

[72] Inventor Louis R. Focht
Huntington Valley, Pa.
[21] Appl. No. 862,101
[22] Filed July 7, 1969
[45] Patented Sept. 7, 1971
[73] Assignee Philco-Ford Corporation

[54] TIME-DOMAIN PITCH DETECTOR AND CIRCUITS
FOR EXTRACTING A SIGNAL REPRESENTATIVE
OF PITCH-PULSE SPACING REGULARITY IN A
SPEECH WAVE
9 Claims, 10 Drawing Figs.

[52] U.S. Cl. 179/1 SA
[51] Int. Cl. G101 1/00
[50] Field of Search 179/1 SA, 1
N, 1 VC, 1 SB, 1 VS; 328/132, 140

[56] References Cited

UNITED STATES PATENTS

3,463,885 8/1969 Upton 179/1

Primary Examiner—Kathleen H. Claffy
Assistant Examiner—Horst F. Brauner
Attorney—Robert D. Sanborn

ABSTRACT: A time-domain detector circuit for producing a signal representative of the pitch pulses of a speech wave. The circuit comprises two peak detectors having two different time constants to which the input speech wave is supplied. The output of the shorter-time-constant detector is supplied to an isolating stage (e.g., an emitter follower network). The output of the longer-time-constant peak detector is connected, via a zener diode, to the output terminal of the emitter-follower network. A differentiator circuit is connected to the output terminal of the longer-time-constant detector. When the potential difference across the zener diode rises to a sufficiently high value, the zener diode conducts, enabling the emitter follower network to load the longer-time-constant peak detector. The shorter-time-constant peak detector detects the pitch pulses and supplies them to the output terminal of the longer-time-constant detector. When the amplitude level of the speech wave is relatively constant, the zener diode isolates the output of the shorter-time-constant detector from the output terminal of the other detector and only the long-time-constant detector supplies an output signal thereto.

One circuit includes an input speech wave, a clipper, ramp generator, peak detector and low-pass filter connected in series. Another circuit includes pitch pulses, a pulse width-to-amplitude converter, a sample-and-hold circuit, a differentiator and a second sample-and-hold circuit connected in series with delayed pitch pulses timing each sample and hold circuit.

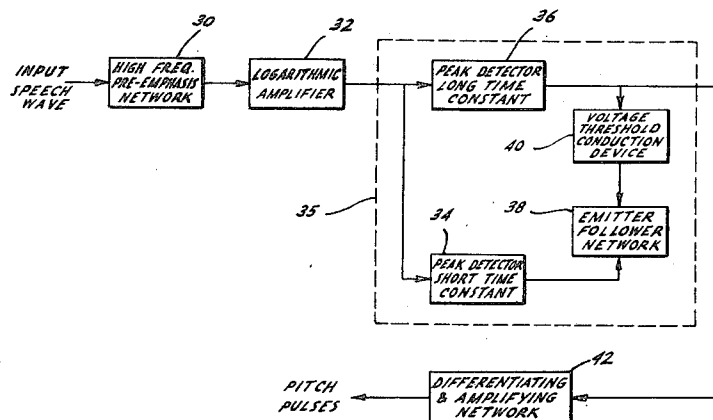
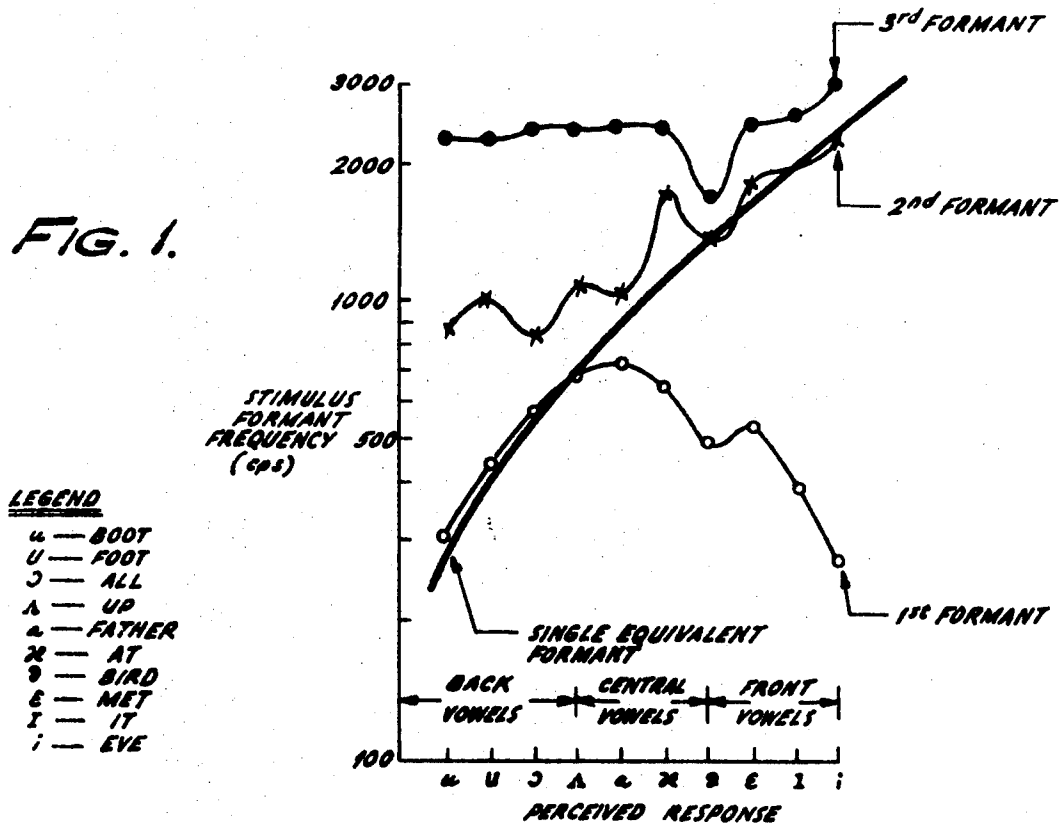


FIG. 1.



RELATIVE FORMANT AMPLITUDE IN dB AFTER 9dB PER OCTAVE HIGH FREQUENCY EMPHASIS

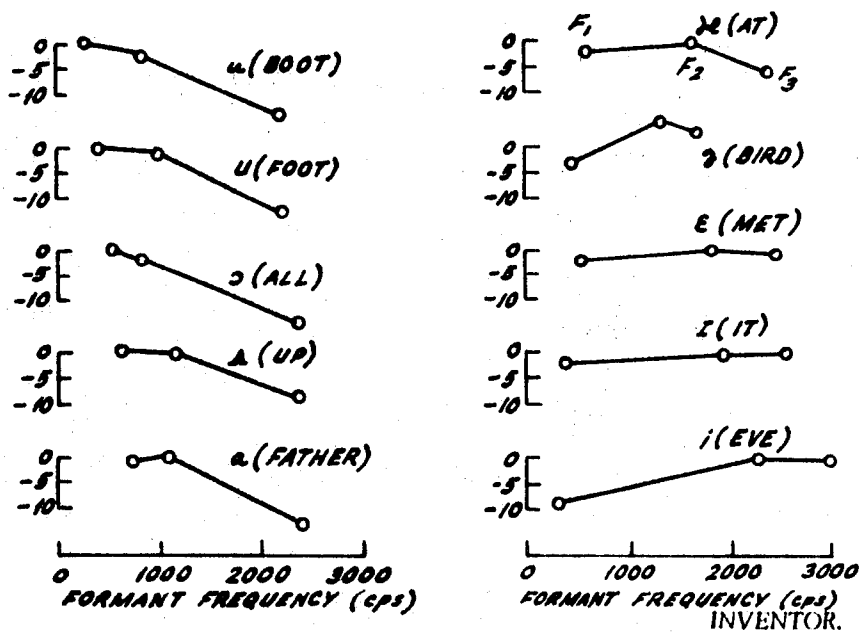


FIG. 2.

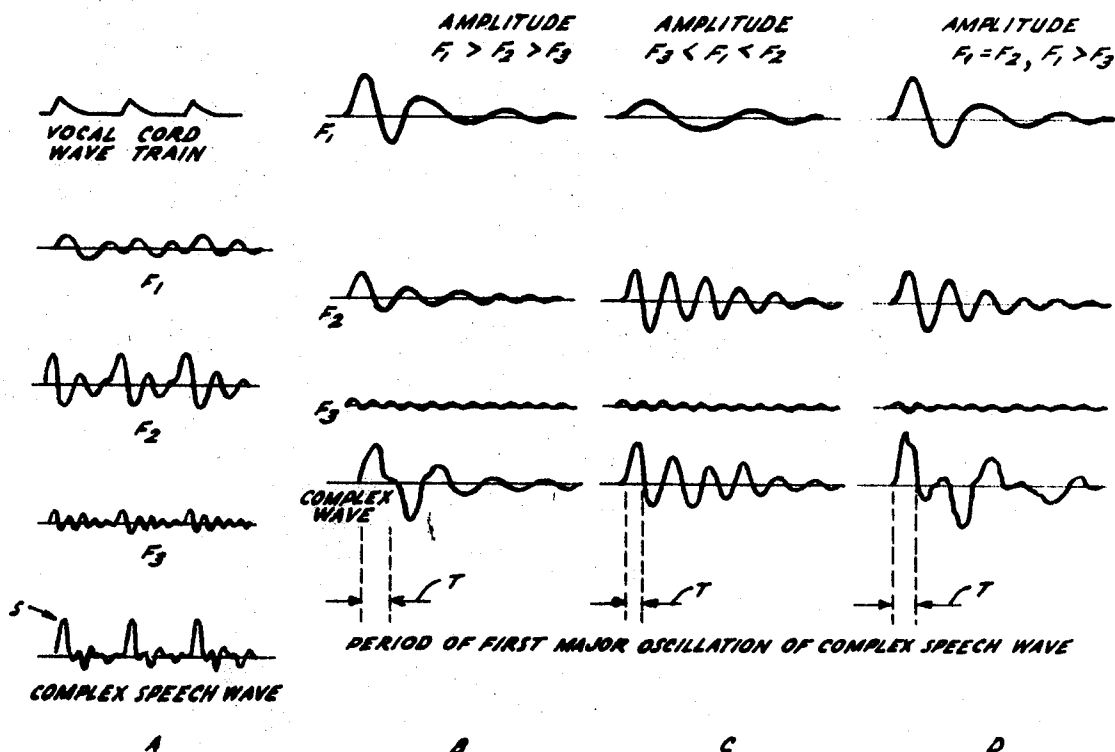


FIG. 3.

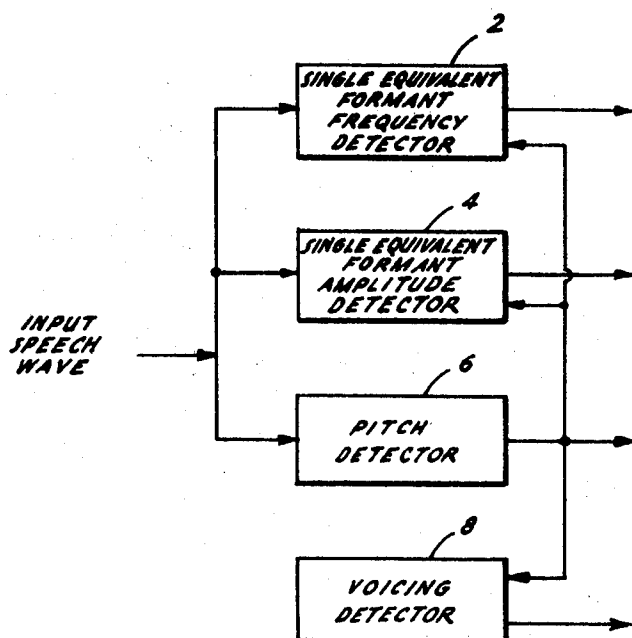


FIG. 4.

INVENTOR.
LOUIS R. FOCHT

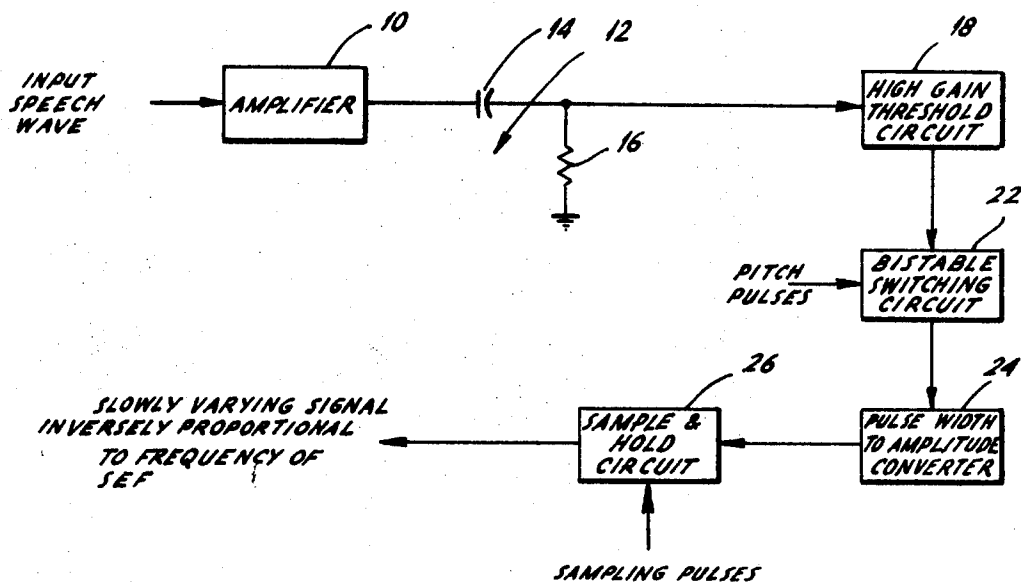


FIG. 5.

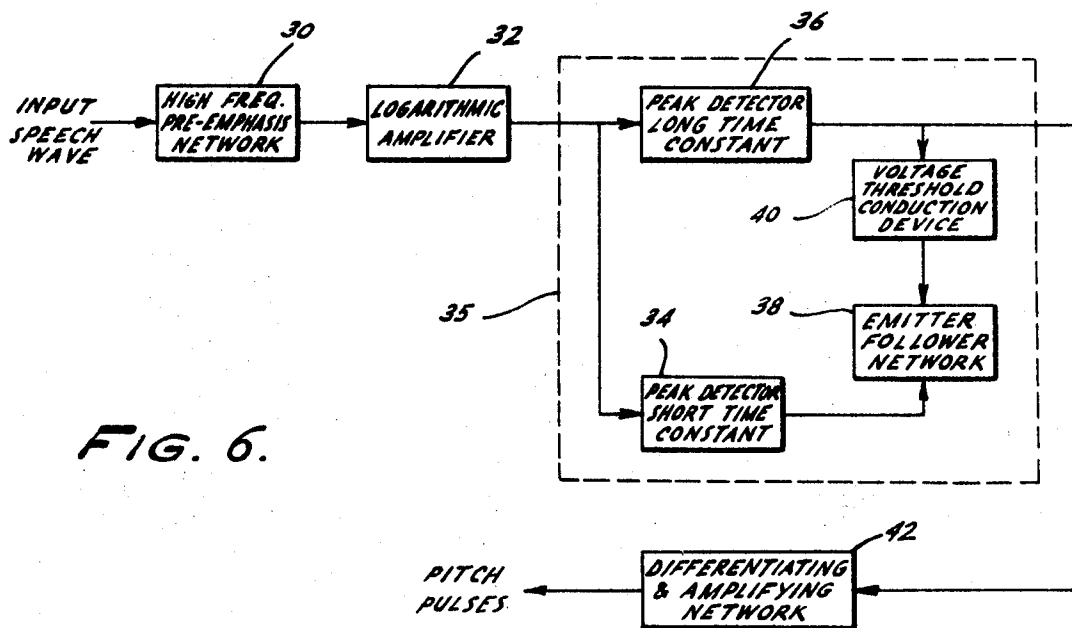


FIG. 6.

INVENTOR.
LOUIS R. FOCHT

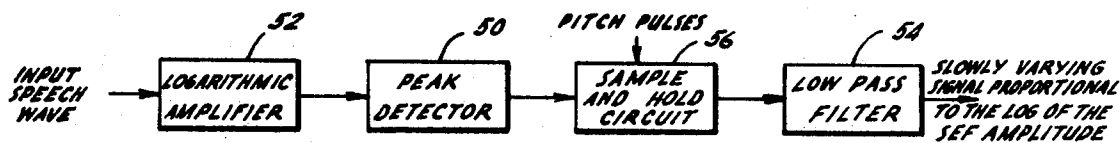


FIG. 7.

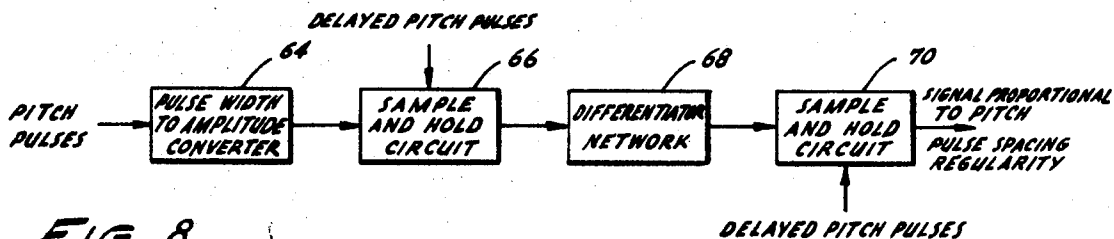


FIG. 8.

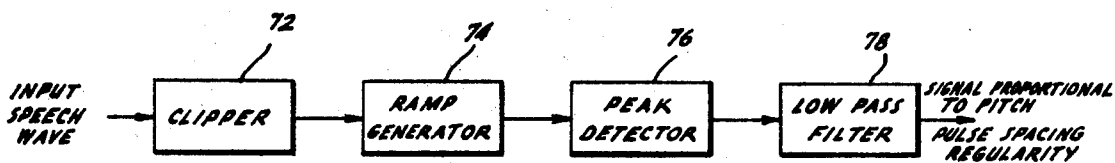


FIG. 9.

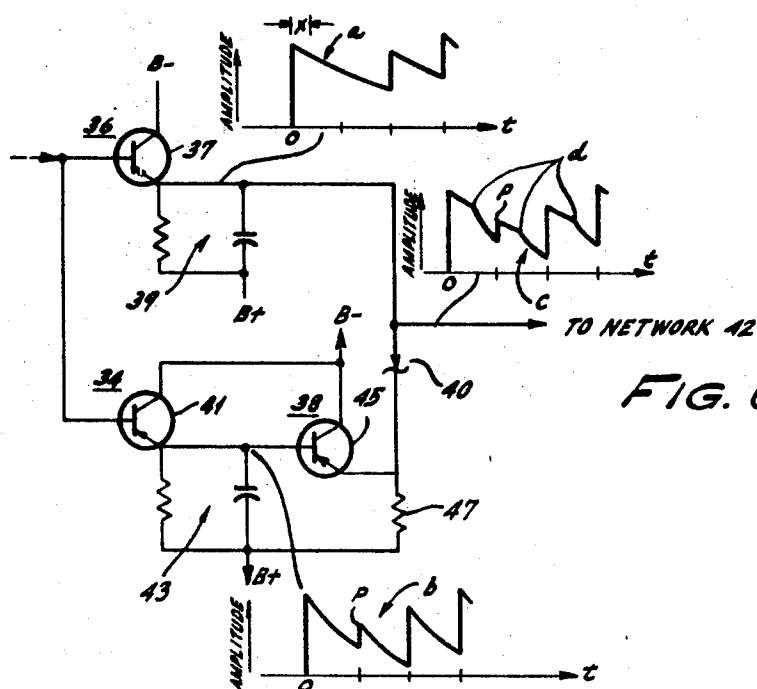


FIG. 6a.

INVENTOR.
LOUIS R. FOCHT

TIME-DOMAIN PITCH DETECTOR AND CIRCUITS FOR EXTRACTING A SIGNAL REPRESENTATIVE OF PITCH-PULSE SPACING REGULARITY IN A SPEECH WAVE

This is a division of application Ser. No. 582,605, filed Sept. 28, 1966, now U.S. Pat. No. 3,488,442.

The human speech mechanism can be described as an acoustic cavity which is bounded by the larynx at one end and by the lips and teeth at the other end. In the production of speech, the acoustic cavity is varied by movement of the tongue and jaws. The tongue and jaws divide the acoustic cavity into resonant cavities which produce the succession of sounds which makeup the speech wave.

Speech waves are a series of damped sinusoids rich in harmonic content. When such waves are analyzed on a frequency basis, it is found that there are a number of local resonance points which correspond to the resonant frequencies of the cavities in the speech mechanism. These resonant frequencies are referred to as formants. Although a speech wave may contain upward of five formants, the first three formants are the principal factors in determining sound color.

Because the speech wave is rich in harmonic content, it is highly redundant and contains more information than is needed to control a speech recognition system or a speech communication system. A bandwidth of approximately 3,000 cycles per second is normally required for voice communication by the conventional transmission systems. Transmission of the speech waveform in a less redundant form makes it possible to maintain communication over channels having a bandwidth of less than 300 cycles per second.

Prior art systems have extracted several parameters characteristic of the speech wave for speech communication and speech recognition systems. The most promising of the parameters that have been used are the frequencies of the first three formants of a speech sound, the respective amplitudes of the first three formants, a voiced-unvoiced sound decision, and pitch. The voiced-unvoiced sound decision and the pitch are used to specify the harmonic content of the complex speech wave. Information may be transmitted by means of these eight apparently independent parameters whose pattern of movement and position are ultimately recognized as representing words. However, it is obvious that from the standpoint of bandwidth compression and simplicity of the ultimate communication or recognition system, a speech representation system requiring fewer speech representative parameters would be preferred.

It is, accordingly, an object of the present invention to provide means for and a method of generating a novel parameter representative of a speech wave.

It is another object of the present invention to provide means for and a method of generating a plurality of novel parameters representative of a speech wave.

According to the present invention, six parameters of the prior art comprising the frequencies of the first three formants and the amplitudes of these formants 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 frequency of the single equivalent formant and the amplitude of the single equivalent formant. According to the single equivalent formant concept, a sound can be represented by the frequency and amplitude of a signal which may or may not correspond to one of the formants of the sound. By using this concept it is possible to replace three formant speech with its single formant equivalent and thereby reduce the information needed to specify the content of speech. When pitch and voicing parameters are used in conjunction with the single equivalent formant frequency parameter and the single equivalent formant amplitude parameter, only four parameters rather than the eight parameters of the prior art are required to specify the content of speech.

In a preferred embodiment of the present invention, the single equivalent formant frequency is extracted by measuring

the period of the first major oscillation of the complex speech wave.

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 frequency of the single equivalent formant for 10 vowel sounds;

FIG. 2 is a graph showing the relative formant amplitudes for the 10 vowel sounds of FIG. 1;

FIG. 3 is a diagram showing the formation of a complex speech wave;

FIG. 4 is a block diagram of the single equivalent formant speech analyzer of the present invention;

FIG. 5 is a block diagram of a circuit for producing a signal representative of the frequency of the single equivalent formant;

FIG. 6 is a block diagram of a circuit for extracting pitch pulses;

FIG. 6a is a schematic diagram of a portion of the circuit of FIG. 6;

FIG. 7 is a block diagram of a circuit for extracting the log of the amplitude of the single equivalent formant; and

FIGS. 8 and 9 are block diagrams of circuits for extracting the voicing parameter.

To understand the concept of the single equivalent formant and the apparatus for extracting the single equivalent formant from a complex 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 perceived phonetic values slightly away from that of the dominant formant. It is further postulated that formant amplitude is the principal factor determining formant dominance and hence the frequency of the single equivalent formant. More specifically, it is postulated that the frequency of the single equivalent formant is primarily dependent upon the frequency of the formant of largest amplitude. The foregoing postulates for determining the frequency of the single equivalent formant were confirmed by psychoacoustic testing. That is, a burst of a single frequency damped sinusoidal sound was presented to a test group and the group indicated what phonetic pronunciation (phoneme) corresponded to the burst of sound. The testing showed that the postulates are correct.

Referring to FIG. 1, ten phonetic vowel sounds are plotted on the horizontal axis and their corresponding first three formant frequencies are plotted on the vertical axis. Each of the three formant frequencies is plotted against its corresponding perceived vowel response. The vowel sounds are grouped as back, central, and front vowels. The back, central and front vowel groups are articulated in the rear, central and front portions of the acoustic cavity, respectively. The pronunciation of the ten phonetic vowel sounds is shown in the legend of FIG. 1, the lower frequencies represent the first formant, the intermediate frequencies represent the second formant, and the highest frequencies represent the third formant. The heavy line superimposed on the graph represents the single equivalent formant frequency for the ten vowels shown on the horizontal axis.

An examination of FIG. 1 shows that for the back vowels the first formant frequency nearly equals the frequency of the single equivalent formant. For the front vowels, the frequency of the single equivalent formant is nearly equal to the second formant frequency. For the central vowels, however, the frequency of the single equivalent formant does not correspond to either the first or second formant frequencies. For these vowels, the frequency of the single equivalent formant appears to be an average of the first and second formant frequencies.

The correlation between the frequency of the single equivalent formant and the formant frequencies for the ten

vowels illustrated in FIG. 1 can best be explained by reference to FIG. 2. FIG. 2 shows the frequencies of the first three formants of the ten vowels illustrated in FIG. 1 plotted against their relative formant amplitudes in decibals after a 9 db. per octave high frequency emphasis. The 9 db. per octave high frequency emphasis is necessary to illustrate accurately the effect of the formants on the human hearing system because it is believed that a high frequency emphasis of approximately 9 db. per octave is performed in the human hearing mechanism. FIG. 2 shows that the amplitudes of the first formants for the back vowels are larger than the amplitudes of the second or third formants for the back vowels and that the amplitudes of the second formants for the front vowels are larger than the amplitudes of the first or third formants for the front vowels. FIG. 2 also shows that the amplitudes of the first and second formants of the central vowels are approximately equal and larger than the third formant.

The extraction of the single equivalent formant is based upon the characteristics of the speech wave and the psychological factor of dominance just described. FIG. 3A shows the conceptual formation of a three-formant speech sound. The shock of the vocal cord wave train excites the various resonant cavities of the speech mechanism, producing a series of damped sinusoids, F_1 , F_2 and F_3 . The ringing frequencies of the damped sinusoids, F_1 , F_2 and F_3 , are the first, second, and third formant frequencies, respectively. The damped sinusoids F_1 , F_2 and F_3 combine to form the complex speech wave S.

FIGS. 3B, C and D show how the complex speech wave S is affected by the relative amplitudes of the damped sinusoids F_1 , F_2 and F_3 . When the first formant F_1 is larger in amplitude than the second formant F_2 and much larger in amplitude than the third formant F_3 (FIG. 3B), the period "T" of the first major oscillation of the complex speech wave 3, produced as a result of vocal cord wave train excitation, is approximately equal to the period of the first major oscillation of the largest or first formant F_1 . When the second formant F_2 is larger than the first formant F_1 and much larger than the third formant F_3 (FIG. 3C), the period "T" of the first major oscillation of the complex speech wave 5, produced as a result of vocal cord wave train excitation, is approximately equal to the period of the first major oscillation of the largest or second formant F_2 . However, when both formants, F_1 and F_2 , are of approximately equal amplitude and larger than the third formant F_3 (FIG. 3R), the resultant period "T" of the first major oscillation of the complex speech wave differs from the period of the first major oscillation of either the first or second formant. Equal amplitude formants produce a speech wave having a first major oscillation period approximately equal to the average value of the first major oscillation periods of the two equal formants. FIG. 3, therefore, shows that the period of the formant of largest amplitude of a complex speech wave will primarily determine the period of the first major oscillation of the wave.

Since the frequency of the largest amplitude formant of a sound is the primary factor determining the frequency of the single equivalent formant of the sound (FIGS. 1 and 2) and the period of the first major oscillation of the complex speech wave at each shock of the vocal cords is approximately equal to the period of the first major oscillation of the formant of largest amplitude (FIG. 3), the period of the first major oscillation of the complex speech wave at each shock of the vocal cords will approximately represent the reciprocal of the frequency of the single equivalent formant. More particularly, since the period of the first major oscillation of a complex speech wave is approximately inversely proportional to the frequency of the largest amplitude formant, the period of the first major oscillation of the complex speech wave will be approximately inversely proportional to the frequency of the single equivalent formant.

The block diagram of FIG. 4 shows the speech analyzer or speech parameter generator 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 pitch detector 6 has its output terminal coupled to the detectors 2 and 4, and to a voicing detector 8. The operations of the detectors 2, 4, 6, and 8 will be explained presently.

Although the single equivalent formant frequency parameter and the single equivalent formant amplitude parameter provide sufficient information for the identification of some speech sounds, additional information is required when a large vocabulary of sounds are to be identified. The voicing signal generated by voicing detector 8 will supply the additional information that is required when a large vocabulary of sounds are to be identified.

Theoretically, the sounds emitting from the acoustic cavity can be designated as either voiced or unvoiced sounds. If the acoustic cavity is excited by a series of pulses of nearly constant frequency generated by the vocal cords, the sound waves from the acoustic cavity contain harmonically related energy and the sounds are designated as voiced. In the case of unvoiced sounds, excitation is provided by passing air turbulently through constrictions in the acoustic cavity and the speech waves produced contain nonharmonically related energy. Theoretically, voiced sounds are designated as vowels and unvoiced consonants and unvoiced sounds are designated as unvoiced consonants. If two sounds have similar single equivalent formant frequency and amplitudes, it will be important to know whether the sound is voiced or unvoiced so that a determination can be made as to whether the sound is a vowel or an unvoiced consonant.

In actual speech, however, sounds do not fall ideally into the voiced and unvoiced categories. The most obvious discrepancy occurs in the voiced fricative sounds, such as occur in the pronunciation of the letters, *f*, *v*, *s*, and *z*, which are a mixture of harmonic and nonharmonically related energy. Furthermore, vowels are rarely characterized by purely harmonic energy, since they also contain a small amount of nonharmonically related energy. This is the result of a small amount of turbulence produced by the air stream passing through constrictions in the mouth. Similarly, unvoiced consonants are not necessarily characterized by purely nonharmonically related energy because the vocal cords do not stop vibrating instantaneously when the human rapidly changes from vowel articulation to unvoiced consonant articulation. However, unlike voiced sounds, the excitation pulses produced during unvoiced sounds occur in a random manner.

Since the detection of just two voicing states, harmonic or nonharmonically related energy, will not convey sufficient information to distinguish between sounds when a large vocabulary of sounds are to be identified, it is desirable to have a voicing parameter that specifies the ratio of harmonic to nonharmonic related energy in the speech wave. The voicing detector 8 of the present invention measures the degree of regularity between adjacent pitch pulses and thereby specifies the ratio of harmonic to nonharmonically related energy in the speech wave.

FIG. 5 is a block diagram of the single equivalent formant frequency detector 2 of FIG. 4. It comprises a circuit for measuring the period of the first major oscillation of the complex speech wave and, hence, the frequency of the single equivalent formant. The electrical representation of the input speech wave is coupled through an amplifier 10 and a high frequency preemphasis network generally indicated as 12 to the input of a high gain threshold circuit 18, such as a Schmitt trigger. The high frequency preemphasis network 12 comprises a series capacitor 14 and a shunt resistor 16. Network 12, acting as a differentiator, emphasizes the high frequency components of the input speech wave. High gain threshold circuit 18 is set to give an output only for one polarity of the differentiated input speech wave.

The output of circuit 18 is supplied to one input terminal of a bistable switching circuit 22, such as a flip-flop circuit. The output of time domain pitch detector 6, whose construction will be explained presently, is supplied to a second terminal of

circuit 22. Bistable switching circuit 22 is coupled by means of a pulse width-to-amplitude converter 24, which may take the form of a ramp generator, to the input of a sample and hold circuit 26. The output of the sample and hold circuit 26 is a signal of slowly varying amplitude, the instantaneous amplitude of which is inversely proportional to the frequency of the single equivalent formant.

The function of the time domain pitch detector 6 will now be explained. As previously stated, the speech waveforms of voiced sounds are produced by a periodic excitation of the vocal cords. A close examination of voiced speech waveforms makes evident a point in time at which the vocal cords are excited (FIG. 3). The point where the discontinuity occurs in the speech wave is an indication of the initiation of the vocal cord excitation function. The time domain pitch detector 6 indicates each of these points of discontinuity, which are referred to hereinafter as pitch pulses.

Referring to FIG. 6, the construction of the time-domain pitch detector 6 of FIG. 4 is shown. The input speech wave amplifier is coupled through a high frequency preemphasis network 30 to a nonlinear or logarithmic amplifier 32. Logarithmic amplifier 32 operates on the speech signal so that it occupies a relatively constant dynamic range. Amplifier 32 has its output coupled to a peak detector 34 and to a peak detector 36. Peak detector 36 is coupled by a voltage threshold conduction device 40, such as a zener diode, and an emitter follower network 38 to the output of the peak detector 34 and to a differentiating and amplifying network 42 the output of which is a signal having pulses at the pitch rate of said speech wave.

The time-domain pitch detector of FIG. 6 functions in the following manner. The input speech wave is preferentially amplified above the threshold voltage of the peak detectors 34 and 36 in the logarithmic amplifier 32. That is, the amplifier 32 amplifies the low level signals of the input speech wave to a greater degree than it amplifies the high-level signals of the input speech wave, thus compressing the dynamic range of the signal and counteracting changes in voice inflection.

Peak detectors 34 and 36 generate a resultant signal that emphasizes the peak amplitudes of the signal from amplifier 32. That is, peak detectors 34 and 36 generate signals which have amplitude peaks corresponding to an input signal amplitude greater than a predetermined value. Each generated signal decays exponentially after each amplitude peak until the occurrence of another input signal amplitude peak greater than the predetermined value. The signal being generated increases to the now input signal amplitude and then decreases exponentially until the occurrence of another input pulse of at least the predetermined value. Since vocal cavity excitation by the vocal cords produces a damped sinusoidal like wave that has its maximum amplitude at the point of excitation, the resultant output waveform of the peak detectors will have its peak amplitudes at the excitation or pitch pulse and hence will indicate the pitch pulses.

Conventional peak detectors have not satisfactorily indicated pitch pulses under conditions of rapidly falling speech amplitude. Long time constant peak detectors have a sufficiently long time constant to eliminate harmonic peaks that may occur between the fundamental peaks of the input speech wave. However, because the amplitude of the output signal of such a detector decays so slowly, a rapid drop in speech amplitude produces a loss of pitch pulses, amplitude of the output signal of a short time constant peak detector decays rapidly enough to permit such a detector to respond to all fundamental pitch pulses even though there is a rapid drop in speech amplitude. However, such a detector also responds to undesirable pulses, i.e. harmonics of the pitch frequency occurring between the pitch pulses. The deficiencies of the long and short time constant peak detectors are overcome by combining the two different time constant detectors into a dual time constant peak detector.

Referring to FIG. 6a, which is a schematic circuit diagram of section 35 of the block diagram of FIG. 6, the peak detector

36 of FIG. 6 comprises a transistor 37 and an emitter follower network 39 and the peak detector 34 of the FIG. 6 comprises a transistor 41 and an emitter follower network 43. Each emitter follower network 39 and 43 comprises a shunt connected resistor and capacitor. The ends of the emitter follower networks 39 and 43 remote from the transistors 37 and 41, respectively, are connected to a positive source of bias potential. The values of the respective resistors and capacitors of the emitter follower networks 39 and 43 are chosen so that the network 39 has a longer time constant than the network 43.

The emitter of transistor 41 is connected to the base of a transistor 45 and the emitter of transistor 37 is connected through voltage threshold conduction device 42, shown as a zener diode, to the emitter of transistor 45. The respective collector electrodes of transistors 37, 41, and 45 are connected to a negative source of bias potential. Zener diode 40 is poled to conduct in the forward direction when the potential at network 39 is more positive than the potential at the emitter of transistor 45. A resistor 47 is connected between the emitter of transistor 45 and the positive source of bias potential. Resistor 47 and transistor 45 comprise the emitter follower network 38 of FIG. 6. The base electrodes of transistors 37 and 41 are coupled to the output of amplifier 32 of FIG. 6.

In the absence of conduction of zener diode 40, peak detector 36 peak detects the output waveform supplied by amplifier 32 to produce the waveform "a" shown in FIG. 6a and peak detector 34 peak detects the output waveform supplied by amplifier 32 to produce waveform "b" shown in FIG. 6a. Since peak detector 34 has a shorter time constant than peak detector 36, waveform "b" decreases from the peak amplitude points more rapidly than waveform "a" decreases from the peak amplitude points. Zener diode 40 conducts whenever the potential difference between waveforms "a" and "b" exceeds the zener breakdown voltage. Due to emitter follower 38, network 39 is heavily loaded when zener diode 40 is conducting. Therefore the discharge characteristics of network 39 will follow the discharge characteristics of network 43 during the time that the zener diode 40 is conducting. Waveform "c" of FIG. 6a shows the output of the dual time constant peak detector. In waveform "c," points "d" indicate the initiation of conduction by zener diode 40.

Since the potential difference between waveforms "a" and "b" is small immediately after a fundamental peak amplitude point, region "x" of waveform "a," zener diode 40 will not conduct and therefore harmonic peaks that occur immediately after fundamental peaks will not result in undesirable pitch pulses. Since fundamental peak pulses do not usually occur in rapid succession, nonconduction of diode 40 immediately after a fundamental peak amplitude point will not suppress a desired pitch pulse. As the time after the occurrence of a pitch pulse increases beyond region "x" of waveform "a," a point is reached where there is a sufficient potential difference between waveforms "a" and "b" to initiate conduction of zener diode 40. Since the dual peak detector will now follow the discharge characteristic of peak detector 34, the dual peak detector will detect lower amplitude pitch pulses, such as pitch pulse "p" of waveform "b," and hence be able to follow rapid changes in speech amplitude.

Since the peak detected wave rises rapidly in amplitude at the occurrence of each amplitude peak and, as previously stated, the speech wave has its maximum amplitude at the excitation or pitch pulses, a circuit which emphasizes the points of rapidly increasing amplitude will produce a signal representative of the pitch pulses. In the circuit of FIG. 6, differentiating and amplifying network 42 emphasizes, i.e. preferentially transmits, the high frequency or rapidly varying components of the peak detected wave to produce a signal representative of the pitch pulses. This signal is an input to the bistable switching circuit 22 of the single equivalent formant frequency extractor circuit of FIG. 5.

Referring again to FIG. 5, a pulse from the pitch detector 6 sets the bistable switching circuit 22 in a first stable state. Circuit 22 remains in the first state until a first pulse is received

from the high gain threshold circuit 18. The pulse from circuit 18 reacts the circuit 22 to the second stable state. Since, as previously stated, circuit 18 is set to give an output only when the input speech wave is of one polarity, the pulse from circuit 18 will indicate when the input speech wave has completed its first major oscillation. If the output of bistable switching circuit 22 is taken across a load in which current flows only when the circuit 22 is in the first stable state, the output of circuit 22 will be a pulse length modulated signal having components equal to the period of the first major oscillation of the complex speech wave at each shock of the vocal cords and therefore can be used to measure the frequency of the single equivalent formant. The pulse width-to-amplitude converter 24 converts the pulse length modulated signal from the bistable switching circuit 22 into a series of amplitude modulated pulses. The amplitude of each pulse generated by converter 24 is proportional to the duration of the corresponding pulse from circuit 22. Sample and hold circuit 26 periodically samples the peak amplitude of the pulses from converter 24 and produces an output signal of constant amplitude between samples, this amplitude being equal to the amplitude of the converter 24 signal at the time of sampling. The amplitude varying signal from sample and hold circuit 26 is a slowly varying signal having an instantaneous amplitude proportional to the period of the first major oscillation of the sounds incorporated in the speech wave and hence is a slowly varying signal having an instantaneous amplitude inversely proportional to the frequency of the single equivalent formant.

Although only a dual peak detector network and a single differentiator and amplifier network have been shown as components of the time domain pitch detector, it is obvious that a plurality of serially connected peak detector and differentiator networks could be used to assure that all harmonic amplitude peaks are eliminated. In lieu of the dual peak detector circuit shown in FIG. 6, one or more single peak detector networks having a time constant intermediate the time constants of peak detector 34 and 36 could be used. If a single peak detector is used the voltage threshold conduction device 40 and the emitter follower 38 will be eliminated.

Another novel parameter, the single equivalent formant amplitude, is also useful in speech recognition and communication systems. Because the amplitude of the first major oscillation of the complex speech wave envelope is proportional to the amplitude of the single-equivalent formant (FIG. 3), a sample and hold circuit gated by the pitch detector output suffices to extract this parameter.

FIG. 7 shows the circuitry for extracting the log of the amplitude of the single equivalent formant. The complex speech input waveform is supplied to a peak detector 50 by means of a logarithmic amplifier 52. A sample and hold circuit 56 is coupled to peak detector 50 and to low pass filter network 54. Pitch pulses from the pitch detector 6 gate the sample and hold circuit 56 in order to measure the log of the peak amplitude of the complex speech wave.

Referring again to FIG. 7, the complex speech wave is preferentially amplified by the logarithmic amplifier 52 to compress the dynamic range of the speech wave. Peak detector 50 functions in the same manner as peak detectors 34 and 36 of FIG. 6 to detect the log of the peak amplitude points of the output from amplifier 52. Circuit 56 samples the log of the amplitude of the signal from detector 50 at each pitch pulse and maintains the amplitude of its output signal at the amplitude at the instant of sampling until the occurrence of another pitch pulse. Since, the amplitude of the speech wave is maximum at the occurrence of the pitch pulses (FIG. 3), the signal from sample and hold circuit 56 will be proportional to the log of the amplitude of the single equivalent formant. Low pass filter 54 removes the high frequency components of the amplitude modulated waveform to produce a slowly varying signal the amplitude of which is proportional to the log of the amplitude of the single equivalent formant.

The block diagram of FIG. 8 shows a circuit for extracting the voicing parameter from the output of the pitch detector 6

of FIG. 4. The output signal from the pitch detector 4 is supplied through a pulse width-to-amplitude converter 64, such as a ramp generator, to the input of a first sample and hold circuit 66. A differentiator network 68 couples the first sample and hold circuit 66 to a second sample and hold circuit 70. Slightly delayed pitch pulses from the pitch detector 6 control the amplitude of the signal generated by converter 64. That is, one pitch pulse initiates a ramp waveform signal from converter 64 which is terminated by the next pitch pulse. Sample and hold circuit 66 samples the ramp waveform from converter 64 at the pitch rate and maintains the output signal amplitude at the instantaneous sampling amplitude until the occurrence of the next sampling pulse. The signal generated by sample and hold circuit 66 is differentiated in differentiator 68 to obtain the time difference between adjacent pitch pulses. Smoothing of the differentiator 68 output signal by the second sample and hold circuit 70, which also samples at the pitch rate, produces a signal proportional to the regularity of the spacing of the pitch pulses. The input pitch pulses supplied to sample and hold circuits 66 and 70 are slightly delayed, for example by a series of one shot multivibrators, so that the sample and hold circuits 66 and 70 will sample the waveforms from the converter 64 and the network 68, respectively, at their points of maximum amplitude.

During voiced portions of an utterance, pitch periods will be of approximately the same duration and the signal from the sample and hold circuit 70 will be near zero. As the sounds change to voiced fricatives and to unvoiced sounds, the pitch periods will not be of approximately the same duration and sample and hold circuit 70 will produce a greater output signal. If it is desirable to produce an output signal that merely produces a voiced-unvoiced decision, a threshold circuit could be coupled to the output of the second sample and hold circuit 70.

A second circuit for extracting the voicing parameter measures the low frequency components of the complex speech wave by measuring the zero-crossing rate for voiced and unvoiced sounds rather than measuring differences of excitation frequency. FIG. 9 illustrates a block diagram for this type of circuit. Clipper 72 clips the positive portions of an input speech wave to determine the period of the zero-crossings of the wave. The positive portions are used to drive a ramp generator 74 and the output of the ramp generator 74 is peak detected by a peak detector 76. Peak detector 76 is coupled to a low pass filter network 78. Low pass filter network 78 removes most of the variations in the signal produced by the decay of the peak detectors. Since voiced sounds have high frequency, relatively high energy first formants, voiced sounds produce speech waves that have long periods between zero-crossings. However, unvoiced sounds have little or no first formant energy and do not produce long periods between the zero-crossings of a speech wave. Peak detector 76 has a sufficient long time constant so that it only detects the highest peaks of the waveform produced by ramp generator 74. Since the peaks of the ramp generator 74 are determined by the periods between zero-crossings of the wave, peak detector 76 only indicates the minimum frequency of the zero-crossings. As previously stated, voiced and unvoiced sounds have different zero-crossing frequencies. Since the output of peak detector 76 indicates the zero-crossing frequency it can be used to distinguish between voiced and unvoiced sounds.

The single equivalent formant groups together all speech sounds that have the same phonetic meaning regardless of variations in the acoustic spectrum. Thus, the single equivalent formant signal is invariant under the conditions of different speaker sex, speaker fatigue, pitch variations, speech rate and amplitude variations. The use of the single equivalent formant concept results in two major advantages in a speech recognition or communication system. First, it reduces the number of parameters that must be extracted and analyzed. This has a direct bearing on the size of the ultimate speech recognition logic and the bandwidth needed for speech communication. Second, it simplifies the extraction process itself.

To date, extracting the location of each of the individual formants of speech has been a difficult and complicated task. However extracting the single equivalent formant has been shown to be simple and economical.

The single equivalent formant parameters extracted by the previously discussed circuitry can be used in all types of speech communication and speech recognition systems. For example, the parameters can be quantized and used in a word recognition logic. The word recognition logic may consist of a set of generalized gates, such as AND, OR, NOR, and NAND gate combinations, for extracting the parameters characteristic of a sound vocabulary. Such a speech recognition logic would be simpler to implement than prior art speech recognition logics, since it would use fewer acoustic parameters, and could therefore make use of binary logic rather than analog weighted resistor threshold circuits.

The novel parameters can also be encoded and transmitted by conventional wire facilities and electromagnetic systems to a decoder and synthesizer network.

Although the foregoing specification has described only four speech recognition and communication parameters (single equivalent formant frequency, single equivalent formant amplitude, voicing and pitch) and apparatus for extracting these parameters, other parameters derived from the four parameters can be used. For example, the single equivalent formant amplitude, the derivative of the single equivalent formant amplitude, the derivative of Log of the single equivalent formant amplitude, and the derivative of the single equivalent formant frequency can be used as parameters.

I claim:

1. In a circuit for extracting a signal representative of the pitch pulse spacing regularity of a speech wave, a signal representative of a speech wave, first means for generating a signal representative of the low frequency components of said signal when said signal is supplied to said first means, and second means coupled to said first means for generating a signal proportional to the peak voltage points of the output signal from said first means.

2. The circuit of claim 1 in which said first means comprises a peak clipping network and a pulse length to pulse amplitude converter network and said second means is a peak detector network.

3. The circuit of claim 2 in which a low pass filter network is coupled to said peak detector network.

4. In a circuit for extracting a signal representative of the pitch pulse spacing regularity of a speech wave, a source of pitch pulses, a first sample and hold circuit coupled to said source, a differentiator network, a second sample and hold circuit, and means for coupling said differentiator network in cascade with said first and second sample and hold networks.

5. The circuit of claim 4 in which said sample and hold circuits sample the maximum amplitude points of their respective input signals.

6. In a circuit for producing a signal representative of the pitch pulses of a speech wave, a source of a signal representative of said speech wave, a first peak detector stage and a

second peak detector stage, each of said stages having an input terminal and an output terminal, means for supplying said signal representative of said speech wave to said input terminal of each of said stages, an emitter follower network having an input terminal and an output terminal, means connecting said output terminal of said second peak detector stage to said input terminal of said emitter follower network, a voltage threshold conduction device, means connecting said device between said output terminal of said emitter follower network and said output terminal of said first peak detector stage, and means, coupled to said output terminal of said first peak detector stage, for preferentially transmitting the high frequency components of the signal produced by said circuit at said output terminal of said first peak detector stage in response to said signal representative of said speech wave.

7. The circuit of claim 6 in which said first peak detector stage has a longer time constant than said second peak detector stage and said voltage threshold conduction device is a zener diode poled to conduct when the potential difference between said output terminal of said first peak detector stage and said output terminal of said emitter-follower stage exceeds a given amount.

8. A circuit according to claim 7, wherein

said first peak detector stage comprises a first transistor having an emitter, a collector and a base of given conductivity type, a first time-constant network comprising first resistive means and first capacitive means connected in parallel relationship, said first network having a given time constant, means including said first network for applying a first operating voltage to said emitter, and means for applying a second operating voltage to said collector, said second peak detector stage comprises a second transistor having an emitter, a collector and a base of said given conductivity type, a second time-constant network comprising second resistive means and second capacitive means connected in parallel relationship, said second network having a time constant shorter than said given time constant, means including said second network for applying said first operating voltage to said emitter of said second transistor, and means for applying said second operating voltage to said collector of said second transistor,

said emitter-follower stage comprises a third transistor having an emitter, a collector and a base of said given conductivity type, means connecting said emitter of said second transistor to said base of said third transistor, and third resistive means for applying said first operating voltage to said emitter of said third transistor, and means connecting said zener diode between said emitter of said first transistor and said emitter of said third transistor.

9. A circuit according to claim 8, wherein each of said first, second and third transistors has an N-type base and said zener diode has its anode connected to said emitter of said first transistor and its cathode connected to said emitter of said third transistor.

UNITED STATES PATENT OFFICE
CERTIFICATE OF CORRECTION

Patent No. 3,603,738 Dated September 7, 1971

Inventor(s) Louis R. Focht

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

Column 2, line 44: "wheat" should read --what--

Column 2, line 56: After "1" the comma should be changed to a period; before "lower" "the" should be changed to --The--

Column 3, line 34: "3" should be changed to --S--

Column 3, line 40: "5" should be changed to --S--

Column 3, line 45: "3R" should be changed to --3D--

Column 4, line 26: "frequency" should read --frequencies--

Column 5, line 20: "amplifier" should be deleted

Column 5, line 47: "now" should read --new--

Column 5, line 53: "pulse" should read --pulses--

Column 5, line 62: A period should be inserted after "pulses";
--The-- should be inserted before "amplitude"

Column 5, line 63: "peck" should be changed to --peak--

Column 6, line 13: Figure "42" should be changed to --40--

Column 8, line 47: "high" should be changed to --low--

Signed and sealed this 6th day of June 1972.

(SEAL)
Attest:

EDWARD M. FLETCHER, JR.
Attesting Officer

ROBERT GOTTSCHALK
Commissioner of Patents