July 23, 1968 J. L. FLANAGAN ETAL 3,394,228

APPARATUS FOR SPECTRAL SCALING OF SPEECH

Filed June 3, 1965 5 Sheets-Sheet 1

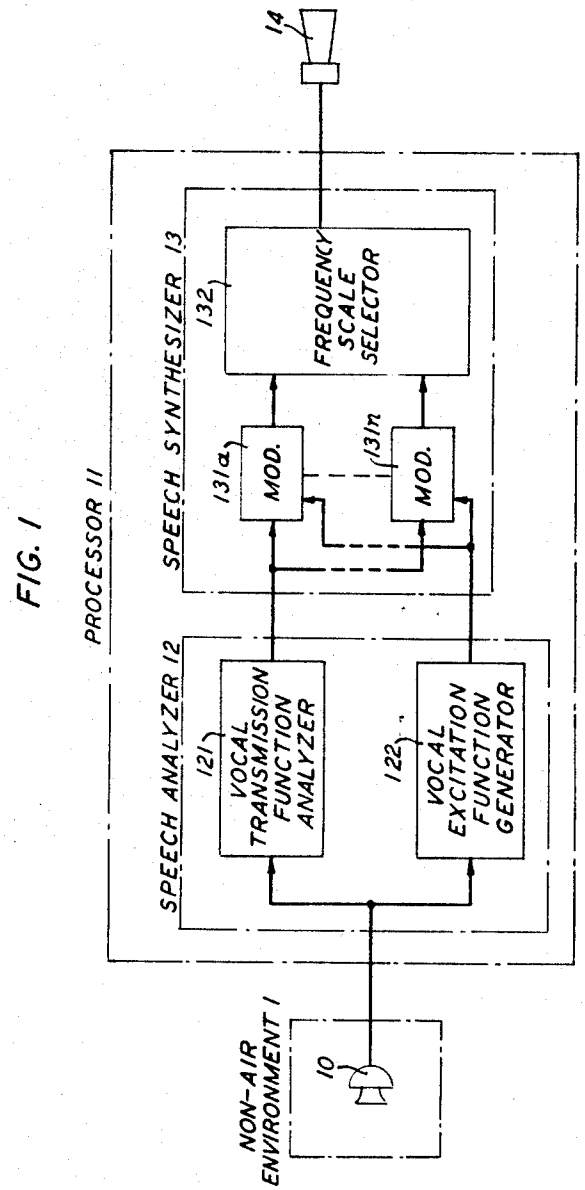


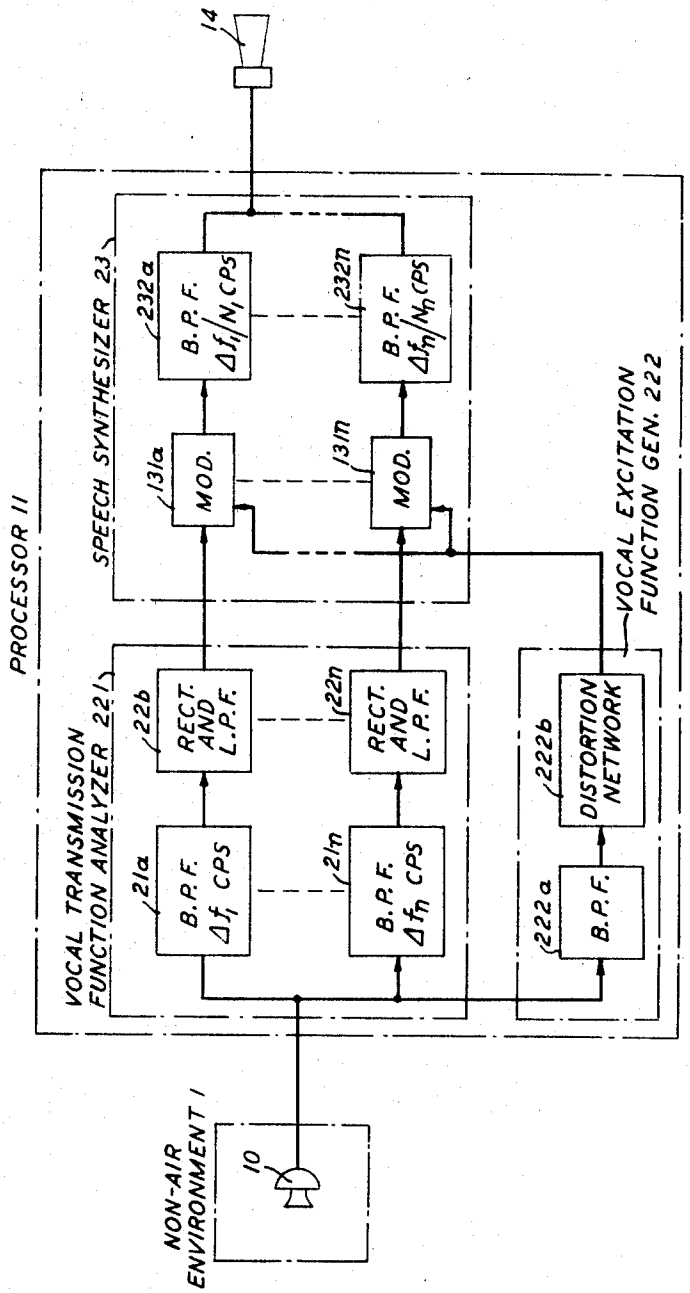
FIG. 1

INVENTORS: J. L. FLANAGAN
M. R. SCHROEDER
BY Jules P. Kirsch
ATTORNEY

APPARATUS FOR SPECTRAL SCALING OF SPEECH

Filed June 3, 1965 5 Sheets-Sheet 2

FIG. 2A



APPARATUS FOR SPECTRAL SCALING OF SPEECH

Filed June 3, 1965 5 Sheets-Sheet 3

FIG. 2C

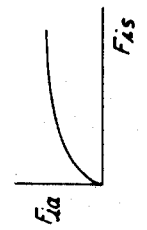
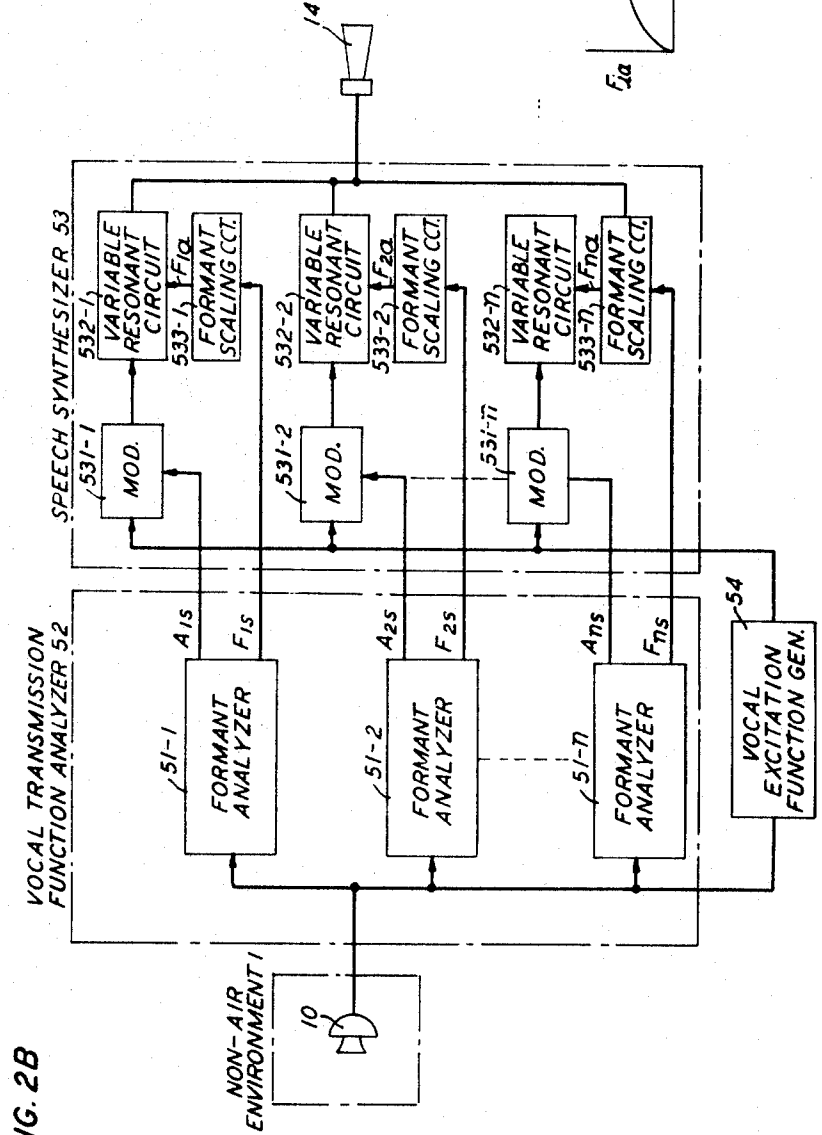


FIG. 2B



APPARATUS FOR SPECTRAL SCALING OF SPEECH

Filed June 3, 1965 5 Sheets-Sheet 4

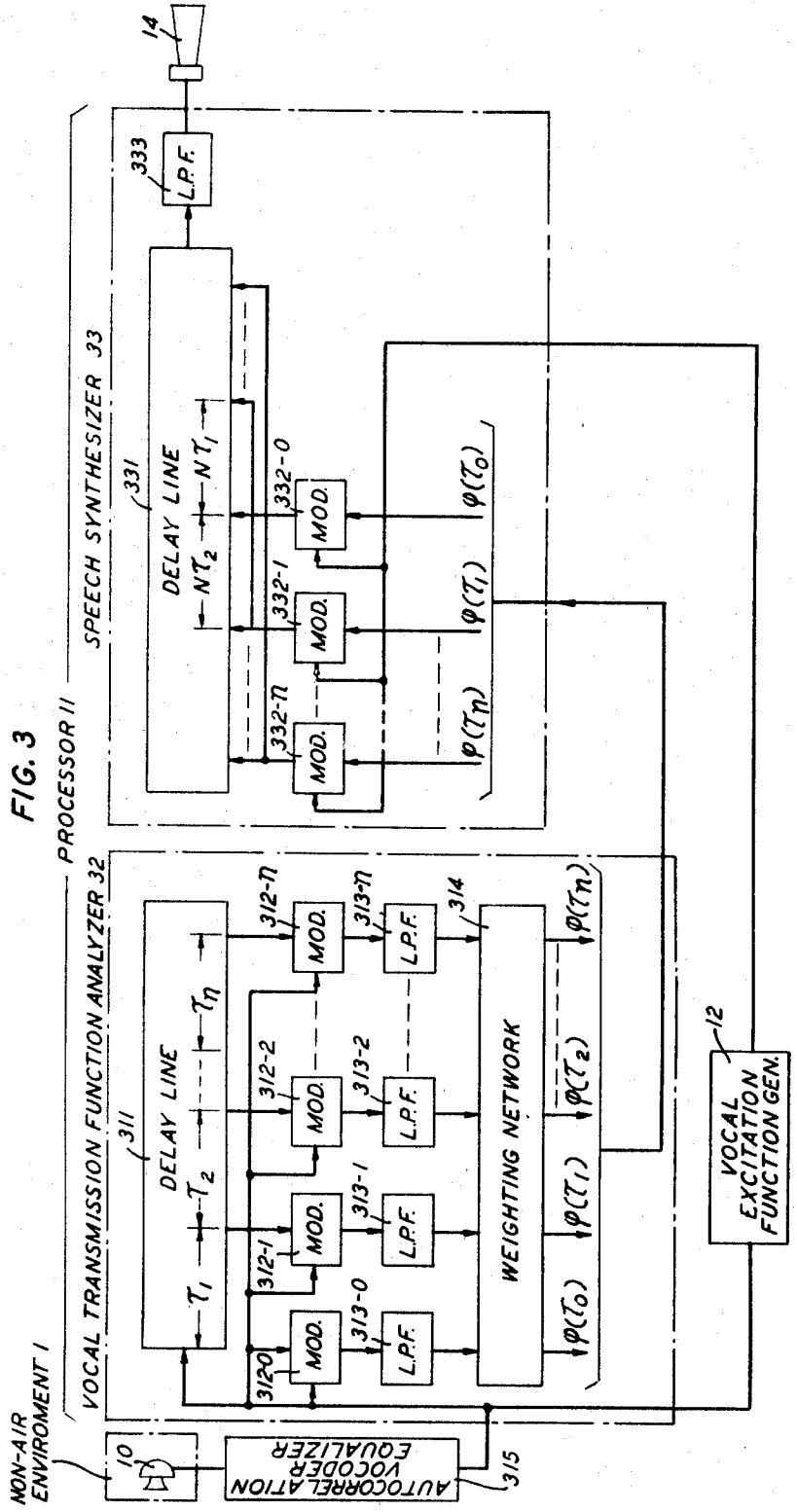


FIG. 3

FIG. 4A

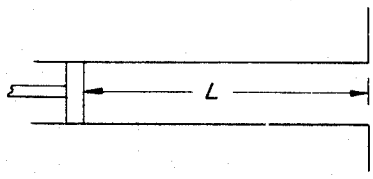


FIG. 4B

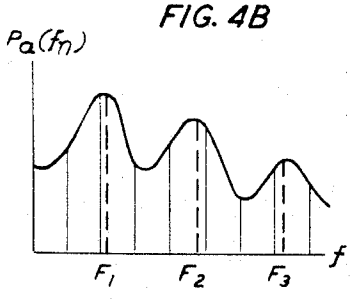


FIG. 4C

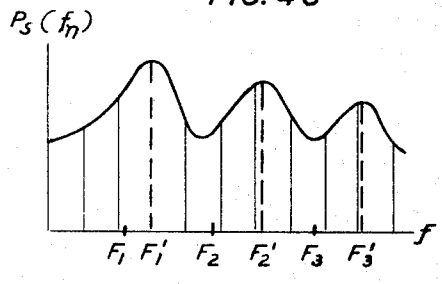


FIG. 5A

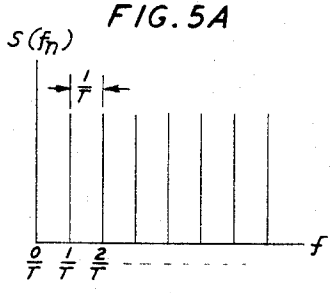


FIG. 5B

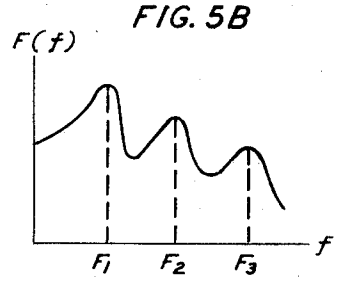


FIG. 5C

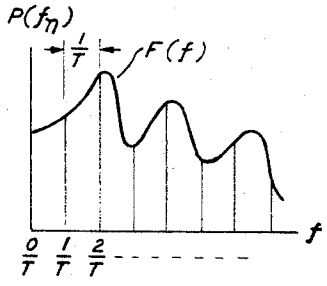
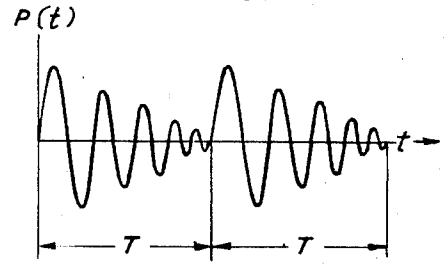


FIG. 5D



United States Patent Office 3,394,228 Patented July 23, 1968

3,394,228

APPARATUS FOR SPECTRAL SCALING OF SPEECH

James L. Flanagan, Warren Township, Somerset County, and Manfred R. Schroeder, Gillette, N.J., assignors to Bell Telephone Laboratories, Incorporated, New York, N.Y., a corporation of New York

Filed June 3, 1965, Ser. No. 460,930
8 Claims. (Cl. 179—1)

ABSTRACT OF THE DISCLOSURE

The intelligibility of speech uttered in an "altered atmosphere" environment, such as helium under pressure, may be improved by dividing a speech signal into its excitation function and its transmission function, altering the formants in the transmission function, and synthesizing the altered speech signal. The alteration of the formants may be linear or nonlinear and the degree of alteration required is empirically determined.

This invention relates to speech transmission systems, and in particular to systems for providing intelligible, natural sounding speech communication in an environment where the speaker is breathing a gas characterized by a different sound velocity than that of air at temperatures and pressures customarily found at the surface of the earth.

It is well known that human speech is produced by a broad-spectrum excitation of the human vocal tract, with voiced sounds being produced by quasi-periodic puffs of air released from the lungs into the vocal tract by the glottis or vocal cords, and unvoiced sounds being produced by the turbulent flow of air from the lungs through constrictions in the vocal tract. Thus in system-function terms, human speech is the product of a vocal excitation function describing the characteristics of the air released into the vocal tract, and a vocal transmission function describing the characteristics of the vocal tract. There are a number of characteristics that distinguish different speech sounds for each other, but one of the most important in recognizing the same sound enunciated by different talkers is the so-called formant pattern associated with each sound. The formant pattern refers to the resonances or poles of the vocal transmission function, and one of the ways in which a speaker produces a variety of speech sounds is by varying the shape and constrictions of the vocal tract to change the resonant frequencies. Since the vocal tract resonances are manifested in the amplitude spectrum of a voice wave by several principal peaks or maxima in the spectral envelope, one of the distinguishing characteristics of various sounds is the set of frequencies at which the principal spectral peaks associated with each speech sound occur.

The importance of formant patterns in speech recognition has been observed in measurements of speech produced in a non-air environment such as that required in deep-sea diving. In a letter entitled "Helium Speech," by K. Holywell and G. Harvey, vol. 36, Journal of Acoustical Society of America, page 210 (1964), it was pointed out that a talker breathing a mixture of oxygen and helium produces speech with severely impaired intelligibility. Analysis of the distorted speech produced in this environment showed that the formant patterns had been shifted from their normal frequencies in air, but that pitch is relatively unaffected. Several techniques were suggested in the above-mentioned letter for processing the impaired speech to restore the formant patterns to their normal locations and substantially improve intelligibility. However, the suggested processing techniques require recording and subsequent playback of the distorted speech, hence they do not provide "real time" or instantaneous speech communication for a talker in a non-air environment. Further, these processing techniques were based on the assumption of a linear shift in formant frequencies, but a purely linear processing does not remove other significant nonlinear distortions of formant patterns which may occur in a non-air environment. In addition, the suggested processing introduced unwanted distortion by altering the pitch of the talker's speech.

The present invention provides an arrangement for restoring formant patterns in distorted speech to their normal frequency locations while a talker is speaking in a non-air environment so that instantaneous speech communication may be maintained. Further, this arrangement is not limited to removing unwanted linear distortion of the frequency scale of the speech spectrum, but may be adapted to remove unwanted nonlinear distortion as well. Moreover, the arrangement provided by this invention maintains the original pitch of the talker's speech.

These features are realized in the present invention by separating a distorted speech wave into its vocal excitation and vocal transmission characteristics, following which there is synthesized from the two characteristics a synthetic speech wave having an amplitude spectrum in which each of the principal spectral peaks is displaced in frequency by a predetermined amount from the corresponding spectral peak of the distorted speech wave. The predetermined amount by which each synthetic spectral peak is displaced in frequency from the corresponding spectral peak of the distorted speech wave is selected to take into account both linear and nonlinear shifts in resonant frequencies of the human vocal tract, so that the synthetic spectral peaks occur at frequencies that are normal to an air environment. In the present invention, therefore, the synthetic speech wave delivered to a listener is an intelligible replica of the talker's distorted speech wave, and since the synthetic speech wave is developed almost instantaneously, a talker in a non-air environment is able to maintain immediate speech communication with others. In addition, because the original excitation characteristic is separated from the vocal transmission characteristic in the present invention, the correction of formants does not affect the original excitation characteristic, and hence the original pitch is preserved in the synthetic speech wave.

The invention will be more fully understood from the following detailed description of illustrative embodiments thereof taken in connection with the appended drawings, in which:

FIG. 1 is a schematic block diagram illustrating in outline form the principal features of the principles of this invention;

FIG. 2A is a schematic block diagram illustrating in detail certain features of a particular frequency domain arrangement embodying the principles of this invention;

FIG. 2B is a schematic block diagram illustrating in detail a frequency domain embodiment alternative to that shown in FIG. 2A;

FIG. 2C is a graph of assistance in explaining the operation of the apparatus shown in FIG. 2B;

FIG. 3 is a schematic block diagram illustrating a time domain arrangement embodying the principles of this invention;

FIG. 4A is a drawing of a simplified mechanical model of the human vocal mechanism;

FIGS. 4B and 4C are diagrams of speech amplitude spectra which are of assistance in explaining the kind of

distortion that is corrected by this invention; and

FIGS. 5A, 5B, 5C, and 5D are graphs which are of assistance in explaining the system-function analysis of speech.

THEORETICAL CONSIDERATIONS

Turning first to FIGS. 5A, 5B, 5C, and 5D, these drawings illustrate graphically the idealized system-function analysis of voiced speech sounds, in which human speech is specified as the product of a voiced vocal excitation function $S(f_n)$, shown in FIG. 5A, and a vocal transmission function, $F(f)$, shown in FIG. 5B, that is

$$P(f_n)=S(f_n)\cdot F(f) \quad (1)$$

The speech wave $p(t)$ resulting from the product of these two functions has the amplitude spectrum, $P(f_n)$, shown in FIG. 5C, in which it is observed that the uniformly spaced individual harmonic components of the excitation function, which are indicated by the vertical lines in FIG. 5A are adjusted in amplitude as a result of multiplication by the transmission function to become the harmonic components of $P(f_n)$ with the same uniform spacing. Although the drawings and the analysis illustrate the principles of this invention in terms of voiced speech sounds, it is to be understood that these principles are equally applicable to unvoiced speech sounds produced by a broad spectrum noise excitation of the vocal tract.

The principal resonances or formants of the vocal tract are manifested by peaks in the vocal transmission function at frequencies denoted F_1, F_2, and F_3 in FIG. 5B, and these formant peaks are carried over into the speech amplitude spectrum as peaks in the spectral envelope. As described in the appendix, the amplitude spectrum $P(f_n)$ and the vocal transmission function $F(f)$ are proportional to one another at the harmonic frequencies, $f_n=n/t$. Thus in the frequency domain the vocal transmission function is identified with the envelope of the speech spectrum, while the vocal excitation function is identified with the frequency spacing between the individual frequency components of the speech spectrum.

The vocal excitation and transmission function characteristics also manifest themselves in the time domain, as shown by the speech time waveform in FIG. 5D which has the spectrum shown in FIG. 5C. The time waveform $p(t)$ shown in FIG. 5D is periodic with period T, the period being determined by the spacing between successive harmonic components, $1/T$, of the excitation function shown in FIG. 5A.

Referring now to FIG. 4A, this drawing shows a simplified approximation of the human vocal mechanism in which the vocal tract is represented by a uniform diameter hollow pipe or tube of length L, having at one end a vibrating piston representing the glottis or vocal cords, and an opening at the other end representing the mouth. It is well known that the natural frequencies or resonances, F_n, of such a pipe are, ideally, those frequencies for which the length L, is an odd multiple of a quarter wavelength of sound waves in the medium filling the pipe,

$$L=\frac{(2n+1)\lambda}{4},\ n=0,\ 1,\ 2\ \ldots \quad (2)$$

where λ denotes sound wavelength. Similarly, it is well known that the sound wavelength, λ, is a function of both the velocity of sound, c, in the medium, and the frequency of the sound wave, f, according to the following relation:

$$\lambda=\frac{c}{f} \quad (3)$$

Combining Equations 2 and 3, the resonant frequencies may be expressed in terms of the sound velocity and the length of the pipe,

$$F_n=\frac{(2n+1)c}{4L},\ n=0,\ 1,\ 2\ \ldots \quad (4)$$

From Equation 4, it is evident that the resonant frequencies are directly proportional to the sound velocity in a particular medium, so that for mediums having different sound velocities, the resonances of the pipe will occur at different frequencies.

Further, the velocity of sound is not a constant even for a specific gas medium, but is a nonlinear function of the ambient pressure, p_0, and the ambient density, ρ_0, as given by the relationship

$$c=\sqrt{\frac{V\cdot p_0}{\rho_0}} \quad (5)$$

where V is the adiabatic constant for the gas.

From the above discussion it is clear that the formant patterns or resonances of human speech sounds are dependent upon the characteristics of the medium in which a talker is speaking, that is, the characteristics of the gas which is being breathed at the time that a talker is speaking. This is illustrated graphically in FIGS. 4B and 4C, in which a speech sound uttered in air has an amplitude spectrum $P_a(f_n)$ with an envelope having formant peaks at one set of frequencies, F_1, F_2, F_3, as shown in FIG. 4B, while the same sound uttered in helium has an amplitude spectrum $P_s(f_n)$ with an envelope having formant peaks at a different set of frequencies, F_1', F_2', F_3', as shown in FIG. 4C. It is this shift in formant locations which is thought to be primarily responsible for the loss of intelligibility that occurs in a non-air medium.

Although Equation 4 specifies that a change in sound velocity affects all resonant frequencies by the same factor, it was pointed out in the "Helium Speech" reference mentioned above that a linear factor alone does not account for the changes in formant frequencies of speech sounds uttered in a non-air environment. Investigations have shown that non-linear factors such as the side-wall impedance of the vocal tract and the vocal cord source also influence the resonant frequencies of the vocal tract; for example, see Quarterly Progress Report, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden, page 9 (July 15, 1964), and vol. 36 Journal of the Acoustical Society of America, page 2,001 (1964).

Therefore, in order to improve the intelligibility of a speech sound uttered in a non-air environment, it is necessary to take into account both the linear distortion of formant frequencies attributable to the difference in sound velocity specified by Equation 4 and the nonlinear distortion of formant frequencies attributable to other factors. Further, it is recognized in the present invention that both the linear and the nonlinear distortions that are caused by a non-air environment affect primarily the formant frequency locations and have little, if any, effect on the pitch or frequency of vibration of the human vocal cords. In terms of system-function analysis, a non-air environment distorts human speech by distorting the frequency scale of the speech spectrum, thereby altering the locations of the resonances or formant peaks in the vocal transmission function, $F(f)$. However, a non-air environment does not significantly affect the vocal excitation function, $S(f_n)$. The present invention corrects the distortion of the speech spectrum by separating a speech sound into its vocal excitation function and vocal transmission characteristics, and synthesizing from the separated vocal excitation and transmission functions a synthetic speech wave having the same vocal excitation function as the original speech wave but a synthetic vocal transmission function having formant peaks shifted in frequency to the positions normally occupied in air.

Method and apparatus

Turning now to FIG. 1, this drawing illustrates both the method provided by this invention to improve the intelligibility of speech uttered in a non-air environment and one of several alternative apparatuses for performing this method. A talker in a non-air environment, indicated by the dashed enclosure 1, utters speech sounds

that are converted into an electrical wave by transducer **10**, which may be a conventional microphone. By hypothesis, the speech sounds represented by the electrical wave from transducer **10** have spectra with distorted formant patterns of the type shown in FIG. 4C, and this wave is passed to processor **11** of this invention in order to restore the distorted formant patterns to the frequency positions that are normally occupied by the formants of speech sounds uttered in an air environment.

Within processor **11**, the incoming wave from transducer **10** is applied in parallel to vocal transmission function analyzer **121** and vocal excitation function generator **122**. Analyzer **121** and generator **122** may be any one of a number of well known time or frequency domain devices for separating a speech wave into its transmission and excitation characteristics, respectively. For example, analyzer **21** may be a vocoder analyzer of the type shown in H. W. Dudley Patent 2,151,091, issued Mar. 21, 1939, in which the amplitude of the spectral envelope within selected frequency sub-bands $\Delta f_1'$, $i=1 \ldots, n$, of the speech wave from transducer **10** are represented by a corresponding plurality of n unidirectional control signals. Similarly, vocal excitation function generator **122** may be any one of a number of well known devices, for example, a conventional pitch detector, for deriving a so-called excitation signal having the characteristics of the excitation function of the incoming speech wave from transducer **10**.

The output signals of analyzer **12** are passed to speech synthesizer **13**, in which each of the control signals from analyzer **121** is applied to the control terminal of a corresponding modulator **131***a* through **131***n*, and the excitation signal from generator **122** is applied in parallel to the input terminals of all of the modulators **131***a* through **131***n*. Each of the modulators **131***a* through **131***n*, which may be variable gain amplifiers, adjusts the amplitude of the excitation signal by an amount specified by the control signal, and the amplitude adjusted output signal of each modulator **131***a* through **131***n* is delivered to a frequency scale selector **132**.

The amplitude-adjusted output signals of modulators **131***a* through **131***n* represent the amplitudes of selected samples of either the time domain or frequency domain versions of the vocal transmission function of the incoming speech wave at a given instant, and frequency scale selector **132** individually arranges each of these samples in either the time or frequency domain so that the formants of the synthetic speech wave formed at the output terminal of selector **132** occur at the normal frequency positions for speech uttered in a normal air environment. The arrangement of these samples by selector **132** for a particular non-air medium may be determined by experimental measurement. From the output signal developed by selector **132** natural sounding speech may be reproduced by transducer **14**, for example, a conventional loudspeaker.

Referring next to FIG. 2A, this drawing illustrates a specific frequency domain embodiment of the apparatus show in FIG. 1. An incoming distorted speech wave from source **10** is divided within analyzer **221** into a plurality of contiguous frequency sub-band Δf_1 through Δf_n by bandpass filters **21***a* through **21***n*. The widths of these frequency sub-bands depend upon the manner in which it is desired to represent the transmission characteristic in terms of samples of the spectral envelope of the incoming speech wave. For example, the pass bands of filters **21***a* through **21***n* may be made relatively narrow so that each frequency sub-band will contain only one harmonic component, in which case the spectral envelope will be represented by a relatively large number of frequency domain samples.

Each of the filters **21***a* through **21***n* is followed by a corresponding rectifier and low-pass filter **22***a* through **22***n*, and each of the elements **22***a* through **22***n* derives a so-called control signal representative of the amplitude of the spectral envelope within the frequency sub-band passed by the preceding filter.

The vocal excitation generator **222** may be designed in accordance with the distortion network disclosed in M. R. Schroeder Patent 3,030,450, issued Apr. 17, 1962. Within generator **222** the speech wave from transducer **10** is applied to bandpass filter **222***a* which passes a selected portion of the incoming speech wave to distortion network **222***b*. If desired, the entire speech wave may be passed to network **222***b*, and network **222***b* may comprise the elements shown in FIGS. 3 and 7 of the above-mentioned Schroeder patent, thereby to generate an excitation signal having a relatively broad spectrum that faithfully preserves the periodic or random nature of the source that excited the human vocal tract.

The n control signals from analyzer **221** and the excitation signal from generator **222** are applied to synthesizer **23**, in which each of the n modulators **131***a* through **131***n* is respectively controlled by the corresponding control signal to adjust the amplitude of the excitation signal by an amount specified by the control signal. Thus, each of the amplitude-adjusted output signals of modulators **131***a* through **131***n* has a relatively broad frequency range identical with that of the incoming excitation signal, but each amplitude-adjusted output signal has an amplitude corresponding to a sample of a particular frequency sub-band Δf_i of the spectral envelope of the incoming distorted speech wave. Because of the distortion of the incoming speech wave, each sub-band Δf_i of the original spectral envelope is displaced from its normal, air environment position on the frequency scale by a factor N_i, $i=1, 2 \ldots, n$, where N_i is a factor that may be determined from experimental observations. To compensate for this displacement and thereby remove distortion, filters **232***a* through **232***n* are respectively provided with contiguous pass bands $\Delta f_1/N_1$ through $\Delta f_n/N_n$ so that there is selected from each of the amplitude-adjusted output signals an individual synthetic frequency sub-band that is displaced in frequency from its corresponding sub-band of the spectral envelope of the distorted speech wave, but has the same amplitude as its corresponding sub-band of the distorted spectral envelope. The synthetic frequency sub-bands selected by filters **232***a* through **232***n* are therefore reconstituted samples of the amplitude of the spectral envelope of the original speech wave, but these reconstituted samples have been arranged in the frequency domain by filters **232***a* through **232***n* to define a synthetic vocal transmission function with resonances that occur at normal, air environment frequencies. Hence by combining the individual reconstituted samples from filters **232***a* through **232***n* there is formed a synthetic speech wave having the same vocal excitation function as the original speech wave, and a synthetic vocal transmission function with resonances which occur at frequencies normally associated with speech sounds uttered in an air environment.

Another frequency domain embodiment of the principles of this invention is shown in FIG. 2B, in which the vocal transmission function of an incoming speech wave from transducer **10** is represented in terms of samples of the principal formants in the spectral envelope. The speech wave from transducer **10** is applied to analyzer **52**, in which formant analyzers **51–1** through **51–***n* determine the amplitude and frequency of n selected principal formant peaks in the warped spectral envelope, where n is a positive integer having a typical value of 3. The amplitude and frequency of the ith formant in a non-air environment are represented by a pair of control signals respectively denoted A_{is} and F_{is},

$$i=1, 2 \ldots, n,$$

and it is these signals, together with the excitation signal from generator **54**, that are delivered to synthesizer **53** in order to synthesize a natural sounding replica of the original distorted speech wave.

Within synthesizer **13**, the frequency control signals F_{1s} through F_{ns} are respectively passed through a formant scaling circuit **533–1** through **533–*n*** having an input-output characteristic that is the inverse of the frequency scale distortion caused by the particular non-air environment. This characteristic may be linear or non-linear depending upon the nature of the distortion; FIG. 2C illustrates a non-linear characteristic for converting a distorted formant signal F_{is} into a corrected formant signal F_{ia}. The corrected formant signals F_{1a} through F_{na}, respectively, tune variable resonant circuits **532–1** through **532–*n***. The input signal to each variable resonant circuit is supplied from a corresponding modulator or variable gain device **531–1** through **531–*n***, each modulator being controlled by the corresponding amplitude control signal A_{is} to adjust the amplitude of the excitation signal from generator **54**. In general, correction of the amplitude control signals is not required, but in special circumstances a correction may be applied, if desired, before the amplitude control signals are applied to modulators **531**. Because of the compensating frequency scale factor built into each formant scaling circuit, the output signal of each circuit **532–1** through **532–*n*** is a reconstituted sample of the amplitude of a corresponding formant peak in the spectral envelope of the original speech wave, but each formant sample has been arranged on the frequency scale to define a synthetic vocal transmission function with resonances that occur at normal, air environment frequencies. Hence by adding the reconstituted formant samples from variable resonant circuits **532–1** through **532–*n*** there is formed a synthetic speech wave with the same vocal excitation function as the original speech wave, and a synthetic vocal transmission function with resonances that occupy their normal frequency positions for speech sounds uttered in an air environment.

In situations where there is a linear distortion of the frequency scale of the speech spectrum, the following time domain embodiment of the principles of this invention, shown in FIG. 3, may be employed to restore the formants to their proper positions. In the apparatus shown in FIG. 3, analyzer **32** derives from an incoming speech wave from source **10** a set of samples of each period of the autocorrelation function of the speech wave in the manner shown in E. E. David Patent 3,069,507, issued Dec. 18, 1962. However, in order to prevent distortion due to squaring of the speech amplitude spectrum, the spectrum of the speech wave is square-rooted by passing the speech wave from source **10** through the autocorrelation vocoder equalizer **315** before being applied to analyzer **32**, where equalizer **315** may be of the type shown in M. R. Schroeder Patent 3,091,665, issued May 28, 1963. Thus the signals denoted $\varphi(\tau_0)$, $\varphi(\tau_1)$ . . . , $\varphi(\tau_n)$ represent samples of the autocorrelation function, $\varphi_{h'h'}(\tau)$ of the spectrally square rooted speech wave at selected delay times, $\tau_0=0$, τ_1 . . . , τ_n, as specified by the locations of the taps on delay line **311**; for example, the taps on delay line **311** may be spaced at the Nyquist interval

$$\tau_i=\frac{i}{2W}$$

$i=1, 2 \ldots, n$, where W is smaller than the bandwidth of the incoming speech wave.

Within synthesizer 33, the control signals are applied to the control terminals of modulators **332–0** through **332–*n***, and the excitation signal from generator 34 is applied in parallel to the input terminals of the modulators. The excitation signal from generator **12** comprises a series of uniform amplitude pulses having a repetition rate corresponding to the fundamental pitch frequency of the incoming speech wave, and modulators **332–0** through **332–*n*** adjust the amplitude of each successive pulse by an amount specified by the incoming control signals. For every excitation pulse from generator **12**, the amplitude-adjusted output signals simultaneously developed by modulators **332–0** through **332–*n*** represent reconstituted samples of the speech autocorrelation function of a speech wave having an unsquared amplitude spectrum. As shown in the appendix, the spectral envelope of the original speech wave may be represented in the time domain by samples of such an autocorrelation function.

The samples of the autocorrelation function represented by the amplitude-adjusted output signals of modulators **332–0** through **332–*n*** are then passed in parallel to delay line **331** via pairs of taps which are symmetrically disposed about a center tap. The amplitude-adjusted signal from modulator **332–0** is applied to the center tap, and each of the other amplitude-adjusted signals is applied to a pair of taps symmetrically disposed about the center tap. Delay line **331** converts the parallel signals into a serial succession of signals which are transformed by low-pass filter **333** into a time waveform representing the reconstituted autocorrelation function as the signals emerge from the delay line. A similar time waveform is generated for each excitation pulse to produce a periodic function at the output terminal of filter **333**. The time scale of each period developed by filter **333** depends upon the spacing of the taps on delay line **331**, which must be uniformly spaced, and since the time scale of each period correspondingly determines the scaling in the frequency domain of the spectral envelope, the spacing of the taps on delay line **331** is chosen to compensate for the observed distortion of the original spectral envelope. Thus if the influence of the non-air environment causes the peaks in the spectral envelope to occur at frequencies higher by a constant factor than those associated with a normal air medium, then the uniform spacing for the taps of delay line **331** is chosen to arrange the incoming time domain samples at an appropriate interval greater by an appropriate constant factor N than the Nyquist interval utilized at the analyzer in order to effect a division or reduction of the frequency scale of the spectral envelope of the synthesized time waveform developed by low-pass filter **333**.

Appendix

In situations where a speech wave is distorted by a linear shift in formant frequencies, the time domain embodiment shown in FIG. 3 and described above may be utilized to restore the formants of the distorted speech wave to their normal frequencies, based upon the following analysis. The vocal tract has been previously described in frequency domain terms as a function $F(f)$, and this function may be considered as the Fourier integral of the response, $f(t)$, of the vocal tract to a single impulse. A periodic portion of the speech wave $p(t)$ may then be written as a succession of impulse responses,

$$p(t)=\sum_n f(t-nT) \tag{6}$$

where T is the period of $P(t)$. The amplitude spectrum, $P(f_n)$, of $p(t)$ consists of the coefficients of the Fourier series expansion of $p(t)$, and it can be shown that the amplitude spectrum, $P(f_n)$, is related to the Fourier integral, $F(f)$, in the following manner:

$$P(f_n)=\frac{1}{T}F(f_n) \tag{7}$$

where

$$f_n=n\cdot\frac{1}{T}$$

Now a single pitch period of $p(t)$, denoted $h(t)$, has a Fourier integral, $H(f)$, which is equal to the amplitude spectrum of $p(t)$ at the harmonic frequencies, f_n, that is,

$$H(f_n)=P(f_n) \tag{8}$$

It is well known that the autocorrelation function of $h(t)$, denoted $\varphi_{hh}(\tau)$, has a Fourier transform which is equal to the power spectrum, $|H(f)|^2$ of $h(t)$, hence Equation 8 specifies that at the harmonic frequencies, f_n, the Fourier

transform of the autocorrelation function, $\varphi_{hh}(\tau)$, is equal to the power spectrum of $p(t)$.

$$|H(f_n)|^2=|P(f_n)|^2 \quad (9)$$

If it be assumed that the spectrum of the speech wave $p(t)$ is square-rooted, for example, in the manner described above in connection with FIG. 3, to produce a speech wave $p'(t)$ having an amplitude spectrum $|P(f_n)|^{1/2}$, then a single period $h'(t)$ of $p'(t)$ has a Fourier transform $H'(f)$ which is equal to $|H(f)|^{1/2}$. Hence the autocorrelation function of $h'(t)$, denoted $\varphi_{h'h'}(\tau)$, has a Fourier integral equal to $|H'(f)|^2$, where

$$|H'(f)|^2=|H(f)|$$

and from Equation 9,

$$|H'(f_n)|=|P(f_n)|$$

Therefore the periodic time function generated by repetition of $\varphi_{h'h'}(\tau)$ with period T is the sum

$$\sum_n \varphi_{h'h'}(t-nT)$$

which has an amplitude spectrum $|P(f_n)|$; that is, the periodic autocorrelation function of each period $h'(t)$ of the spectrally square-rooted speech wave $p'(t)$ has the same amplitude spectrum as the original speech wave $p(t)$, so that samples of each period of

$$\sum_n \varphi_{h'h'}(t-nT)$$

constitute an equivalent time domain representation of the spectral envelope $F(f)$.

Restoration of the formants to their proper positions on the frequency scale can be accomplished through proper adjustment of the time scale of each period of

$$\sum_n \varphi_{h'h'}(t-nT)$$

by virtue of the well known relation between the time scale and the frequency scale of a Fourier transform pair. Thus if $g(t)$ and $G(f)$ are a Fourier transform pair, then increasing the time scale of $g(t)$ by a factor k to produce $g(kt)$ results in a decrease of the frequency scale of $G(f)$ by a factor k, $G(f/k)$, and vice versa; for example, see A. A. Kharkevich, Spectra and Analysis, page 62 (1960). Hence if the formants of $p(t)$ are linearly shifted upward in frequency by a factor k, they may be restored to their proper positions by increasing the spacing between samples of each autocorrelation function sample by a factor k.

Although this invention has been described in terms of removing distortion that affects only the vocal transmission characteristic of speech, it is apparent that the separation of the vocal transmission and vocal excitation characteristics during the processing provided by this invention permits distortion in the vocal excitation characteristic to be removed also if desired. In addition, it is to be understood that applications of the principles of this invention are not limited to the particular embodiments illustrated and described, but include other speech processing equipment such as speech spectrum analyzers, vobancs, vocal tract analogs, and other arrangements that may be devised for the principles of this invention by those skilled in the art without departing from the spirit and scope of the invention.

What is claimed is:

1. Apparatus for improving the intelligibility of a speech wave characterized by a vocal transmission function and a vocal excitation function in which the resonances of said vocal transmission function occur at frequencies other than those normally associated with speech sounds uttered in air at normal pressures and temperatures at the surface of the earth, which comprises speech analyzer means for obtaining from said speech wave a first signal representation of said vocal transmission function of said speech wave and a second signal representation of said vocal excitation function of said speech wave, and speech synthesizer means for obtaining from said first and second signal representations a synthetic speech wave characterized by the same vocal excitation function as said original speech wave and a synthetic vocal transmission function, said synthesizer means including scale selector means for adjusting said synthetic vocal transmission function to create resonances which are proportional in amplitude to the resonances in the vocal transmission function of said original speech wave but which are displaced in frequency by a predetermined amount from the resonances of the vocal transmission function of said original speech wave so that the resonances of said synthetic vocal transmission function occur at said normal frequencies.

2. Apparatus for improving the intelligibility of a speech wave uttered in a medium that causes the formant peaks of the spectral envelope of said speech wave to occur at frequencies other than those normally associated with said formant peaks, which comprises:

a speech analyzer for dividing said speech wave into its spectral envelope and its excitation function, including:

means for deriving from said speech wave a plurality of control signals representative of samples of the amplitude of said spectral envelope within each of a corresponding plurality of selected frequency sub-bands, and

means for generating from said speech wave an excitation signal representative of said excitation function,

and a speech synthesizer for reconstructing an intelligible replica of said speech wave from said control signals and said excitation signal, including:

a plurality of modulating means each of which is individually controlled by a corresponding one of said plurality of control signals for adjusting the amplitude of said excitation signal to obtain a plurality of amplitude-adjusted excitation signals corresponding in amplitude to said plurality of frequency sub-bands of said spectral envelope,

means for deriving from each of said amplitude-adjusted excitation signals a frequency sub-band that is displaced in frequency by a predetermined amount from the corresponding frequency sub-band of said spectral envelope, and

means for combining each frequency sub-band derived from said amplitude-adjusted excitation signals to form said replica.

3. Apparatus for improving the intelligibility of an incoming speech wave having an amplitude spectrum which is the product of a vocal excitation function and vocal transmission function, wherein the formant peaks of the envelope of said spectrum occur at frequencies other than those normally associated with speech sounds uttered in air at normal pressures and temperatures at the surface of the earth, which comprises:

vocal transmission analyzer means for deriving from said speech wave a plurality of control signals representative of the amplitude of said spectral envelope within each of a corresponding plurality of selected frequency sub-bands,

vocal excitation analyzer means for deriving from said speech wave an excitation signal indicative of the excitation function of said speech wave,

a plurality of amplitude adjusting means in one-to-one correspondence with said plurality of control signals, wherein each of said amplitude adjusting means is supplied with said excitation signal and controlled by said corresponding control signal to obtain a corresponding ampiltude-adjusted excitation signal having

an amplitude proportionate to the amplitude of said corresponding frequency sub-band of said spectral envelope,

a plurality of filter means each following a corresponding one of said amplitude adjusting means for selecting from each of said amplitude-adjusted excitation signals a frequency sub-band that is displaced in frequency by a predetermined amount from said corresponding frequency sub-band of said spectral envelope,

and means for combining said frequency subbands selected from said amplitude-adjusted excitation signals to form a synthetic speech wave that is an intelligible replica of said incoming speech wave.

4. Apparatus for improving the intelligibility of an incoming speech wave characterized by a vocal transmission function and a vocal excitation function in which the resonances of said vocal transmission function occur at frequencies other than those normally associated with speech sounds uttered in air at normal pressures and temperatures at the surface of the earth, which comprises:

a vocal transmission function analyzer for deriving from said speech wave a plurality of control signals representative of the amplitudes and frequencies of selected resonances in the vocal transmission function of said speech wave,

a vocal excitation function analyzer for deriving from said speech wave an excitation signal representative of the vocal excitation function of said speech wave, and

a synthesizer for obtaining from said plurality of control signals and said excitation signal a synthetic speech wave characterized by the same vocal excitation function as said incoming speech wave and a synthetic vocal transmission function having resonances which are proportional in amplitude to the resonances in the vocal transmission function of said incoming speech wave but which are individually displaced in frequency from the resonances in the vocal transmission function of said incoming speech wave by predetermined amounts so that the resonances in said synehetic vocal transmission function occur at said normal frequencies.

5. Apparatus for improving the intelligibility of an incoming speech wave characterized by a vocal transmission function and a vocal excitation function in which the resonances of said vocal transmission function occur at frequencies other than those normally associated with speech sounds uttered in air at normal pressures and temperatures at the surface of the earth, which comprises:

first analyzer means for deriving from said speech wave a plurality of control signals representative of time domain samples of said vocal transmission function,

second analyzer means for deriving from said speech wave an excitation signal representative of said vocal excitation function,

means for reconstituting said time doman samples from said plurality of control signals and said excitation signals, and

means for arranging said reconstituted time domain samples on the time scale so that said reconstituted samples form a synthetic speech wave having a synthetic vocal transmission function with resonances that are displaced in frequency by predetermined amounts from the resonances in the vocal transmission function of said incoming speech wave.

6. Apparatus as defined in claim 5 wherein said first analyzer means comprises:

means for deriving from said speech wave a first plurality of samples of each period of the autocorrelation function of said speech wave,

means for unsquaring the amplitude spectrum of the autocorrelation function represented by said first plurality of samples to form a second plurality of samples of an unsquared autocorrelation function having a Fourier transform equal to said vocal transmission function, and

means for deriving said plurality of control signals from said second plurality of samples.

7. Apparatus for improving the intelligibility of an incoming speech wave characterized by a vocal transmission function and a vocal excitation function in which the resonances of said vocal transmission function occur at frequencies other than those normally associated with speech sounds uttered in air at normal pressures and temperatures at the surface of the earth, which comprises:

first analyzer means for deriving from said speech wave a plurality of control signals representative of frequency domain samples of said vocal transmission function,

second analyzer means for deriving from said speech wave an excitation signal representative of said vocal excitation function,

means for reconstituting said frequency domain samples from said plurality of control signals and said excitation signal, and

means for arranging said reconstituted frequency domain samples on the frequency scale so that said reconstituted samples form a synthetic speech wave having a synthetic vocal transmission function with resonances that are displaced in frequency by predetermined amounts from the resonances in the vocal transmission function of said incoming speech wave.

8. The method of improving the naturalness and intelligibility of a speech wave having a spectrum in which the formant peaks of the spectral envelope occur at frequencies other than those normally associated with speech sounds uttered in air at normal pressures and temperatures at the surface of the earth, which comprises the steps of:

separating said speech wave into its spectral envelope and its excitation function,

dividing said spectral envelope into selected frequency sub-bands,

shifting the frequency of each of said sub-bands by a predetermined amount so that said formant peaks of the spectral envelope are restored to their normal frequencies, and

combining said frequency-shifted sub-bands and said excitation function to form a natural sounding replica of said speech wave.

References Cited

UNITED STATES PATENTS

2,183,248	12/1939	Riesz.
2,903,521	9/1959	Ellison.
3,071,652	1/1963	Schroeder.

KATHLEEN H. CLAFFY, *Primary Examiner.*

R. P. TAYLOR, *Assistant Examiner.*