

June 13, 1967

H. J. MANLEY

3,325,596

SPEECH COMPRESSION SYSTEM

Filed July 3, 1963

2 Sheets-Sheet 1

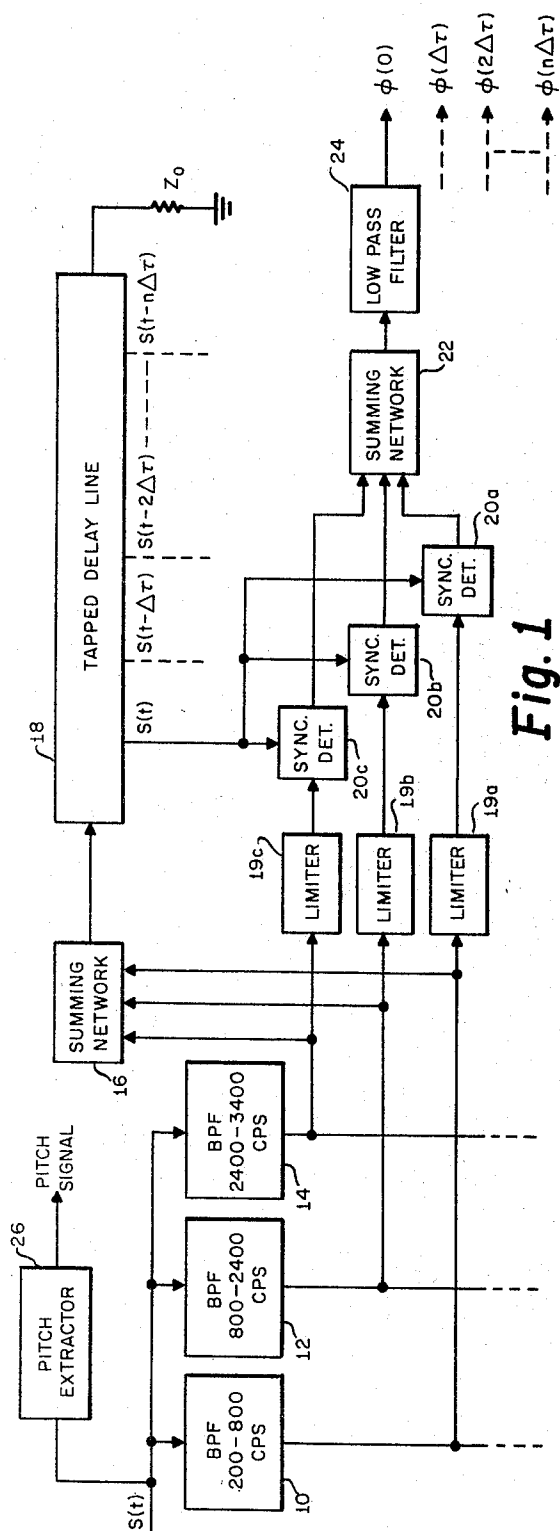


Fig. 3a

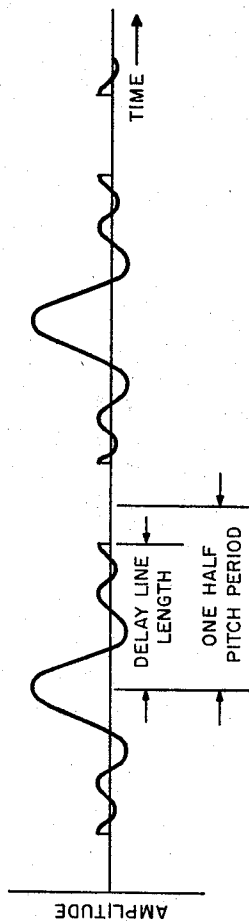
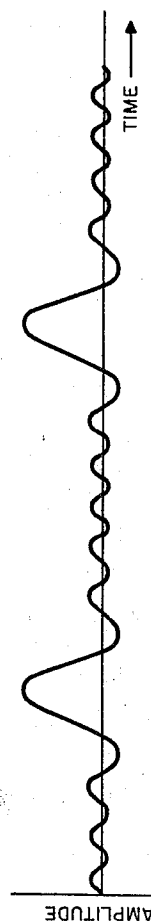


Fig. 3b



INVENTOR.
HAROLD J. MANLEY

BY

Y *James E. Olson*
ATTORNEY

ATTORNEY

June 13, 1967

H. J. MANLEY

3,325,596

SPEECH COMPRESSION SYSTEM

Filed July 3, 1963

2 Sheets-Sheet 2

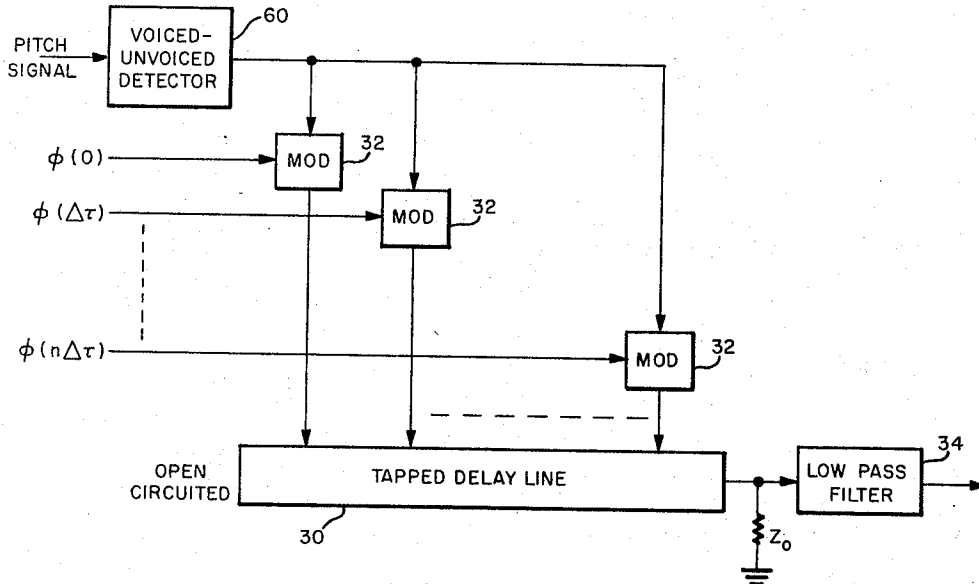


Fig. 2

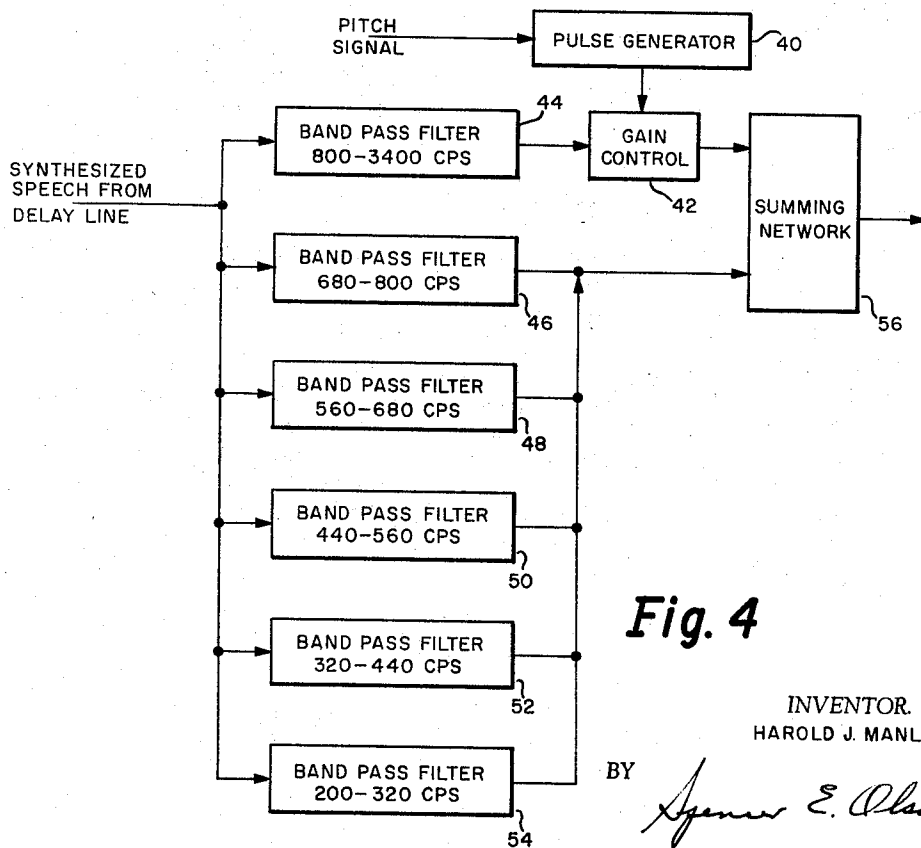


Fig. 4

INVENTOR
HAROLD J. MANLEY

BY

James E. Olson
ATTORNEY

1

3,325,596

SPEECH COMPRESSION SYSTEM

Harold J. Manley, Sudbury, Mass., assignor to Sylvania Electric Products Inc., a corporation of Delaware
Filed July 3, 1963, Ser. No. 292,656
14 Claims. (Cl. 179-1)

This invention relates to speech compression systems, and more particularly to a system utilizing autocorrelation techniques.

It is well known that a speech signal can be represented by an orthogonal function series of its autocorrelation function, since the expansion of the autocorrelation function of a signal into an orthogonal function series is equivalent to an orthogonal function expansion of the power spectrum of a signal. Inherent in autocorrelation analysis, however, is the fact that the signal representation describes the power spectrum rather than the absolute magnitude of the Fourier transform of the speech signal, which results in the resynthesized signal having an amplitude spectrum which is the square of the spectrum of the original signal. This effect, known as spectrum squaring, results in the synthesized speech having a distorted and unnatural sound, and must be eliminated or minimized if natural sounding speech is to be achieved.

A known means of implementing autocorrelation analysis and synthesis of speech is the autocorrelation vocoder in which a speech wave is correlated with delayed versions of itself to obtain an autocorrelation function which is then transmitted over a narrow-band channel to a speech synthesizer which reconstructs the speech. Since the autocorrelation function varies very little from one period to the next, the bandwidth required to transmit this signal is much less than the bandwidth required to transmit the original speech. The reconstructed speech, however, suffers from the spectrum squaring distortion alluded to above. Heretofore, such distortion has been overcome, generally, by two methods. One method described in "New Methods for Speech Analysis-Synthesis and Bandwidth Compression," M. R. Schroeder, B. F. Logan and A. J. Prestigiacomo, Fourth International Congress on Acoustics, Copenhagen, Aug. 21-28, 1962, paper G41, employs a variable equalizer which performs a square rooting operation on the spectrum of the speech signal so that spectrum squaring due to autocorrelation processing results in a signal having the original amplitude spectrum rather than a quadratic amplitude spectrum. While the variable equalizer does reduce spectrum squaring distortion, it is by nature an extremely complex circuit, and is difficult and costly to implement. Another means of overcoming spectrum squaring distortion is to cross correlate the speech signal with a speech-derived signal having a flat spectral envelope as described in Patent No. 3,071,652. Several factors prevent this technique from providing high quality speech. For example, the cross correlation function is not symmetric thus adding to the distortion, and in addition, the reference signal does not have an absolutely flat spectrum thereby causing the synthesized speech spectrum to deviate from the original speech spectrum. It is therefore, an object of the present invention to provide a speech compression system having no spectrum squaring distortion.

Another object of the invention is to provide autocorrelation analysis by means of synchronous detection.

Another object of the invention is to provide an autocorrelation vocoder of reduced cost and complexity.

A further object of the invention is to eliminate the truncation error in a synthesized speech signal.

Still another object of the invention is to provide simple means of processing unvoiced sounds.

Briefly, the invention comprises separating a speech

2

wave into its major formants and then synchronously detecting the formant signals with delayed versions of the original speech wave to produce slowly varying sample signals which are representative of the original speech signal and which have no spectrum squaring distortion. A particular feature of the invention is that truncation error associated with autocorrelation analysis is eliminated by use of a unique time-varying filter technique. Another feature of the invention is that unvoiced sounds are processed expediently by providing a tapped delay line in which the first several of a succession of taps are afforded twice the bandwidth than that afforded to the remaining taps.

The invention will be more fully understood from the following detailed description taken in conjunction with the drawings, in which:

FIG. 1 is a schematic block diagram of a speech analyzer in accordance with the present invention;

FIG. 2 is a schematic block diagram of a speech synthesizer in accordance with the invention;

FIGS. 3a and 3b are waveform diagrams useful in explaining the operation of the invention; and

FIG. 4 is a schematic block diagram of a time-varying filter used in the invention to eliminate truncation error.

At the outset, it will be helpful to consider the operation of a synchronous detector and how it can be used to eliminate spectrum squaring distortion. It is well known that the output signal of a synchronous detector is the difference of the magnitudes of the sum and difference of the two input signals; or mathematically, the output signal can be expressed as,

$$f(x, y) = \frac{|x+y| - |x-y|}{2} \quad \text{Eq. 1}$$

where x and y are the input signals.

Multiplying both numerator and denominator by

$$f(x, y) = \frac{|x+y| + |x-y|}{2 \times y} \quad \text{Eq. 2}$$

If one input signal, say x , is made larger with respect to the other, the output signal is approximately,

$$f(x, y) = \frac{xy}{|x|} \quad \text{Eq. 3}$$

Thus, the output signal is approximately equal to the product of the input signals divided by the magnitude of the larger input signal.

If the input signal x is a speech signal $S(t)$ and the input signal y is a delayed speech signal $S(t-\tau)$, then Eq. 3 can be rewritten as

$$f[S(t), S(t-\tau)] = \frac{S(t)S(t-\tau)}{|S(t)|} \quad \text{Eq. 4}$$

The speech signal $S(t)$ is composed of three major spectral components, known as speech formants, and can be expressed as

$$S(t) = A_1 e^{S_1 t} + A_2 e^{S_2 t} + A_3 e^{S_3 t} \quad \text{Eq. 5}$$

where A =complex amplitude and S =complex frequency. Similarly, the delayed version of the speech signal $S(t-\tau)$ can be expressed as

$$S(t-\tau) = A_1 e^{S_1(t-\tau)} + A_2 e^{S_2(t-\tau)} + A_3 e^{S_3(t-\tau)} \quad \text{Eq. 6}$$

It will be noted that the multiplication indicated in the numerator of Eq. 4 produces an expression in which the terms are squared. It will also be noted that the magnitude of the speech signal $S(t)$ in the denominator of Eq. 4 is essentially equal to the magnitude of the largest component of the signal, which is the first formant $A_1 e^{S_1 t}$. Since they are of different frequencies, the first formant in the numerator will not correlate with the second and

third formants in the numerator; consequently, the resulting expression has a quadratic character. The spectrum of the output signal of the synchronous detector thus has a quadratic character and spectrum squaring still results.

If, however, a formant signal, for example the first formant $A_1 e^{S_1 t}$, is correlated with the delayed speech signal, then Eq. 3 can be written

$$f[A_1 e^{S_1 t}, S(t-\tau)] = \frac{A_1 e^{S_1 t} S(t-\tau)}{|A_1 e^{S_1 t}|} \quad \text{Eq. 7}$$

An expansion of Eq. 7 produces three terms identical to the three terms of Eq. 6, except, of course, that each term is multiplied by $A_1 e^{S_1 t}$. The effect of this multiplier is negated, however, by the presence of $|A_1 e^{S_1 t}|$ in the denominator. The resulting expression, therefore, has a first power nature. The output of a synchronous detector, then, in which the input signals are a formant signal and a delayed speech signal, is a signal which is representative of the formant signal and which has an amplitude spectrum identical to the amplitude spectrum of the formant signal. Autocorrelation analysis of a speech wave can thereby be carried out by synchronously detecting each formant with a delayed speech signal and then combining the detected signals to produce a slowly varying sample of the speech wave, as will be further explained hereinafter. It should be remembered that the formant signal is kept large with respect to the delayed speech signal in order for Eq. 7 to be valid.

An alternate means of implementing Eq. 3 is suggested if it is recognized that the quantity $x/|x|$ is equivalent to the algebraic sign of x for all non-zero values of x . Designating the sign of x as $\text{SGN } x$, Eq. 3 can be rewritten as,

$$f(x,y) = y \text{SGN } x \quad \text{Eq. 8}$$

It is known that a hard limiter produces a signal which is indicative of the sign of x . Accordingly, Eq. 8 can be instrumented by hard limiting the signal x , and then multiplying the signal thus obtained with the signal y .

A speech analyzer, in accordance with the above-discussed principles, is shown in FIG. 1, wherein a speech signal $S(t)$ is applied to bandpass filters 10, 12 and 14 which separate the speech signal into three major speech formants. Filters 10, 12 and 14 have bandwidths of 200–800 c.p.s., 800–2400 c.p.s., and 2400–3400 c.p.s., respectively. These bandwidths are merely representative, however, as some variation is permissible since the limits of the formant bands are not strictly defined. Although three formant filters are shown in the illustrated embodiment, more or less than three could be used if desired. Reasonable speech quality can be obtained by using only two formants; however, it has been found that three formants provide speech of superior quality, while the use of more than three formants does not significantly improve the speech quality.

The formant signals are combined in a summing network 16, the output of which is applied to the input of a tapped delay line 18 which is terminated in its characteristic impedance Z_0 to prevent reflections. The input to the delay line is obtained from the summation of the formant signals rather than from the unfiltered speech signal to insure that the relative phase shifts of the formant frequencies remain unchanged. The formant signals from filters 10, 12, and 14 are also respectively applied to limiters 19a, 19b, and 19c, and thence to one input terminal of respective synchronous detectors 20a, 20b, and 20c. It will be recalled that the formant signal must be large with respect to the delayed speech signal in order for Eq. 7 to be valid. In keeping with this requirement, limiters 19a, 19b, and 19c provide amplification of the formant signals to maintain their amplitude of the order of ten times that of the delayed speech signals. In addition, the limiters also limit formant signals above a given amplitude to prevent overdriving the synchronous

detectors. Although only one group of detectors is illustrated, it is to be understood that a similar group of detectors are provided for each tap of delay line 18. As already mentioned, the output of each synchronous detector is a signal which is representative of the formant signal and which has an amplitude spectrum identical to the amplitude spectrum of the formant. These detected signals are combined by a summing network 22, and then applied to a low-pass filter 24 to produce slowly varying sample parameters $\phi(o)$, $\phi(\Delta\tau)$, . . . $\phi(n\Delta\tau)$ of the autocorrelation function of the speech signal. Pitch information is provided by a pitch extractor 26 which provides a reference pulse at the start of each pitch period. The reference pulse is transmitted to the synthesizer where it provides synchronizing signals to properly assemble the sample parameters at the correct pitch frequency. Alternatively, pitch information can be provided by well known voice excitation techniques wherein a low frequency portion of the speech, of the order of 80–350 c.p.s., is transmitted to the synthesizer to synchronize the reconstruction of the speech using the sample parameters.

To transmit the speech samples to a remote synthesizer, they can be multiplexed on a transmission line by any one of several well known means. Since the slowly varying sample parameters are transmitted, rather than the entire speech wave, the bandwidth required for such transmission is only of the order of 350 c.p.s., as compared to a bandwidth of the order of 3500 c.p.s. for uncompressed speech, a ten-to-one reduction.

The synthesizer, shown in FIG. 2, receives the sample parameters $\phi(o)$, $\phi(\Delta\tau)$. . . $\phi(n\Delta\tau)$ from the analyzer and assembles them in the proper sequence to produce a replica of the speech spectrum. Appropriate excitation must, of course, be provided to properly reproduce voiced sounds or unvoiced sounds, as the case requires. It is known that a noise source provides correct excitation for the synthesis of unvoiced sounds, while voiced sounds are suitably excited by pitch signals which convey the appropriate pitch and time information. Accordingly, the voiced-unvoiced detector 60 receives signals from pitch extractor 26 and provides pitch information in response to voiced sounds, and noise excitation in response to unvoiced sounds. It will be appreciated that voiced fricatives can also be reproduced by designing detector 60 to provide a combination of both types of excitation.

To reconstruct the speech, the sample parameters are applied to respective modulators 32, which are gated by excitation signals from voiced-unvoiced detector 60 to simultaneously apply the sample parameters to appropriate taps of a delay line 30 which correspond to similar taps in delay line 18 used in the analyzer of FIG. 1. The delay line is terminated at one end in a matched impedance Z_0 to prevent reflection, while the opposite end is open-circuited to provide in-phase reflection. It is clear that two output signals from the delay line will occur for each input pulse applied to the taps; one output due to propagation of a signal from a tap to the output terminal, and the second due to propagation of a signal from a tap to the open-circuited end of the delay line and thence to the output. Signals which are symmetrical about the time instant when modulators 32 are gated are thereby produced at the output of delay line 30 to produce an output signal representative of the autocorrelation function of the speech wave. A low-pass filter 34 is provided at the output of the delay line to time average the reproduced wave. Audible speech is available at the output of filter 34, and can be reproduced by any well known transducer, such as a loudspeaker.

A difficulty associated with autocorrelation analysis is truncation of the output signal due to the delay line length being less than one half pitch period. As seen in FIG. 3a, the output signal is abruptly terminated when the end of the delay line is reached, causing distortion

which results in the reconstructed speech having a rough or nasal character. This deleterious effect is overcome by the use of a novel time varying filter, illustrated in FIG. 4, which is connected in the system to receive the synthesized speech signals and also pitch pulses from the pitch extractor. In response to a pitch pulse from pitch extractor 26, a pulse generator 40 applies a control pulse to a gain control circuit 42 which varies the gain of a high frequency band-pass filter 44. Gain control 42 is varied such that filter 44, which is capable of reproducing the abrupt discontinuities in the output signal, is off during the period of discontinuity. The lower frequency bandpass filters 46, 48, 50, 52, and 54, which convey the first formant frequencies, which are below approximately 800 c.p.s., are on continuously. The second and third formant frequencies, passed by filter 44, have relatively wide bandwidths and are, therefore, damped out more rapidly than the first formant frequencies; consequently, little energy is lost by omitting them from the output signal during the period of discontinuity. In operation, when a synthesized speech wave is applied to the time varying filter, one of the bandpass filters 46, 48, 50, 52 and 54 responds to the major oscillation of the speech signal, and its output continues to oscillate during the time the high frequency filter 44 is off, thereby maintaining the waveform oscillation during the period of discontinuity, as illustrated in FIG. 3b. The signals from the bandpass filters are then combined by a summing network 56, the output of which is the synthesized speech signal having no truncation distortion. This filter technique is applicable to any speech compression system in which means are available to define the period of discontinuity. An alternative construction which provides nearly as good performance with less equipment is obtained by simply omitting pulse generator 40 and gain control 42 and applying the signal from filter 44 directly to summing network 56. Since filter 44 is relatively broadband, its output signal will be negligible during the period of discontinuity, while the narrow band filters will contribute to the output signal, as before.

As a further feature of the invention, it has been found that unvoiced sounds can be easily processed by the expedient of providing several delay line taps having half the delay of the remaining taps. Unvoiced sounds, by their nature, do not require the high spectral resolution needed to suitably represent voiced sounds. The unvoiced sounds can be represented by a relatively few spectral components spaced over a wide bandwidth of, say, five or six kilocycles. If the inter-tap delay between the first several of a succession of taps of delay line 18 is made half the inter-tap delay between the remaining taps, these first several taps are thereby afforded twice the bandwidth of the remaining taps. As a consequence, the first several taps are available to provide spectral information spaced over a relatively wide bandwidth suitable to correctly represent the unvoiced sounds. As a typical example, seven taps of a twenty-one tap delay line have been used to satisfactorily represent the unvoiced sounds, the inter-tap delay between the first seven taps being 83 microseconds, while the delay between the remaining taps is 167 microseconds. Accordingly, the bandwidths of the first seven taps and of the balance of the taps are 6 kc. and 3 kc., respectively. Of course, the frequency response of filter 14 must be increased accordingly so that signals up to 6 kc. can be applied to the delay line. Synthesis of the unvoiced sounds can be accomplished by apparatus of the type shown in FIG. 2 and described hereinbefore. It will be recalled that the taps of delay line 30 in the synthesizer correspond to similar taps of delay line 18 in the analyzer, and in accordance with the instant technique, the inter-tap delay between the first several taps of delay line 30 is made half the inter-tap delay of the remaining taps to correspond to the inter-tap delay configuration of delay line 18 in the analyzer. By this technique, unvoiced sounds are synthesized more realistically since the repre-

sentative signals derived by the analyzer and presented to the synthesizer convey more information characteristic of these sounds.

From the foregoing, it is evident that a relatively simple speech compression system has been provided in which reconstructed speech is produced having no spectrum squaring distortion and no truncation error. In addition, the speech compression system lends itself to realistically reproducing unvoiced sounds.

While there have been described what are now thought to be preferred embodiments of the present invention, many modifications and alternative implementations will occur to those skilled in the art without departing from the true spirit and scope thereof. In addition, while the invention is primarily of interest in the analysis and synthesis of speech, it is not limited to such use. The signal processing techniques of the invention can be used in analyzing a complex signal produced by a variety of means; for example, in analyzing seismological signals. Accordingly, it is not intended to limit the invention by what has been particularly shown and described, except as indicated in the appended claims.

What is claimed is:

1. A speech analyzer comprising means for separating a speech signal into its major formants, means for correlating each of said formants with said speech signal, and means for time averaging the correlated signals to produce a slowly varying representation of said speech signal.

2. A speech compression system comprising means for separating a speech signal into its major formants, means for correlating each of said formants with said speech signal, means for time averaging the correlated signals to produce a slowly varying representation of said speech signal, and means for reconstructing said speech signal in response to said slowly varying representation.

3. A speech analyzer comprising a source of speech waves, means for deriving the major speech formants from said speech waves, and means for correlating said formants with said speech waves to produce signals representative of the autocorrelation function of said speech waves.

4. A speech analyzer comprising means for separating a speech signal into its major formants, means for providing delayed versions of said speech signal, means for correlating each of said formants with each of said delayed versions, and means for time averaging the correlated signals to produce a slowly varying representation of said speech signal.

5. A speech analyzer comprising three bandpass filters for separating a speech signal into three major speech formants, means for providing delayed versions of said speech signal, means for correlating each of said formants with each of said delayed versions of said speech signal to produce signals which are representative of said formants, and means responsive to said representative signals to produce signals which are representative of said speech signal.

6. A speech analyzer comprising a source of speech signals, means for deriving major speech formants from said speech signals, a delay line having an input terminal and a plurality of output terminals, means for applying said speech signals to the input terminal of said delay line, a plurality of synchronous detectors each having a pair of input terminals and an output terminal, means for applying each of said formants to one of the input terminals of respective ones of said detectors, means for applying a signal from an output terminal on said delay line to the other input terminal of said respective ones of said detectors, means for summing the output signals of said detectors, and means for time averaging said summed signals.

7. A speech analyzer comprising a source of speech signals, means for deriving three major speech formants from said speech signals, a delay line having an input

terminal and a plurality of output taps, summing means for combining said formants and applying the combined signal to the input terminal of said delay line, three synchronous detectors for each of said delay line output taps, each of said detectors having two input terminals and one output terminal, means for applying, in parallel, a signal from one of the output taps of said delay line to one of the input terminals of said three detectors, means for applying one of said major speech formants to the other input terminal of a respective one of said detectors, means for summing the output signals of said three detectors, and a low-pass filter for averaging said summed signal.

8. In a speech compression system which includes a delay line for providing delayed versions of a speech signal, means for eliminating truncation error in a synthesized speech signal comprising, a plurality of filters for separating said speech signal into a like plurality of narrow frequency band signals, means operative in response to a pitch signal to render selected ones of said filters inoperative during a period of signal discontinuity, and means for summing said narrow band signals to produce a signal representative of said synthesized speech.

9. In a speech compression system which includes a delay line for providing delayed versions of a speech signal, means for eliminating truncation error in a synthesized speech signal comprising, a plurality of band-pass filters each covering, in succession, a portion of the speech spectrum, means for controlling the gain of selected ones of said filters to render said selected ones inoperative during a period of signal discontinuity, and means for summing the output signals of said filters to produce a signal representative of said synthesized speech.

10. In a speech compression system which includes a delay line for providing delayed versions of a speech signal, means for eliminating truncation error in a synthesized speech signal comprising a plurality of band-pass filters each covering, in succession, a portion of the speech spectrum, one of said filters covering a frequency band equivalent to the second and third formant bands, means for controlling the gain of said one filter to render said filter inoperative during a period of signal discontinuity, and means for summing the output signals of said filters to produce a signal representative of said synthesized speech.

11. In a speech analyzer which includes a delay line for providing delayed versions of a speech signal and means for correlating said speech signal with a reference signal to provide a slowly varying representation of said speech signal; means for processing unvoiced speech sounds comprising, a delay line having an input terminal, a first succession of output taps, and a second greater succession of output taps consecutive with said first succession, said first succession of output taps having an inter-tap delay which is half the inter-tap delay of said second succession of output taps.

12. A speech analyzer comprising a source of speech signals, means for deriving major speech formants from

said speech signals, a delay line having an input terminal, a first succession of output taps, and a second greater succession of output taps consecutive with said first succession and having an inter-tap delay which is twice the inter-tap delay of said first succession of output taps, means for applying said speech signals to said input terminal, a plurality of synchronous detectors for each of said output taps, each of said detectors having a pair of input terminals and one output terminal, means for applying each of said formants to one input terminal of respective ones of said detectors, means associated with each of the output taps of said delay line for applying a signal from said output taps to the other input terminal of corresponding ones of said detectors, means for summing the signals appearing at the output terminal of said detectors, and means for time averaging said summed signals.

13. A speech analyzer comprising a source of speech signals, means for deriving three major speech formants from said speech signals, a delay line having an input terminal, a first succession of output taps, and a second greater succession of output taps consecutive with said first succession, said first succession of output taps having an inter-tap delay which is half the inter-tap delay of said second succession of output taps, summing means for combining said formants and applying the combined signal to the input terminal of said delay line, three synchronous detectors associated with each of the output taps of said delay line, each of said detectors having first and second input terminals and an output terminal, means for applying, in parallel, signals from each of the output taps of said delay line to the first input terminals of corresponding ones of said three detectors, means for applying each of said formants to the second input terminal of a respective one of said three detectors, means for summing the output signals of said three detectors, and a low-pass filter for averaging said summed signal.

14. Apparatus for synthesizing speech which comprises a source of a plurality of signals each representative of a correlation function of a speech signal, a delay line open-circuited at one end and having an output terminal at the other end, a first succession of input taps, and a second greater succession of input taps consecutive with said first succession, the sum of said first and second successions equalling said plurality, said first succession of input taps having an inter-tap delay which is half the inter-tap delay of said second succession of input taps, means operative in response to a pitch signal to produce excitation signals, means responsive to said excitation signals for applying said representative signals to respective ones of said input taps, and means for averaging a signal appearing at said output terminal to produce artificial speech.

No references cited.

55 KATHLEEN H. CLAFFY, *Primary Examiner*.
R. MURRAY, *Assistant Examiner*.