

[54] SYNTHESIS OF PURE SPEECH FROM A REVERBERANT SIGNAL

[75] Inventor: Jont Brandon Allen, Fair Haven, N.J.

[73] Assignee: Bell Telephone Laboratories, Incorporated, Murray Hill, Berkeley Heights, N.J.

[22] Filed: Dec. 7, 1972

[21] Appl. No.: 311,731

[52] U.S. Cl. 179/1 SA

[51] Int. Cl. G101 1/00

[58] Field of Search 179/1 SA, 1 J, 1 P, 179/15.55 R; 84/DIG. 26

[56] References Cited

UNITED STATES PATENTS

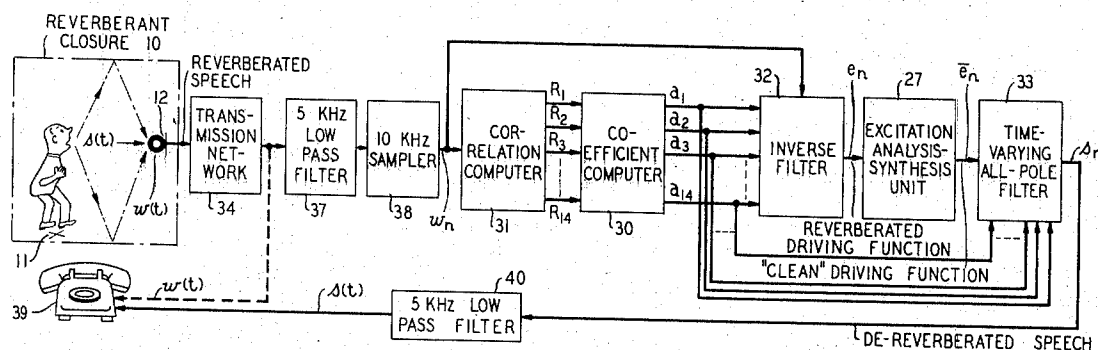
3,440,350	4/1969	Flanagan.....	179/1 SA
3,542,954	11/1970	Flanagan.....	179/1 J
3,662,108	5/1972	Flanagan.....	179/1 SA
3,715,512	2/1973	Kelly.....	179/1 SA

Primary Examiner—Kathleen H. Claffy
Assistant Examiner—Jon Bradford Leahey
Attorney—C. E. Graves

[57] ABSTRACT

Speech that has been reverberated by the transfer function of a reverberant enclosure is analyzed to detect parameters from which an unreverberative synthetic version of the original speech may be constructed. The process involves continuously approximating the vocal tract transfer function of the speaker. The effect of this transfer function is then removed from the reverberant speech by inverse filtering, the residual signal being the glottis excitation signal reverberated by the room. The reverberant excitation function is then analyzed to determine when the speaker's driving function is voiced or unvoiced, the periodicity when voiced, and a unique gain factor. Then clean speech is synthesized using the foregoing three parameters operating on an all-pole filter that is continuously adapted to approximate the vocal tract transfer function.

10 Claims, 9 Drawing Figures



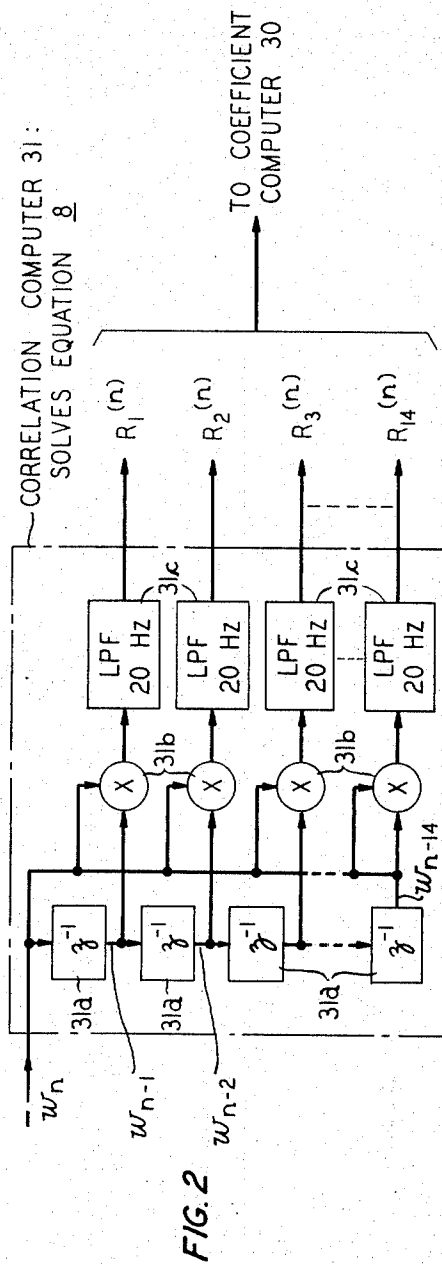
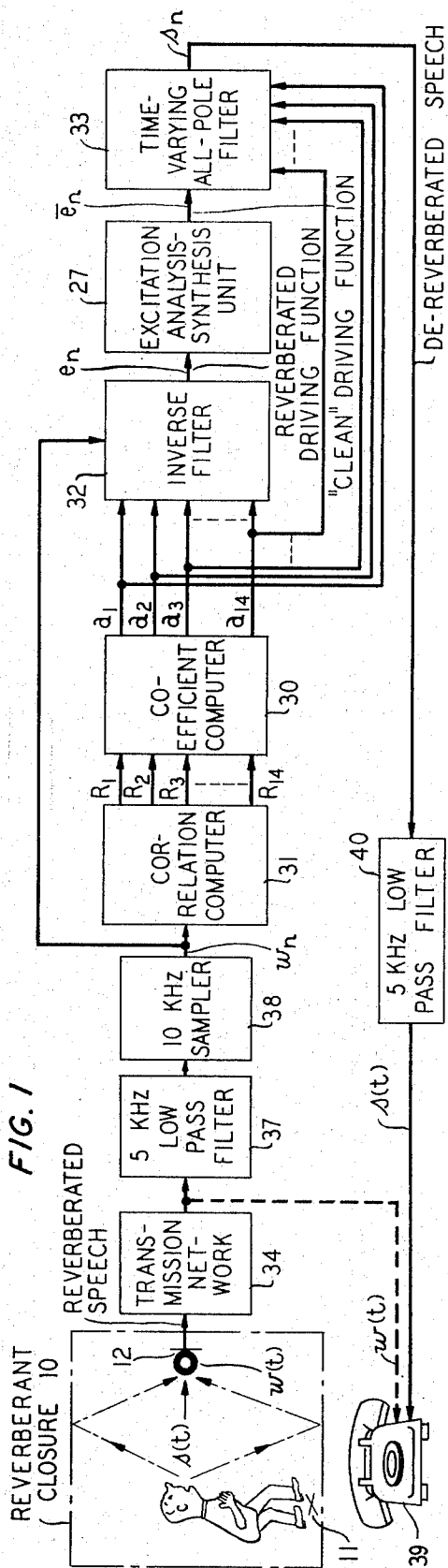


FIG. 3

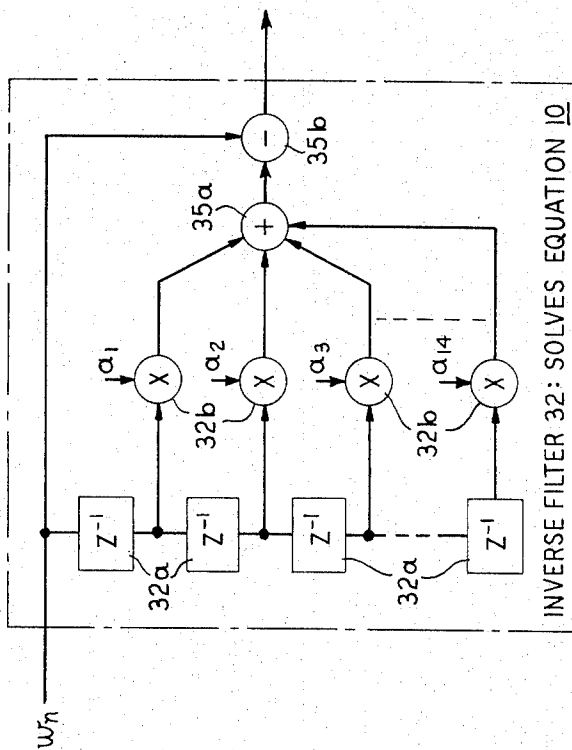


FIG. 8

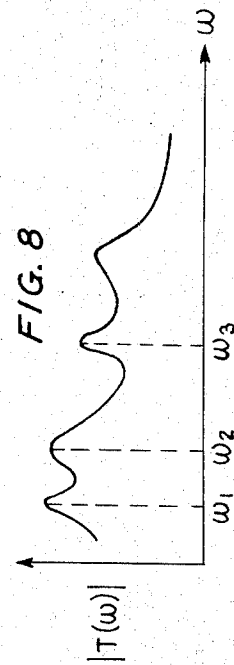


FIG. 5

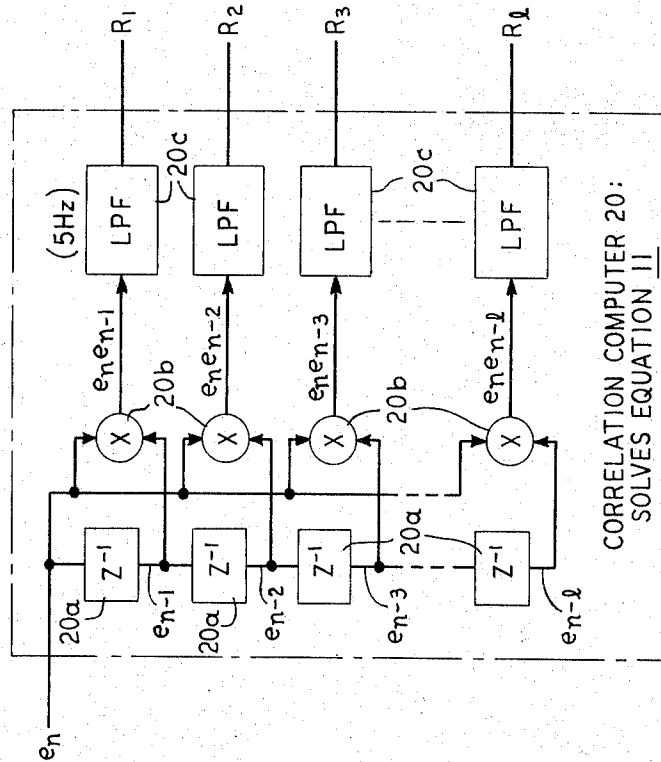


FIG. 4

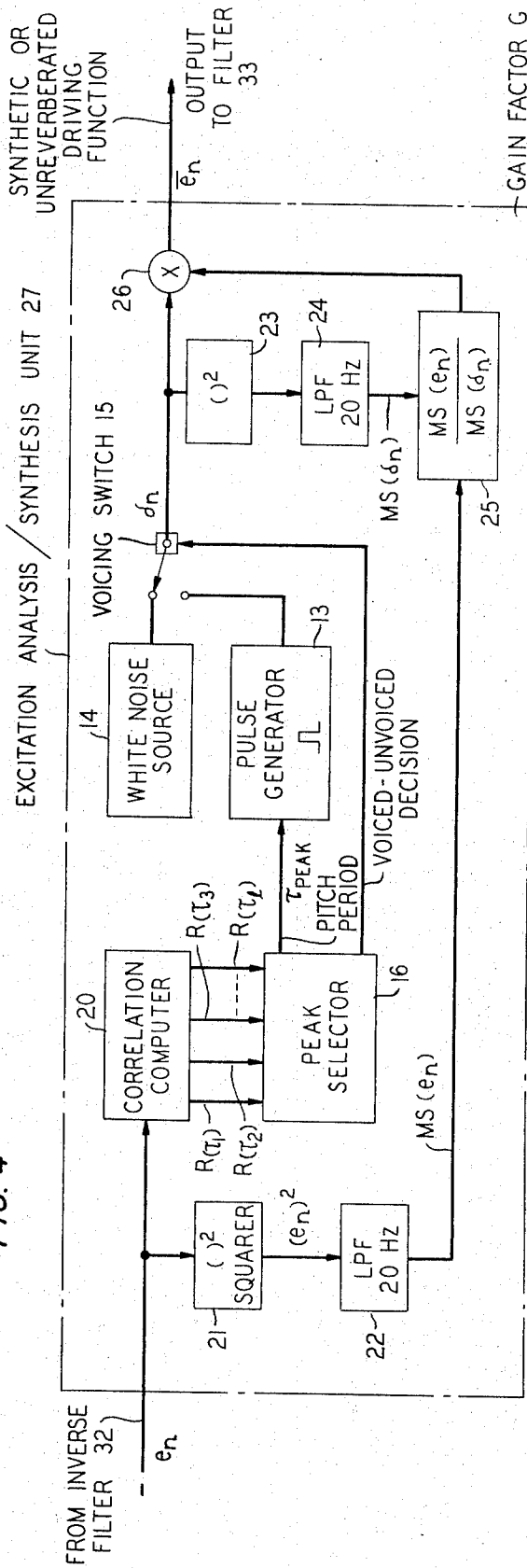
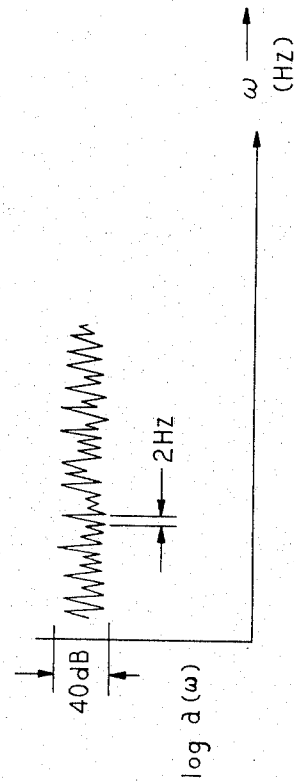
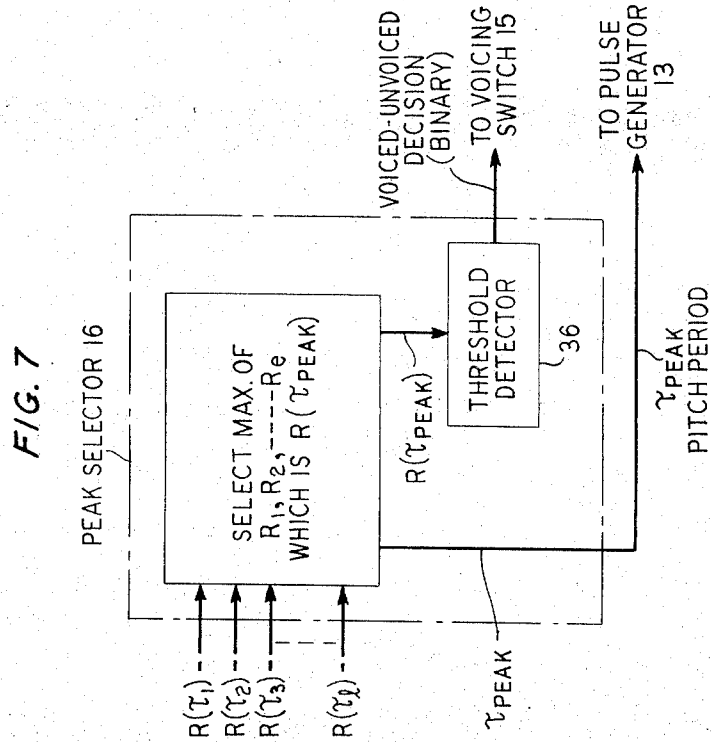
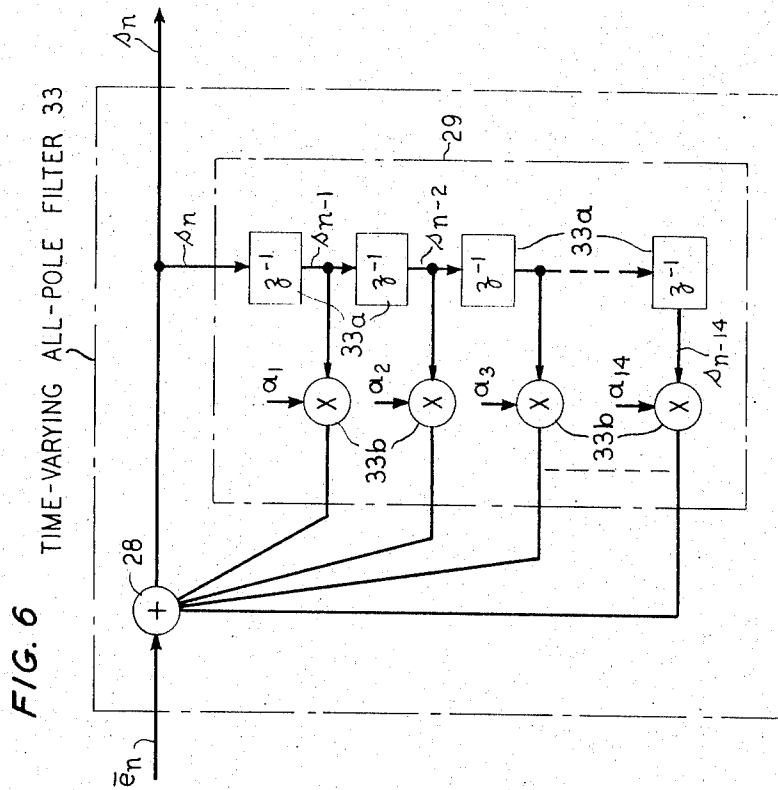


FIG. 9





SYNTHESIS OF PURE SPEECH FROM A REVERBERANT SIGNAL

FIELD OF THE INVENTION

This invention relates to the removal of distortion from a speech signal. In particular, this invention relates to the synthesizing of a distortion-free speech signal from a signal originating in a reverberative enclosure.

Background of the Invention

It is well known that speech, when produced as an acoustic signal in a reverberative chamber, reaches a remotely located microphone in that chamber at different times via a large number of paths of differing lengths. The signal received at the microphone will in general consist of the direct path energy, which is received first, followed closely by infinitely many delayed and filtered replicas of varying amplitudes. As perceived by the human ear, the effect is reverberative.

There are believed to be two separate effects present. The first effect is the "coloration" or spectral distortions due to the summation occurring at the microphone of the directly received signal and its many delayed dispersive reflections from the numerous walls and surfaces in the room 10. The second effect, the "echo", are the temporal or time distortions arising from the slow decay of energy typically encountered in any moderately lossless room or cavity. These time distortions are closely related to the reverberation time of the room. For the subject who is not physically present in the chamber but listens through a connection to the microphone placed in the chamber, the effects of coloration and echo on intelligibility of the received signal are often severe. This condition is, unfortunately, frequently characteristic of "hands-free" telephonic transmissions.

Numerous schemes have been proposed to remove these degradations perceived in reverberant speech signals. Examples of two such schemes are found in U.S. Pat. Nos. 3,440,350 and 3,662,108 issued to J. L. Flanagan. One drawback of prior art schemes is their lack of facility to adapt to a room transfer function that is continually time-varying. A second drawback is an inability to rely on only the reverberant speech signal itself as a source of information with which to reconstruct the "clean" speech.

Accordingly, it is one object of this invention to remove from reverberant speech the spectral distortions altogether, and also those temporal distortions which are equal to or less than the articulation times of the original speech of interest.

A further object of the invention is to realize a way of reconstructing an original speech signal by analysis of the reverberated speech. This object seeks to overcome prior art schemes wherein the parameters which control the synthesis of the undistorted signal are derived under unrealistic conditions, or are contingent on a stationary room transfer function.

Another inventive object is to devise a speech processing system of the type alluded to in the foregoing object, that has the property of sensing or detecting the parameters that characterize the original speech, so that by other aspects of the inventive process, undistorted speech will be synthesized from a knowledge of these parameters.

A still further object of the invention is to enlist and adapt the speech reconstruction method known generally as linear predictive filtering to novel use under reverberant conditions.

The processes of the present invention are based on known properties of human speech and the theory of linear prediction as expounded, for example, in the article "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," B. S. Atal and S. L. Hanauer, Journal of the Acoustic Society of America, Vol. 50, pages 637-655 (1971); and in U.S. Pat. No. 3,624,302, issued to B. S. Atal on Nov. 30, 1971, both of which are hereby incorporated by reference. By way of understanding the general relevance of applicant's invention with respect to the cited prior work, the following brief review of this background art is in order.

Synthesis, or the production of an original or pre-existing speech signal from a set of more basic parameters, depends in general upon activating some device whose basic transfer properties are akin to those of the human vocal tract, by some excitation signal which is akin to the excitation which drives the human vocal tract. For ongoing real time speech synthesis, Atal and others have recognized that a short time spectral analysis of the original speech signal does not readily yield control signal information for this excitation signal or driving function. Atal has realized more reliable control signals by modeling the human vocal tract as an acoustic tube of variable dimensions. In the Atal model, the vowel and vowel-like sounds of the output at any instant of time are a weighted sum of a discrete number of recent past values of the output plus the value of the input or driving function at that instant of time. Thus:

$$s_n = \sum_{k=1}^P a_k s_{n-k} + e_n \quad (1)$$

where P is nominally 14,
 s_n are speech samples,
 e_n are excitation signal samples, and
 a_k are the 14 vocal tract filter coefficients.

This type of linear combination may be shown to be equivalent to a linear all-pole filter. The latter can be made to behave like the human vocal tract by the proper choice of filter parameters. One may produce speech wave forms by exciting the all-pole filter with the proper combinations of quasi-periodic pulses and white noise, referred to herein as the excitation function e_n . The parameters of this filter are the weighting coefficients alluded to above, and termed a_k , where a_k is the gain applied to the speech sample delayed by k samples.

One inventive embodiment of Atal involves bandwidth reduction. The parameters are derived in the Atal approach from an undistorted original or preexisting speech signal which is to be reproduced at some remote location. Inherent in the reverberation reduction situation, however, is the availability of only reverberant speech as a source from which to derive parameters. It is not apparent that pure speech can be synthesized using only a reverberative speech as a parameter source.

SUMMARY OF THE INVENTION

The present invention in its broadest sense lies in the

recognition that the time-varying vocal tract transfer function of a subject speaking within an enclosure, can indeed be sufficiently determined even after the speech has undergone severe reverberative distortion. This is the case, whether or not the room transfer function is also varying or is altogether unknown.

The speech signal $w(t)$, which pursuant to the present invention is to be dereverberated, results from an as yet unknown excitation signal $e(t)$ driving a vocal tract as described above with transfer function $T(\omega)$ (where $\omega = 2\pi \times \text{frequency}$). The speech so produced, $s(t)$, is then reverberated by the room's transfer function $H(\omega)$ to produce a reverberative speech signal $w(t)$. The problem is to extract, from the reverberated speech $w(t)$ information which can be used to reconstruct or synthesize the original speech signal $s(t)$.

Pursuant to a prime aspect of the invention, it has been recognized that any practical or typical room transfer function $H(\omega)$ has certain properties that make it possible to accurately determine the speaker's vocal tract transfer function $T(\omega)$, from the reverberative speech signal $w(t)$. The principal property that makes the foregoing possible is that the mode structure, i.e., mode density, is almost always sufficiently great that the modes are closer than their bandwidths over the frequency range of useful speech information. Further, the reverberation times, i.e., the 60 dB energy delay time, of the vast majority of office or room size reverberant enclosures are less than those which would damage the articulation because of echos. In contrast, articulation damage could be expected to occur in the case of a large auditorium with hard walls.

By analysis of the reverberant speech, the vocal tract transfer function $T(\omega)$ of the speaker is continuously approximated. Then, the effect of the vocal tract transfer function is removed from the reverberant speech by inverse filtering, leaving only the spectrally flattened glottis excitation signal $e(n)$ reverberated by the room transfer function $H(\omega)$. Pursuant to the invention, analysis is then performed at this point on the reverberant excitation function to determine when the driving function $e(t)$ of the speaker is quasiperiodic (which is the voiced condition) or white noise (which is the unvoiced condition). The gain of the driving function $e(t)$ and the period of the quasiperiodic source during voicing are also derived.

Then, pursuant to the invention, clean speech is synthesized using:

1. $T(\omega)$, the vocal tract transfer function;
2. a binary parameter denoting voiced or unvoiced information;
3. a parameter denoting the period of the voiced part of the speaker's vocal tract driving function $e(t)$;
4. and a gain parameter denoting the mean-squared level of the driving function $e(t)$.

Advantageously, this process is continuously performed digitally by a sampling of the reverberated speech at, nominally, a 10 kHz rate. In a given communications link, the sampling and processing can occur at any point, such as at the transmitting station, the receiving station, or at some central point such as a central office if the system is telephonic. In the latter case, one speech processor pursuant to the present invention can be constructed to process a multiplicity of reverberative speech signals that are routed through the office.

THE DRAWING

FIG. 1 is a schematic block diagram of the entire inventive process in combination with a communications transmission network.

FIG. 2 is a schematic circuit diagram of a unique computer.

FIG. 3 is a schematic circuit diagram of an inverse filter.

FIG. 4 is a schematic circuit diagram of an excitation analysis/synthesis unit.

FIG. 5 is a schematic circuit diagram of a second correlation computer.

FIG. 6 is a schematic circuit diagram of a synthetic speech generator.

FIG. 7 is a schematic circuit diagram of a peak selector portion of said excitation analysis/synthesis unit.

FIG. 8 is a graph depicting resonant frequencies.

Fig. 9 is a graph depicting an aspect of a typical room transfer function.

THEORY OF THE INVENTION

A greater understanding of the illustrative embodiment will be gained by first more fully considering the theory of the invention and definitions of certain terms.

Excitation Signal or Driving Function $e(t)$

In order to cause an output at the mouth from the human vocal tract, the vocal cords of the glottis are excited to produce pulses recurring at a quasiperiodic rate. The sounds so produced are voiced. Other sounds are unvoiced, such as *sss*, *fff*, *p*, and *k*. The latter are formed by turbulent air at the mouth, throat, and lips without vocal cord excitation. The voiced and unvoiced sounds in total are the signal source from which human speech originates, and are called the excitation signal $e(t)$. In order to generate an output from a model of the vocal tract, such as a filter with transfer function $T(\omega)$, an excitation signal must be applied. Speech so produced is of course synthetic. The excitation signal, which in sampled form is herein denoted $e(n)$, may consist of a pulse generator with a variable pulse period and a white noise source, selectively applied to a variable gain amplifier. The pulse generator supplies the excitation signal for voiced sounds, and the "white" noise source supplies the excitation signal for unvoiced sounds.

Pursuant to this invention, a variable pulse period, a voiced-unvoiced decision and a gain factor may all be derived from the reverberated excitation function such as $e(n)$. These data will provide all the information necessary to drive a vocal tract filter and hence synthesize speech.

Vocal Tract Transfer Function $T(\omega)$

It has been demonstrated by Atal and Hanauer that the human vocal tract may be accurately modeled as an all-pole filter $T(\omega)$ which closely approximates the transfer properties of the vocal tract. Such a filter has a transfer function in the frequency domain given by:

$$T(\omega) = 1/1 - a_1z^{-1} - a_2z^{-2} - a_3z^{-3} \dots - a_{14}z^{-14} \quad (2)$$

where

$$z = \exp(i\omega/\omega_s) \quad (2a)$$

in which

$$i = \sqrt{-1}$$

ω_s = the radian sampling frequency

ω = the radian frequency

The number 14 in Equation (2) is a typical value. Equation (2) is the reciprocal of a polynomial where the zeros formed for a given set of coefficient values $a_1, a_2, a_3, \dots, a_{14}$ determine the frequencies where $T(\omega)$ has its maximum values. The latter frequencies are the resonant frequencies or poles of the filter shown in FIG. 8 as ω_1, ω_2 , etc.

If a driving function $e(t)$ comprising a specified combination of periodic impulses and white noise is applied to such a filter, a speech signal $s(t)$ will result. In sampled form, we denote s_n (n an integer) as the speech samples, and e_n as the driving signal samples. Then, as in Equation (1), since:

$$s_n = \sum_{k=1}^{14} a_k s_{n-k} + e_n \quad (3)$$

(where s_n is the output of this filter and e_n is the input), the resulting s_n is the output of the vocal tract further defined by the a_k coefficients and being driven by e_n . Equation (3) is an application of the method of linear digital filtering, and it states that the present output value of s_n can be estimated from a weighted sum of past (or delayed) output values (s_{n-k}) plus the new input value e_n of the driving function.

Room Transfer Function $H(\omega)$

It is well known that an enclosure such as a room is a linear system. This means that the effect a room has on a signal such as speech is to cause numerous filtered delays of the signal in its travel to a stationary microphone, via many diverse path lengths. All the delayed signals are additively combined by a microphone placed in the enclosure. For example, two unit amplitude sinusoidal signals $\cos \omega_1 t$ and $\cos \omega_2 t$ launched in an enclosure will be recovered by a microphone, or perceived by a listener, with an altered amplitude b_1, b_2 ; and each will have been delayed by an amount expressible as respective phase angles ϕ_1 and ϕ_2 . Thus:

$$\begin{aligned} \cos \omega_1 t &\rightarrow b_1 \cos (\omega_1 t + \phi_1) \\ \cos \omega_2 t &\rightarrow b_2 \cos (\omega_2 t + \phi_2) \end{aligned} \quad (4)$$

where ϕ and b are functions of the frequency ω and the location of the microphone and loudspeaker, but not otherwise a function of time.

Thus, a room transfer function is an expression that in one respect describes how signals of various given frequencies will be relatively affected in amplitude and phase by being propagated in the room. FIG. 9 depicts a typical room transfer function. FIG. 9 illustrates the fact that propagated frequencies differing by as little as 2 Hz may differ substantially in power at a stationary point remote from the source by 40 dB. It can be seen that a room is a filter, and its transfer function is that of a filter.

The problem of describing how an enclosure affects propagated acoustic wave signals may be approached analytically either in terms of the frequency response $H(\omega)$, as depicted in FIG. 9; or in terms of impulse response $h(t)$ where H is a complex number and $h(t)$ is a real time-varying signal amplitude.

Given $H(\omega)$, one can, by a Fourier transform, derive the function $h(t)$. Also, given the problem of a time-

varying input signal of amplitude $s(t)$ passing through a reverberant enclosure having a known impulse response $h(t)$, the output signal $w(t)$ may be predicted as:

$$w(t) = s(t) * h(t) \quad (6)$$

where the symbol $*$ denotes convolution. Likewise, given the enclosure input signal $s(t)$ by its Fourier transform $S(\omega)$ and the frequency response $H(\omega)$ of the enclosure, the output frequency response $W(\omega)$ is:

$$W(\omega) = H(\omega) \cdot S(\omega) \quad (7)$$

DETAILED DESCRIPTION OF AN ILLUSTRATIVE EMBODIMENT

Theory will now be applied by reference to a typical reverberant chamber shown as a four-sided room 10 in FIG. 1 wherein a speaker is speaking from location 11 and his speech is received by a microphone 12 remotely placed. In general, as the distance between a speaker and the receiving microphone is increased, beginning at a point typically beyond a few inches a rain-barrel-like quality to the speech will be increasingly evident at the microphone location and hence, of course, also at any receiver connected thereto.

If the speaker at location 11 is characteristic of most male and female adults speaking English, his articulation times are commonly of the order of less than 1 second. For the typical reverberant closure such as room 10, the 60 dB energy decay time is less than one second; and room dimensions are at least 100 times greater than the dimensions of a human vocal tract.

Summary of Process

FIG. 1 depicts the continuous operation of the inventive process, in a sequence of stages. First, the values of the a_k terms of Equation (1) are calculated from w_n by the correlation computer 31 and the coefficient computer 30. The a_k terms, which number typically 14, constitute an estimate of the time-varying filter which approximates, in its transfer function $T(\omega)$, the speaker's vocal tract. Then, the reverberated signal w_n and the varying values of a_k are applied to inverse filter 32 which determines the reverberated driving function e_n of the speech w_n . In excitation analysis-synthesis unit 27, the driving function e_n is analyzed and then synthesized, thereby rendered clean, or dereverberated. The resulting function $\bar{e}_n$ is a characterization of e_n in terms of three slowly varying band-limited functions of time: pitch, RMS level, and voicing. Finally, unreverberant speech signals s_n are synthesized in filter 33 from the driving function $\bar{e}_n$ and from the parameters a_k computed in coefficient computer 30.

Process Details

The reverberant speech is forwarded via a transmission network 34 to the intended receiving point such as telephone 39. For simplicity, the network 34 is shown as separate from the speech processor; but obviously the process could be located within network 34, such as in a central office. Similarly, telephone 39 includes a direct connection to network 34 and an indirect connection thereto via the processor, thus to indicate that the processor could be an add-on feature located at the telephone.

Advantageously, at or near the point where the processing is to occur, the reverberant speech $w(t)$ is first low-pass filtered in filter 37. The latter is a 5 kHz filter, designed to the proposition that human speech infor-

mation is sufficiently specified within the frequencies below 5 kHz. The low-pass filtered speech signal is then sampled in sampler 38 at a 10 kHz rate, in keeping with the Nyquist sampling theorem. The output of sampler 38 is a stepwise succession of voltages whose amplitudes are indicative of the low-pass filtered speech signal strength at times corresponding to the sampling times. This output is w_n , n being an integer analogous to time.

The sampler 38 output w_n is fed to correlation computer 31. On a continuous (every sample) or periodic (for example every 6 ms) basis, computer 31 forms the following combinations of the input:

$$R_l = \sum_{r=1}^{\infty} w_r \cdot w_{r-1} \cdot h_r; l=1, 2, 3, \dots P \quad (8)$$

where P is typically 14 and h_r is the impulse response of each of the filters 31c.

FIG. 2 depicts the structure of correlation computer 31. The sampled signal w_n is introduced through a shift register consisting of p stages. The samples are delayed one sample value per stage, z^{-1} being notation signifying a delay of one sample. Thus, signals w_{n-1} , w_{n-2} , w_{n-3} , $\dots$, w_{n-14} (if $p=14$) are present at a given time as outputs of successive delay stages of register 31a. These signals are each multiplied separately in respective multipliers 31b, by the present speech sample w_n giving $w_n \cdot w_{n-r}$, $r=1, 2, \dots, P$. The outputs of amplifiers 31b are fed respectively to low-pass filters 31c which are for example 20 Hz filters and which average over the respective filter inputs with a weighting defined by their impulse response h_n . The outputs of filters 31c, designated $R_1^{(n)}$, $R_2^{(n)}$, $\dots$, $R_{14}^{(n)}$, are fed to coefficient computer 30.

Coefficient computer 30 sets up and solves a set of linear simultaneous equations for the values ($a_1, a_2, a_3, \dots, a_{14}$) of a_k . These equations are:

$$R_0 a_1 + R_1 a_2 + \dots + R_{13} a_{14} + R_1$$

$$R_1 a_1 + R_0 a_2 + \dots + R_{12} a_{14} = R_2$$

$$R_{13} a_1 + R_{12} a_2 + \dots + R_0 a_{14} = R_{14} \quad (9)$$

A particular suitable computation apparatus to solve the equations denoted (9) is described in Atal U.S. Pat. No. 3,631,520 issued Dec. 28, 1971. The derivation and significance of these equations has been discussed by Atal.

The output values ($a_1, a_2, a_3, \dots, a_{14}$) of computer 30 are fed to inverse filter 32 shown in FIG. 3, as is the sampled reverberant speech signal w_n . In inverse filter 32, the reverberant excitation signal is formed pursuant to the following equation:

$$e_n = w_n - \sum_{k=1}^P a_k w_{n-k} \quad (10)$$

It will by now be appreciated that e_n is the result of the clean speech excitation function reverberated by the room transfer function $H(\omega)$, since the vocal tract has been removed by inverse filter 32.

FIG. 3 depicts the structure of inverse filter 32 as consisting of a delay line or shift register 32a of p stages

where p equals the number of stages, for example, 14 of register 31a. The input to shift register 32 is the sampled signal w_n . As in shift register 31a, the samples are delayed one sample value per stage. The delayed samples are picked off successive stages of register 32a and respectively led to multipliers 32b. The inputs to respective multipliers 32b are the values ($a_1, a_2, a_3, \dots, a_{14}$) calculated in computer 30. The outputs of all multipliers 32b are combined in adder 35a; and from this sum, the value of sampled speech w_n is subtracted in subtractor 35b. The subtractor 35b output, e_n , is a driving function in sampled form of the original unreverberant speech signal s_n , reverberated by the effects of enclosure 11.

The next step in the inventive process involves "dereverberating" the excitation function e_n . The unit which performs the preceding is excitation analysis/synthesis unit 27, shown in FIG. 4. Its purpose is to synthesize, from a revamping of the driving function e_n , a "clean" driving function $\bar{e}_n$. Driving function e_n is first autocorrelated in correlation computer 20 to determine any dominant periodicities or lack thereof. This process involves the apparatus of FIG. 5 which computes the result:

$$R_l^{(n)} = \sum_{p=n}^{\infty} e_p e_{p-r} \cdot y_{-p} l=1, 2, \dots l \quad (11)$$

y_{-p} is the impulse response of the low-pass filters 20c; r runs over the range of possible pitch periods, i.e., 3-13 ms.

As seen in FIG. 5, the sampled driving function e_n is introduced through a shift register 20a consisting of l stages where l corresponds to delays of up to 13 ms in keeping with the largest pitch periods which may be encountered. The samples are delayed one sample value per stage. Thus, signals e_{n-1} , e_{n-2} , e_{n-3} , $\dots$, e_{n-l} are present at a given time at the output of the successive stages of shift register 20a. These signals are each multiplied separately in respective multipliers 20b by the quantity e_n . The outputs of the respective multipliers 20b are low-pass filtered in respective filters 20c which are 20 Hz, for example, selected because of the inherent slowly varying nature of the correlation.

The outputs of the respective filters 20c for each l sample are a set of numbers $R(\tau_1), R(\tau_2), R(\tau_3), \dots, R(\tau_l)$ which are a measure of the degree of correlation for delays τ_i . The delay τ_i corresponding to the maximum of the just-performed autocorrelation is now ascertained in peak picking selector 16 seen in FIG. 7. The maximum $R(\tau_i)$ value among $R(\tau_1), R(\tau_2), \dots, R(\tau_l)$ is selected, and the delay τ_i associated with that largest value of R , denoted τ_{peak} , is used as the pitch period parameter required by pulse generator 13.

Additionally, selector 16 includes a threshold detector 36, which inspects the values of each signal $R(\tau_{peak})$ to determine whether the driving function at that time is voiced or unvoiced.

The output of threshold detector 36 is a binary level signal which is fed to voicing switch 15. Also, the output τ_{peak} of selector 16, which represents the pitch period, is fed to pulse generator 13. The latter can, for example be an astable oscillator of variable period well known to the state of the art.

Pulse generator 13 waits τ_{peak} samples with zero output, then produces a unit amplitude output. The output of generator 13 is connected to voicing switch 15.

White noise generator 14 is a conventional noise generator creating power of all frequencies at equal levels. Its output is also connected to voicing switch 15. When threshold detector 36 determines that a voiced excitation has occurred, it supplies an order to voicing switch 15 to effect a connection to pulse generator 13. Otherwise, when detector 36 identifies presence of an unvoiced excitation, it causes a connection of voicing switch 15 to white noise source 14. The output side of voicing switch 15, denoted δ_n , is a fixed amplitude source signal which is either a sequence of pitch pulses at the given pitch period, or a burst of white noise.

The third parameter derived in unit 27 is a gain factor, denoted G in FIG. 4, which amplitude modulates or multiplies the fixed amplitude source signal δ_n by an amount that makes the result $\bar{e}_n$ identical in mean-squared (MS) level to the reverberant driving signal e_n . The latter is the quotient, calculated in divider 25, of a dividend MS (e_n) and a divisor MS (δ_n). The value MS (e_n) is generated by feeding the sampled signal e_n through squarer 21 and thence 20 Hz low-pass filter 22. The value MS (δ_n) is generated by feeding the sampled signal δ_n through squarer 23 and 20 Hz low-pass filter 24.

The quotient of these two, namely gain factor G , is continuously applied to the signal δ_n through variable gain amplifier 26. The output of the latter is $\bar{e}_n$ which approximates the driving function of the original unreverberant speech $s(t)$ in 10. It remains now to synthesize the clean speech; and this is accomplished in all-pole filter 33.

The all-pole filter 33 seen in FIG. 6, is a vocal tract model such as taught by Atal in his U.S. Pat. No. 3,624,302. Filter 33 consists of the delay line shift register 33a having for example 14 stages, each stage causing a delay z^{-1} ; and a corresponding number of multipliers 33b connected between the respective stages. The coefficients $a_1, a_2, a_3, \dots, a_{14}$ derived in coefficient computer 30 are supplied to the respective multipliers 33b. The combination of delay line shift register 33a and multipliers 33b, designated 29 in FIG. 6 are known in the art as a transversal delay line.

In transversal delay line 29, the terms $a_k s_{n-k}$ are calculated. They are then summed by summer 28 along with the clean driving function $\bar{e}_n$ giving the output stated in Equation (1). The result is the synthesized speech signal s_n , free of reverberative effects.

The digital signal s_n at the output of all-pole filter 33 may be converted to an analog version by the conventional technique of low-pass filtering at half the sample frequency for use in driving the receiver of telephone 39, for example.

Multiple-Microphone Signal Pickup

Although the invention has so far been described as operating with a single microphone 12, arrays of plural microphones can also be used to advantage. The benefit of microphone arrays is understood by recognizing that a better estimate of the parameters is attained through the availability of more data. For this case, each new microphone requires its own correlation computer 31. The new outputs from this computer $R'(\tau_1), R'(\tau_2), \dots, R'(\tau_{14})$ are added to the other $R(\tau)$'s of other microphones thus giving more accurate data

for coefficient computer 30. The signal from each microphone is then inverse filtered and applied to its own autocorrelation computer 20 to derive signals $R''(\tau_1), R''(\tau_2), \dots$ etc. These signals are then averaged with those from other microphones to produce a composite error autocorrelation function to be analyzed as before by peak selector 16.

It is to be understood that the embodiments described herein are merely illustrative of the principles of the invention. Various modifications may be made thereto by persons skilled in the art without departing from the spirit and scope of the invention.

What is claimed is:

1. Apparatus for synthesizing speech comprising:
 - transducer means located within a reverberant enclosure remotely from a speaker therein, for receiving reverberated speech signals from said speaker;
 - means for continuously deriving, from said reverberated speech signals, first signals representative of the vocal tract transfer function of said speaker;
 - means for developing, from said reverberated speech signals and said first signals, second signals representing the reverberated excitation source of said speaker;
 - means for dereverberating said second signals; and
 - means for developing, from said first signals and said dereverberated second signals, synthetic speech signals substantially approximating said speaker's original speech.
2. Apparatus for constructing an undistorted replica of a speaker's original speech uttered in a reverberant enclosure comprising
 - means for continuously extracting from the reverberant speech signal an approximation of the vocal tract transfer function of said speaker;
 - means for removing from said reverberant speech the effect of said vocal tract transfer function, the resulting residual signal being substantially the speaker's reverberated excitation function;
 - means for deriving, from said reverberated excitation function,
 - a first parameter denoting the voiced or unvoiced nature of said excitation function;
 - a second parameter denoting the pitch period of voiced portions of said excitation function; and
 - a third parameter denoting the mean-squared level of said excitation function; and
 - means for combining said first, second, and third parameters with said vocal tract transfer function to produce said undistorted replica.
3. Apparatus pursuant to claim 2 wherein said extracting means further comprises means for recurrently estimating a sequence of weighting coefficients a_k which constitute a unique estimate of a time-varying filter that approximates in its transfer function $T(\omega)$ said speaker's vocal tract.
4. Apparatus pursuant to claim 3 wherein said removing means comprises an inverse filter having as its inputs said weighting coefficients a_k and said reverberant speech signal.
5. Apparatus pursuant to claim 4 wherein said means for deriving said second parameter comprises means for autocorrelating said excitation function to determine a maximum value, and means for ascertaining a unique delay associated with said maximum value, said delay constituting the said pitch period parameter.

11

6. Apparatus pursuant to claim 5 wherein said means for deriving said first parameter comprises a fixed threshold detector for inspecting the level of each of said maximum values resulting from said autocorrelation of said excitation function, and means for producing a voiced- unvoiced decision based on whether a given said maximum value exceeds or falls below said fixed threshold.

7. Apparatus pursuant to claim 6 further comprising:
a white noise source,
a variable period pulse generator,
means for transmitting said pitch period parameter to said pulse generator to control the period of said pulses, and

a voicing switch responsive to the output of said threshold detector for selecting a its output either said white noise source or said pulse generator, thereby to produce a fixed amplitude source signal.

8. Apparatus pursuant to claim 7, wherein said means for generating said third parameter comprises means for multiplying said fixed amplitude source signal by an amount that renders the speaker's said reverberated excitation function identical in mean-squared level with said fixed amplitude source signal, the result being a synthetic unreverberated excitation function.

9. Apparatus pursuant to claim 8, wherein said combining means comprises an all-pole filter having as its inputs said weighting coefficients a_k and said unrever-

12

berated excitation function, the output of said combining means constituting the speaker's synthesized speech signal free of reverberative effects.

10. A speech dereverberation system for an enclosure characterized by a fixed transfer function $H(\omega)$, comprising:

a source of white noise;

a source of electrical impulses;

a recursive filter having a variable transfer function;
means for receiving a reverberative speech signal generated from within said enclosure;

means for deriving, from successive discrete sample sets of said speech signal, a specific said filter setting representing a currently valid vocal tract transfer function;

means for deriving an indicia of the pitch period and an indicia of the ratio of voiced-to-unvoiced intervals in said speech signal;

means for applying said pitch period indicia to said impulse source to control the impulse generation rate;

a voiced-unvoiced interval ratio detector having a fixed threshold level; and

means including said detector for applying said impulses to said recursive filter when said threshold is exceeded and otherwise for applying said noise to said filter.

* * * * *

30

35

40

45

50

55

60

65