March 10, 1970 L. R. FOCHT 3,499,986

SPEECH SYNTHESIZER

Filed Sept. 28, 1966 2 Sheets-Sheet 1

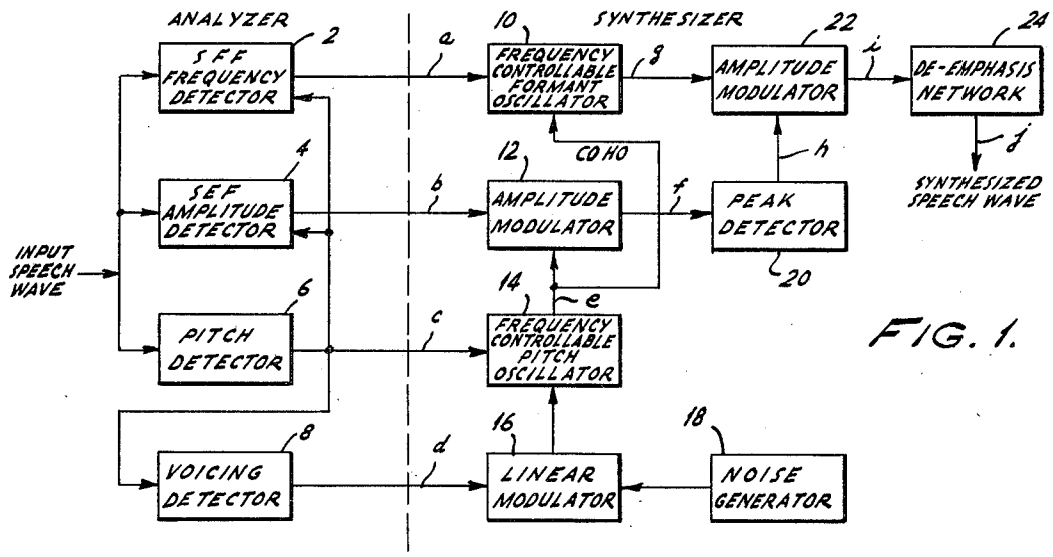

FIG. 1.

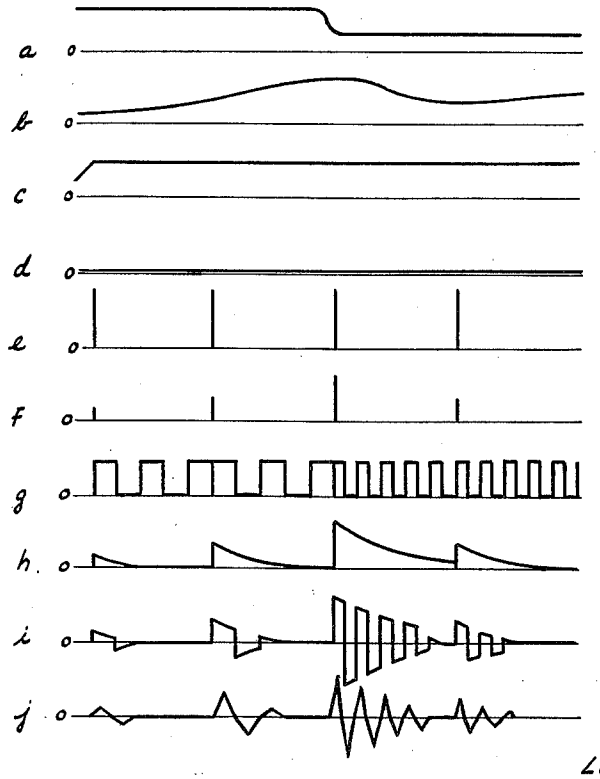

FIG. 2.

INVENTOR.
LOUIS R. FOCHT
BY Leonard Zalman
ATTORNEY

March 10, 1970 L. R. FOCHT 3,499,986

SPEECH SYNTHESIZER

Filed Sept. 28, 1966 2 Sheets-Sheet 2

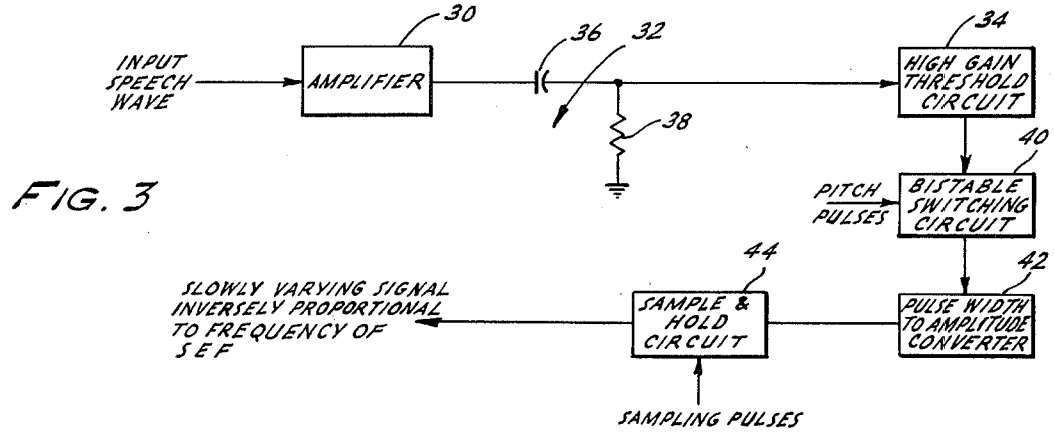

FIG. 3

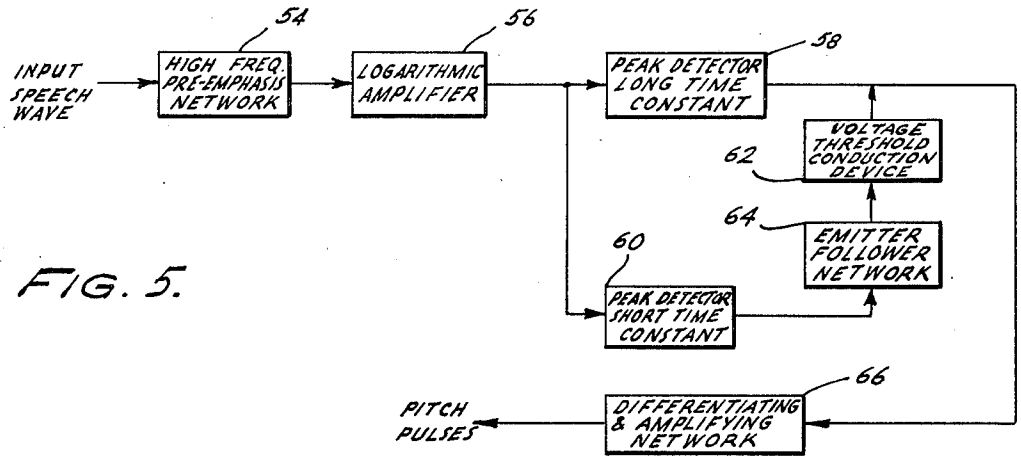

FIG. 5.

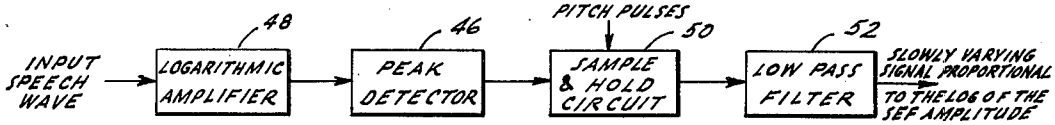

FIG. 4.

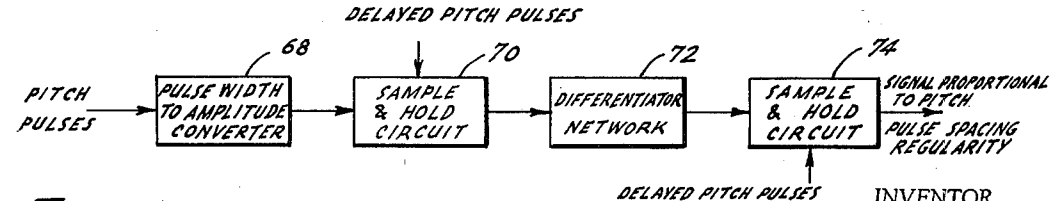

FIG. 6.

INVENTOR.
LOUIS R. FOCHT

BY Leonard Zalman

ATTORNEY

United States Patent Office 3,499,986 Patented Mar. 10, 1970

3,499,986

SPEECH SYNTHESIZER

Louis R. Focht, Huntingdon Valley, Pa., assignor to Philco-Ford Corporation, Philadelphia, Pa., a corporation of Delaware

Filed Sept. 28, 1966, Ser. No. 582,573
Int. Cl. H04m 1/60
U.S. Cl. 179—1 6 Claims

ABSTRACT OF THE DISCLOSURE

A speech synthesizer responsive to no more than four input control signals transmitted thereto to synthesize a speech wave. Said control signals include a signal representative of the single equivalent formant frequency of a sound and a signal representative of the single equivalent formant amplitude of that sound. The single equivalent formant is a time-varying signal the amplitude of which is directly dependent on the peak amplitude of the first major oscillation occurring after each pitch pulse, of the electrical signal representative of the speech wave, and the frequency of which is approximately equal to the frequency of said first major oscillation after pre-emphasis of high-frequency components of that electrical signal. The synthesizer includes apparatus for generating a signal representative of the single equivalent formant frequency of the sound and apparatus for amplitude modulating that signal in accordance with a signal representative of the single equivalent formant amplitude of the sound.

The ever-increasing volume of communication traffic severely strains existing wire facilities and the useful radio spectrum. Running additional cables and expanding the RF spectrum to and beyond light frequencies offer only partial solutions to this problem. Cabling can be extremely expensive and the higher RF frequencies are limited to line-of-sight and usually to good weather conditions. The more practical and economical approach to the problem of communication traffic is to make efficient use of existing channels by reducing signal redundancy and by employing advanced modulation techniques.

Because of fundamental considerations, speech waves are highly redundant and considerable saving in bandwidth can be realized by proper processing of the signal to eliminate components not required for intelligible speech communication. A bandwidth of approximately 3,000 cycles per second is required to transmit directly an intelligible voice communication. This bandwidth can be reduced by a factor of 10 or more by proper signal processing. However prior art signal processing systems are subject to numerous disadvantages.

The most common speech bandwidth compression systems are the channel vocoder and the formant tracking vocoder. A channel vocoder typically consists of an analyzer having a plurality of bandpass filter stages covering the frequency range from 200 to 3,200 cycles per second and a synthesizer composed of electrical filters and modulators that reconstructs speech from the signals appearing at the analyzer output. Channel vocoders have several major drawbacks that restrict their use in speech bandwidth compression systems. For acceptable speech transmission, they require a transmission channel bandwidth of over 300 cycles per second. Furthermore, since channel vocoders use nearly 100 filters, the circuits for such vocoders are very complex.

The formant vocoder method of speech communication and bandwidth compression is based on transmitting signals representative of the formants or vocal tract resonances of the speech wave. The system requires the

transmission of signals representative of the frequency and amplitude of the three principal formants in the speech wave as well as signals representative of voicing and pitch information. Thus, such a system requires the transmission of eight independent parameters which convey the intelligibility of speech. Format vocoders also have several major drawbacks that restrict their use in speech communication systems. For standard formant vocoder transmission, a transmission channel bandwidth of approximately 200 cycles per second is required. Furthermore, since the tracking of the formant frequencies may require filters using inductors, the size of the system makes it incompatible with present microminiaturization techniques.

It is, accordingly, an object of the present invention to provide a novel speech communication system.

It is another object of the present invention to provide a speech communication system that uses fewer speech parameters than prior art communication systems.

It is a further object of the present invention to provide a speech communication system having reduced bandwidth requirements.

According to the present invention the three formant frequency parameters and the three amplitude parameters of the prior art formant vocoder are replaced by two new parameters. These two new parameters contain most of the phonetic information of the original six parameters and of the original speech wave. The two new parameters are the single equivalent formant frequency and its amplitude. According to the single equivalent formant concept, a sound can be represented by a single frequency signal which may or may not correspond to one of the formant frequencies of the sound. By using this concept, a speech communication system can be built that is less complicated than prior art systems and also capable of transmitting a speech signal within a smaller bandwidth than can prior art speech communication systems.

The above objects and other objects inherent in the present invention will become more apparent when read in conjunction with the following specification and drawings in which:

FIG. 1 is a block diagram showing the analyzer and synthesizer of the system of the present invention;

FIG. 2 is a graph showing waveforms at various locations in the synthesizer of FIG. 1; and

FIGS. 3 through 6 are block diagrams of components of the system of FIG. 1.

To understand the concept of single equivalent formant speech communication, the apparatus for analyzing the speech wave to extract the single equivalent formant frequency, and the apparatus for synthesizing the speech wave, it is necessary to describe the factors involved in single equivalent formant speech. It is postulated that when a human hears a multiformant sound, as in human speech, his attention focuses upon only one formant, called the dominant formant. The presence of any other formants, called recessive formants, serve only to shift the phonetic values slightly away from that of the dominant formant. It is further postulated that after a high frequency pre-emphasis of 6 db per octave is applied to a speech wave, the formant amplitude is the principal factor determining formant dominance and hence the single equivalent formant frequency. Since the formant of largest amplitude is the primary factor determining the period of the first major oscillation of a high frequency pre-emphasized speech wave, the period of the first major oscillation of the high frequency pre-emphasized speech wave at each shock of the vocal cords will approximately represent the single equivalent formant period.

Reproduction of human speech requires that the synthesized speech wave have a damped sinusoid characteristic. The repetition rate of the synthesized damped sine wave must be controlled by the pitch frequency and the ringing frequency of the synthesized damped sine wave must be the single equivalent formant frequency. Furthermore, the amplitude of the synthesized damped sine wave must be modulated in accordance with the single equivalent formant amplitude.

The block diagram of FIG. 1 shows the analyzer and synthesizer of the single equivalent formant communication system of the present invention. An electrical representation of a speech wave, such as produced by a standard telephone carbon microphone, is supplied to a single equivalent formant frequency detector **2**, a single equivalent formant amplitude detector **4**, and a pitch detector **6**. The output of pitch detector **6** is supplied to the detectors **2** and **4** and to a voicing detector **8**.

FIG. 3 is a block diagram of a preferred form of the single equivalent formant frequency detector **2** of FIG. 1. It comprises a circuit for measuring the period of the first major oscillation of a complex speech wave after each pitch pulse thereof and, hence, the inverse of the frequency of the single equivalent formant. The electrical signal representative of the input speech wave is supplied through an amplifier **30** and a high frequency pre-emphasis network **32** to the input of a high gain threshold circuit **34**, such as a Schmitt trigger. Network **32**, which includes a capacitor **36** and a resistor **38**, acts as a differentiator, emphasizing the high frequency components of the input speech wave. High gain threshold circuit **34** is set to produce an output signal only in response to one polarity of the differentiated input speech wave. The output signal of circuit **34** is supplied to one input terminal of a bistable switching circuit **40**. The output of pitch detector **6**, the construction of which is explained hereinafter, is supplied to a second input terminal of circuit **40**. Bistable switching circuit **40** is coupled by means of a pulse width-to-amplitude converter **42**, which may take the form of a ramp generator, to the input of a sample and hold circuit **44**. The output of sample and hold circuit **44** is a signal of slowly varying amplitude, the instantaneous amplitude of which is inversely proportional to the frequency of the single equivalent formant.

FIG. 4 is a block diagram of a preferred form of the single equivalent formant amplitude detector **4** of FIG. 1. The input speech wave is supplied to a peak detector **46** by means of a logarithmic amplifier **48**. A sample and hold circuit **50** is coupled to peak detector **46** and to a low pass filter **52**. Pitch pulses from pitch detector **6** gate the sample and hold circuit **50** to effect measurement of the log of the peak amplitude of the complex speech wave. Filter **52** removes the high frequency components from the output signal of circuit **50** thereby providing a slowly varying signal proportional to the log of the amplitude of the single euqivalent formant.

FIG. 5 is a block diagram of a preferred form of the pitch detector **6** of FIG. 1. The input speech wave is supplied via a high frequency pre-emphasis network **54** to a nonlinear or logarithmic amplifier **56**. The output of amplifier **56** is coupled to a peak detector **58** which has a long time constant and to a peak detector **60** which has a short time constant. Peak detector **60** is coupled by a voltage threshold conduction device **62**, such as a Zener diode, and an emitter follower network **64** to the output of peak detector **58** which is coupled to a differentiating and amplifying network **66**. Since the potential difference between the output signals of detectors **58** and **60** is small immediately after a pitch pulse, voltage threshold conduction device **62** does not conduct immediately after the occurrence of a pitch pulse. Hence those harmonic peaks in the input speech wave which occur immediately after a pitch pulse are not detected. When the potential difference between the output signals of detectors **58** and **60** is sufficient to initiate conduction of device **62**, the peak detector follows the discharge characteristics of short time constant detector **60**. Hence the peak detector detects pitch pulses even when there is a rapid decrease in the amplitude of the input speech wave. Accordingly, the output signal of network **66** comprises pulses the repetition rate of which is the same as the pitch rate of the input speech wave.

FIG. 6 is a block diagram of a preferred form of the voicing detector **8** of FIG. 1. Pitch pulses from the pitch detector **6** are supplied via a pulse width-to-amplitude converter **68**, such as a ramp generator, to the input of a first sample and hold circuit **70**. A differentiator network **72** couples sample and hold circuit **70** to a second sample and hold circuit **74**. Since the output signal of differentiator network **72** has amplitude peaks only when the repetition rate of the pitch pulses is irregular, the value of the output signal of circuit **74** is zero when the repetition rate of the pitch pulses is regular (voiced sounds) and other than zero when the repetition rate of the pitch pulses is irregular (unvoiced sounds).

The construction and operation of detectors **2**, **4**, **6** and **8** are described in more detail in my copending U.S. patent application Ser. No. 582,605, filed concurrently herewith.

The signals generated by the detectors **2**, **4**, **6**, and **8**, shown as waveforms *a*, *b*, *c*, and *d*, respectivley, of FIG. 2, are transmitted by conventional wire facilities or electromagnetic systems to a synthesizer network. For example, the detector signals can be transmitted by continuously varying the amplitude of a RF carrier signal in accordance with the amplitude of the detector signals. If a constant amplitude signal is being transmitted, an amplitude voltage reference level could be established at the receiver and the amplitude of the transmitted signal compared therewith.

The output waveform of the single equivalent formant frequency detector **2** is supplied to a frequency-controllable oscillator **10**, such as a multivibrator, the frequency of which is a function of the amplitude of the control signal from the detector **2**. If a transistorized multivibrator is used as the frequency-controllable oscillator **10**, the oscillation frequency can be controlled by changing the base voltage of both multivibrator transistors.

The output waveform of the single equivalent formant amplitude detector **4** is supplied to an amplitude modulator **12** of the synthesizer. Amplitude modulator **12** may be a diode switch modulator. The output waveform of pitch detector **6** is supplied to a frequency-controllable pitch oscillator **14**. Oscillator **14** generates a signal having a frequency which is a function of the amplitude of the control signal from detector **6**. Since, as previously described, the amplitude of the pitch detector signal is a function of the pitch frequency, the signal generated by oscillator **14** has a frequency equal to the pitch of the speech wave transmitted.

A linear modulator **16** is coupled to oscillator **14**. Modulator **16** is supplied by a first signal from voicing detector **8** and by a second signal from a noise generator **18**. Modulator **16** functions when unvoiced speech is being transmitted to assure that the randomly spaced pitch pulses characteristic of unvoiced speech are present in the synthesized speech wave.

Oscillator **14** is coupled to amplitude modulator **12** for regulating the amplitude of the synthesized speech wave and to oscillator **10** for assuring that the repetition rate of the synthesized speech wave is at the pitch rate. Modulator **12** is coupled through a peak detector **20** to an amplitude modulator **22**. Oscillator **10** is also coupled to amplitude modulator **22**. Amplitude modulator **22** and peak detector **20** assure that the synthesized speech wave is properly damped. Modulator **22** is coupled to a 6 db per octave de-emphasis network **24** for suppressing distortion components present in the modulator **22** output signal. De-emphasis network **24** has an output waveform representative of the synthesized speech wave.

The operation of the system of FIG. 1 will now be explained by reference to the waveforms of FIG. 2, of the accompanying drawing.

The frequency of the output waveform of the frequency-controllable pitch oscillator **14** is controlled during purely voiced sounds solely by the amplitude of the pitch detector signal supplied by detector **6**, modulator **16** being deactivated by the voicing detector signal supplied by detector **8** during voiced sounds. During purely unvoiced sounds the modulator **16** is activated by the voicing detector signal and the output of the noise generator **18** is supplied through the modulator **16** to the pitch oscillator **14** so as to "vary" the frequency of oscillator **14** in a random fashion. If the output waveform from the voicing detector **8** is a signal proportional to the amount of voiced and unvoiced energy in the sound wave to be synthesized, the signal being generated in the manner described in the aforementioned U.S. patent application, the output signal of modulator **16** will vary in accordance with the magnitude of the waveform from the voicing detector **8**. Thus, for voiced sounds the output of the pitch oscillator **14** is periodic and equal to the pitch of the speaker. During unvoiced sounds, the output of the pitch oscillator **14** is a series of randomly spaced pulses as a result of the signal generated by modulator **16** in response to the noise signal supplied thereto by noise generator **18**. This random excitation provides a perceptally satisfactory unvoiced sound. Voiced fricative sounds, usually identified with consonants such as f, v, s, z, etc., are produced with less noise introduced into the pitch oscillator **14** by the modulator **16** than for unvoiced sounds. Waveform *d* of FIG. 2 indicates that the voicing signal for the sound presently being transmitted is approximately zero and therefore the modulator **16** is deactivated. The pitch pulses from the oscillator **14** are therefore evenly spaced at a frequency controlled by the amplitude of wavform *c* of FIG. 2. The pitch pulses are shown as waveform *e* in FIG. 2.

The regenerated pitch pulses or the randomly spaced pulses of the oscillator **14** are amplitude modulated by the single equivalent formant amplitude signal in the modulator **12** to adjust the amplitude of the pitch pulses in accordance with the amplitude of the original speech wave. Waveform *f* of FIGURE 2 shows the output waveform of the modulator **12** when waveforms *b* and *e* are the outputs of the single equivalent formant amplitude detector **4** and oscillator **14**, respectively.

The output of the oscillator **14** also coheres the oscillator **10**. That is, the signal from the oscillator **14** stops and restarts the oscillator **10** every time a pitch pulse occurs. The variable frequency output of the oscillator **10** is shown as waveform *g* in FIG. 2 when the signal from the detector **2** has the variable amplitude output shown by waveform *a* of FIG. 2.

Waveform *f*, the output of the amplitude modulator **12**, is peak detected by the peak detector **20** to produce the damped characteristic of the natural speech wave. The time constant and hence the decay rate of the peak detector is made equal to the required decay of the damped sinusoid of the wave to be synthesized. Waveform *h* of FIG. 2 shows the exponentially decaying output of the peak detector **20**. The cohered variable frequency signal from oscillator **10**, waveform *g* of FIG. 2, is then modulated with the peak detector output waveform, waveform *h* of FIG. 2 to produce the damped square wave of waveform *i* of FIG. 2.

Waveform *i* contains many of the required characteristics of the human speech wave. It has a damped characteristic, a repetition rate controlled by the pitch frequency, and a ringing frequency controlled by the single equivalent formant frequency.

However waveform *i* contains a large third-harmonic distortion component. The amplitude of the formants of natural human speech fall off at a rate of approximately 10 db per octave while the synthesizer apparatus just described produces a flat response. High frequency de-emphasis network **24** is coupled to the amplitude modulator output signal for adjusting the formant amplitude of the synthesized waveform *i* to match those of human speech. Since an octave is the interval between two sounds having a frequency ratio of two to one, the third harmonic of the speech wave is spaced from the fundamental of the speech wave by one and one-half octaves. Network **24** reduces the third harmonic by 6 db per octave and thereby reduces the third harmonic by an additional 9 db. This results in the damped triangular wave shown as waveform *j* in FIG. 2. In waveform *j* the difference between the first and third harmonics is now approximately 19 db. This reduction in third harmonic content improves the quality of the synthesized speech wave.

The use of the single equivalent formant concept results in several major advantages over prior art communication systems. First, it reduces the number of speech parameters that must be extracted and transmitted. This feature substantially reduces the size of the ultimate speech compression circuitry and requires a transmission channel bandwidth of only 80 to 120 cycles per second.

Second, it simplifies the extraction process itself. To date, extracting the location of the three individual formants has been a difficult and complicated task. However, extracting the single equivalent formant has been shown to be simple and economical. Furthermore, the extraction process does not use filter banks or inductors; for this reason, it is a process that lends itself to microminiaturization techniques.

While the invention has been described with reference to certain preferred embodiments thereof, it will be apparent that various modifications and other embodiments thereof will occur to those skilled in the art within the scope of the invention. Accordingly, I desire the scope of my invention to be limited only by the appended claims.

I claim:

1. A speech synthesizer responsive to no more than four input signals derived from a speech wave, said signals including a first signal representative of the pitch pulses of said speech wave, a second signal representative of the period of the first major oscillation of said speech wave occurring after each of said pitch pulses, and a third signal representative of the maximum amplitude of each of said first major oscillations, said synthesizer comprising first means supplied with and responsive to said first signal for producing a pitch signal, second means coupled to said first means and supplied with and responsive to said third signal for amplitude modulating said pitch signal, third means coupled to said first means and supplied with and responsive to said second signal for producing a frequency modulated signal the periodicity of which equals the pitch rate of said pitch signal, and fourth means coupled to said second and third means for amplitude modulating said frequency modulated signal so as to produce a synthesized speech wave.

2. The synthesizer of claim **1** wherein said first means comprises a first signal controlled oscillator; said second means comprises a first amplitude modulator; said third means comprises a second signal controlled oscillator; and said fourth means comprises a second amplitude modulator, a peak detector and means coupling the output of said peak detector to an input of said second amplitude modulator.

3. The synthesizer of claim **2** further comprising fifth means coupled to said first oscillator and supplied with and responsive to the fourth of said input signals for varying randomly the pitch rate of said pitch signal.

4. The synthesizer of claim **3** wherein said fifth means comprises a linear modulator, a noise generator, and means for coupling the output of said noise generator to an input of said linear modulator.

5. The synthesizer of claim **4** including in addition a high frequency de-emphasis network coupled to the output of said second amplitude modulator.

6. The synthesizer of claim 5 wherein an input of said second amplitude modulator is connected directly to the output of said second oscillator, and the input of said peak detector is connected directly to the output of said first amplitude modulator.

References Cited

UNITED STATES PATENTS

2,635,146	5/1953	Steinberg.
2,824,906	2/1958	Miller.
3,087,989	5/1963	Nagata.
3,190,963	6/1965	David.
3,335,225	8/1967	Campanella.
3,387,090	6/1968	Bridges.

KATHLEEN H. CLAFFY, Primary Examiner

CHARLES JIRAUCH, Assistant Examiner

U.S. Cl. X.R.

179—15.55; 324—77