Jan. 20, 1970 L. R. FOCHT ET AL 3,491,205

PLURAL FORMANT SPEECH SYNTHESIZER

Filed Sept. 29, 1966 3 Sheets-Sheet 1

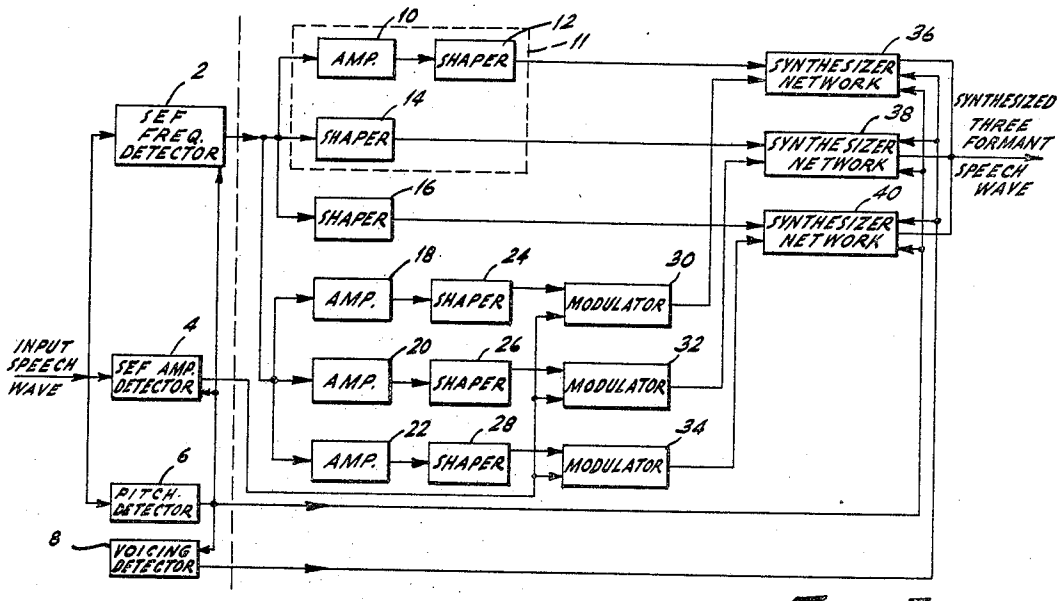


FIG. 3.

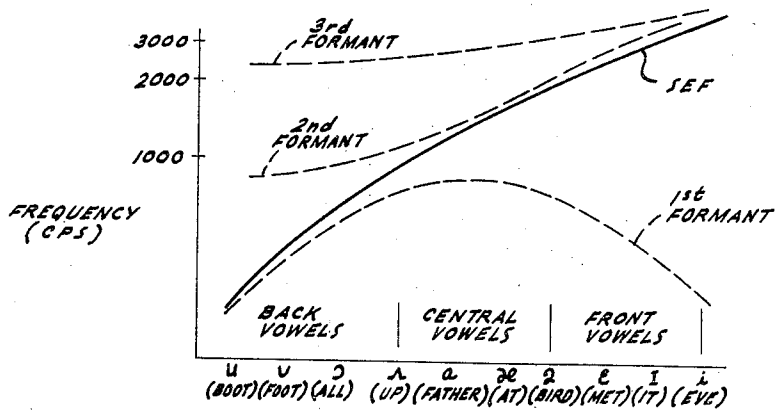


FIG. 1.

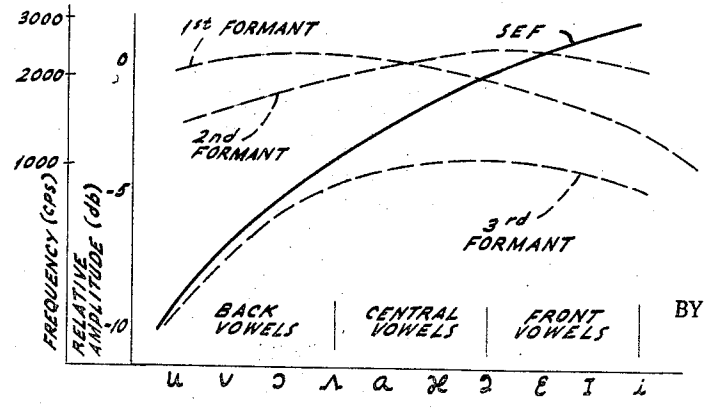


FIG. 2.

INVENTORS
LOUIS R. FOCHT
CHARLES F. TEACHER
BY Leonard Zalman
ATTORNEY

PLURAL FORMANT SPEECH SYNTHESIZER

Filed Sept. 29, 1966 3 Sheets-Sheet 2

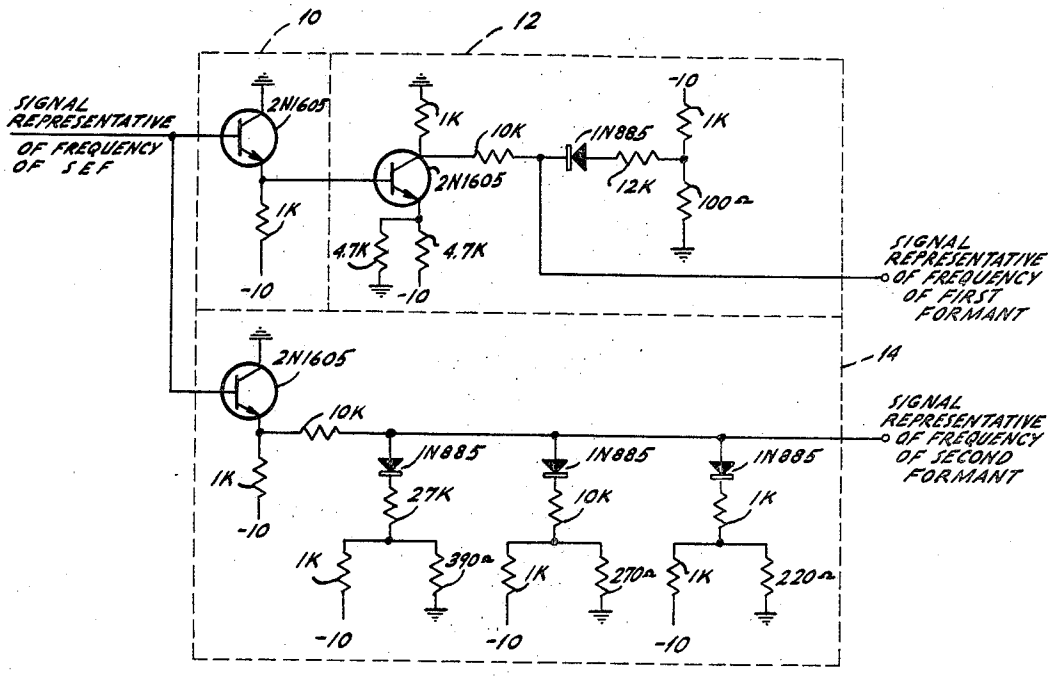


FIG. 4.

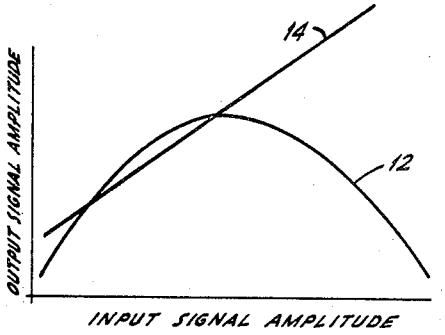


FIG. 4a.

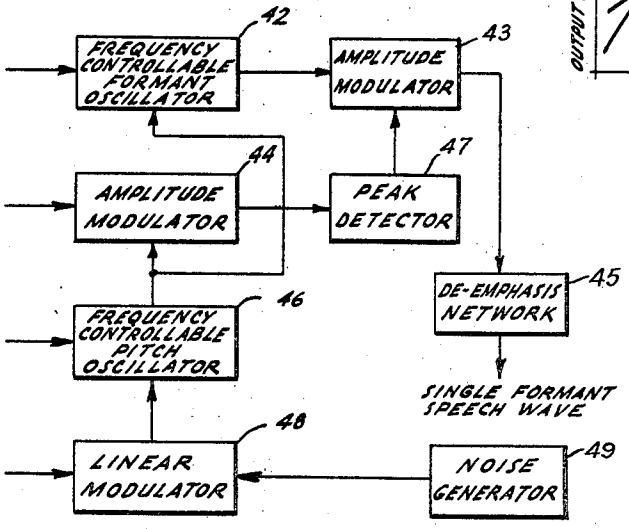


FIG. 5.

INVENTORS
LOUIS R. FOCHT
CHARLES F. TEACHER
BY
Leonard Zalman
ATTORNEY

Jan. 20, 1970 L. R. FOCHT ET AL 3,491,205

PLURAL FORMANT SPEECH SYNTHESIZER

Filed Sept. 29, 1966 3 Sheets-Sheet 3

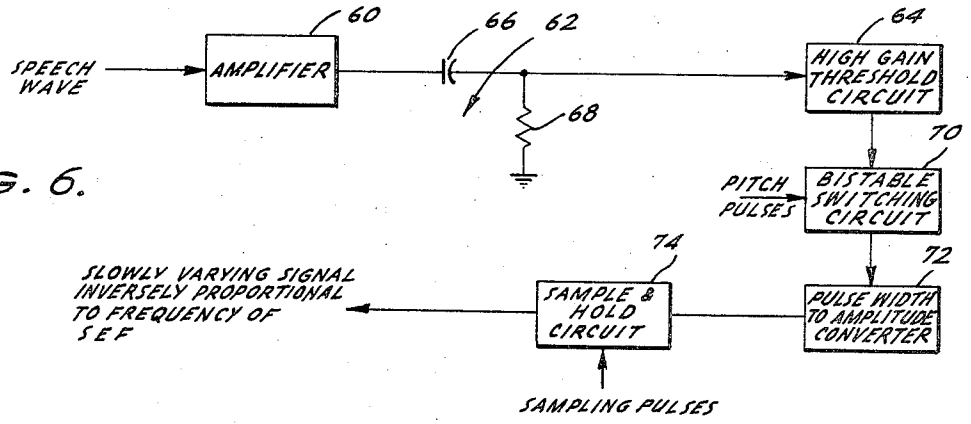


FIG. 6.

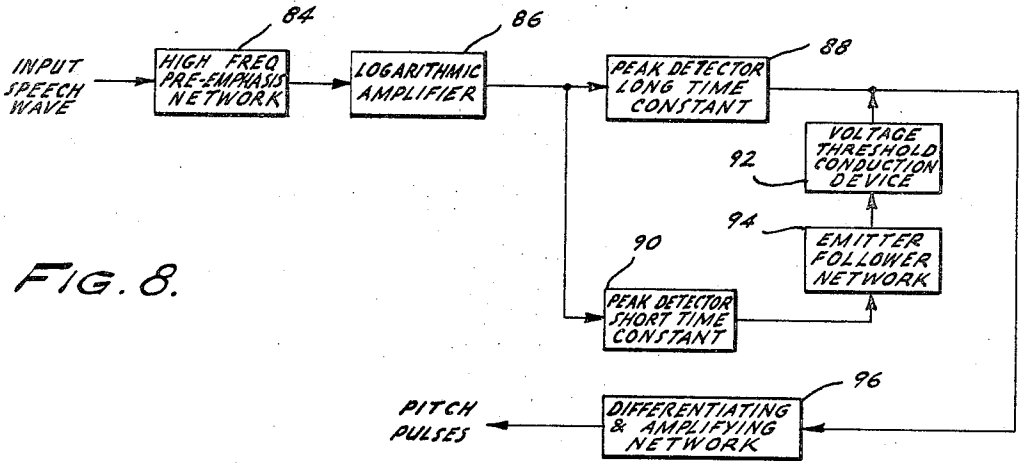


FIG. 8.

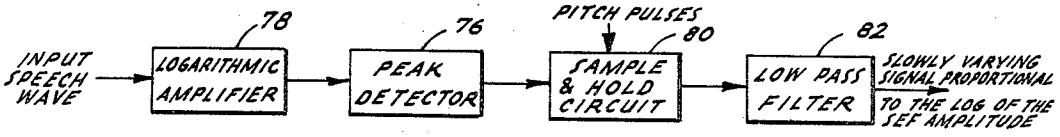


FIG. 7.

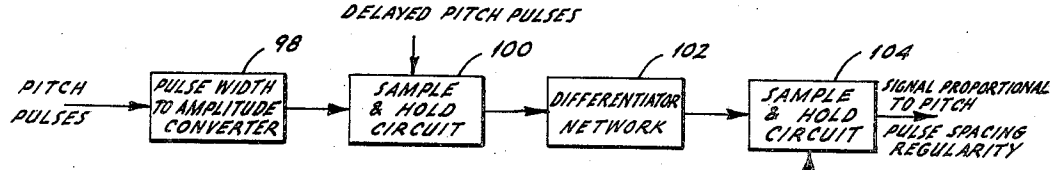


FIG. 9.

INVENTORS
LOUIS R. FOCHT
CHARLES F. TEACHER
BY
Leonard Zalman
ATTORNEY

United States Patent Office

3,491,205

Patented Jan. 20, 1970

3,491,205

PLURAL FORMANT SPEECH SYNTHESIZER

Louis R. Focht, Huntington Valley, and Charles F. Teacher, Philadelphia, Pa., assignors to Philco-Ford Corporation, a corporation of Delaware

Filed Sept. 29, 1966, Ser. No. 582,898

Int. Cl. H04m 1/19

U.S. Cl. 179—1 4 Claims

ABSTRACT OF THE DISCLOSURE

A speech synthesizer which generates three-formant speech from a first control signal representative of the period of the first major oscillation of a speech wave following each pitch pulse of the speech wave, a second control signal representative of the maximum amplitude of each such oscillation, a pitch signal, and a voicing signal. The synthesizer includes first and second groups of signal shaping networks, each group comprising three signal shaping networks. The first control signal is supplied to the first group to produce three signals each of which has an amplitude representative of the frequency of a different formant of a speech wave, and to the second group to produce three signals each of which has an amplitude proportional to the amplitude of a different formant of the speech wave. The synthesizer also includes three modulators the input of each of which is supplied with the output signal of a different one of the shaping networks forming the second group and with the second control signal. Each modulator is responsive to the signals supplied thereto to produce a signal having an amplitude representative of the amplitude of a different formant of the speech wave. The synthesizer also includes formant synthesizers each supplied with the output of a different one of said modulators, the output signal of a different one of the shaping networks of the first group, and the pitch and voicing signals. Each synthesizer produces, in response to the four input signals, a different one of the three formant signals, and those three formant signals are added to produce a three-formant speech wave.

Speech waves are highly redundant and considerable saving in bandwidth can be realized by proper processing of a speech representative 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.

A common type of speech bandwidth compression system is the formant tracking vocoder. The formant vocoder type of speech bandwidth compression system is based on the transmission of signals representative of the formants or vocal tract resonances of the speech wave. The conventional formant tracking vocoder 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. Recently it has been discovered that the three formant amplitude parameters and the three formant frequency parameters of the prior art formant vocoder can be replaced by two new parameters. The two new parameters are the single equivalent formant frequency and its amplitude. These two new parameters contain most of the phonetic information of the original six parameters and of the original speech wave. According to the single equivalent formant concept, a sound can be represented at any instant 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 at a smaller bandwidth than prior art speech communication systems. The concept of the single equivalent formant and a speech communication system that utilizes the concept are described in detail in co-pending U.S. patent application Ser. Nos. 582,605, filed Sept. 28, 1966 and 582,573, also filed Sept. 28, 1966 by L. Focht.

The synthesis at the receiving location of a single equivalent formant speech wave described in aforementioned U.S. patent application Ser. No. 582,573, is simple in implementation; however, the speech reconstructed in this manner has a nasal quality that may be distracting to the uninitiated listener. Therefore, it is sometimes desirable to synthesize from the single equivalent formant speech information the plural formant speech wave that the listener is accustomed to hearing.

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 in which the synthesized speech wave is the type that the listener is accustomed to hearing.

According to the present invention, the single equivalent formant signal extracted at the analyzer of a single equivalent formant communication system is transmitted to the synthesizer of the communication system at the receiving location and there converted to plural formant speech. That is, plural formant speech is reconstructed at the synthesizer from the single equivalent formant representative signals transmitted from the analyzer.

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 graph showing the frequencies of the first three formants and the corresponding frequency of the single equivalent formant for each of ten vowel sounds;

FIG. 2 is a graph showing the relative amplitudes of the first three formants and the frequency of the single equivalent formant for each of the ten vowel sounds of FIG. 1;

FIG. 3 is a block diagram of a communication system according to the present invention;

FIG. 4 is a schematic circuit diagram of a signal shaper portion of the system of FIG. 3;

FIG. 4*a* is a plot of the signal input-output characteristics for the signal shaper circuits employed in the system of FIG. 3;

FIG. 5 is a block diagram of a component of the system of FIG. 3, and

FIGS. 6 to 9 are block diagrams illustrative of components of the block diagram of FIG. 3.

In order to understand the concept of the generation of three formant speech from single equivalent formant speech, it is necessary to know the relationship between the frequency of the single equivalent formant and the frequencies and amplitudes of the first three formants of a particular sound. Referring to FIGURE 1, the frequency of the single equivalent formant and the frequencies of the first three formants for ten vowel sounds are graphically shown. The vowel sounds are grouped as back, central, and front. The back, central, and front vowels are articulated in the back, central, and front portions of the vocal tract, respectively.

FIGURE 1 shows that for each value of the single equivalent formant frequency, there is a corresponding value for each of the first three formants. The frequency

of the single equivalent formant is lowest for the vowel sound U (boot) and progressively higher in the order of the vowell sounds shown in FIG. 1. The frequency of the first formant for the ten vowel sounds also is low for the vowel sound U (boot) and increases for the back vowels until a maximum value is reached in the region of the central vowels and then decreases for the front vowels. The frequencies of the second and third formants again are lowest for the vowel sound U (boot) and progressively higher in the order of the vowels shown in FIG. 1. The rate of increase is less for the central and front vowels (i.e. higher single equivalent formant frequencies) than it is for the back vowels. It can be shown that a similar relationship between the frequency of the single equivalent formant and the frequencies of the first three formants holds for other speech sounds. Since in the transmission of speech by means of single equivalent formant parameters the frequency of the single equivalent formant is represented by the amplitude of an electrical signal, it is possible to develop from this latter signal signals having amplitudes proportional to the frequencies of the first three formants of the speech wave. All that is required are three amplifiers each having a gain at any input signal amplitude level which is proportional to the ratio of the frequency of the single equivalent formant to the frequency of a selected one of the first three formants at the frequency of the single equivalent formant represented by that amplitude level. The resultant signals are signals having amplitudes at any instant which are representative of the frequencies of the first three formants of the sound being transmitted at that instant. These amplitude varying signals may be employed to control the frequency of three oscillators which regenerate signals having instantaneous frequencies equal to the frequencies of the first three formants of the sound to be represented at that instant.

Although, as previously described, signals at the frequencies of the first three formants of a sound can be produced when the single equivalent formant frequency is known, additional information must be conveyed by the single equivalent formant frequency signal if three formant speech is to be synthesized. The additional information that must be conveyed is the amplitude of each of the first three formants of a sound. Since the relative amplitude of the first three formants of a sound is the principal factor determining the single equivalent formant frequency of the sound (attention is directed to aforementioned U.S. patent application Ser. No. 582,605), knowledge of the single equivalent formant frequency of a sound conveys sufficient information for determining the amplitudes of the first three formants of the sound relative to the amplitude of the single equivalent formant.

The manner in which the relative amplitudes of the first three formants of a sound can be determined by knowledge of the single equivalent formant frequency of the sound will be explained in conjunction with FIGURE 2. FIGURE 2 graphically shows the relative formant amplitude in db after a 9 db per octave high frequency emphasis for the ten vowels sounds shown in FIGURE 1. The single equivalent formant frequency for the ten vowels is also superimposed on the graph of FIGURE 2.

FIGURE 2 shows that for each value of the single equivalent formant frequency there is a corresponding value for the ratio of the amplitude of each of the first three formants to the amplitude of the single equivalent formant. The graph of the amplitude of each of the three formants for the ten vowels has a slight foldover characteristic for increasing values of the single equivalent formant frequency. By foldover characteristic it is meant that the magnitude of a formant increases with an increasing single equivalent formant frequency until a maximum value of the formant amplitude is reached; beyond the maximum value the formant amplitude decreases in magnitude even though the frequency of the single equivalent formant continues to increase. Thus, by using amplifiers in the synthesizer having gains controlled by the amplitude of the signal representative of the frequency of the single equivalent formant; the amplitudes of each of the three formants of three formant speech can be derived from the signal representative of the amplitude of the single equivalent formant for each frequency value of the single equivalent formant.

The block diagram of FIG. 3 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 (not shown) 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. 6 is a block diagram of a preferred form of the single equivalent formant frequency detector **2** of FIG. 3. It comprises a circuit for measuring the period of the first major oscillation of the 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 **60** and a high frequency pre-emphasis network **62** to the input of a high gain threshold circuit **64**, such as a Schmitt trigger. Network **62**, which includes a capacitor **66** and a resistor **68**, acts as a differentiator, emphasizing the high frequency components of the input speech wave. High gain threshold circuit **64** is set to produce an output only in response to one polarity of the differentiated input speech wave. The output signal of circuit **64** is supplied to one input terminal of a bistable switching circuit **70**. The output of pitch detector **6**, whose construction is explained hereinafter, is supplied to a second input terminal of circuit **70**. Bistable switching circuit **70** is coupled by means of a pulse width-to-amplitude converter **72**, which may take the form of a ramp generator, to the input of a sample and hold circuit **74**. The output of the sample and hold circuit **74** is a signal of slowly varying amplitude, the instantaneous amplitude of which is inversely proportional to the frequency of the single equivalent formant.

FIG. 7 is a block diagram of a preferred form of the single equivalent formant amplitude detector **4** of FIG. 3. The input speech waveform is supplied to a peak detector **76** by means of a logarithmic amplifier **78**. A sample and hold circuit **80** is coupled to peak detector **76** and to a low pass filter **82**. Pitch pulses from the pitch detector **6** gate the sample and hold circuit **80** to effect measurement of the logarithm of the peak amplitude of the complex speech wave. Filter **82** removes the high frequency components from the output signal of circuit **50**, thereby providing a slowly varying signal proportional to the logarithm of the amplitude of the single equivalent formant.

FIG. 8 is a block diagram of a preferred form of the pitch detector **6** of FIG. 3. The input speech wave is supplied via a high frequency pre-emphasis network **84** to a non-linear or logarithmic amplifier **86**. The output of amplifier **86** is coupled to a peak detector **88** which has a long time constant and to a peak detector **90** which has a short time constant. Peak detector **90** is coupled by a voltage threshold conduction device **92**, such as a Zener diode, and an emitter follower network **94** to the output of peak detector **88**, which is coupled to a differentiating and amplifying network **96**. Since the potential difference between the output signals of detectors **88** and **90** is small immediately after the occurrence of a pitch pulse, voltage threshold conduction device **92** does not conduct immediately after such occurrence. 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 **88** and **90** is sufficient to initiate conduction of

device 92, the peak detector follows the discharge characteristics of the short time constant detector 90. 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 96 comprises pulses the repetition rate of which is the same as the pitch rate of the input speech wave.

FIG. 9 is a block diagram of a preferred form of the voicing detector 8 of FIG. 3. Pitch pulses from the pitch detector 6 are supplied via a pulse width-to-amplitude converter 98, such as a ramp generator, to the input of a first sample and hold circuit 100. A differentiator network 102 couples sample and hold circuit 100 to a second sample and hold circuit 104. Since the output signal of differentiator network 102 has amplitude peaks only when the repetition rate of the pitch pulses is irregular, the value of the output signal of circuit 102 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 the aforementioned copending U.S. patent application Ser. No. 582,605

The signals generated by the detectors 2, 4, 6 and 8 are transmitted in any convenient manner, for example 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 an RF carrier signal in accordance with the amplitude of the detector signals. If the signals from the detectors are transmitted directly rather than as a modulation of a carrier wave, an amplitude voltage reference level could be established at the receiver and the amplitude of the transmitted signal compared therewith.

The signal from the single equivalent formant frequency detector 2 is supplied through an amplifier circuit 10 and a shaper circuit 12 to the input of a first formant synthesizer network 36, through a shaper circuit 14 to a second formant synthesizer network 38, and through a shaper circuit 16 to a third formant synthesizer network 40. Synthesizer networks 36, 38 and 40 have their output terminals coupled together.

FIG. 4 is a typical schematic circuit diagram of section 11 of the block diagram of FIG. 3. The input-output response characteristics of amplifiers 12 and 14 are shown in FIG. 4*a*. A circuit similar to the circuit diagram of shaper 14 could be used as the circuit for shaper 16. The values of the load resistors of the circuit of shaper 16 would be chosen to produce the desired non-linear input-output signal relationship required of shaper 16.

The signal from the detector 2 is also supplied to amplifier networks 18, 20, and 22. Amplifier networks 18, 20, and 22 are coupled by means of shaper networks 24, 26, and 28, respectively, to modulators 30, 32, and 34, respectively. The circuitry of amplifier networks 18, 20, and 22 may be similar to the circuitry of amplifier network 10 and the circuitry of shaper networks 24, 26, and 28 may be similar to the circuitry of shaper network 12.

The signal from the single equivalent formant amplitude detector 4 is also supplied to amplitude modulators 30, 32, and 34. Modulators 30, 32, and 34 are coupled to snythesizer networks 36, 38, and 40, respectively.

Each of synthesizer networks 36, 38 and 40 can have the structure shown in block diagram in FIG. 5. For synthesizer network 36, shaper 12 is connected to the input of an oscillator 42 the output signal of which is supplied via an amplitude modulator 43 to a deemphasis network 45. The output signal of modulator 30 is supplied to one input of an amplitude modulator 44 the output signal of which is supplied to modulator 43 via a peak detector 47. The pitch and voicing signals from the detectors 6 and 8 are supplied respectively to oscillator 46 and linear modulator 48. Modulator 48 also receives a signal from noise generator 49. The output signal of linear modulator 48 is supplied to frequency-controllable pitch oscillator 46, and the output signal of oscillator 46 is supplied to frequency-controllable formant oscillator 42 and to amplitude modulator 44. Synthesizer networks 38 and 40 are similarly connected, with shapers 14 and 16 respectively substituted for shaper 12 and modulators 32 and 34 respectively substituted for modulator 30.

The circuit of FIG. 3 functions in the following manner. Amplifier circuit 10 and the shaper circuits 12, 14, and 16 modify the input signal which has an amplitude representative of the frequency of the single equivalent formant to produce three control waveforms at the inputs to the networks 36, 38, and 40, respectively, having instantaneous amplitudes corresponding to the frequencies of the first, second, and third formants (FIG. 1). Amplifier circuits 18, 20, and 22 function in conjunction with shaper circuits 24, 26, and 28 to modify the input signal from detector 2 to produce waveforms at the inputs to the modulators 30, 32, and 34, respectively, proportional to the amplitudes of the first, second, and third formats (FIG. 2). The waveforms proportional to the relative values of the amplitudes of the first, second, and third formants modulate the signal from the single equivalent formant amplitude detector 4 to produce control signals at the inputs to the networks 36, 38, and 40, respectively, representative of the absolute amplitudes of the first, second, and third formants.

The inputs to synthesizer networks 36, 38, and 40 contain all the phonetic information needed to produce the first, second, and third formants of human speech, respectively. Referring specifically to network 36, oscillator 46 (FIG. 5) produces a pitch signal which is supplied to amplitude modulator 44 and to frequency controllable formant oscillator 42. Modulator 44 amplitude modulates the pitch signal in response to the output signal of modulator 30. The pitch signal supplied to oscillator 42 controls the repetition rate of the frequency-modulated signal produced by oscillator 42. The frequency of the latter signal between successive pitch signals is determined by the amplitude of the output signal of shaper 12. The amplitude-modulated pitch signal produced by modulator 44 undergoes peak-detection by detector 47, and the signal produced by detector 47 in response to the amplitude-modulate the frequency-modulated output signal supplied thereto by oscillator 42. The resultant amplitude-modulated, frequency-varying signal is smoothed by de-emphasis network 45 to produce a signal representative of the first formant of the speech wave being synthesized.

In a similar manner networks 38 and 40 produce signals respectively representative of the second and third formants of the speech wave being synthesized. The signals representative of the first, second and third formants are summed to obtain a synthesized three-formant speech wave. The operation of the network of FIG. 5 is described in detail in the aforementioned copending U.S. patent application Ser. No. 582,573.

Although the system for generating three formant speech from single equivalent formant speech has been described as using particular amplifying and shaping circuits, it is obvious that other circuits that will produce the same values of amplitude and frequency control signal for a particular single equivalent formant frequency can be used.

The system of the present invention provides a major advantage over prior art communication systems because it permits plural formant speech to be synthesized from a transmitted single equivalent formant signal. This allows the listener to hear the plural formant speech that he is accustomed to hearing while taking advantage of the decreased data rate and bandwidth of the transmitter characteristic of single equivalent formant speech transmission.

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 we desire the scope of our invention to be limited only by the appended claims.

We claim:

1. A speech synthesizer for synthesizing a multi-formant speech wave in response to a first input signal representative at any given time of the period of the first major oscillation of a speech wave occurring after that pitch pulse of said wave which immediately precedes said given time, a second input signal representative at said given time of the maximum amplitude of said first major oscillation, a pitch signal and a voicing signal, said synthesizer comprising a first group of signal shaping networks supplied with and responsive to said first input signal to produce a first plurality of signals each of which is representative of the frequency of a different formant of said speech wave; a second group of signal shaping networks supplied with and responsive to both said first and second input signals to produce a second plurality of signals each of which is representative of the amplitude of a different formant of said speech wave; plurality of formant synthesizer networks supplied with and responsive to said pitch signal, and having their outputs coupled together; and means for supplying those signals of said first and second plurality of signals that are representative of the same formant of said speech wave to only one of said plurality of synthesizer networks.

2. The synthesizer according to claim 1 wherein each of said second group of signal shaping networks includes an amplifier, a signal shaper supplied with and responsive to the output of the amplifier, and a modulator supplied with and responsive to the output of the signal shaper; said synthesizer also comprising means for supplying said first input signal to each of said amplifiers, and means for supplying said second input signal to each of said modulators.

3. The synthesizer according to claim 2 wherein each of said formant synthesizer networks includes a first signal controlled oscillator supplied with and responsive to a signal representative of the frequency of one formant, an amplitude modulator supplied with and responsive to a signal representative of the amplitude of said one formant, a second signal controlled oscillator supplied with and responsive to said pitch signal, and means coupling the output of said second signal controlled oscillator to an input of said first signal controlled oscillator and an input of said amplitude modulator.

4. The synthesizer according to claim 3 wherein each of said formant synthesizer networks further includes a noise generator, a linear modulator supplied with and responsive to both said voicing signal and the output of said noise generator, means coupling the output of said linear modulator to an input of said second signal controlled oscillator, a peak detector supplied with the output of said first amplitude modulator, and a second amplitude modulator supplied with both the output of said first signal controlled oscillator and the output of said peak detector.

References Cited

UNITED STATES PATENTS

2,819,341	1/1958	Barney.
2,824,906	2/1958	Miller.

KATHLEEN H. CLAFFY, Primary Examiner

CHARLES JIRAUCH, Assistant Examiner

U.S. Cl. X.R.

179—15.55; 324—77