinct regions of the human vocal tract;

developing parametric signals in response to input data specifying selected phonetic symbols, stress and pause information;

modifying said parametric signals according to stored rules derived from analyses of human vocal tract configurations;

developing from said modified parametric signals a sequence of smoothly changing analog signals sufficient to define changing areas of the vocal tract, stress and pause in excitation; and

utilizing said analog signals to control a speech synthesizer to produce audible speech sounds.

#### References Cited

#### UNITED STATES PATENTS

3,158,685 11/1964 Gerstman et al. ----- 179—1

KATHLEEN H. CLAFFY, Primary Examiner

J. B. LEAHEEY, Assistant Examiner