

[54] **SPEECH SYNTHESIS BY CONCATENATION OF FORMANT ENCODED WORDS**

[75] Inventors: **James Loton Flanagan**, Warren; **Lawrence Richard Rabiner**, Berkeley Heights; **Ronald William Schafer**, New Providence, all of N.J.

[73] Assignee: **Bell Telephone Laboratories Incorporated**, Murray Hill, N.J.

[22] Filed: **Oct. 30, 1970**

[21] Appl. No.: **85,660**

[52] **U.S. Cl.** **179/1**
 [51] **Int. Cl.** **G101 1/00**
 [58] **Field of Search**.... 179/1 S, 1 A, 1 SB, 15.55 R, 179/15.55 T; 324/77; 340/148

[56] **References Cited**

UNITED STATES PATENTS

2,860,187	11/1958	David.....	179/15.55 R
3,158,685	11/1964	Gerstman.....	179/1 SA
3,319,002	5/1967	DeClerk.....	179/1 SA
3,369,077	2/1968	French.....	179/1 SA
3,532,821	10/1970	Nakata.....	179/1 SA
3,588,353	6/1971	Martin.....	179/1 SA

OTHER PUBLICATIONS

Rabiner, A Model for Synthesizing Speech by Rule,

IEEE Transactions AU-17 3/69, pp. 7-13.

J. L. Flanagan et al. Synthetic Voices for Computers, IEEE Spectrum, pp. 22-45, October 14, 1970.

Primary Examiner—Kathleen H. Claffy

Assistant Examiner—Jon Bradford Leaheey

Attorney, Agent, or Firm—A. E. Hirsch; G. E. Murphy

[57]

ABSTRACT

Audio response units that select speech sounds, stored in analog or coded digital form, as the excitation for a speech synthesizer are widely used, for example in telephone audio announcement terminals. The speech produced by most units is noticeably artificial and mechanical sounding.

According to this invention, human speech is analyzed in terms of formant structure and coded for storage in the unit. As the individual words are called for, a stored program assembles them into a complete utterance, taking into account the durations of the words in the context of the complete utterance, pitch variations common to the language, and transitions between voiced portions of the speech. The result is a more natural sounding synthetic utterance.

13 Claims, 8 Drawing Figures

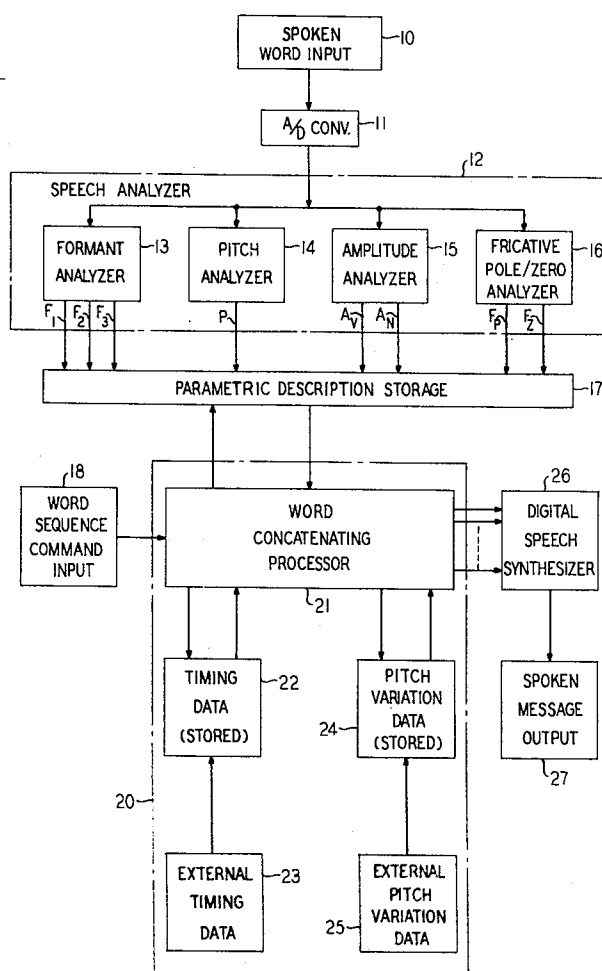
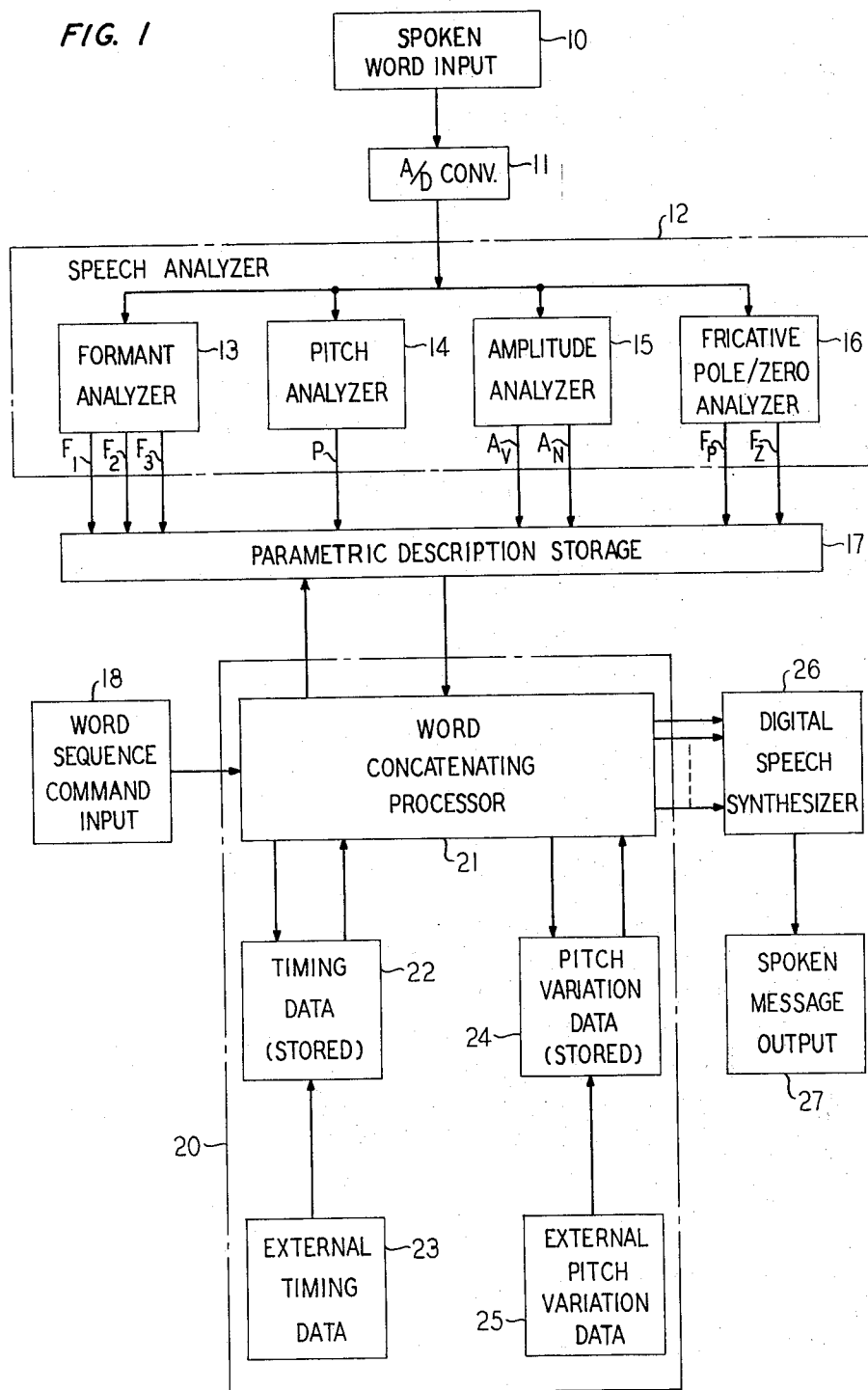


FIG. 1



J. L. FLANAGAN
INVENTOR L. R. RABINER
R. W. SCHAFER
BY *G. E. Hirsch Jr.*
ATTORNEY

FIG. 2

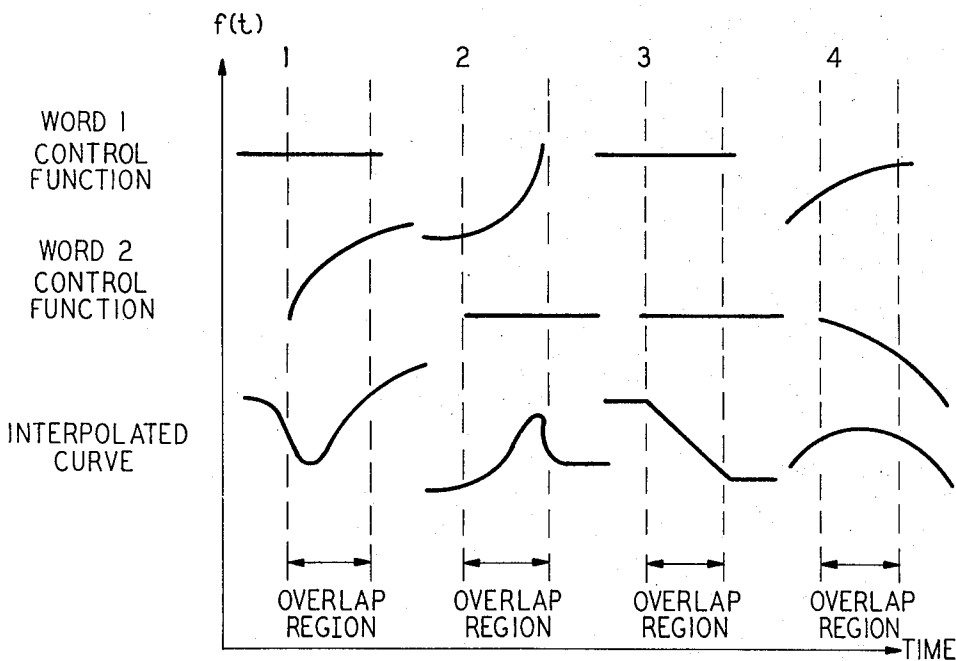


FIG. 3

	PHONEMES / DIGIT			
	1	2	3	4
1	250	330	410	490
2	280	330	390	450
3	450	500	560	610
4	260	300	340	380
5	340	370	410	440
6	230	280	340	390
7	290	380	460	550

TIME (m SEC.)

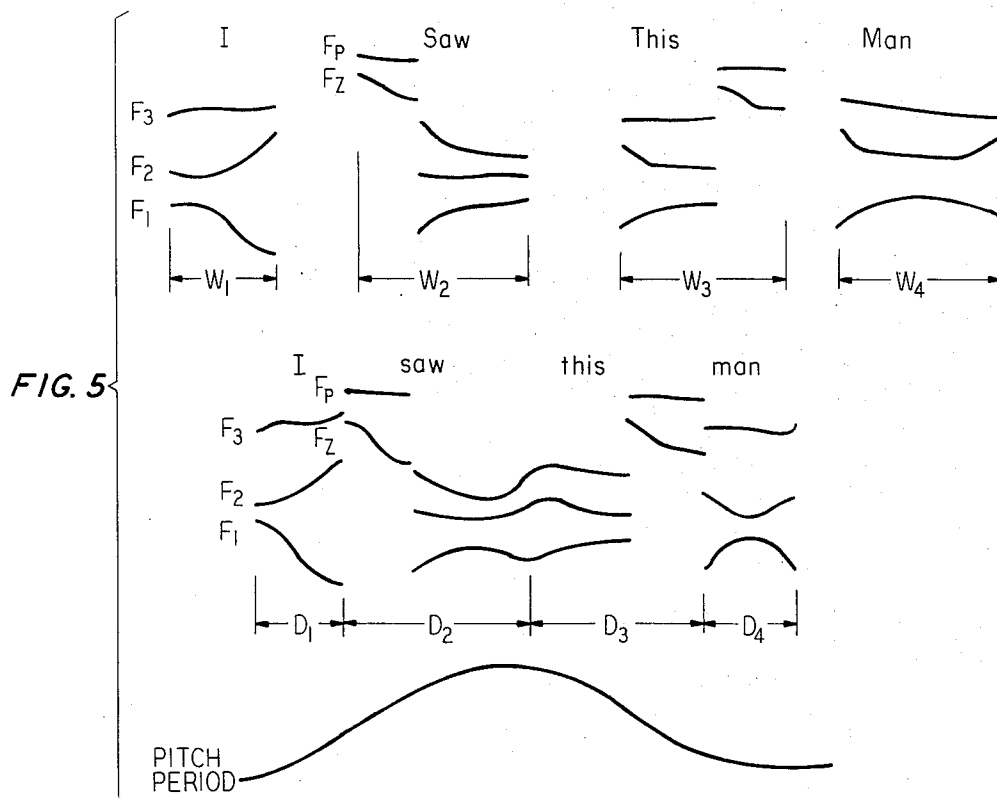
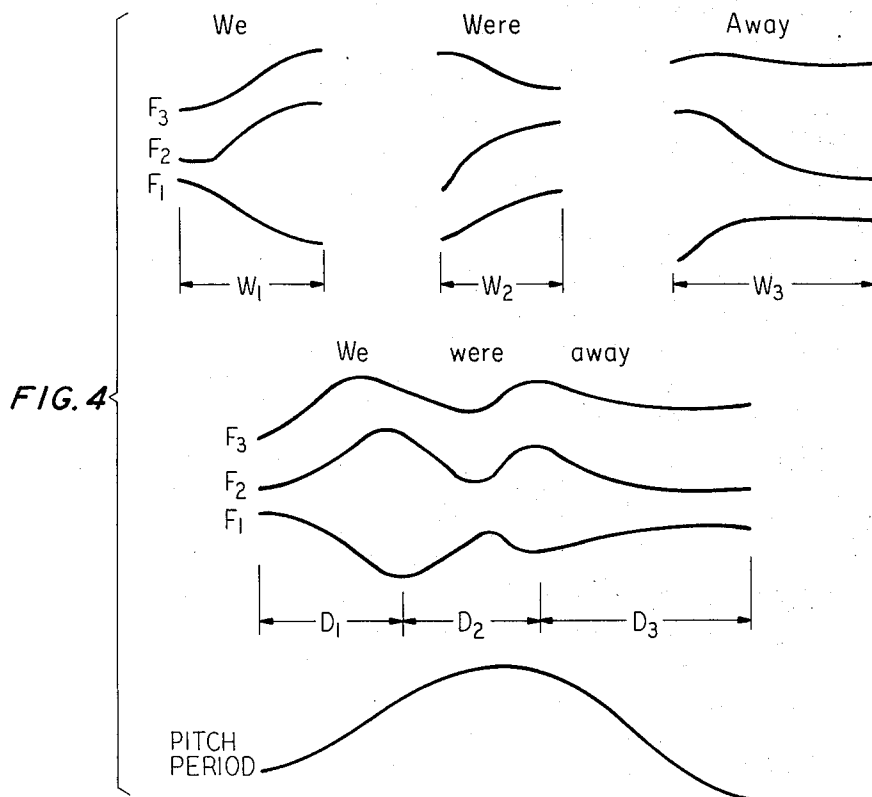


FIG. 6A

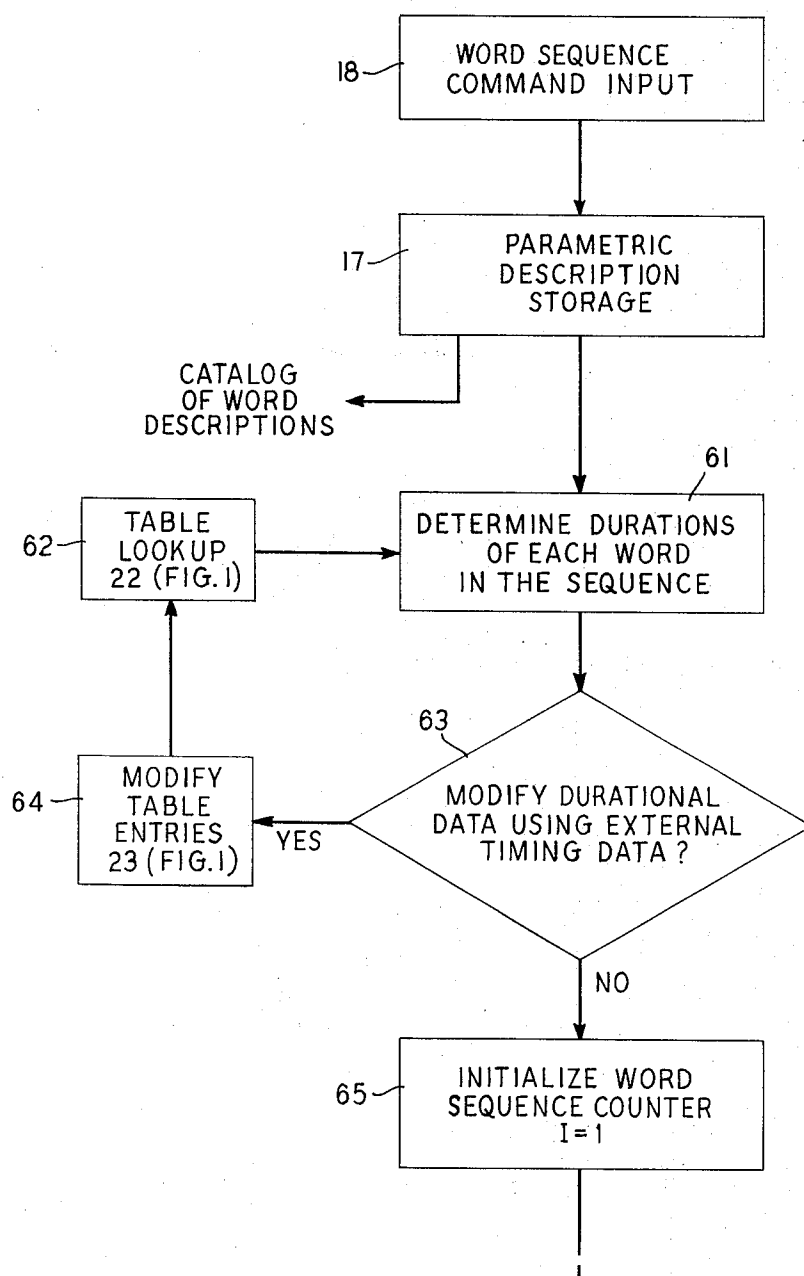


FIG. 6B

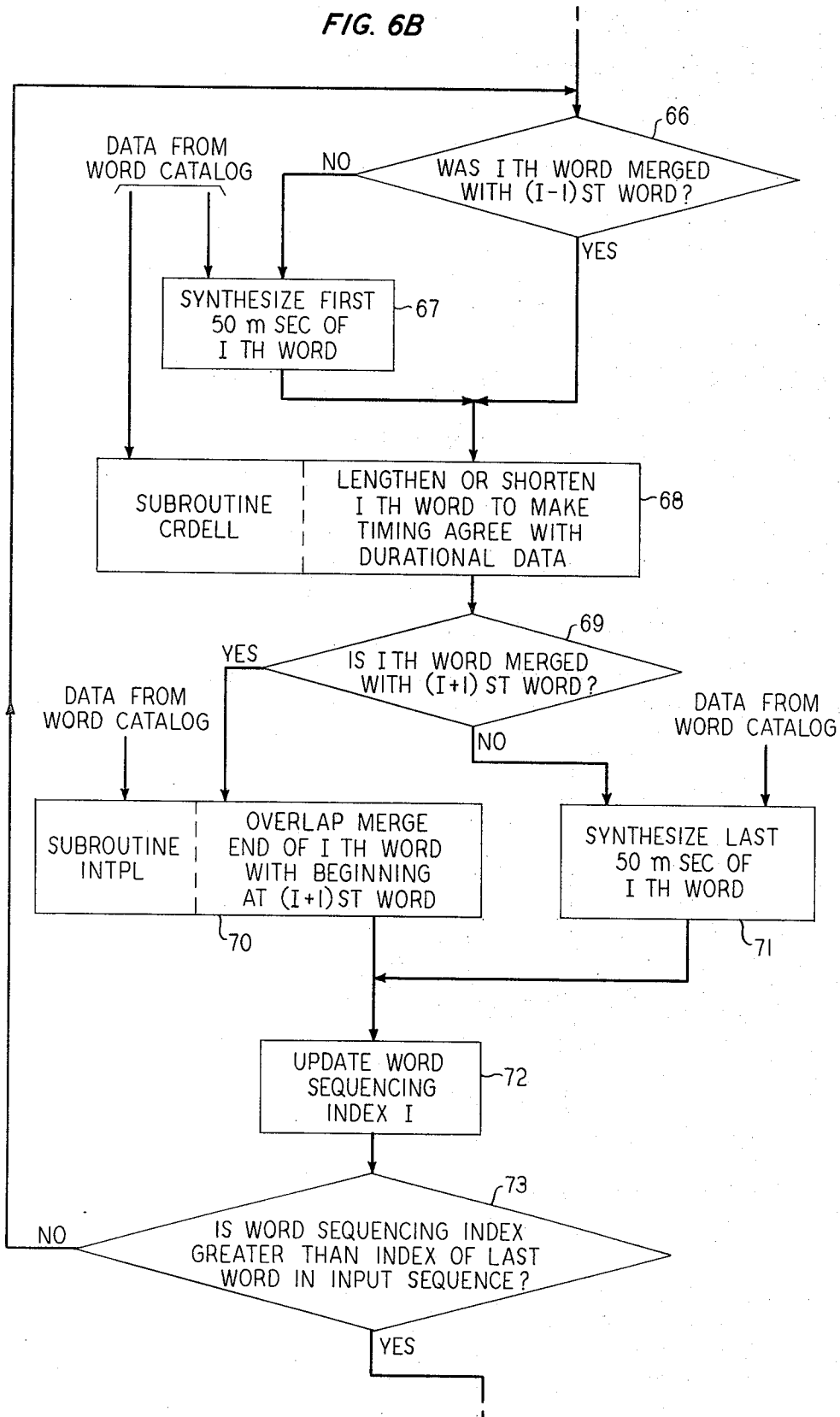
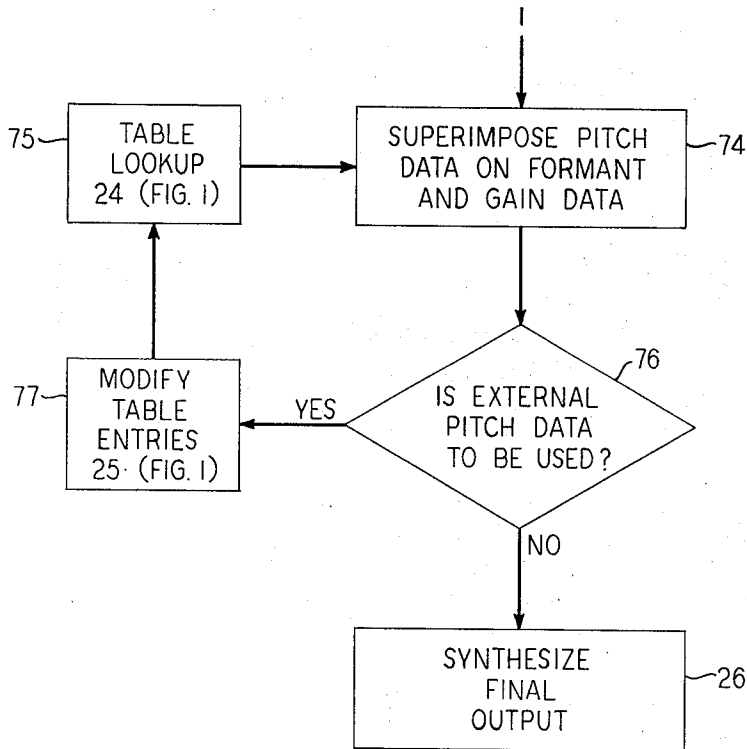


FIG. 6C



SPEECH SYNTHESIS BY CONCATENATION OF FORMANT ENCODED WORDS

This invention relates to the synthesis of limited context messages from stored data, and more particularly to processing techniques for assembling stored information into an appropriate specification for energizing a speech synthesizer.

BACKGROUND OF THE INVENTION

Much effort has been expended recently toward the generation of artificial speech from data supplied to or stored in a digital computer. Artificially generated speech messages are used, for example, in telephone automatic intercept and information systems, in automatic inventory reporting, for weather forecasting, and for the oral delivery of stock market quotations. In each of these applications, an inquiry is directed to a system, the system responds to the inquiry, develops the required answer, and calls upon a stored assemblage of spoken language data to formulate sufficient information with which to energize a speech synthesizer. The response to the question is eventually provided in the form of complete spoken utterances.

For such a service, it is evident that the system must have a large and flexible vocabulary. The system, therefore, must store sizable quantities of speech information and it must have the information in a form amenable to the production of a great variety of messages. Speech generated by the system would be as intelligible as natural speech. Indeed, the possibility exists that it might be made more intelligible than natural speech. It need not, however, sound like any particular human and may even be permitted to have a machine accent.

DESCRIPTION OF THE PRIOR ART

One technique for the synthesis of messages is to store individually spoken words and to select the words in accordance with the desired message output. Words pieced together in this fashion yield intelligible but highly unnatural sounding messages. One difficulty is that word waveforms cannot easily be adjusted in duration. Also, it is difficult to make smooth transitions from one word to the next. Nonetheless, such systems are relatively simple to implement and afford a relatively large vocabulary with simple storage apparatus.

To avoid some of the difficulties of word storage and to reduce the size of the store needed for a reasonable variety of message responses, individual speech sounds may be stored in the form of phoneme specifications. Such specifications can be called out of storage in accordance with word and message assembly rules and used to energize a speech synthesizer. However, speech at the acoustic level is not particularly discrete. Articulations of adjacent phonemes interact, and transient movements of the vocal tract in the production of any phoneme last much longer than the average duration of the phoneme. That is, the articulatory gestures overlap and are superimposed on one another. Hence, transient motions of the vocal tract are perceptually important. Moreover, much information about the identity of a constant is carried, not by the spectral shape at the steady-state time of the consonant but by its dynamic interactions with adjacent phonemes.

Speech synthesis, therefore, is strongly concerned with dynamics. A synthesizer must reproduce not only the characteristics of sounds when they most nearly represent the ideal of each phoneme, but also the dynamics of vocal tract motion as it progresses from one phoneme to another. This fact highlights a difference between speech synthesis from word or phrase storage and synthesis from more elementary speech units. If the library of speech elements is a small number of short units, such as phonemes, the linking procedures approach the complexity of the vocal tract itself. Conversely, if the library of speech elements is a much larger number of longer segments of speech, such as words or phrases, the elements can be linked together at points in the message where information in transients is minimal.

Thus, although phoneme synthesis techniques are attractive and sometimes adequate, the intermediate steps of assembling elementary speech specifications into words and words into messages according to prescribed rules requires complicated equipment and, at best, yields mechanical sounding speech.

SUMMARY OF THE INVENTION

These shortcomings are overcome in accordance with the present invention by storing representations of spoken words or phrases in terms of individual formant and other speech defining characteristics. Formants are the natural resonances of the vocal tract and they take on different frequency values as the vocal tract changes its shape during talking. Typically, three such resonances occur in the frequency range most important to intelligibility, namely, 0 - 3 kHz. Representation of the speech wave as a set of slowly varying excitation parameters and vocal tract resonances is attractive for at least two reasons. First it is more efficient for data storage than, for example, a pulse code modulation (PCM) representation of the speech waveform. Secondly, a formant representation permits flexibility in manipulation of the speech signal for the concatenation of words or phrases.

Thus, in accordance with the invention individual, naturally spoken, isolated words are analyzed to produce a word library which is stored in terms of formant frequencies. In the formant representation of an utterance, formant frequencies, voice pitch, amplitude and timing, can all be manipulated independently. Thus in synthesizing an utterance, an artificial pitch contour, i.e., the time course of the relevant parameters, can be substituted for the natural contour. A steady-state sound can be lengthened or shortened, and even the entire utterance can be speeded up, or slowed down with little or no loss in intelligibility. Formants can be locally distorted, and the entire formant contour can be uniformly raised, or lowered, to alter voice quality.

Upon program demand, word length formant data are accessed and concatenated to form complete formant functions for the desired utterance. The formant functions are interpolated in accordance with spectral derivatives to establish contours which define smooth transitions between words. Speech contour and word duration data are calculated according to stored rules. Following the necessary processing and interpolation, concatenated formant functions are used to synthesize a waveform which approximates a naturally spoken message. As an added advantage, economy in storage is achieved because formant and excitation parameters

change relatively slowly and can be specified by fewer binary numbers per second (bits) than can, for example, the speech waveform.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention will be fully apprehended from the following detailed description of illustrative embodiments thereof taken in connection with the appended drawings in which:

FIG. 1 illustrates schematically a suitable arrangement in accordance with the invention for synthesizing message-length utterances upon command;

FIG. 2 illustrates the manner of overlapping individual word formants, in accordance with the invention, for four different combinations of words;

FIG. 3 illustrates timing data which may be used for processing formant data;

FIG. 4 illustrates the processing of voiced formant data for individual words to produce a concatenated formant structure useful for actuating a speech synthesizer;

FIG. 5 illustrates the processing of both voiced and fricative formant data for individual words to produce a concatenated formant structure useful for actuating a speech synthesizer; and

FIGS. 6A, 6B and 6C illustrate by way of a flow chart the operations employed in accordance with the invention, for processing parametric data and for concatenating these data to produce a complete set of control signals for energizing a formant speech synthesizer.

DETAILED DESCRIPTION OF THE INVENTION

A system for synthesizing speech by the concatenation of formant encoded words, in accordance with the invention, is illustrated schematically in FIG. 1. Isolated words spoken by a human being are analyzed to estimate the parameters required for synthesis. Thus, naturally spoken, isolated words originating, for example, in system 10, which may include either studio generated or recorded words, are converted, if desired, to digital form in converter 11. The individual words, in whatever format, are supplied to speech analyzer 12, wherein individual formants, amplitudes, pitch period designations, and fricative pole and zero identifications are developed at the Nyquist rate. A suitable speech analyzer is described in detail in a copending application of Rabiner-Schafer, Ser. No. 872,050, filed Oct. 29, 1969, now U. S. Pat. 3,649,765, granted Mar. 14, 1972. In essence, analyzer 12 includes individual channels, including analyzer 13 for identifying formant (voiced) frequencies F_1 , F_2 , F_3 , analyzer 14 for developing a pitch period signal P , analyzer 15 for developing buzz, A_v , and hiss, A_h , level control signals, and analyzer 16 for developing fricative (unvoiced) pole and zero signals, F_p and F_z .

These control parameter values are delivered to parametric description storage unit 17, which may take any desired form. Both analog and digital stores, which may be accessed upon command, are known in the art. When completed, storage unit 17 constitutes a word catalog which may be referenced by the word concatenation portion of the system. The parameter values maintained in catalog 17 may be revised from time to time by the addition or deletion of new words.

INPUT COMMAND

An input command from word sequence input 18 ini-

tiates the necessary operations to synthesize a message composed of words from catalog 17. The exact form of input 18 depends upon the particular application of the word synthesis system. Typically, an inquiry of some form is made to the system embodied by unit 18, the necessary data for a response is formulated, and the appropriate word designations for the response, for example, in the English language, are assembled in code language and delivered to the synthesis system as the output signal of unit 18. Such response units are known to those skilled in the art and are described in various patents and publications. The output developed by such a responsive unit may thus be in the form of machine code language, phoneme or other linguistic symbols, or the like. Whatever the form of the output signal, it is delivered, in accordance with this invention, to word processing System 20, wherein required word data is assembled, processed, and delivered to speech synthesizer 26.

To synthesize a message composed of words from storage unit 17 requires the generation of timing contours, a pitch contour, and formant and amplitude contours. Processor 20, in accordance with the invention, employs separate strategies for handling the "segmental" features of the message, such as formant frequencies, unvoiced pole and zero frequencies and amplitudes, and the "prosodic" features, such as timing and pitch. Program strategy for treating the segmental features is self-stored in the processor. The prosodic feature information needed for processing is derived in or is supplied to processor 20. It is this flexibility in manipulating formant-coded speech that permits the breaking of the synthesis problem into two parts.

TIMING DATA

Timing information may be derived in one of several ways. For limited vocabulary applications, such as automatic intercept services, the timing rules need be nothing more complicated than a table specifying word duration as a function of position in an input string of data and as a function of the number of phonemes per word. Timing data for a typical seven number digit string is illustrated in the table of FIG. 3 and is normally stored in timing unit 22. For more sophisticated applications, word duration is determined from rules which take into account the syntax of the specific message to be produced, i.e., rules based on models of the English language. Such data also is stored in timing store 22. It is also possible to specify the duration of each word in the input string to be synthesized from external timing data supplied from unit 23. In this case, word duration is chosen according to some external criterion, for example, or measured from a naturally spoken version of the message to be synthesized, and is not necessarily a typical duration for that word, independent of context. Thus, external timing data may be supplied from stored data or from real time adjustments made during synthesis.

PITCH DATA

Synthesis also requires the determination of the appropriate pitch contour, i.e., pitch period as a function of time, for the message being synthesized. Pitch information can be obtained in several ways. For example, the pitch character of the original sequence of spoken words may be measured. Alternatively, a monotone or an arbitrarily shaped contour may be used. However,

in practice both of these have been found to give unacceptable, unnatural results. Accordingly, it is in accordance with this invention to use a time-normalized pitch contour, stored in unit 24, and to modify it to match the word portions as determined from the timing rules. Thus, pitch data stored in unit 24 are supplied to concatenating processor 21 wherein the contour is locally lengthened or shortened as required by the specific utterance timing as specified by the timing data. If desired, pitch variation data may be supplied from external source 25, either in the form of auxiliary stored data, or as real time input data. For example, a pitch contour extracted from a naturally spoken version of the message may be used. Such data would normally be used when word durations have been obtained in a similar manner, i.e., from external timing unit 23.

Pitch and timing information obtained externally in this manner provide the most natural sounding synthesized speech. It is also possible to calculate pitch contour information by rule. Thus, there are many ways in which the prosodic information for a message can be obtained, and the choice depends strongly on the desired quality of the synthetic speech and the specific application for which it is to be used.

WORD DURATION ADJUSTMENT

Once the timing pattern for the message is established, isolated words in word catalog 17 can be withdrawn and altered to match the specified timing. Thus, formant data for a word in the catalog may be either lengthened or shortened. The formant contours for successive voiced words are smoothly connected together to form continuous transitions and continuous formant contours for the message. The choice of place in a word to alter duration is based on the dynamics of the formant contours. For each subinterval of a voiced sound, typically 10 msec in duration, a measure of the rate of change of formant contours is computed in processor 21. This measure is called the "spectral derivative." Regions of the word where the spectral derivative is small are regions where the word can be shortened or lengthened with the least effect on word intelligibility. Thus, to shorten a word by a given amount, an appropriate number of 10 msec intervals are deleted in the region of the smallest spectral derivative. To lengthen a word, the region of the lowest spectral derivative is lengthened by adding an appropriate number of 10 msec intervals. Unvoiced regions of words are never modified.

In practice, the measure of spectral derivative, SD_i , is calculated where

$$SD_i = \sum_{j=1}^3 |F_j(i) - F_j(i-1)|, \quad (1)$$

$i(1, 2, \dots)$ is the i^{th} 10 msec interval and $F_j(i)$ is the value of the j^{th} formant in the i^{th} time interval. To determine how many 10 msec intervals must be added to (or subtracted from) the isolated word controls, an equation is used based on desired word duration, isolated word duration, and some simple contextual information concerning how the current word is concatenated with its preceding and following neighbors. By defining the symbols:

$I_{PM} = 1$ — if the end of the preceding word is voiced, and the beginning of the current word is also voiced; 0 — otherwise

$I_{NM} = 1$ — if the end of the current word is voiced, and the beginning of the following word is also voiced; 0 — otherwise

W_i = duration of current word spoken in isolation

W_D = duration of current word spoken in context (as determined from timing rules)

W_C = number of 10 msec intervals to be added if $W_C > 0$ (or subtracted if $W_C < 0$)

then

$$W_C = W_D - W_i + 5 \times (I_{PM} + I_{NM}) \quad (2)$$

The reason for the last term in the above equation is that whenever either $I_{PM} = 1$ or $I_{NM} = 1$, it means that the two words must be smoothly merged together, and will overlap each other by 100 msec. However, this 100 msec region is shared by the two words; hence 50 msec (5 intervals) are allotted to each word separately in terms of the overall timing. The technique by which the W_C additional 10 msec intervals are inserted, or removed, is based entirely on the spectral derivative measurement. As noted above, for each 10 msec voiced interval of the isolated word, the spectral derivative is calculated. To shorten a word, the W_C intervals having smallest spectral derivatives are removed. To lengthen a word, the region of the word having smallest spectral derivative is located and W_C intervals are inserted at the middle of this region. Each of the W_C intervals is given the control parameters of the center of the interval i.e., a steady-state region of W_C intervals is added.

OVERLAP OF WORD DESCRIPTIONS

Except for the case when the end of the current word, as well as the beginning of the following word, are both voiced, the control data from word to word are simply abutted. Whenever the end of one word is voiced and the beginning of the next word is also voiced, a smooth transition is thus made from the formants at the end of one word to those at the beginning of the next word. This transition is made, for example, over the last 100 msec of the first word and the first 100 msec of the second. The transition rate depends on the relative rates of spectrum change of the two words over the merging region.

To perform this transition task, an interpolation function is used whose parameters depend strongly on the average spectral derivatives of the two words during the merging region. If the spectral derivative symbols are defined as:

$$\overline{SD1} = \sum_{i=n_0}^{n_0+9} SD1_i$$

$$\overline{SD2} = \sum_{i=1}^{10} SD2_i$$

n_0 = starting interval of merging region for current word

$F_j(l)$ = Value of formant j of the message contour at time l during the merger region, $l = 0, 1, \dots, 9$,

then the interpolation function used is $F_j(l) = F_j^1(n_0+l) \cdot (9-l) \cdot \overline{SD1} + F_j^2(l) \cdot l \cdot \overline{SD2} / (9-l) \cdot \overline{SD1} + l \cdot \overline{SD2}$

where $F_j^k(l)$ = value of the j^{th} formant, at time l for word

k ($k = 1$ is current word, $k = 2$ is following word).

FORMANT INTERPOLATION

FIG. 2 illustrates the type of interpolation performed for four simple cases in accordance with these considerations. Although all three formants of a sound are interpolated, only one formant is illustrated for each word to simplify the presentation. For the words in column 1, word 1 (the top spectrum) exhibits a very small change over its last 100 msec of voicing, whereas word 2 (middle spectrum) exhibits a large change. The interpolated curve shown at the bottom of the first column, although beginning at the formants of word 1, rapidly makes a transition and follows the formants of word 2. Column 2 shows the reverse situation; word 2 exhibits little spectrum change whereas word 1 has a large spectrum change. The interpolated curve, therefore, follows the formants of word 1 for most of the merging or overlap region and makes the transition to the formants of word 2 at the end of the region. Columns 3 and 4 show examples in which spectrum changes in both words are relatively the same. When they are small, as in column 3, the interpolated curve is essentially linear. When they are large, as in column 4, the interpolated curve tends to follow the formants of the first word for half of the overlap region, and the formants of the second word for the other half.

The interpolated curve thus always begins at the formants of word 1 (the current word) and terminates with the formants of word 2 (the following word). The rate at which the interpolated curve makes a transition from the formants of the first word to those of the second is determined by the average spectral derivatives $SD1$ and $SD2$. In the example of column 1, the spectral derivative of the second word is much greater than that of the first so the transition occurs rapidly at the beginning of the overlap region. For example of the second column the spectral derivative of the first word is the greater so that the transition occurs rapidly at the end of the overlap region. As indicated above, the spectral derivatives for both words in the examples of columns 3 and 4 are much the same so that no rapid transitions take place in the overlap region.

EXAMPLES OF CONCATENATION

FIGS. 4 and 5 illustrate the manner in which these rules and considerations are turned to account in the practice of the invention. FIG. 4 illustrates the manner in which three voiced words, "We," "Were," and "Away" are linked together to form the sentence "We were away." As spoken, the words have durations W_1 , W_2 , W_3 , as indicated, and through analysis have been determined to have formants F_1 , F_2 , and F_3 . These formant data are stored in storage unit 17 (FIG. 1) for the individual words, as discussed above. Upon an input command from word sequences unit 18 to assemble the three words into the sentence "We were away," the formant data is drawn from storage unit 17 and delivered to word concatenating processor 21. Timing data from storage 22 (or alternatively from external unit 23) and pitch variation data from store 24 (or alternatively from external source 25) are supplied to the processor. It is initially determined that the words "We" and "Were" are normally linked together in speech by a smooth transition and uttered as one continuous

phrase, "Wewere." Hence, the two voiced words are adjusted in duration to values D_1 , D_2 in accordance with the context of the utterance, and the formants of the words are overlapped and interpolated to provide the smooth transition. Similarly, the words "were" and "away" are normally spoken as "wereaway" with time emphasis on "away". Hence, the duration of "away" is lengthened to D_3 and the formants for the two words are overlapped and interpolated.

The resulting smoothly interpolated formant specification is further modified by superimposing the pitch period contour illustrated in the figure. The resultant is a contiguous formant specification of the entire utterance. These formant data as modified, together with the pitch period contour, and voiced-unvoiced character data A_V and A_N , are delivered to speech synthesizer 26 (FIG. 1).

FIG. 5 illustrates the concatenation of the words "I," "Saw," "This," and "Man," to form the phrase "I saw this man". In this case the words "I" and "Saw" are not overlapped because of the intervening fricative at the beginning of "Saw." However, the words "Saw" and "This" are generally spoken with a smooth transition. Hence, these words are overlapped and the formants are interpolated. Since the word "This" ends in a fricative, the words "This" and "Man" are not overlapped. In accordance with the context of the expression, the individual word lengths W are each modified to the new values D . Finally, a stored pitch period contour is superimposed according to a stored rule. The resultant specification of the phrase "I saw this man" is thus delivered together with voiced-unvoiced character data, A_V , A_N , and fricative pole-zero data, F_P and F_Z , to the speech synthesizer.

INTENSITY DATA

The unvoiced intensity parameters, A_N , is obtained directly from the stored controls in word catalog 17 when the interval to be synthesized is unvoiced. The voiced intensity parameter, A_V , is similarly obtained directly from word catalog 17, except during a merging region of two voiced intervals, in which case it is obtained by interpolation of the individual voiced intensities of the two words in a fashion similar to that described for the interpolation of formants.

CONCATENATION PROCESSOR IMPLEMENTATION

Although the operations described above for processing word formant data to form word sequence information may be carried out using any desired apparatus and techniques, one suitable arrangement used in practice relies upon the high-speed processing ability of a digital computer. In practice a general purpose digital computer, namely, the Honeywell DDP-516 or the GE-635, have been found to be satisfactory. The two machines and their software systems are equally adaptable for receiving a program prepared to convert them from a general purpose machine to a special purpose processor for use in the practice of the invention.

A flow chart of the programming steps employed to convert such a machine into special purpose processing apparatus which turns to account the features of the invention, is shown in FIGS. 6A, 6B, and 6C, taken together as one complete description. Each step illustrated in the flow chart is itself well known and can be reduced to a suitable program by any one skilled in the

programming art. The unique subroutines employed in the word length modification operation and in the overlapping operation are set forth in Fortran IV language in Appendices A and B attached hereto.

Although any general purpose digital computer may be adapted to perform the operations required by the flow chart of FIG. 6, a unit with characteristics similar to that of the DDP-516 is preferred. The DDP-516 includes 16 k of core memory, hardware, multiply and divide, direct multiplex control with 16 data channels (0.25 mHz each), and a direct memory access channel (1.0 mHz). Input is by way of a teletypewriter. A Fortran IV compiler, DAP-16 machine-language assembler, match libraries, and various utility software are standard items supplied by the manufacturer and delivered with the machine. If desired, a number of peripheral units may be interfaced with the computer for convenience. This may include auxiliary word stores, card readers, display scopes, printers, tape readers, registers, and the like. Such units are well known to those skilled in the art and are generally available on the open market. They may be interconnected with the basic computer as required by the specific application to which the processor of this invention is to be used.

PROCESSOR OPERATIONS

In the portion of the flow chart shown at the top of FIG. 6A there is indicated schematically the parametric description storage unit 17 of FIG. 1 which contains a catalog of formant pitch amplitude and fricative specifications for each of the words in the catalog. Upon command from word sequence input 18, these data are transferred to word concatenating processor system 20, which is illustrated by the remainder of the flow chart.

Initially, the duration of each word in the connected sequence is determined, as indicated in block 61, for example, by examining a stored table of timing data 62, of the sort illustrated in FIG. 3 and by unit 22 in FIG. 1. If a timing change is necessary, the program statements of unit 63 determines whether data in store 62 is sufficient of whether external timing data from unit 64 (block 23 of FIG. 1) should be used. In either event, the duration of each commanded word is established and a word sequence counter, in unit 65, is initialized by setting $I=1$.

It is then necessary to modify the parametric description of the first word in accordance with timing data and other stored rules. Accordingly, it is determined whether the I^{th} word was merged with the $(I-1)^{\text{st}}$ word. This determination is represented by block 66. If it was not, information for the I^{th} word is withdrawn from word catalog 17 and the first 50 msec of the I^{th} is synthesized by unit 67. If the I^{th} word was so merged, the I^{th} word is lengthened or shortened to make timing agree with durational data supplied as above. This operation takes place in unit 68 in conjunction with subroutines CRDELL, a listing for which appears in Appendix A.

It is then ascertained whether the I^{th} word is to be merged with the $(I+1)^{\text{st}}$ word via the steps of block 69. If there is to be a merger, the operations of block 70 are carried out to overlap the end of the I^{th} word with the beginning of the $(I+1)^{\text{st}}$ word. This operation is carried out in conjunction with subroutine INTPL, a listing for which appears as Appendix B. If it is determined in

block 69 that there is to be no merging, the operations of block 71 synthesize the last 50 msec of the I^{th} word using data for that word supplied from store 17.

It is then necessary in unit 72 to update the word sequencing of index I and, in operation 73, to determine if the word sequencing index is greater than the index of the last word in the input sequence. If it is not, control is returned to block 66, and the next word is composed in the fashion just described. The operations are thus iterated until the index is equal to the index of the last word in the input sequence, at which time the data from block 73 is transferred to block 74.

Pitch data is then superimposed on the formant and gain structure of each word in the utterance in the fashion described in detail above. These data are available in pitch variation data store 75 (store 24 of FIG. 1). It is next determined by the steps indicated in block 76 whether external pitch data is to be used. If it is, such data from unit 77 (unit 25 in FIG. 1) is supplied by way of data store 75 to the operations of unit 74.

When the pitch contour operation has been completed, all of the data in the word concatenating processor 20 as modified by the program of FIG. 6, is transferred, for example, to speech synthesizer 26 of FIG. 1.

FORMANT SYNTHESIS

When all of the control parameter contours of the commanded utterance have been generated, they may, if desired, be smoothed and band-limited to about 16 Hz. They are then used to control a formant synthesizer which produces a continuous speech output. Numerous systems, both analog and digital, have been described for synthesizing speech from formant data. One suitable synthesizer is described in J. L. Flanagan Pat. No. 3,330,910, another in David-Flanagan, Pat. No. 3,190,963, FIG. 5, and another is described in Gerstman-Kelly Pat. No. 3,158,685. The abovesited Rabiner-Schafer application illustrates a typical formant synthesizer and relates the exact parameters described hereinabove to the input of the synthesizer described in the Flanagan patent. Very simply, a formant synthesizer includes a system for producing excitation as a train of impulses with a spacing proportional to the fundamental pitch of the desired signal. The intensity of the pulse excitation is controlled and the signal is applied to a cascade of variable resonators.

Suffice it to say, speech synthesizer 26 generates a waveform which approximates that required for the desired utterance. This signal is utilized in any desired fashion, for example, to energize output unit 27 which may be in the form of a loudspeaker, recording device, or the like.

APPENDIX A

```

C **** SUBROUTINE TO ELIMINATE OR ADD 10
      MSEC
C **** INTERVALS TO MAKE WORD DURATION
      MATCH
60 C **** DESIRED DURATION
      SUBROUTINE CRDELL(IST,ISP,NEL)
      COMMON JPTCH(300),JADR(8),LSTA,LNO,LTYPE,
      LCALL,ILLENG
      COMMON JAP(720,8),JAD(25),JSEQ(25),
      JIWTIM(25),J2,J3,JRD2,JRD4,JRD8,JAV
      COMMON IFCT,JF1(500),JF2(500),JF3(500),JF4(500),
      JF5(500),JF6(500),JF7(500)
      IF(TS7,GE1,ISP) RETURN
      ILOC=IST
      NC=MAXO(NEL,-NEL)

```

```

      IF(NC.EQ.0)GOTO22
C **** LOCATE REGION OF MINIMUM SPECTRAL
C **** DERIVATIVE OVER NEL INTERVALS
      IS=30000
      DO20I=IST.ISP
      ITEST=0
      DD10J=1.NC
      II=T+J-1
      IF(JAR(II,1),EQ.O)GOTO20
      IF(II.GT.ISP)GOTO22
10    ITEST=ITEST+JAR(II,8)
      IF(ITEST.GT.IS)GOTO20
      IS=ITEST
      ILOC=I
20    CONTINUE
22    CONTINUE
      IF(NEL.GT.O)GOTO35
C **** ADD NEL INTERVALS IN MIDDLE OF
C **** REGION OF
C **** MINIMUM SPECTRAL DERIVATIVE
C **** ADDITION OF CARDS
      II=TLLOC+NC/2
      IC=II-IST
      IF(IC.NE.O)CALL CRDSYN(IST,IC)
      IF(NEL.EQ.O)GOTO30
      NEL=-NEL
      DO25I=1.NEL
25    CALLCRDSYN(T1,I)
30    IC=ISP-IST+I-IC
      RETURN
35    CONTINUE
C **** FIND SPECTRAL DERIVATIVE LEVEL SUCH
C **** THAT
C **** NEL INTERVALS HAS THIS LEVEL OR
C **** SMALLER
      IDIF=500
      ITHR=500
37    JC=0
      DO40I=IST.ISP
      IF(JAR(T,1),EQ.O)GOTO40
      IF(JAR(I,8),LE.ITHR)JC=JC+1
40    CONTINUE
      IDIF=IDIF/2
      IF(IDIF.EQ.O)GOTO55
      IF(JC.GT.(NEL+1))GOTO45
      IF(JC.LT.NEL)GOTO50
      GOTO55
45    ITHR=ITHR-IDIF
      GOTO37
50    ITHR=ITHR+IDIF
      GOTO37
55    ICNT=0
C **** FLIMINATE THE INTERVALS WITH
C **** SPECTRAL
C **** DERIVATIVES LESS THAN THIS LEVEL
      DO60I=IST.ISP
      IF(JAR(I,1),EQ.O)GOTO56
      IF(JAR(I,8),LE.ITHR)GOTO57
56    CONTINUE
      CALLCRDSYN(I,1)
      GOTO60
57    ICNT=ICNT+1
      IF(ICNT.GE.NEL)ITHR=-1
60    CONTINUE
      RETURN
      END

```

APPENDIX B

```

C **** SUBROUTINE TO MERGE WORDS AND
C **** INTERPOLATE
C **** THEIR CONTROL SIGNALS
      SUBROUTINE TPL(IST,JST,IW,LST,IL1,IL2,ITPANS,
      INUM)
      COMMON JPTCH(300),JADR(8),LSTA,LNO,LTYPE,
      LCALL,ILLENG
      COMMON JAR(720,8),IA(25),IAD(25),ISEQ(25),
      IWTIM(25),J2,J3,JRD2,JRD4,JRD8,IAV
      COMMON IFCT,JF1(500),JF2(500),JF3(500),JF4(500),
      JF5(500),JF6(500),JF7(500)
      DIMENSION JFIN(7)
      DIMENSION JSAT(7)
C **** CALCULATE AVERAGE SPECTRAL
C **** DERIVATIVES OF BOTH WORDS OVER THE
C **** MERGING REGION WHICH CONSIST OF IW
C **** INTERVALS
2    CONTINUE
      JS1=0
      JS2=0
      DO51I=1.IW

```

```

      II=IST+I-1
      I2=JST+I-1
      JS1=JS1+JAR(II,8)
5    JS2=JS2+JAR(I2,8)
      IND=1
5    DO30I=1.IW
C **** GET STARTING ADDRESSES OF DATA FOR
C **** BOTH WORDS
      IL=IST+I-1
      JL=JST+I-1
      KL=LST+I-1
      LL=IND+I-1
10    LM=IW+1-LL
      NORM=JS1+LM+JS2+LL
C **** MERGE AND INTERPOLATE CONTROL
C **** SIGNALS OVER THESE IW INTERVALS
      DO20J=1,7
      JL1=JAR(IL,J)+LM
      JL2=JAR(JL,J)+LL
15    JAR(KL,J)=JKOL(JL1,JS1,NORM)+JKOL(JL2,JS2,
      NORM)
20    CONTINUE
30    CONTINUE
      CALLCRDSYN(LST,IW)
      RETURN
20    END

```

What is claimed is:

1. A system for composing speech messages from sequences of prerecorded words, which comprises:
 - 25 means for analyzing each word of a vocabulary of spoken words to produce a separate parametric description of each;
 - means for storing said parametric descriptions;
 - means under control of an applied command signal for sequentially withdrawing from storage those descriptions required to assemble a desired spoken message;
 - means for individually altering the duration of the description of each word of said message in accordance with prescribed timing rules;
 - 35 means for merging consecutive word descriptions together on the basis of the respective, voice-unvoiced character of the merged word descriptions;
 - 40 means for altering the pitch characteristic of said continuous message description in accordance with a prescribed contour; and
 - means for utilizing said continuous description to control a speech synthesizer.
- 45 2. A system for composing speech messages as defined in claim 1, wherein,
 - said parametric description of each word in said vocabulary comprises:
 - a representation of the formants, voiced and unvoiced amplitudes, and fricative pole-zero characteristics of said spoken word.
- 50 3. A system for composing speech messages as defined in claim 2, wherein,
 - said representations are in a coded digital formant.
- 55 4. Apparatus for processing parametric descriptions of selected prerecorded spoken words to form a continuous description of a prescribed message suitable for actuating a speech synthesizer, which comprises:
 - 60 means for deriving a spectral derivative function for each word description of said message;
 - means for individually altering the durations of selected word descriptions in accordance with stored timing information;
 - means operative in response to said spectral derivative functions for developing parametric descriptions of transitions between voiced word regions scheduled to be merged to form said message;
 - means for concatenating said altered word descrip-

tions with said transition descriptions in accordance with said prescribed message to form a continuous parametric message description; and means for altering the pitch characteristic of said message description in accordance with prescribed rules.

5. Apparatus for processing parametric descriptions as defined in claim 4, wherein:

said stored timing information comprises a schedule of word durations as a function of position in an input string of words, and of the number of phonemes per word.

6. Apparatus for processing parametric descriptions as defined in claim 4, wherein, said stored timing information comprises:

a schedule of word durations derived from rules based on common language usage.

7. Apparatus for processing parametric descriptions as defined in claim 4, wherein, said stored timing information comprises:

a schedule of word durations assembled from measurements of a naturally spoken version of said prescribed message.

8. Apparatus for processing parametric descriptions of selected prerecorded words, as defined in claim 4, wherein,

said parametric descriptions of transitions are developed for the last 100 msec of the first of two words to be merged and the first 100 msec of the second of said two words to be merged.

9. Apparatus as defined in claim 8, wherein, the rate of transition between said two words is proportional to the average of said spectral derivatives

for said two words.

10. Apparatus for processing parametric descriptions of selected words as defined in claim 4, wherein said means for altering the pitch characteristic of said message description comprises:

a stored, time-normalized pitch contour for a selected number of different messages; and means for modifying said contour in accordance with said altered word description durations.

11. Apparatus for developing control signals for a speech synthesizer, which comprises:

means supplied with word length segmental and prosodic functions of each individual word of a desired message for deriving the spectral derivatives of each of said functions;

means responsive to said spectral derivatives for interpolating said segmental functions to establish contours which define smooth transitions between the words of said message;

means for concatenating said segmental functions in accordance with said transition contours, and,

means for utilizing said prosodic functions to alter said concatenated segmental functions to develop control waveform signals which approximate the waveform of said desired message.

12. Apparatus as defined in claim 11, wherein, said segmental functions include the formant frequencies, unvoiced pole and zero frequencies and amplitudes of each of said words.

13. Apparatus as defined in claim 11, wherein, said prosodic functions include timing and pitch variations for said words as a function of message syntax.

* * * * *

35

40

45

50

55

60

65

UNITED STATES PATENT OFFICE
CERTIFICATE OF CORRECTION

Patent No. 3,828,132 Dated August 6, 1974

Inventor(s) James L. Flanagan-Lawrence R. Rabiner-
Ronald W. Schafer

It is certified that error appears in the above-identified patent and that said Letters Patent are hereby corrected as shown below:

- Col. 1, line 31, "should" would; read --should--;
line 66, "constant" should read --consonant--.
- 2, line 49, "parameters" should read --parameter--.
- 6, line 49, "spectram" should read --spectrum;
line 64, " $F_j(1)$ " should read -- $F_j(l)$;
- line 67, equation (3) should read
- $$-F_j(l) = F_j^1(n_0+l) \cdot (9-l) \cdot \overline{SD1} + F_j^2(l) \cdot l \cdot \overline{SD2} \cdot \frac{(9-l)\overline{SD1} + l \cdot \overline{SD2}}{(3)} \quad (3)$$
- The bar should be only over $\overline{SD1}$ and $\overline{SD2}$ and the numeral (3) should be separated from the equation by spaces as this numeral is not part of the equation but is only intended to identify same.
- Col. 8, line 13, "continguous" should read --contiguous--;
- Col. 10, line 67, "IF(TS7)" should read --IF(TST)--.
- Col. 11, line 1, "IF(NC.EQ.0)GOTO22" should read
--IF(NC.EQ.01)GOTO22--;
- line 7, "I1=T+J-1" should read --I1=I+J-1--;
- line 18, "I1 = TLOC+NC/2" should read --I1=ILOC+NC/2--;
- line 20, "IF(NEL.EQ.0)GOTO30" should read
--IF(NEL.EQ.0)GOTO30--;
- line 22, "CALLCRDSYN(T1,1)" should read
--CALLCRDSYN(I1,1)--;
- Following line 23, add:
--IF(IC,NE,0)CALLCRDSYN(I1.IC);
- line 40, "FLIMINATE" should read --ELIMINATE--;
- line 55, "SUBROUTINE NTPL" should read
--SUBROUTINE INTPL--;
- line 62, "DIRENSIONJSAT(7)" should read
--DIMENSIONJSAT(7).

Signed and sealed this 1st day of April 1975.

(SEAL)
Attest:

RUTH C. MASON
Attesting Officer

C. MARSHALL DANN
Commissioner of Patents
and Trademarks