Nov. 24, 1964 L. J. GERSTMAN ETAL 3,158,685

SYNTHESIS OF SPEECH FROM CODE SIGNALS

Filed May 4. 1961 4 Sheets-Sheet 1

FIG. 1

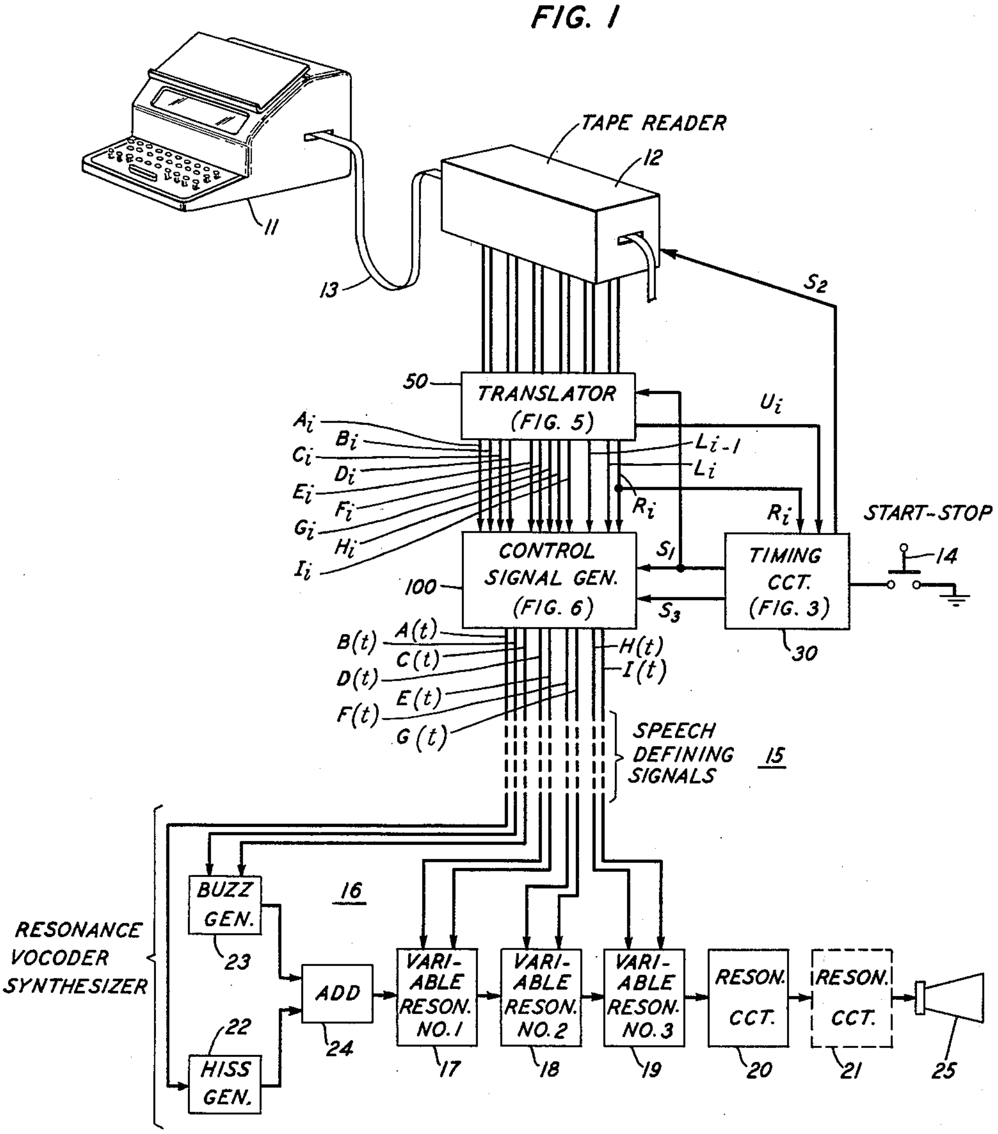
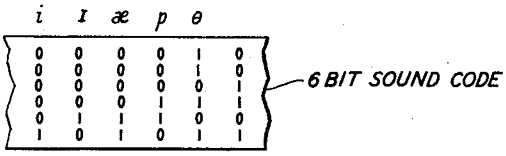


FIG. 1A



i	I	æ	p	θ	
0	0	0	0	1	0
0	0	0	0	1	0
0	0	0	0	0	1
0	0	0	1	1	1
0	1	1	1	0	0
1	0	1	0	1	1

INVENTORS: L. J. GERSTMAN
J. L. KELLY, JR.

BY G. E. Hirsch Jr.
ATTORNEY

SYNTHESIS OF SPEECH FROM CODE SIGNALS

Filed May 4. 1961

4 Sheets-Sheet 2

FIG. 2

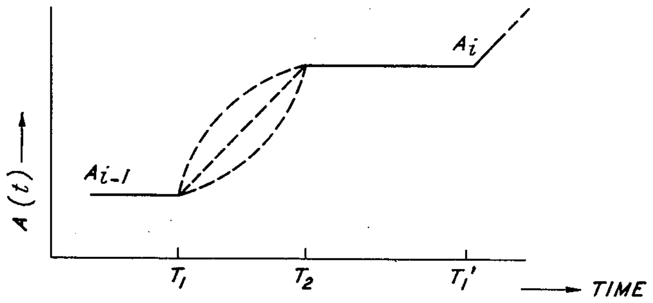


FIG. 3

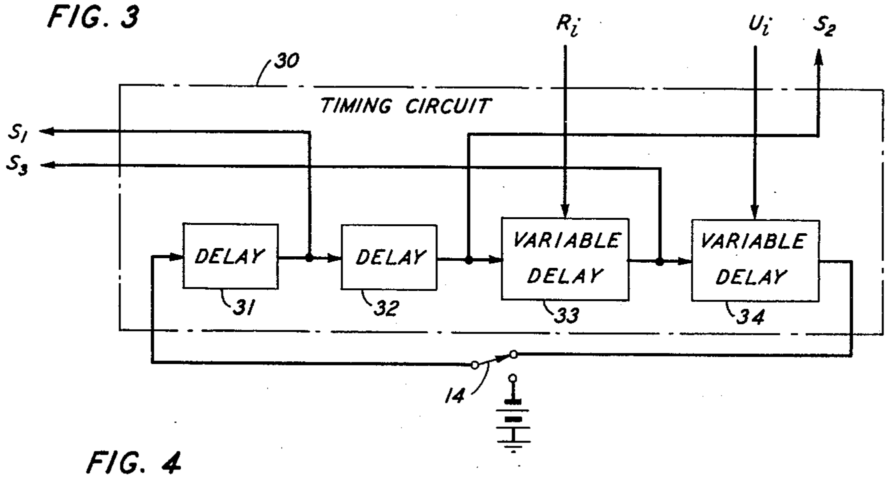
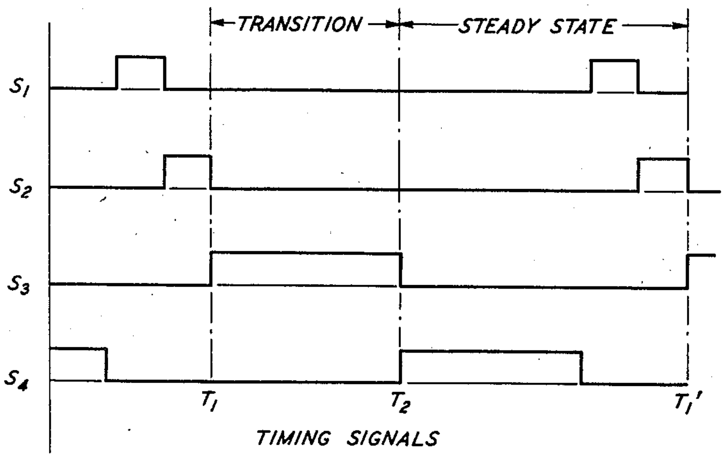


FIG. 4



INVENTORS: L. J. GERSTMAN
J. L. KELLY, JR.

BY A.E. Hirsch Jr.
ATTORNEY

Nov. 24, 1964 L. J. GERSTMAN ETAL 3,158,685

SYNTHESIS OF SPEECH FROM CODE SIGNALS

Filed May 4. 1961 4 Sheets-Sheet 3

FIG. 5

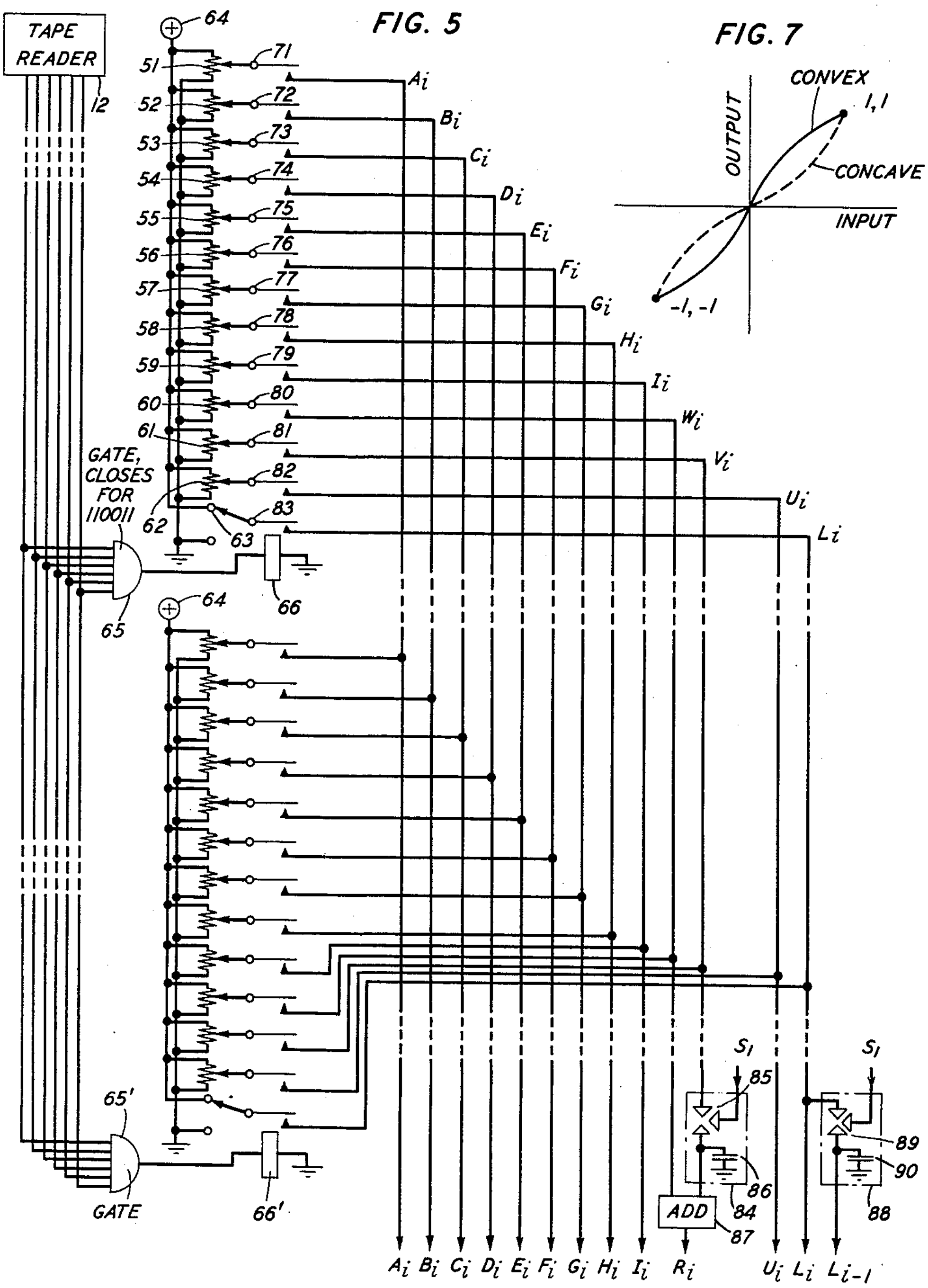
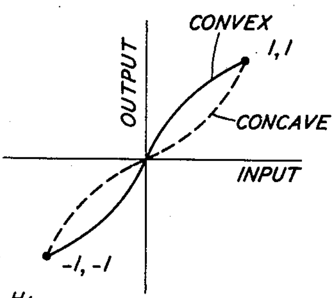


FIG. 7



INVENTORS: L. J. GERSTMAN
J. L. KELLY, JR.

BY A. E. Hirsch Jr.

ATTORNEY

FIG. 6

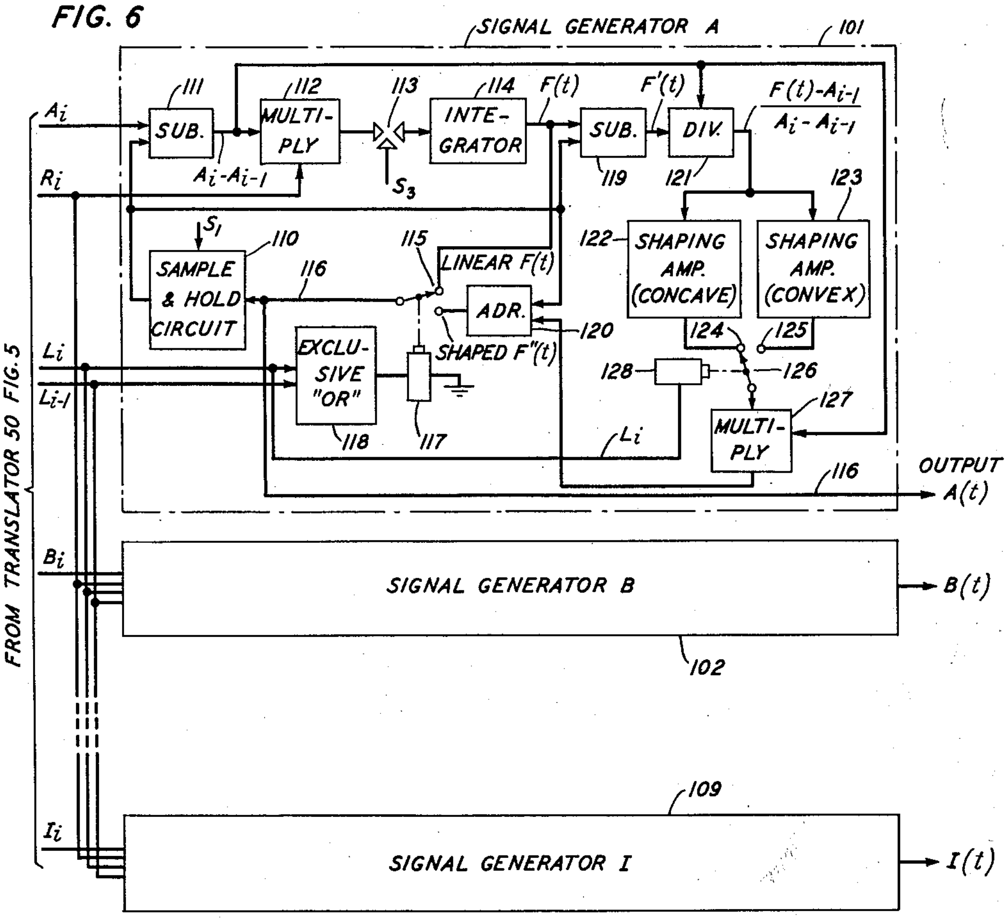
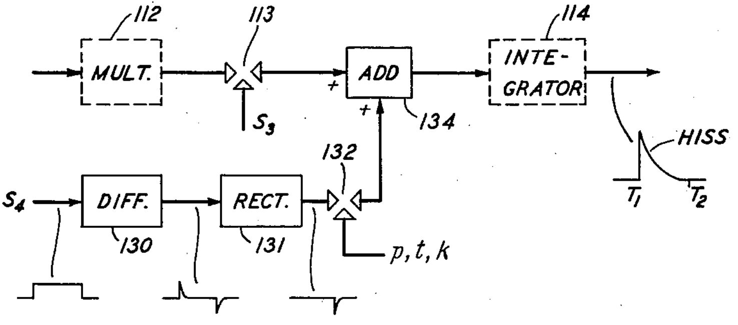


FIG. 8



INVENTORS: L. J. GERSTMAN
J. L. KELLY, JR.

BY A. E. Hirsch Jr.
ATTORNEY

United States Patent Office

3,158,685

Patented Nov. 24, 1964

3,158,685

SYNTHESIS OF SPEECH FROM CODE SIGNALS

Louis J. Gerstman, Bernardsville, and John L. Kelly, Jr., Berkeley Heights, N.J., assignors to Bell Telephone Laboratories, Incorporated, New York, N.Y., a corporation of New York

Filed May 4, 1961, Ser. No. 107,697

8 Claims. (Cl. 179—1)

This invention relates to speech producing systems and particularly to the production of speech with smooth gradations from one speech sound to another.

An object of the invention is to produce speech in response to the typing of phonetic symbols on a keyboard by an unskilled and untrained operator.

Another object of the invention is to convert transmitted telegraph signals to understandable speech thus realizing for speech the natural advantage of the telegraphy over telephony in reduced frequency bandwidth required for transmission, and in improved signal-to-noise ratio.

Another object of the invention is to translate from the printed word to the spoken word.

It has long been recognized by phoneticians that speech may be thought of as being composed of a sequence of fundamental speech units, called phonemes, much as written text is composed of a sequence of alphabet letters. Experiments show, however, that the mere juxtaposition of these various sounds, such as would be obtained, for example, by splicing together segments of magnetic tape each containing one sound segment, does not produce speech. The difficulty is not with the concept of phonemes as building blocks of speech but with their interpretation as segments of speech sharply bounded in time. Stated in another way, while most phoneticians agree on the number of phonemes in a particular utterance, they seldom agree on the proper segmentation of the utterance into phonemes.

Perhaps the most successful method of synthesizing speech from recorded segments of human speech consists of recording pairs of phonemes in all possible combinations and joining the middle of one phoneme to the middle of the next. That is, the steady part of continuants and the silent parts of stop consonants are connected. This method produces quite natural transitions but gives rise to some discontinuities since a person rarely says the sound in the same way twice. The main fault, however, is that an excessive number of sounds must be recorded and stored. Another approach is to store acoustic parameters, such as the frequencies of vocal tract resonances, i.e., formants, and fundamental frequency, i.e., pitch, rather than samples of speech waveforms for each phoneme. Acoustic parameters are in close correspondence with articulatory parameters, such as tongue position, lip closure, and the like. From acoustic parameters of this sort control signals may be generated that are appropriate for controlling speech synthesis apparatus, for example, apparatus known in the art as resonance vocoder synthesizers. Apparatus has been described in the literature that generates control signals according to discrete phonemes and which move from one set of stored values to another with a fixed mode of transition determined by some preselected smoothing circuit. Experiments show, however, that the detailed nature and duration of these transitions vary from one pair of phonemes to another, and that to produce intelligible speech these variations must be carefully preserved.

In the present invention these difficulties are overcome by storing for each phoneme not only the values of acoustic parameters but also the duration of the steady-state period of the parameters and auxiliary parameters which influence the duration and the mode of the adjacent transitions. From these parameters a set of control signals is developed that may be used to control the operation of resonance vocoder synthesizer apparatus. Further improvement is obtained by providing a greater number of controllable parameters in the synthesizing apparatus than is normal and by tailoring the synthesizer apparatus more closely to correspond to the human vocal tract.

In the practice of the invention, a message is first converted into a sequence of electric code signals through the action of appropriate apparatus, for example, that commonly known as a teletypewriter. This apparatus may be provided with a keyboard whose keys are identified by phonetic character labels instead of the customary alphabetic letters. The message may be what a sender wishes to "speak" at the moment, or it may be matter that he reads from a printed page. Subsequently, the single sets of stored parameters for discrete phonemes are read out at a variable rate determined by the particular phonemes involved and used to generate several time-varying control signals. Each of these signals is characterized by alternate periods of constancy (steady-state) and motion (transition). The values assumed during the steady-state period, as well as the duration of the steady-state period, are determined entirely by the parameters stored for the current phoneme. The shapes and durations of the transitions are determined by certain parameters associated with each of the bounding phonemes. In this manner, the transitions, which constitute a most important part of the reconstructed speech, are closely keyed to the structure of the word under analysis. It has been found that for transitions from vowel to vowel or from consonant to consonant, a substantially linear transition is sufficient. It is necessary only to specify the duration of the transition interval and the range of the transition. To the contrary, for transitions from vowel sounds to consonant sounds, a nonlinear transition whose duration is a function of the bounding phonetic parameters is most appropriate, and for transitions from consonants to vowels, a different nonlinear transition of specified duration is preferred. Control signals are thus generated in accordance with the vowel-consonant structure of a word and then utilized to control the synthesis of speech in apparatus which corresponds closely to the synthesizer portion of a tandem-connected resonance vocoder.

It has been found that a satisfactory analysis of individual phonemes, one sufficient to specify adequately a transition with suitable contextual variation, requires the specification of thirteen quantities for each phoneme. Nine of these quantities specify the control signals during the steady-state portion of the phoneme, namely; three specify the center frequencies of the first three formants of speech, three specify the bandwidths of the formants, one specifies intensity of buzz, one intensity of hiss, and one specifies the fundamental frequency (pitch). The other four quantities control timing and transition shaping: one specifies steady-state duration, two contribute to the adjacent transition times, and one denotes the vowel-consonant nature of the phoneme. Once the nature and duration of the transition is specified, each of the nine control signals is made to change according to a selected one of a number of preassigned modes from one set of steady-state values to the next.

It is perhaps helpful to compare the synthesis of a number of representative words in order to emphasize the gravity of the problem encountered in synthesizing connected speech from single sets of stored phoneme parameters. Consider the synthesis of the word "we" with that for the word "sea." In the word "we," three vowel formants must be set to the center frequencies appropriate for the phoneme /w/, and they must be sustained at those frequencies for approximately fifty milliseconds. They must then proceed lineary during the next seventy-five milliseconds or so to the center frequencies appropriate for the phoneme /i/. The new frequencies are then sustained for a duration appropriate to /i/, characteristically about one hundred milli seconds in a stressed syllable. If an attempt is made to synthesize the word "sea" in similar fashion, several obstacles are encountered which, if left unresolved, cause the resultant synthesized sound to be unintelligible to a listener. Several of these obstacles and the methods by which they are overcome in the present invention are described below.

The transitions between /s/ and /i/ in the word "sea" are not linear as they are in the case of the word "we." The formant frequencies change rapidly at first in "sea" and then more slowly as they approach the vowel. When the phonemes are reversed in speech, /is/, the transition shape is also reversed and the format frequencies change slowly is also reversed and the formant frequencies change slowly at first and then more rapidly. An examination of sound spectrograms reveals that these nonlinear transitions occur between vowels and all consonants other than the glides /w, r, l, y/. If the glides are considered to be vowels for this purpose, transitions should as a general rule describe a convex curve from consonants to vowels, a concave curve from vowels to consonants, but move linearly between two vowels or two consonants. In the present invention this criterion is achieved by classifying each phonemes as a vowel or as a consonant, storing this information in the system, and subsequently using it in the generation of transitions. Alternatively, this information is specified on the tape along with the parameters defining the individual phonemes. For example, one bit of a binary code may be assigned to this role.

The durations of transitions between /s/ and /i/ in the word "sea" are also shorter, i.e., the transitions proceed at a faster rate, than those between /w/ and /i/. The transition rate is still faster between /p/ and /i/ in the word "pea." It follows that information regarding the transition rate is also a function of both bounding phonemes and is not an immutable function of one alone. Accordingly, in the practice of the present invention, the appropriate transition rate between two phonemes is expressed as the sum of two rates, one assignable to the preceding phoneme and the other assignable to the succeeding phoneme. Each phoneme is thus identified by an appropriate incoming rate and an appropriate outgoing rate; the two rates being suitable for combination with a large number of other phonemes. The exact rates for each phoneme are selected according to classical articulatory classes of speech. Fortunately there are only about seven of these in American English; namely, voiceless stops, /p, t, k/, voiced stops /b, d, g/, nasals /m, n, ŋ/, voiceless fricatives /f, θ, s, ʃ, h/, voiced fricatives /v, ð, z, 3/, glides /w, r, l, y/, vowels /i, ɪ, e, ε, æ, ɑ, ɔ, o, u, ʊ, ʌ, ə/ and silence (no phoneme). Thus, for each phoneme the rates appropriate for its articulatory class are stored. If, in actual practice, it is found that the transition rates for a given phoneme are consistently different from those stored in its class, provision is made for individual adjustment of the stored rates for that phoneme.

Unlike the case of /w/, spectrograms of /s/ reveal different starting frequencies for the phoneme in dependence on the following vowel. For example, in the word "sea" the second and third formant transitions begin at higher frequencies than the "sue." In the spectrograms of these two words the transitions for each appear to have originated from the same, though invisible, starting frequencies. It is in accordance with the present invention to store, as a part of the parameters defining each phoneme, information indicative of the virtual loci of such starting frequencies while simultaneously preventing the transitions from becoming explicit on the basis of this information until a portion of the transitional movement occurs. Since the control signals generated from the stored parameter values of successive phonemes are utilized in accordance with the invention to energize resonance vocoder apparatus, this information is utilized in altering the operation of the synthesizer thereby to take into account the virtual loci information. Specifically, the vocoder synthesizer is programmed so that it has a D.-C. gain equal to unity. This causes the frequency of the first formant to control the over-all amplitude of all three formants to the extent that when the first formant is set at essentially zero cycles per second, very little energy is present in the second and third formants. Transitions of consonants possessing virtual loci are then started at the second and third formant levels appropriate to the consonant whereas the first formant is started at zero cycles per second. As a consequence, the resulting spectrograms for words like "see" and "sue" look remarkably like those for human speech.

The greatest difference between "we" and "sea" is that the /s/ transitions begin not with steady-state vowel formants but with steady-state noise. In the vocoder synthesizing apparatus the vowel formants require initial hiss energy. However, if the hiss energy is merely passed through resonators set at the starting frequencies of the formant transitions, an intelligible /s/ is not produced. In a spectrogram of the word "sea" there is no observable hiss energy at the beginning of the first formant and indeed the presence of low frequency hiss apparently harms intelligibility. One solution to this problem is to by-pass the formant resonators to supply hiss of the required frequency separately at the output of the vocoder. However, such a procedure requires two sets of stored parameters for each of the phonemes that require hiss; one set pertaining to the formant transitions and the other pertaining to the hiss itself. Accordingly, in the present invention the hiss energy is selectively preshaped so that its amplitude rises, for example, at a rate of 6 db per octave. Hence, there is no available hiss energy at the level of the first formant so that the hiss energy may be applied at the input of the tandemly connected resonators and passed through the same resonators used for the formants. As a consequence all phonemes have a steady-state. However, for some phonemes the steady-state is preshaped hiss rather than buzz energy.

One final difference between "we" and "sea" concerns the steady-state portions of the /w/ and the /s/ in the two words. Whereas the bandwidths of the resonances in /w/ are approximately one hundred fifty cycles per second each, the bandwidth of the hiss in /s/ is about one thousand cycles per second. Consequently, differential bandwidth information is supplied in accordance with the invention for the various phonemes. Since spectrograms of the sounds indicate that the bandwidths change smoothly from one phoneme to another, it is sufficient simply to specify the bandwidths of the steady-state formants, whether produced as buzz or hiss, and to allow them to change in synchrony with the formant transitions. This class of variations is also effective to handle nasals. As a final measure of variation, the bandwidths of all of the other phonemes are suitably adjusted to produce spectrograms similar to those produced by human talkers. Specifically, the bandwidths are assigned in accordance with articulatory classes of speech as in the case of transition rates. Provision is also made for changing the stored specification of bandwidths for each of a selected number of phonemes, as required.

In the apparatus to be described hereinafter by which the foregoing considerations are turned to account, additional flexibility is provided by provision of means for independently manipulating selected ones of the individual synthesizer control signals. For example, if a monotone of speech is sufficient for a particular purpose, the fundamental frequencies stored for the several phonemes may be held constant. If, however, natural inflections of speech are desired, the pitch controls may be ganged together and externally controlled, for example, by an operator of the apparatus. This information may also be precoded on the input tape by adding extra bits to the input code. Or, alternatively, the information may be supplied on a separate "inflection" tape which is then read in synchrony with the phoneme tape. Different dialects may be produced by continuously adjusting the steady-state formant frequencies of the vowels and the steady times of the vowels in order to produce lengthening and shortening of the time thus to provide stress. As before, such auxiliary control signals may be supplied on a separate inflection tape.

The objects and advantages of the invention will be more fully understood from the following detailed description of an illustrative embodiment thereof taken in connection with the appended drawings in which:

FIG. 1 shows diagrammatically the relation of the various circuit units employed in the practice of the invention;

FIG. 1A shows a portion of a tape perforated to specify discrete phonetic elements in accordance with the present invention;

FIG. 2 shows diagrammatically the mode of transition between two steady-state phonemes;

FIG. 3 shows in block schematic form the details of a timing circuit suitable for use in the apparatus of FIG. 1;

FIG. 4 shows a sequence of timing signals produced by the apparatus of FIG. 3;

FIG. 5 shows in schematic form the details of the translator apparatus of FIG. 1;

FIG. 6 shows in block schematic diagram form details of the control signal generator used in the embodiment of the invention shown in FIG. 1;

FIG. 7 shows the characteristics of shaping amplifiers **111** and **112** in the control signal generator of FIG. 6; and

FIG. 8 shows in block schematic form, apparatus suitable for addinng a "burst" of energy to the apparatus of FIG. 6 for voiceless stops.

The System

FIG. 1 shows in block schematic form apparatus for translating from printed intelligence to spoken intelligence. Intelligence in the form of printed phonetic symbols ordered in a fashion required by the corresponding spoken word is supplied to the apparatus by way of a punched tape or the like. Since the rate at which phonemes are required in the composition of speech is generally at a substantially different rate from that at which a person can operate a keyboard device, some form of storage is preferably employed. Accordingly, a teletypewriter apparatus **11**, or the like, is utilized to enter the successive phonemes on a storage medium such as a punched tape. The keyboard of the teletypewriter is provided with one key for each of the phonemes commonly found in speech. For English speech approximately 35 to 40 keys are required. As the successive phonemes are punched out by the machine, the corresponding characters are entered on the tape in the form of a binary code or the like. For example, a code, derived from the International Phonetic Alphabet is quite satisfactory. An example of a portion of such a code is shown below.

Keyboard symbol:	Code group
i	000001
ɪ	000010
æ	000011
p	000110
θ	110101

A sample of tape coded in accordance with a number of successive phonemes is shown by way of example in FIG. 1A. This tape is supplied as required to tape reading apparatus **12** wherein the individual code symbols are read out, e.g., as ones and zeroes, and supplied to translator **50**.

The details of operation of the teletypewriter, tape storage system and tape reader may be of any desired form. For example, this equipment may be of the type described in detail in Dudley-Harris Patent 2,771,509, November 20, 1956. If desired, the exact code described in the Dudley-Harris patent may be utilized in the practice of the present invention.

After a message, in the form of successive phonemes, has been stored on the tape, it may at any time thereafter be read out and utilized to produce connected speech. Once the read-out process is initiated, it is under the control of a sequence of timing pulses generated in timing circuit **30**. Thus, since discrete phonemes require different intervals in speech, the read-out of successive phonemes from tape **13** is in strict accordance with the context of the message entered on the tape, that is, the phonemes are read out in accordance with their position in the encoded message.

This implies of course that the message is continuously analyzed for message sound content as the reading process continues. Translator **50** is responsible for this logical analysis of succesive phonemes and, in dependence upon succesive phoneme pairs, two signals are supplied to timing circuit **30** by way of conductors U_1 and R_1 to influence the timing circuit. Reciprocally, tape reader **12** is advanced by a pulse on conductor S_2 from timing circuit **30** in dependence on the decisions made in translator **50**.

Translator **50**, in addition to providing a measure of stored logic, contains for each of the phonemes in the speech alphabet thirteen stored quantities. Nine of these quantities, labeled for convenience A_1 through I_1, produce for each phoneme a steady-state function indicative of the corresponding phoneme. Three of the stored quantities W_1, V_1, and U_1 are related to the appropriate duration of the steady-state period and transition betwen the pheneme and other phonemes, and one quantity L_1 classifies the phoneme as a vowel or consonant so that the mode of transition between the phoneme and another one may be established. In addition to the stored quantities, a quantity L_{1-1} is produced within the translator to indicate the classification of the last phoneme read out of apparatus **12**.

Thus, as a phoneme is read out of reader **12**, that set of stored quantities corresponding to the indicated phoneme is supplied simultaneously by the translator **50** to control signal generator **100**. Signal generator **100** ingests the thirteen stored quantities and computes an appropriate set of nine control signals which specify the acoustic parameters as functions of time. Signal generator **100** is, in addition, under control of timing circuit **30**. The nine control signals generated in apparatus **100**, i.e., signals $A(t)$ through $I(t)$ are sufficient to energize the synthesizer portion of a resonance vocoder.

Resonance vocoder synthesizer **16** may be of any desired form. Typically, synthesizer apparatus of this sort is energized by a sequence of control signals produced by a corresponding resonance vocoder analyzer. As is well known in the art, these vocoder control signals specify at every instant the pitch, buzz intensity, hiss intensity, and the center frequency and bandwidths of a selected number of formant resonating circuits. Preferably, a so-called tandem resonance vocoder is employed in the practice of the present invention. It includes a number of resonating circuits **17** through **21**; the first three of which are variable and are proportioned to the first three formants of speech. Resonators **20** and **21** need not be adjustable. Hiss or noise signals responsible for the fricatives of speech are generated in conventional hiss generator **22** while buzz energy for the reconstruction of voicd sounds is produced in generator **23**. Control singals $A(t)$ and $B(t)$ supervise, respectively, the pitch and intensity of the buzz energy, and signal $C(t)$

controls the intensity of the hiss energy. By carefully shaping the spectrums of both the hiss and buzz energy signals, the two may be combined directly in adder **24** to produce a composite excitation signal. This signal is supplied to serially connected resonance circuits **17** through **21** wherein it is suitably shaped and supplied to loudspeaker **25**.

The nine control singals $A(t)$ through $I(t)$ produced by signal generator **100** are supplied by way of channel **15** to vocoder synthesizer **16**. Since the control signals are of extremely narrow bandwidth, on the order of 0 to 25 cycles per second each, they may be combined in any desired fashion, e.g., by well known multiplex techniques or the like, and conveyed over any distance to the synthesizer apparatus. In practice, the entire apparatus including the teletypewriter **11** and synthesizer **16** are located near one another. However, if desired, the control singal energy may be conveyed over a transmission path of considerable length to produce, at a receiving station, speech in accordance with the written intelligence produced at a transmitter station.

Tandem connected resonance vocoder synthesizer **16** is a good analog to the vocal tract. The buzz generator corresponds to the glottis. The hiss generator provides the energy for unvoiced sounds. For most unvoiced sounds. e.g., Voiceles Fricatives and Voiceless Stops, difficulty is ordinarily experienced in a tandemly connected synthesizer since these sounds are not produced at the back of the vocal tract and passed through the oral cavity. Thus the analogy fails. However, if certain precautions are taken, hiss energy may nevertheless be combined with buzz energy and applied to a point in the synthesizer that correpsonds to the glottis. Accordingly, provision is made in the present invention for controlling the shape of the spectrum of suond produced by the synthesizer in response to the combined hiss and buzz excitation. By changing the bandwidths of the resonant circuits of the synthesizer slightly as required by this class of sounds, e.g., fricatives and the like, a good replica of all sounds made in portions of the vocal tract other than the glottis region may be generated satisfactorily. In essence, fricatives are generated in accordance with the invention by changing the bandwidths of selected resonance circuits from their standard vowel positions. Sufficient information is available in the translator **50** to enable suitably altered bandwidth control signals to be produced by generator **100**. This effectively compensates for the somewhat arbitrary insertion point of the hiss energy. Moreover, the "shape," i.e., frequency spectrum, of the hiss and buzz energy is preshaped to afford optimum synthesis of speech. It has been found that pre-emphasis of hiss at approximately a 6 db per octave rate, and de-emphasis of the buzz energy at the same or slightly greater rate is helpful.

The several resonant circuits together constitute a linear circuit which corresponds in large measure to the modulating portion of the human vocal tract. The transfer characteristic of each of the several circuits is determined, for example, by a simple LCR network. Each is sufficiently broad to pass excitation energy, and is suitably shaped to yield, in response to excitation, the frequencies appropriate for one formant only. The actual transfer characteristic of the entire series circuit is the product of the transfer characteristics of the individual resonant circuits. The order in which the series circuits are connected is therefore immaterial.

In the human vocal tract, the long time average of input and output volume flows are equal. In similar fashion, the vocoder synthesizer apparatus of the present invention is adjusted to an over-all D.-C. gain of one, where D.-C. means long time average, so that the D.-C. gain of the over-all circuit remains one regardless of the instantaneous value of any of the control signals. This has been found to be most helpful in synthesizing natural sounding speech from phonetic symbols.

Vocoder synthesizer **16** may, of course, be controlled by substantially fewer externally applied control signals than shown in FIG. 1. The nine signals indicated, however, provide a considerable degree of flexibility.

Mode of Transition

Since transitions between successive phonemes are of extreme importance in the determination of the sound reproduced in artificial speech apparatus, it is in accordance with the present invention to specify both the duration and range of each transition and its shape, i.e., its mode. This dual specification is, in large measure, responsible for the excellent speech naturalness obtained with the apparatus of the invention. Since the transition specification for each possible phoneme pair encountered in speech is in some measure different from all others, an alphabet of forty phonemes would require approximately 1600 different transition mode specifications. Statistically, a separate transition mode for each of the possible pair combinations is unnecessary. In the present invention the number of required modes has been substantially reduced by additionally classifying stored phonemes according to the class of indicated sounds, that is, by specifying them as voiced stops, as unvoiced fricatives, or the like. Accordingly, two additional parameter values are stored with each phoneme, one of which specifies the desired rate of transition following a steady-state indication of the phoneme, and the other of which specifies transition rate preceding the steady-state value of the phoneme. The actual transition mode between phoneme pairs is determined by the sum of the two stored parameter values.

These parameters, generated in translator apparatus **50** (FIGS. 1 and 5), are identified as V_i for the period following the phoneme and W_i for the transition preceding each phoneme. For each phoneme, an adjustment may be made in the translator apparatus to control, at least partially, the nature of the transition both preceding and following it in its context in connected speech. Thus, for example, if a relatively fast transition following a phoneme is desired, a potentiometer, or the like, controlling the V_i signal is advanced, whereas if a relatively slow transition is desired the potentiometer value is decreased. Similarly, adjustments of the W_i signal may be made to enhance or retard the transition preceding the phoneme.

FIG. 2 shows a typical waveform representing one of the control signals, $A(t)$, as a function of time. As shown in FIG. 1 $A(t)$ controls the amount of hiss in resultant speech. FIG. 2 could equally well apply, however, to any of the other eight control signals. A_{i-1} is the value of parameter A (hiss) for the last phoneme as determined by the translator, and A_i is the corresponding value for the new phoneme. In general the index (i) indicates a parameter belonging to the curren phoneme, i.e., the one which is in the tape reader at the time of the discussion, while $(i-1)$ indicates the phoneme just previously read. The time interval T_1 to T_2 is the transition period during which $A(t)$ moves from its old to its new value. T_2 to T_1' is the steady-state period for the current phoneme. The mode of the transition and its duration are carefully specified in accordance with the context of the phonemes bounding the transition. For an ordinary vowel-to-vowel or consonant-to-consonant transition a smooth linear connection suffices. However, for a consonant-to-vowel transition a substantially convexed path is required, and for vowel-to-consonant transitions, a substantially concave path has been found to be preferred.

Timing Circuit

Timing circuit **30** is the prime mover of the system. Its details are shown in FIG. 3. It comprises a series of connected delay elements **31**, **32**, **33**, and **34**. Delays **31** and **32** are fixed delays and may be of the so-called phantastron type. A delay element of this type is triggered by a sharp signal transition, for example, by a brief negative-going pulse, and generates in response thereto, a new pulse whose negative-going trailing edge occurs at a predetermined time following the initiating pulse. By providing an initiating pulse, for example, by means of switch **14**, delay element **31** produces a new pulse S_1, which is supplied both to signal generator **100** and translator **50** of FIG. 1. The trailing edge (negative) of S_1 initiates, in fixed phantastron delay **32**, a new pulse S_2, which is supplied to tape reader **12** in order to advance the paper tape as required. The trailing edge of S_2 also energizes variable phantastron delay **33** to produce at its output a pulse S_3 used in signal generator **100**. The delay interval of apparatus **33** is controlled by signal R_i; one of the quantities stored in translator **50**. The trailing edge of S_3 subsequently initiates the generation of a pulse S_4 in phantastron delay **34**. The delay interval of apparatus **34** is controlled by signal U_i from translator **50**. As a result, a delayed pulse S_4 is produced, whose trailing edge in turn initiates another chain of events by activating fixed delay **31**.

FIG. 4 illustrates the sequence of pulses produced by the apparatus of FIG. 3. The trailing edge of pulse S_4, or an initial pulse inserted into the system by switch **14**, is responsible for the generation of S_1. The trailing edge of S_1 is responsible for S_2, that of S_2 for S_3 and that of S_3 for S_4. Although not apparent in the drawing, the durations of S_3 and S_4 are variable functions of R_i and U_i, respectively. The interval between the trailing edge of S_2 and the trailing edge of S_3 denotes a transition between successive phonemes, whereas the interval between the trailing edge of pulse S_3 and the trailing edge of the next succeeding S_2 pulse specifies a steady-state signal.

Translator

Translator circuit **50** is shown in detail in FIG. 5. For each phoneme in the selected alphabet, i.e., on the keyboard of apparatus **11** of FIG. 1, thirteen quantities are stored, for example, by means of potentiometers **51** through **62** and switch **63** connected between the positive terminal of a source of potential **64** and its negative terminal, e.g., ground. The stored quantities for the individual phonemes are supplied by way of a number of buses A_i through L_{i-1} each time the corresponding phoneme is detected in the code reader **12** of the apparatus of FIG. 1. Detection of the phoneme may be made in any desired fashion. One convenient one involves generating in tape reader **12** a pair of signals for each bit of the code identifying the phoneme, e.g., true and complementary binary signals. These are supplied to an AND gate **65** and, when the appropriate code condition is detected as by the appropriate code information appearing at the input to the AND gate, a signal is produced at the output of the gate which is sufficient to energize a relay **66**. Alternatively, a bipolar gate responsive to two conditions of each bit of the code from reader **12** may be used to respond to an appropriate code condition. Relay **66** then closes a series of switches **71** through **83** which connect the individual stored quantities to the buses A_i through I_i and W_i, V_i, U_i, and L_i, respectively. The signals appearing on buses A_i through I_i determine the value of the corresponding control signal during steady-state intervals. W_i and V_i are combined to produce signal R_i which determines the duration of transition inversals. To produce R_i, quantity V_i is passed through sample-and-hold circuit **84** which may include a switch **85** under control of signal S_i from timing circuit **30** (of FIG. 1), and shunt capacitor **86**. The "held" value of V_i is added to the current value of V_i in adder **87**; the output of adder **87** consequently is a function of the stored "rate" parameters for two successive phonemes and is designated R_i. R_i is supplied both to timing circuit **30** and to signal generator **100**. Signal U_i determines the extent of the steady-state interval by controlling the duration of pulse S_4 in timing circuit **30**. Quantity L_i classifies the phoneme as a vowel or consonant and is used in the signal generator to establish the mode of transition between phonemes. It is preset, e.g., to the value of source **64**, or to ground, by means of switch **63** to indicate its condition. As a convenience, an additional signal L_{i-1}, representing the quantity L for the last encountered phoneme, is generated in the translator circuit, for example, by passing the signal L_i through sample-and-hold circuit **88** under control of signal S_1 from timing circuit **30**. The resultant "held" value is passed to bus L_{i-1}.

The circuit described above generates for one phoneme, e.g., that represented by code 110011 at gate **65**, signals sufficient to enable generator **100** to form a sequence of vocoder control signals. An identical circuit is required for each phoneme in the alphabet. One additional circuit is shown, by way of example, in FIG. 5. Others are connected as required to the several output buses.

While it is, of course, true that a great many sounds may require identical settings of the potentiometers and switches employed in the translator apparatus and, in actual practice many of the potentiometers may be preset or eliminated, the arrangement shown is preferred because of its great versatility. With it, it is possible to accommodate an alphabet of considerable proportions. Furthermore, information regarding the nature of the several phonemes may, if desired, be encoded directly on the tape **13** as an addition to, or in place of, information stored in the potentiometer bank of the translator. For example, data regarding the vowel-consonant nature of the phoneme may be encoded directly on the tape and passed through gate **65** to bus L_i directly. However, this precludes possible changes.

Signal Generator

Signal generator **100** utilizes the stored information provided by translator **50** to generate control signals for energizing resonance vocoder synthesizer **16**. It includes nine individual portions, one for the generation of each of the nine control signals. One portion, **101**, is shown in detail in FIG. 6. The others, **102** through **109** may be idenical in construction. The nine generator portions are supplied with one each of the stored quantities A_i through I_i from translator **50**, and all are supplied with quantities R_i, L_i, and L_{i-1} from the translator.

With the notation used above, A_i indicates the value of parameter A for the current phoneme (i) and A_{i-1} denotes the last previous value of that parameter. The last previous value, generated in a manner to be described hereinafter, is stored in sample-and-hold circuit **110** wherein it was developed during the last cycle of operation. This condition prevails at the time of occurrence of pulse S_1 from the timing circuit and indicates the time T_1. That is, it prevails at the end of the steady-state period of the last phoneme (i—1). The tape is then advanced by pulse S_2. This causes the signal A_i to be transferred to the input of subtractor **111**. The difference signal is thus a measure of the required change in control signal magnitude between the sound (i—1) and the new sound (i). It is supplied to multiplier **112** wherein it is multiplied by function R_i supplied from translator **50**. R_i specifies the rate of transition, that is, it is proportional to the reciprocal of the duration of the transition between the two phonemes. Since it is derived as an average of two stored quantities, W_i and V_i, representing functions of the instaneous phoneme and its predecessor, the product signal developed at the output of multiplier **112** is intimately related to the context in which the bounding phonemes occur. When the leading edge of pulse S_3 closes switch **113**, by way of a relay or the like, the product signal developed by multiplier **112** is applied to integrator **114**. Switch **113** remains closed for the duration of pulse S_3. Since S_3 is a function of R_i (variable delay **33** in timing unit **30** is controlled by R_i), the integrator will be supplied with the product signal for a period commensurate with the transition interval. During this period the product signal builds up in a substantial linear fashion between a value proportional to quantity A_{i-1} at time T_1 to a quantity proportional to A_i at time T_2. At the end of the transition interval, that is, at time T_2, the value of the function F(t) appearing at the output of integrator **114** has completed a linear transition between the two successive phonemes.

For transitions between like pairs of phonemes, that is, for transitions from vowel to vowel or from consonant to consonant, a linear transition is entirely suitable. Accordingly, the signal F(t) is supplied by way of a switch **115** to an output bus **116** of generator A. The signal appearing at bus **116** constitutes control signal A(t) and may be transmitted to resonance vocoder synthesizer **16**. Switch **115** is under control of relay **117**, or the like, which is energized by the output of an exclusive OR gate **118**. OR gate **118** is supplied at its input terminals with signals L_i and L_{i-1} from translator **50**. L_i, as mentioned above, specifies the nature of the current phoneme, that is, whether it is a vowel or a consonant. L_{i-1} accordingly specifies the vowel-consonant nature of the previous sample. If the two are alike, relay **117** remains unenergized and switch **115** connects the signal F(t) from integrator **114** to output bus **116**. Signal A(t) passed by switch **115** is also supplied to sample-and-hold circuit **110**. At the next sampling interval S_1, it becomes the new value A_{i-1} and is supplied continuously to subtractor **111**.

For transitions between unlike pairs of phonemes, that is, for transition from a vowel to a consonant, or from a consonant to a vowel, a nonlinear transition is preferred. Moreover, the shape of the transition is, in accordance with the invention, different for the two cases. Thus, the linear transition developed for each successive pair of phonemes by integrator **114** is further shaped when an unlike pair of adjacent phonemes is encountered.

To simplify the apparatus required for shaping the transition, it is in accordance with the invention, to normalize the transition functions F(t) to a standard range and datum before shaping them, and to restore the shaped function to its previous range and datum after shaping. Since two variables are involved, the normalizing process involves two separate operations, biasing and scaling. Biasing is necessary to bring one phoneme, preferably the one representing the last one A_{i-1}, bounding the transition to a reference (datum) level. Successive phonemes, of course, may occur at any level within a wide range. Biasing is accomplished simply by subtracting the value of the phoneme A_{i-1} from the transition. The resultant function, $F(t)=F(t)-A_{i-1}$, in relation to a fixed load for all F(t)'s. Accordingly, signal A_{i-1} from sample-and-hold circuit **110** is supplied to subtractor **119** and adder **120**. It is subtracted from F(t) in subtractor **119** and is subsequently added to the shaped function in adder **120**.

Scaling is necessary to restrict the ranges of various pairs of phonemes to a standard range, one that can satisfactorily be accommodated by the shaping apparatus. It is easily accomplished by reducing the range of each transition by a factor equal to the absolute range of the corresponding transition before shaping, i.e., to a standard one volt range, and by subsequently enlarging the shaped transition function by the identical factor. Accordingly, the difference signal (biased value of F(t)) supplied by subtractor **119** is applied to the dividend terminal of divider **121**, and the difference signal

$$A_i-A_{i-1}$$

available at the output of subtractor **111**, is applied to the divisor terminal of the divider. The output of divider **121** is thus the quotient of the two inputs, or

$$\frac{F(t)-A_{i-1}}{A_i-A_{i-1}}$$

The quotient, which represents a suitably normalized value of the linear transition function between phonemes A_i and A_{i-1}, is applied to the inputs of shaping amplifiers **122** and **123**.

Amplifiers **122** and **123**, which may be any form of nonlinear amplifier, respectively impart to the quotient signal, concave or convex characteristics such as, for example, those shown in FIG. 7. Amplifiers with suitable characteristics are well known in the art. The shaped functions are supplied to the terminals **124** and **125** of switch **126** and, in dependence on the vowel-consonant order of the phonemes A_i and A_{i-1}, one of them is supplied by way of the movable arm of switch **126** to multiplier **127**. Switch **126** is actuated by relay **128** which is energized by signal L_i. Relay **128** is polarized so that the convex function (amplifier **123**) is selected if the phoneme (i) is a vowel, as indicated by L_i. Otherwise, a concave function from amplifier **122** is selected. To restore the shaped transition function to its normal value difference signal A_i-A_{i-1}, from subtractor **111**, is applied to the second input of multiplier **127**. Thus, the shaped function is multiplied by the quantity A_i-A_{i-1}, by which function it was reduced in divider **121**. The product signal developed by multiplier **127** is supplied to adder **120**, wherein signal A_i, which was subtracted out to bias F(t) to a base level, is added to it. Adder **120** thus yields a signal F″(t) which in effect is a shaped transition from the value A_{i-1} at time T_1 to the value A_i at time T_2. This signal is applied to the second terminal of switch **115**. If the signals L_i and L_{i-1} applied to exclusive OR gate **118** are different, indicating that successive phonemes are not alike, a signal is developed at the output of OR gate **118** to energize relay **117**. Hence a shaped transition function F″(t) is applied to output bus **116**. This signal will have the appropriate convex or concave shaping in accordance with the transition as previously described. If, to the contrary, signals L_i and L_{i-1} specify that two consecutive sounds are both vowels or are both consonants the linear mode of transition is selected, and transition function F(t) from integrator **114** is applied to bus **116**.

Similar operations take place in each of the nine signal generators **101** through **109** for each of the quantities A_i through I_i from translator **50**. Hence, for each detected phoneme a sequence of nine output signals A(t) through I(t) are produced which are sufficient to control the operation of resonance vocoder synthesizer **16**. The fact that the level A_{i-1} is obtained by sampling (with pulse S_1) the actual output A(t) rather than the signal A_i from the translator gives the entire circuit a self-correcting or feedback property which tends to cancel out any low frequency errors in the integrating and shaping circuits. Any perturbation of signals in the circuit prior to the occurrence of pulse S_1 will be completely eliminated by the end of the following transition. This latter fact is exploited in the design of the burst generator described below.

Burst Addition

A so-called "burst" occurs on all voiceless stops such as p, t, and k. Burst also occurs on some voiced stops, such as b, d, and g, but these are believed not to be of great importance since other vocal cues compensate for their presence. Accordingly, in the present invention, a short burst of hiss energy is inserted into the output sound spectrum following each of the voiceless stops. This is accomplished conveniently by shifting the base level of the product signal passed from multiplier **112** in the A(t) signal generator apparatus **101** of FIG. 6, by a predetermined constant before it is supplied to integrator **114**. This may be done in a variety of ways. One simple one is illustrated in the apparatus of FIG. 8.

The burst energy preferably is applied to integrator **114** of FIG. 8 at an instant slightly before the sampling instant S_1. It is thus operated on by the integrator; i.e., it decays over the period of the transition interval, and, moreover, it is applied to the feedback circuit so that the addition will be canceled out in due course and will not be responsible for a cumulative build up of noise energy. This time corresponds closely to the trailing edge of signal S_4. Accordingly, the burst signal may be generated directly from the S_4 pulse. Pulse S_4 is applied to differentiator **130**, which produces a spike at the leading and trailing edge of each pulse. The negative pulse only is retained by passing the differentiated signal through rectifier **131**. The resultant negative spike is thus applied, as required, by way of switch **132** to adder **134** where it is combined with the product signal from multiplier **112**. This effectively raises the D.-C. level of the product signal from the multiplier. The signal however decays during the transition interval to its normal value. An addition of the burst energy is desirable only for selected voiceless stops. Thus indications of the phonemes p, t, and k are derived from additional resistor-relay elements in the translator **50** shown in FIG. 5. When any one of these sounds is detected, switch **132** is closed and the negative pulse from rectifier **131** is passed to adder **134**. This modification is, of course, employed only in the circuit producing the hiss control signal A(*t*). The other eight circuits are unmodified; i.e., they are used as shown in FIG. 6.

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.

What is claimed is:

1. Apparatus for the production of artificial speech-like sounds which comprises, a source of speech phoneme representations ordered according to a desired phonetic sequence, means for selectively analyzing selected sequences of said phoneme representations to produce a code signal for each individual phoneme in said sequence and for the vowel-consonant structure of said phoneme sequence associated with each inidividual phoneme, means responsive to said code signals representative of successive phonemes and associated vowel-consonant structures for developing a set of speech defining signals which together specify the acoustic parameters of said sequence of phonemes and the transitions between phonemes as a function of time, and synthesizer means continuously supplied with all of said speech defining signals for generating artificial speech-like sounds.

2. In a mechanism for producing speech-like sounds, the combination which comprises: means for storing a set of speech parameters for each of a number of speech sounds; means responsive to a selected sequence of said stored parameters for generating a first sequence of control signals, each of which is representative of one of said speech sounds; means responsive to successive pairs of said stored parameters for generating a second sequence of control signals which vary in a substantially linear fashion between pairs of control signals of said first sequence that represent consecutive vowel or consecutive consonant speech sounds; means responsive to successive pairs of said stored parameters for generating a third sequence of control signals which vary in a substantially nonlinear fashion between pairs of control signals of said first signal sequence that represent consecutive vowel-consonant, or consecutive consonant-vowel speech sounds; the control signals of said first, second, and third sequences thus together representing the acoustic parameters of said speech sounds and the transitions uniquely associated with successive pairs of said speech sounds; and means responsive to all of said control signals together for generating artificial speech.

3. Apparatus for the production of artificial speech-like sounds which comprises: a source of coded representations of phonemes of speech ordered according to a desired phonetic sequence, means responsive to a succession of said representations for generating control signals that persist with a constant, specified, magnitude for intervals in said succession of representations which denote discrete phonemes, means responsive to a succession of said representations for generating control signals that vary both in magnitude and duration according to a prescribed schedule for intervals in said succession which do not denote discrete phonemes but which are bounded by such phoneme intervals, and means for utilizing said control signal in the generation of artificial speech.

4. Apparatus for the production of artificial speech which comprises: a source of coded representations of phonemes of speech ordered according to a desired phonetic sequence; means responsive to successions of said representations which represent discrete phonemes for generating a control signal that persists with a substantially constant magnitude for the duration of each of said phoneme representations; means responsive to successive discrete phoneme representations for generating control signals that vary in a substantially linear fashion between the pairs of said substantially constant magnitude control signals that denote vowel-to-vowel or consonant-to-consonant phoneme representations; means responsive to successive discrete phoneme representations for generating control signals that vary in a substantially nonlinear fashion with a slope that monotonically increases in magnitude between pairs of said substantially constant magnitude control signals that denote a consonant-to-vowel phoneme representative sequence; means responsive to successive discrete phoneme representations for generating control signals that vary in a substantially nonlinear fashion with a slope that monotonically approaches zero between pairs of said substantially constant magnitude control signals that denote a vowel-to-consonant phoneme sequence; and means for utilizing said substantially constant magnitude control signals and said varying control signals together for the generation of artificial speech.

5. In combination, means for storing a succession of coded representations of a selected alphabet of phonemes acording to a desired phonetic order, means for storing for each phoneme a set of analog representations of parameters uniquely associated therewith, signal generator means for developing from sets of said analog representation control signals representative of substantially steady-state phoneme values, means for transferring sets of said analog representations to said signal generator means in accordance with said order of storage of said coded representations, means associated with said signal generator means for analyzing sets of analog representations applied to said generator, means responsive to analyses of successive pairs of analog representations for developing substantially nonlinear control signal segments for interconnecting respectively the control signals corresponding to said sets of analog representations, speech synthesizing means including means for generating hiss energy and buzz energy and for shaping said hiss and buzz energy spectra, and means for utilizing said control signals for controlling the shaping of said energy spectra in said synthesizer to produce intelligible speech.

6. The combination as defined in claim 5 in further combination with means for pre-emphasizing said hiss energy at a rate of substantially 6 db per octave and for de-emphasizing hiss energy at the rate of substantially 6 db per octave, and means for adding the equalized hiss and buzz energy together to form a composite speech excitation signal.

7. Control signal generator apparatus for producing resonance vocoder control signals in response to coded analog representations of ordered speech phonemes that comprises means for developing from said stored analog data a substantially constant signal for each discrete

phoneme and the next consecutive one in said order, said means including means responsive to the vowel-consonant veloping from said analog data a substantially nonlinear control signal portion for each transition between one phoneme and the next consecutive one in said order, said means including means responsive to the vowel-consonant order of successive pairs of phonemes for altering the mode of transition of said nonlinear control signals, and means for supplying one of said control signals for each control function required by speech synthesizing means.

8. In combination with apparatus as defined in claim 7, means operative upon the occurrence of a stored analog representation of one of the voiceless stops, p, t, k, for interposing an abrupt discontinuity in the synthesizer control signal that relates to hiss energy.

References Cited in the file of this patent

UNITED STATES PATENTS

2,595,701	Potter	May 6, 1952
2,771,509	Dudley et al.	Nov. 20, 1956