June 3, 1969 J. M. KELLY 3,448,216

VOCODER SYSTEM

Filed Aug. 3, 1966 Sheet 1 of 2

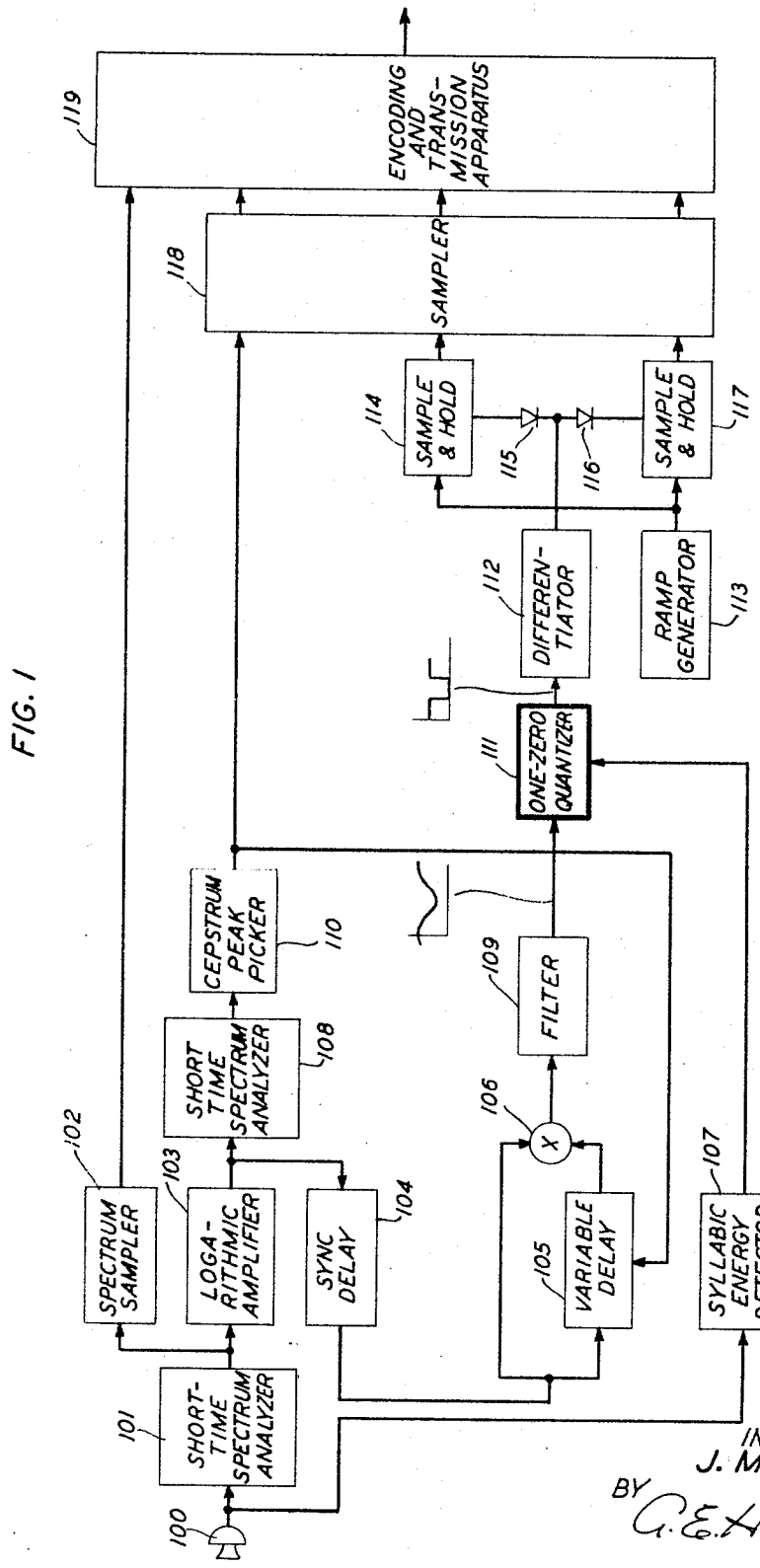

FIG. 1

INVENTOR
J. M. KELLY
BY G. E. Hirsch Jr.
ATTORNEY

June 3, 1969 J. M. KELLY 3,448,216

VOCODER SYSTEM

Filed Aug. 3, 1966 Sheet 2 of 2

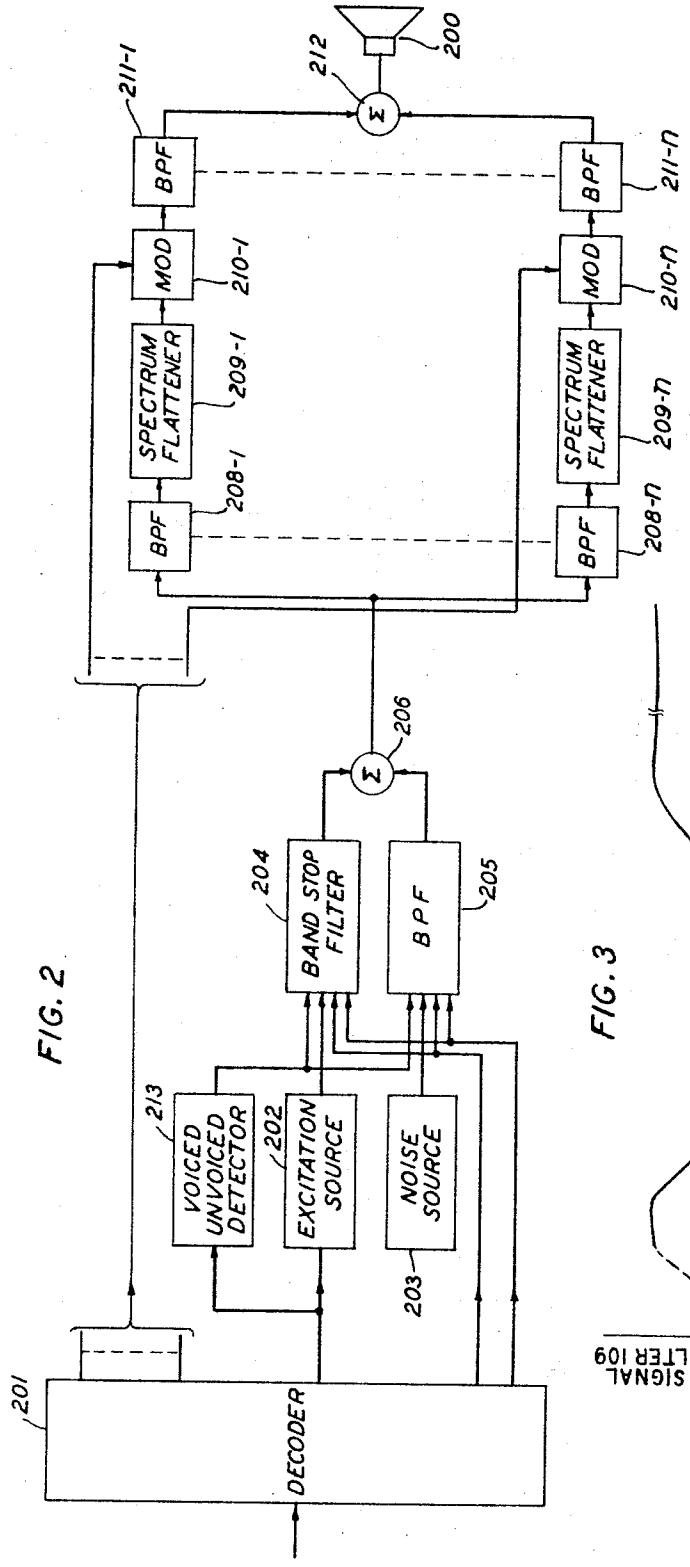

FIG. 2

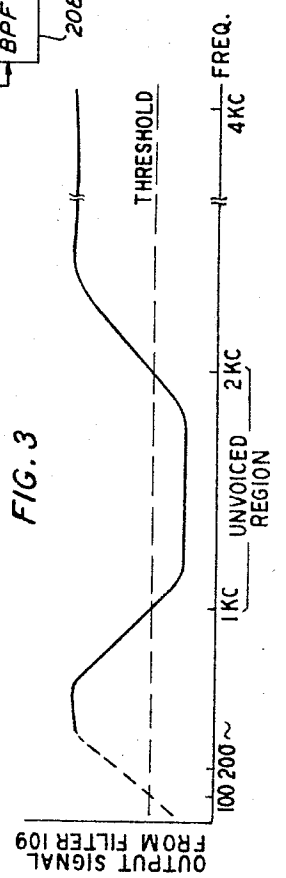

FIG. 3

United States Patent Office

3,448,216

Patented June 3, 1969

3,448,216

VOCODER SYSTEM

James M. Kelly, Morris Plains, N.J., assignor to Bell Telephone Laboratories, Incorporated, Murray Hill and Berkeley Heights, N.J., a corporation of New York

Filed Aug. 3, 1966, Ser. No. 569,898
Int. Cl. H04m 1/24; H04b 1/66
U.S. Cl. 179—1 9 Claims

ABSTRACT OF THE DISCLOSURE

In a channel vocoder, the excitation signals are more accurately characterized as voiced or unvoiced by developing auxiliary control signals responsive to the frequencies at which the speech signals shift from voiced to unvoiced and conversely, and using the control signals to control the pass bands of the channels which transmit the periodic and aperiodic signals, respectively.

This invention relates to the transmission of speech signals and in particular to the transmission of high quality speech signals over a transmission channel of narrow bandwidth. An object of this invention is to improve the quality of speech synthesized by a vocoder by improving the match between the amplitude spectrum of the synthesized speech and the amplitude spectrum of the input speech.

Much attention has been devoted to reducing the transmission channel bandwidth required to transmit speech. One result of this effort is the vocoder. In a vocoder, low frequency control signals representative of an input speech signal are derived at an analyzer. These control signals are then transmitted over a narrow bandwidth transmission channel to a synthesizer where they are used to construct a replica of the input speech signal.

To aid in synthesizing a replica of the input speech, the input speech is usually categorized by the vocoder as either voiced or unvoiced. During voiced speech, most of the speech energy fluctuates periodically due to vocal cord excitation while, during unvoiced speech, most of the speech energy fluctuates aperiodically due to turbulence in the vocal tract. This voiced-unvoiced categorization has long been known to be somewhat arbitrary because, at any given instant, the amplitude spectrum of voiced speech contains some frequency regions characteristic of voiced speech and other frequency regions characteristic of unvoiced speech. Thus voiced speech synthesized by a vocoder on the assumption that input voiced speech contains only periodic or voiced energy is often of poorer quality than the input speech because the amplitude spectrum of the synthesized speech fails to match that of the input speech with respect to voiced and unvoiced characteristics.

This invention overcomes this problem. In this invention the quality of speech synthesized by a vocoder is improved by matching the voiced and unvoiced regions of the amplitude spectrum of the synthesized voiced speech to the voiced and unvoiced regions of the amplitude spectrum of the input voiced speech.

According to one embodiment of this invention, control signals are derived at the vocoder analyzer indicative of the frequencies at which the characteristics of the amplitude spectrum of the input voiced speech shift from those of predominantly voiced speech to those of predominantly unvoiced speech and vice versa. These control signals are then used at the vocoder synthesizer to control the cutoff frequencies of complementary bandstop and bandpass filters which pass periodic and aperiodic excitation signals respectively. Speech synthesized using the sum of these excitation signals possesses a spectrum

which closely matches, with respect to voiced and unvoiced regions, the spectrum of the input voiced speech. The synthesized speech thus is improved in quality over the speech obtained from prior art vocoders.

This invention may be more fully understood from the following detailed description taken together with the attached figures in which:

FIG. 1 is a schematic block diagram of a vocoder analyzer constructed according to this invention;

FIG. 2 is a schematic block diagram of a vocoder synthesizer constructed according to this invention; and

FIG. 3 is a graph of the signal generated at the vocoder analyzer to determine the frequencies at which the characteristics of the input voiced speech spectrum shift from those of voiced speech to those of unvoiced speech and vice versa.

Theory

The described embodiment of this invention uses the so-called "cepstrum" technique to determine the pitch frequency of input voiced speech. This technique involves taking the Fourier transform of the logarithm of the amplitude spectrum of a selected time segment of the input speech signal. Of course, other types of pitch detectors, particularly those capable of detecting the presence of voiced speech despite the absence of the fundamental frequency, are also suitable for use in this invention.

The Fourier transform $F(\omega)$ of a time dependent signal $f(t)$ is defined as

$$F(\omega)=\int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt \tag{1}$$

where ω=radian freqency, $j=\sqrt{-1}$, and t=time. Since $e^{-j\omega t}$ equals the sum of a real part, cos ωt, and an imaginary part, $-j$ sin ωt, the Fourier transform $F(\omega)$ is composed of a real part $R(\omega)$ defined as

$$R(\omega)=\int_{-\infty}^{\infty} f(t)\cos \omega t dt \tag{2a}$$

and an imaginary part $X(\omega)$ defined as

$$X(\omega)=-j\int_{-\infty}^{\infty} f(t)\sin \omega t dt \tag{2b}$$

where

$$F(\omega)=R(\omega)+X(\omega) \tag{3}$$

The amplitude spectrum $A(\omega)$ of $f(t)$ is given by

$$A(\omega)=\sqrt{R^2(\omega)+X^2(\omega)} \tag{4a}$$

while the phase spectrum $\Phi(\omega)$, where $\Phi(\omega)$ is the phase difference between the modulating signal represented in Equation 1 by $e^{-j\omega t}$ and the corresponding frequency component of $f(t)$, is given by

$$\Phi(\omega)=\tan^{-1}X(\omega)/R(\omega) \tag{4b}$$

Of course, it is impossible to obtain the integral of a time function over all time from $-\infty$ to $+\infty$ as required by Equation 1. Thus, as explained in an article by A. Michael Noll entitled "Short-Time Spectrum and 'Cepstrum' Techniques for Vocal-Pitch Detection," published in the February 1964 Journal of the Acoustical Society of America, vol. 36, pp. 296 to 302, a spectrum analyzer yields what is called a "short-time" amplitude or power spectrum. The adjective "short-time" means the spectrum is derived from a signal segment of finite rather than infinite duration.

Further, as stated in the above cited Noll article, the short-time amplitude spectrum of speech $A(\omega)$ equals the product of the amplitude spectra of the speech excitation $|S(\omega)|$ and of the vocal tract $|V(\omega)|$. Thus

$$A(\omega)=|S(\omega)|\cdot|V(\omega)| \tag{5}$$

The speech excitation spectrum $S(\omega)$ represents the frequency characteristics of the vocal cords during periodic speech and of the vocal tract noise sources during aperiodic speech. Tho vocal tract, a series of ducts of variable area which resonate at certain frequencies called formant frequencies, has an impulse responsive with frequency characteristics given by $V(\omega)$. The vocal tract usually amplifies the speech frequency components at or near the formant frequencies, and attenuates all other speech frequency components. As a result, the amplitude spectrum $A(\omega)$ of the speech heard by a listener contains a slow oscillation as a function of frequency representing the contribution of the vocal tract resonances. In addition, during voiced speech, the amplitude spectrum $A(\omega)$ also contains a more rapid fluctuation with frequency, representing the fundamental vocal cord frequency and its harmonics.

By taking the logarithm of the short-time speech spectrum $A(\omega)$, the contributions of the speech excitation spectrum $|S(\omega)|$ and the vocal tract transfer function $|V(\omega)|$ become separate terms in a new logarithmic signal. Thus

$$\log A(\omega)=\log|S(\omega)|+\log|V(\omega)| \tag{6}$$

During periodic speech, the absolute value of $S(\omega)$ is just

$$|S(\omega)|=|S_T(\omega)|\cdot|(1+\sum_{l=1}^{n}e^{-jl\omega T})| \tag{7}$$

where N is a positive integer equal to one less than the number of pitch periods in the short-time speech segment from which $|S(\omega)|$ is derived, $|S_T(\omega)|$ is the absolute value of the speech excitation spectrum $S_T(\omega)$ derived from one (1) pitch period of speech, T is the pitch period and $e^{-jl\omega T}$ represents the phase shift with frequency of the amplitude spectrum $|S_T(\omega)|$ representing the lth repeated pitch period. Equation 7 can be rewritten in closed form as

$$\log|S(\omega)|=\log|S_T(\omega)|+\tfrac{1}{2}\log 2\left[\frac{1-\cos(N+1)\omega T}{(1-\cos\omega T)}\right] \tag{8}$$

The second term on the right-hand side of Equation 8 represents the effect on $\log|S(\omega)|$ of including $N+1$ pitch periods in the finite time segment of the speech signal $f(t)$ used to obtain $S(\omega)$. If $N=1$, then

$$\log|S(\omega)|=\log|S_T(\omega)|+\tfrac{1}{2}\log 2(1+\cos T\omega) \tag{9}$$

and $\log A(\omega)$ clearly contains a ripple,

$$\tfrac{1}{2}\log(1+\cos T(\omega),$$

which oscillates as a function of frequency ω, with a repetition rate of T, the pitch period. The repetition rate T of this "pitch ripple" has units of radians per randian per second, or seconds, and is usually called "quefrency," (a paraphrase of frequency) to avoid confusing this ripple with a time dependent cosinusoidally varying component of the speech signal. For $N>1$, the peaks of this pitch ripple become more pronounced but still occur at a quefrency T.

Thus one technique for detecting the presence of a periodic speech signal is to detect the pitch ripple on the logarithm of the short-time amplitude spectrum $A(\omega)$ of selected segments of the speech signal. This is often done by calculating the short-time spectrum of the logarithm of the short-time amplitude spectrum of the speech signal, called for short, the "cepstrum," (a paraphrase of spectrum) of the speech signal. To obtain the cepstrum, the logarith of the short-time amplitude spectrum $A(\omega)$ is usually converted into a time-dependent waveform, with frequency in the original amplitude spectrum proportional to time. The Fourier transform of this time-dependent waveform, namely the cepstrum, is obtained by conventional methods. The resulting cepstrum is defined in terms of time, rather than frequency as in the case of a spectrum. If the pitch ripple is present in the time-dependent waveform representing $\log A(\omega)$, the cepstrum exhibits a distinct maximum at a time T corresponding to the pitch period. The presence of several different pitch periods in a selected time segment of speech is indicated by a corresponding number of distinct maximums in the cepstrum.

Thus by analogy to Equation 1, the cepstrum $C(\tau)$ is defined as

$$C(\tau)=\int_{-\omega_0}^{\omega_0}\log A(\omega)e^{-i\tau\omega}d\omega \tag{10}$$

where ω_0 is the maximum significant frequency component in the short-time speech spectrum, and τ is called frequency. Note the similarity between τ in the operation indicated by Equation 10 and frequency ω in the transformation to the frequency domain of a time-dependent signal as indicated by Equation 1.

An examination of Equation 4*a* shows that the short-time amplitude spectrum is an even function; that is, $A(\omega)=A(-\omega)$. Therefore, Equation 10 can be rewritten, using only the even cosine component of $e^{-j\tau\omega}$, as

$$C(\tau)=2\int_{0}^{\omega_0}\log A(\omega)\cdot\cos\tau\omega d\omega \tag{11}$$

because the integral of the product $\log A(\omega)\cdot\sin\tau\omega$ from $-\omega_0$ to $+\omega_0$ is zero.

In this invention, the frequencies at which the characteristics of the short-time amplitude spectrum $A(\omega)$ shift from those of voiced speech to those of unvoiced speech and vice versa are determined in the following manner. The pitch period T of a selected segment of the input voiced speech signal is determined from the cepstrum of this segment. A voltage proportional to the pitch frequency is derived from the pitch period T and is used to control the relative delay between two identical time dependent waveforms representing the logarithm of the short-time amplitude spectrum of the speech signal. The integral of the product of these two waveforms over a time proportional to a selected frequency band $\Delta\omega$ gives, as a function of time, the short-time autocorrelation function $\Phi(2\pi/T,\omega-\Delta\omega/2)$ of the logarithm of the short-time amplitude spectrum of the speech signal. That is

$$\Phi\left(\frac{2\pi}{T},\ \omega-\frac{\Delta\omega}{2}\right)=\int_{\omega-\Delta\omega}^{\omega}\log A(\omega)\cdot\log A(\omega-2\pi/T)d\omega \tag{12}$$

When this integral is above a selected threshold, the portions of the amplitude spectrum contributing to this integral are periodic and thus highly correlated, indicating the presence of voiced speech characteristics. When, however, this integral falls below the selected threshold, the portions of the amplitude spectrum contributing to this integral are aperiodic and thus relatively uncorrelated, indicating the presence of unvoiced speech characteristics. The time at which the autocorrelation integral crosses the threshold indicates the frequency at which the characteristics of the amplitude spectrum change from those of voiced speech to those of unvoiced speech or vice versa. The direction of the crossing indicates whether the change is from voiced to unvoiced characteristics or vice versa.

Apparatus

In FIG. 1, an input speech signal $f(t)$ is detected by transducer 100 and converted into an electrical signal. This electrical signal is operated upon by short-time spectrum analyzer **101** to derive a waveform which varies in amplitude as a function of time and which represents the short-time amplitude spectrum $A(\omega)$ of the input speech signal.

In analyzer **101** a selected time segment of the electrical signal representing the input speech signal is compressed in time and stored in a recirculating storage device. This stored signal is continuously updated and repeatedly modulated by sinusoidal and cosinusoidal output

signals from a sweep oscillator contained in analyzer **101**. As dictated by Equations 2*a*, 2*b*, and 3, the sum of the integrals of the modulation products over the duration of the stored, time-compressed segment of the electrical signal is proportional to the short-time frequency spectrum $F(\omega)$ of the stored signal $f(t)$ at a frequency corresponding to the average modulating frequency ω over this time duration. The stored electrical signal $f(t)$ is modulated many times during one sweep of the modulating oscillator over its frequency range. A time series composed of samples proportional to the amplitudes of the resulting integrals is used to generate a waveform $A(t)$ representing the short-time amplitude spectrum $A(\omega)$ of the input speech signal.

The waveform $A(t)$ representing $A(\omega)$ is amplified by logarithmic amplifier **103** to generate a time dependent signal log $A(t)$ proportional to log $A(\omega)$, the logarithm of the short-time amplitude spectrum of the input speech signal. This logarithmic signal log $A(t)$ is in turn compressed in time and stored in a second recirculating storage device constituting part of short-time spectrum analyzer **108** where it is used to obtain the cepstrum $C(\tau)$ of the input speech signal. It remains stored until replaced by a new logarithmic signal representing the short-time amplitude spectrum derived from the next following time segment of the input speech signal.

Spectrum analyzer **108** operates upon log $A(t)$ just as though this logarithmic signal was another time dependent signal similar to the input speech signal. Time of course is really a dummy variable representing frequency since log $A(t)$ represents the variation with frequency of log $A(\omega)$.

The time compressed logarithmic signal, log $A(t)$, is, as shown above, an even function. Thus this signal is modulated repeatedly by a cosinusoidal output signal from a sweep oscillator contained in analyzer **108** to obtain the short-time cepstrum of the input speech signal. As required by Equation 11, this cosinusoidal modulating signal must be carefully synchronized with the waveform representing log $A(\omega)$ so that the cosinusoidal modulating signal has a maximum value precisely at the beginning of the wave form representing log $A(\omega)$, that is, when $\omega=0$.

The integral of the modulation product over the duration of the stored waveform is proportional to the magnitude of the amplitude spectrum of the stored waveform at the average modulating frequency except now real time t is actually proportional to frequency ω. The frequency of the modulating signal in turn is proportional to quefrency τ as indicating by Equations 10 and 11. The resulting integral reaches a maximum value when the quefrency τ of the modulating signal is equal to the pitch period T of the speech signal. A plot of the amplitude of this integral versus the quefrency of the modulating signal is the cepstrum $C(\tau)$ of the input speech signal.

The quefrency τ at which the peak of the cepstrum occurs is a direct measure of the pitch period T of the input speech signal. Thus, during voiced speech cepstrum peak picker **110** generates an output voltage proportional to the pitch frequency.

A vocoder system utilizing this cepstrum technique to detect the pitch of an input speech signal is described in copending application Ser. No. 420,362 filed Dec. 22, 1964, by A. M. Noll and M. R. Schroeder. Thus spectrum analyzer **101**, logarithmic amplifier **103**, analyzer **108**, and cepstrum peak picker **110** will not be described here in further detail.

The output voltage from peak picker **110** is used to control the operation of a short-time autocorrelation function generator consisting of delay **105**, product modulator **106**, and filter **109**. In addition, this voltage is transmitted in coded form to the vocoder synthesizer (FIG. 2) where it is used during voiced speech to generate an excitation signal.

Control signals indicative of the frequencies at which the characteristics of the amplitude spectrum of the input voiced speech signal change from those of voiced speech to those of unvoiced speech and vice versa are derived in the autocorrelation function generator. The waveform log $A(t)$, obtained from the output lead of amplifier **103**, is passed through synchronizing delay **104** to compensate for delays in the derivation of the output signal from peak picker **110**, and is then sent along two paths. The signal sent along one path is delayed in variable delay **105** by a time proportional to the fundamental frequency $2\pi/T$ of the input speech signal. The other signal is transferred undelayed to product modulator **106** where a signal proportional to the instantaneous product log $A(\omega)\cdot\log A(\omega—2\pi/T)$ is obtained. This product signal is passed through filter **109** which essentially integrates the instantaneous value of this product signal over a time period selected to correspond to the frequency band $\Delta\omega$. Thus, the continuous output signal from filter **109**, given mathematically by Equation 12, represents the short-time autocorrelation of the last $\Delta\omega$ cycles per second of the waveform representing log $A(\omega)$ for a delay time proportional to the pitch frequency. The magnitude of $\Delta\omega$ is selected so that the section of log $A(\omega)$ being correlated will usually contain at least two harmonics during voiced speech. For example, $\Delta\omega$ can represent a 500 cycle per second band of log $A(\omega)$.

Variable delay **105** is controlled by the output signal from cepstrum peak picker **110**. As the pitch period changes, the output voltage from cepstrum peak picker **110** varies thereby varying delay **105**. Variable delay **105** is always controlled so that the instantaneous value of the output signal from modulator **106** is the maximum possible value during voiced speech.

A variable delay suitable for use in this invention is described in copending application Ser. No. 538,676, filed Mar. 30, 1966, by R. N. Kennedy. The delay disclosed in the Kennedy application is essentially a domain wall shift register. In this delay, an analog signal is converted into digital form, the bits representing each sample of the signal are driven at a selected rate along magnetic wires, and, upon leaving the wires, are converted back into an analog signal. The delay time is a function of the rate at which the signals are driven along the wires, and this rate in turn is controlled by the signal from peak picker **110**.

A slight error in the magnitude of the autocorrelation signal from filter **109** occurs when variable delay **105** is rapidly changed from one value of delay to another. All bits on the magnetic wires while the delay time is changing emerge from the wires with a slightly erroneous delay. This effect, while small, can be eliminated, if desired, by using a tapped delay line. Tapped delay lines are well known and thus will not be described herein detail.

Thus variable delay **105**, product modulator **106**, and filter **109** work in such a manner that the output signal from filter **109** indicates which sections of the amplitude spectrum of the input voiced speech are characteristic of voiced speech and which sections are characteristic of unvoiced speed. At all frequencies for which the amplitude spectrum of the input speech is characteristic of voiced speech, the output signal from filter **109** is greater than some selected threshold. However, when the amplitude spectrum of the input speech signal resemble more closely the amplitude spectrum of unvoiced rather than voiced speech, the output signal from filter **109** drops below the selected threshold value. FIG. 3 shows the shape of the output signal from filter **109** when the short-time amplitude spectrum of input voiced speech has a region which corresponds to unvoiced speech between 1000 and 2000 cycles per second. The output signal is seen to dip below the threshold at a time corresponding to a frequency of 1000 cycles per second and to rise above the threshold at a time corresponding to a frequency of 2000 cycles per second.

One-zero quantizer **111** generates an output voltage of one during the time the output signal from filter **109** corresponds to voiced characteristics (that is, is above the selected threshold), and an output signal of zero during the time the output signal from filter **109** corresponds to unvoiced characteristics (that is, is below the selected threshold). Quantizer **111** is controlled by a normalizing signal from syllabic energy detector **107** to ensure that low level voiced speech signals are not erroneously categorized as unvoice dsignals. Energy detector **107** can in one embodiment comprise a rectifier in series with a low pass filter.

Differentiator **112** differentiates the output signal from quantizer **111**. Differentiator **112** produces a negative pulse when the characteristics of the amplitude spectrum of the input speech signal shift from those of voiced to those of unvoiced speech. It produces a positive pulse when the characteristics of the amplitude spectrum of the input voiced speech shift from those of unvoiced speech to those of voiced speech.

Ramp generator **113** generates a linearly increasing voltage, the amplitude of which is proportional to the time elapsed from the beginning of the signal representing log $A(\omega)$. The output voltage from ramp generator **113** is set to zero at the beginning of the signal representing log $A(\omega)$. If, for example, each signal representing log $A(\omega)$ has a duration of 10 milliseconds, and represents a frequency range of, for example, 0 to 10,000 cycles per second, generator **113** is reset to zero every 10 milliseconds.

A possible ambiguity occurs during the start of correlation while the output signal from filter **109** is building up to a value representative of the short-time autocorrelation function of log $A(\omega)$. To prevent the low value of the output signal from filter **109** (see FIG. 3) from being interpreted as representing unvoiced speech characteristics, sample and hold circuits **114** and **117** are inhibited from sampling the voltage produced by ramp generator **113** until a time corresponding to the frequency bandwith $\Delta\omega$ has elapsed. It $\Delta\omega$=500 cycles per second, as suggested above, this time is just 0.5 millisecond, using the numbers given in the above example. Thus any pulse from differentiator **112** before 0.5 millisecond has elapsed from the beginning of log $A(\omega)$, fails to actuate sample and hold circuits **114** and **117**.

On the generation of the first pulse by differentiator **112** after, for example, 0.5 millisecond (a negative pulse during voiced speech), sample and hold circuit **114** is activated through the corresponding diode **115** to sample the voltage generated by ramp generator **113** and to hold this voltage for transmission to the vocoder synthesizer. This voltage is proportional to the frequency at which the characteristics of the amplitude spectrum of the input voiced speed shift from those of voiced speech to those of unvoiced speech. On the next pulse from differentiator **112** (a positive pulse during voiced speech), the ramp voltage is sampled again, this time by sample and hold circuit **117** activated through diode **116**. This voltage is proportional to the frequency at which th echaracteristics of the amplitude spectrum of the input voiced speech shift from those of unvoiced speech to those of voiced speech. The voltages held by the sample and hold circuits **114** and **117** are periodically sampled by sampler **118** along with the voltage representing the pitch frequency from peak picker **110**.

In addition, samples of the amplitude spectrum of the input speech signal are periodically and repetitiously obtained by spectrum sampler **102**. The operation of spectrum sampler **102** is described in copending application Ser. No. 557,682, filed June 15, 1966, by J. M. Kelly, A. M. Noll and M. R. Schroeder. The samples obtained by sampler **102** represent the amplitudes of signals in contiguous frequency bands of the input speech signal. These samples, together with the samples from sampler **118** are converted into digital code words and transmitted to the vocoder synthesizer by means of encoding and transmission apparatus **119**. Sampler **118** and transmission apparatus **119** are well known and thus will not be described in detail.

At the synthesizer (FIG. 2), all the transmitted signals are decoded in decoder **201**. Decoder **201** converts the samples representing the amplitudes of signals in selected contiguous frequency bands of the input speech signal into low frequency control signals identical to those obtained in prior art vocoders. Decoder **201** also converts the samples representing the pitch frequency into a replica of the input pitch signal. This pitch signal is used in excitation source **202** during voiced speech to generate an excitation signal containing a fundamental frequency and harmonics similar to those of the input voiced speech signal.

Decoder **201** also produces two voltages, one proportional to the frequency at which the characteristics of the amplitude spectrum of the input voiced speech signal change from those of voiced speech to those of unvoiced speech and the other proportional to the frequency at which the characteristics of the input voiced speech spectrum shift from those of unvoiced speech to those of voiced speech. The voiced-unvoiced control signal is used to set the lower cutoff frequency of bandstop filter **204** and bandpass filter **205** to the frequency at which the characteristics of the amplitude spectrum $A(\omega)$ of the input voiced speech signal shift from those of voiced speech to those of unvoiced speech. The unvoiced-voiced control signal is used to set the upper cutoff frequencies of these two filters to the frequency at which these characterstics shift from those of unvoiced speech to those of voiced speech.

Bandstop filter **204** passes the periodic excitation signal generated by excitation source **202** with the exception of that portion of the excitation signal between the two cutoff frequencies. On the other hand, bandpass filter **205** passes that portion of a noise signal between the same two cutoff frequencies. Thus, the sum of the two signals passed by the two filters, obtained in summing network **206**, has a frequency spectrum which closely matches the frequency spectrum of the input voiced speech signal with respect to voiced and unvoiced characteristics.

The remainder of the vocoder circuit is well known. The signal from summing network **206** is separated into subsignals occupying contiguous frequency bands by bandpass filters **208**. Spectrum flatteners **209** operate on these subsignals to smooth their amplitude spectrums. Modulators **210** each generate an output signal in response to the simultaneous presence of a subsignal and a low frequency control signal from decoder **201**. The signals from the modulators **210** are filtered by corresponding bandpass filters **211** to remove undesired frequency components. The resulting filtered signals are summed in summing network **212** and converted into an acoustic speech signal by loudspeaker **200**.

The amplitude spectrum of this acoustic speech signal faithfully resembles the amplitude spectrum of the input voiced speech signal with respect to voiced and unvoiced characteristics. As a result, the quality of the synthesized voiced speech is improved over the quality of voiced speech synthesized in prior art vocoders.

During unvoiced speech, voiced-unvoiced detector **213** (FIG. 2) generates an output signal which holds the cutoff frequencies of bandstop and bandpass filters **204** and **205** at selected values regardless of any signals which might be generated by sample and hold circuits **114** and **117** (FIG. 1). Excitation source **202** generates an aperiodic signal in response to a signal of constant voltage, for example zero volts, indicating unvoiced speech from cepstrum peak picker **110**. Noise source **203** likewise generates an aperiodic signal. Thus the composite signal from summing network **206** is completely aperiodic, and is characteristic of unvoiced speech.

If during voiced speech no portion of the amplitude spectrum of the input speech is characteristic of unvoiced speech, as indicated by no negative pulse from differentiator **112** (FIG. 1), sample and hold circuit **114** generates a maximum voltage. This voltage drives the lower cutoff frequencies of bandstop filter **204** and of bandpass filter **205** above a selected value, thereby ensuring that the portion of the excitation signal passed by bandpass filters **208** contains only periodic energy components characteristic of voiced speech.

While this invention has been described for the case where the amplitude spectrum of voiced speech contains only one region characteristic of unvoiced speech, the principles of this invention can be extended by one skilled in the art to include cases where the amplitude spectrum of voiced speech contains several noncontiguous regions characteristic of voiced speech.

What is claimed is:

1. In apparatus in which a replica of an input speech signal is produced by modulating an excitation signal with low frequency control signals, that improvement which comprises:

means for analyzing the spectrum of an input speech signal to derive first control signals indicative of the frequencies at which the characteristics of the spectrum shift from those of voiced speech to those of unvoiced speech and second control signals indicative of the frequencies at which the characteristics of the spectrum shift from those of unvoiced speech to those of voiced speech;

means for transmitting said indicative control signals to a speech synthesizer;

and at said synthesizer, means responsive to said indicative control signals for producing an excitation signal containing selected frequency components characteristic of voiced speech and other frequency components characteristic of unvoiced speech.

2. In a vocoder, that improvement which comprises:

means for analyzing the amplitude spectrum of input voiced speech to derive a first control signal indicative of the frequency at which the characteristics of the amplitude spectrum shift from those of voiced speech to those of unvoiced speech and a second control signal indicative of the frequency at which the characteristics of the amplitude spectrum shift from those of unvoiced speech to those of voiced speech;

means for transmitting said control signals to a speech synthesizer;

and at said synthesizer, means responsive to said control signals for producing a periodic excitation signal during intervals of voice speech and an aperiodic excitation signal during intervals of unvoiced speech.

3. Apparatus as in claim **2** in which said analyzing means comprises:

means for generating a first signal proportional to the logarithm of the amplitude spectrum of said input voiced speech;

means for generating a second signal proportional to the pitch frequency of said input voiced speech;

autocorrelation means responsive to said second signal for producing a third signal proportional to the maximum short time autocorrelation function of said first signal;

means responsive to said third signal for producing said first control signal to indicate the frequency at which the characteristics of said amplitude spectrum shift from those of voiced speech to those of unvoiced speech, and said second control signal to indicate the frequency at which said characteristics shift from those of unvoiced speech to those of voiced speech.

4. Apparatus as in claim **3** in which said autocorrelation means comprises:

means controlled by said second signal to delay said first signal an amount proportional to said pitch frequency;

means for obtaining the product of said first signal and said delayed first signal;

and means for integrating said product over a selected time to produce a third signal proportional to the maximum short-time autocorrelation function of said first signal.

5. Apparatus as in claim **3** in which said means responsive to said third signal comprises:

a comparator for comparing the amplitude of said third signal to a selected threshold;

means for generating a constant positive voltage during the time said third signal exceeds said threshold and a zero voltage during the time said third signal is below said threshold;

a differentiator for providing a negative pulse at the instant the output voltage from said generating means drops from said positive voltage to zero voltage and a positive pulse at the instant said output voltage rises from zero voltage to said positive voltage;

a ramp generator for producing a linearly increasing voltage with time;

first sample and hold means responsive to said negative pulse for sampling the voltage produced by said ramp generator to produce said first control signal;

and second sample and hold means responsive to said positive pulse for sampling the voltage produced by said ramp generator to produce said second control signal.

6. Apparatus as in claim **2** in which said means responsive to said control signals comprises:

an excitation source for producing a periodic excitation signal;

a noise source for producing an aperiodic excitation signal;

a bandstop filter with cutoff frequencies controlled by said control signals for passing selected frequency components of said periodic excitation signal;

a bandpass filter with cutoff frequencies controlled by said control signals for passing selected frequency components of said aperiodic excitation signal;

and a summing network for combining said filter periodic and aperiodic excitation signals to produce an excitation signal with an amplitude spectrum containing regions matching the voiced and unvoiced regions of the amplitude spectrum of said input voiced speech.

7. In combination:

means for generating a first signal proportional to the amplitude spectrum of an input speech signal;

means for generating from said amplitude spectrum a second signal proportional to the pitch frequency of said input speech signal during voiced speech;

means controlled by said second signal for generating the maximum autocorrelation function of said first signal;

means for detecting the times at which said maximum autocorrelation function crosses a selected threshold value;

means responsive to said detecting means for generating pulses when said autocorrelation function crosses said threshold;

means responsive to said pulses for generating a first set of control signals proportioanl to the frequencies at which said autocorrelation function crosses said threshold;

means for generating from said first signal a second set of control signals proportional to the energy in selected frequency bands of said input speech signal;

means for transmitting in coded form to a synthesizer, said second signal, said first set of control signals, and said second set of control signals; and

at said synthesizer:

means for converting to analog form said transmitted signals;

voiced-unvoiced detection means responsive to said decoded second signal;

means responsive to said decoded second signal for generating a periodic excitation signal during voiced speech and an aperiodic excitation signal during unvoiced speech;

means for continuously generating an aperiodic noise signal;

bandstop filter means with cutoff frequencies controlled by said first set of control signals during voiced speech and by said voiced-unvoiced detector during unvoiced speech for passing selected frequency components of said periodic excitation signal during voiced speech and said aperiodic excitation signal during unvoiced speech;

bandpass filter means with cutoff frequencies controlled by said first set of control signals during voiced speech and by said voiced-unvoiced detector during unvoiced speech for passing selected frequency components of said aperiodic noise signal;

means for combining said filtered excitation and noise signals;

and means for generating a replica of said input speech signal from said combined excitation and noise signals and said second set of control signals.

8. In combination:

means for generating a first set of control siganls proportional to the energy in contiguous frequency bands of an input speech signal;

means for generating a first signal proportional to the logarithm of the amplitude spectrum of said input speech signal;

means for producing from said first signal a second signal proportional to the pitch frequency of said input speech signal during voiced speech and equal to a constant during unvoiced speech;

means responsive to said second signal for generating a third signal proportional to the maximum autocorrelation function of said first signal;

means responsive to said third signal for generating a second set of control signals indicative of the frequencies at which the characteristics of the amplitude spectrum of said input speech signal shift from those of voiced speech to those of unvoiced speech and from those of unvoiced speech to those of voiced speech;

means for transmitting said first set of control signals, said second signal and said second set of control signals to a synthesizer in coded form;

and at said synthesizer:

means for decoding said transmitted signals;

means responsive to said second signal and said second set of control signals for generating an excitation signal, the amplitude spectrum of which possesses characteristics which closely match those of the amplitude spectrum of said input speech signal with respect to voiced and unvoiced regions;

and means responsive to said excitation signal and said first set of control signals for generating a replica of said input speech signal.

9. In combination:

means for generating a first set of control signals proportional to the energy in selected frequency bands of said input speech signal;

means for generating a first signal proportional to the logarithm of the amplitude spectrum of said input speech signal;

means for generating from said first signal a second signal proportional to the pitch frequency of said input speech signal;

means responsive to said first and second signals for generating a second set of control signals during voiced speech proportional to the frequencies at which the characteristics of the amplitude spectrum of said input speech signal change from voiced to unvoiced and from unvoiced to voiced;

means for transmitting in coded form to a synthesizer said first and second sets of control signals and said second signal;

and at said synthesizer:

means for decoding said transmitted signals;

means responsive to said second signal and said second set of control signals for generating an excitation signal which during voiced speech possesses an amplitude spectrum closely matching the amplitude spectrum of said input voiced speech signal with respect to voiced and unvoiced characteristics;

and means responsive to said excitation signal and said first set of control signals for generating a replica of said input speech signal.

References Cited

UNITED STATES PATENTS

3,030,450	5/1962	Schroeder	179—15.55
3,321,582	5/1967	Schroeder	179—1
3,328,525	6/1967	Kelly	179—1

KATHLEEN H. CLAFFY, *Primary Examiner.*

ROBERT P. TAYLOR, *Assistant Examiner.*

U.S. Cl. X.R.

179—15.55