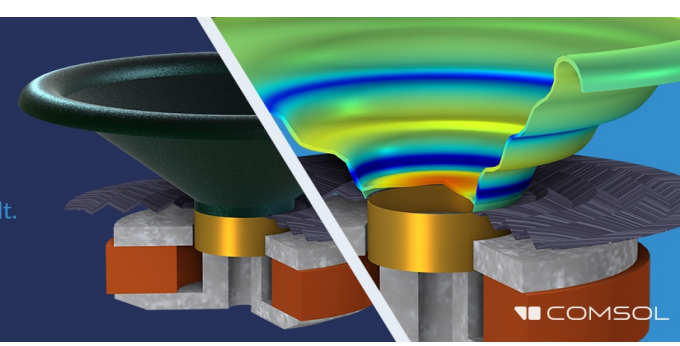
Cepstrum Pitch Determination

A. Michael Noll

Citation: The Journal of the Acoustical Society of America **41**, 293 (1967); doi: 10.1121/1.1910339
View online: https://doi.org/10.1121/1.1910339
View Table of Contents: https://asa.scitation.org/toc/jas/41/2
Published by the Acoustical Society of America



Received 24 August 1966 9.3, 9.8, 9.9

Cepstrum Pitch Determination

A. Michael Noll

Bell Telephone Laboratories, Murray Hill, New Jersey 07971

The cepstrum, defined as the power spectrum of the logarithm of the power spectrum, has a strong peak corresponding to the pitch period of the voiced-speech segment being analyzed. Cepstra were calculated on a digital computer and were automatically plotted on microfilm. Algorithms were developed heuristically for picking those peaks corresponding to voiced-speech segments and the vocal pitch periods. This information was then used to derive the excitation for a computer-simulated channel vocoder. The pitch quality of the vocoded speech was judged by experienced listeners in informal comparison tests to be indistinguishable from the original speech.

INTRODUCTION

VOICED-speech sounds result from the resonant action of the vocal tract on the periodic puffs of air admitted through the vocal cords. For pitch-period determination, the time periodicity of the source signal must be obtained from the observed speech signal. Also, voiced–unvoiced decisions require accurate determination of the presence or absence of such periodic puffs in the source signal. This deceptively simple problem has been the object of considerable research over the past few decades. Aside from its obvious use in analysis of speech sounds from a pure research standpoint, an accurate pitch detector must also perform adequately as an integral part of most speech-bandwidth compression schemes. The design of an accurate pitch detector that works satisfactorily with band-limited, noisy speech signals remains one of the challenging areas of speech processing research.

In a previous paper, a new method for obtaining the fundamental frequency or pitch of human speech was described.[1] Since the logarithm of the amplitude spectrum of a periodic time signal with rich harmonic structure is itself "periodic" in frequency, the new method consisted of spectrum analyzing this log amplitude spectrum. Adopting some new terminology proposed by Tukey, the method was called "cepstrum" pitch detection, where the term cepstrum refers to the spectrum of the log-amplitude spectrum. Computer programs were written to perform short-time cepstrum analyses of speech, and the resultant pitch information was used to obtain the excitation for computer-simulated vocoders. The synthesized speech was quite encouraging as demonstrated by tapes played at the sixty-seventh meeting of the Acoustical Society.[2]

The early computer programs written to simulate the cepstrum analyzer have since undergone a number of changes towards simplicity and efficiency. The results of analyses of speech cepstra were used to design an automatic method for determining the pitch periods from the cepstral peaks. This automatic peak picker, though not previously described, was used to obtain the excitation signals for the computer-simulated vocoders. Some interesting and unexpected pitch fluctuations and pitch doubling have been discovered during the observations of speech cepstra required to develop the algorithms for the cepstral peak picker. These topics and new approaches to explaining and justifying cepstrum pitch determination were not reported in the previous papers; and now is also a good time to present the historical background leading to the concept of short-time cepstrum analysis for vocal-pitch detection. This paper treats all these topics and concludes with descriptions of some possible hardware implementations of cepstrum analyzers.

I. HISTORICAL BACKGROUNDS

In the fall of 1959, Bogert (of Bell Telephone Laboratories) noticed banding in spectrograms of seismic signals. He realized that this banding was caused by

[1] A. M. Noll, "Short-Time Spectrum and 'Cepstrum' Techniques for Vocal-Pitch Detection," J. Acoust. Soc. Am. **36**, 296–302 (1964).

[2] A. M. Noll and M. R. Schroeder, "Short-Time 'Cepstrum' Pitch Detection," J. Acoust. Soc. Am. **36**, 1030 (1964).

"periodic" ripples in the spectra and that this was characteristic of the spectra of any signal consisting of itself plus an echo. The frequency spacing of these ripples equals the reciprocal of the difference in time arrivals of the two waves. Tukey (of both Princeton University and Bell Telephone Laboratories) suggested that this frequency difference might be obtained by first taking the logarithm of the spectrum, thereby making the ripples nearly cosinusoidal. A spectrum analysis of the log spectrum then could be performed to determine the "frequency" of the ripple. In early 1960, Bogert programmed Tukey's suggestion on a computer and proceeded to analyze numerous earthquakes and explosions. Tukey, noticing similarities between time series analysis and log-spectrum series analysis, introduced a new set of paraphrased terms. The spectrum of the log spectrum was called the "cepstrum," and the frequency of the spectral ripples were referred to as "quefrency." Bogert, Tukey, and Healy published their ideas in an article with perhaps one of the weirdest titles ever encountered in the scientific literature: "The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking."[3] In the article, they very clearly expressed a pessimistic view for achieving adequate classification of seismic events by cepstral techniques. In fact, no definitive indication of focal depth was found.

Their article was issued as an internal Bell Laboratories memorandum before publication in Rosenblatt's book. Schroeder read the memorandum and realized that voiced speech spectra also have ripples, and hence cepstrum analysis might be suitable for vocal-pitch determination. In June 1962, Schroeder suggested cepstrum-pitch determination as an area worthy of further study. At that time, he and Atal had just completed a paper on methods for performing short-time spectrum analyses.[4] Thus, the atmosphere was perfect for the concept of short-time cepstrum analysis that then developed.

Seismic signals consist of a single event, and therefore only one cepstrum is obtained. Speech, however, changes with time, and a single cepstrum of a long speech signal would be meaningless. Hence, a series of cepstra for short segments of the speech signal are required—a short-time cepstrum. A scheme was devised for performing such short-time cepstral analyses utilizing delay lines and multipliers as shown in Fig. 1. A computer program was written with a special-purpose block-diagram language to simulate this method, and the short-time spectra and cepstra were automatically plotted by the computer on microfilm.[5] The cepstra for voiced speech intervals had strong peaks corresponding to the pitch period. The conclusion was quite definite: Although a single cepstrum analysis of seismic events was not promising for seismic classification, short-time cepstrum analysis of speech performed excellently as a new means for vocal-pitch determination.

In recent papers, the heutristics of cepstrum analysis for extracting echoed signals from noise has been developed by Bogert and Ossanna.[6] Also, a more general formalism of separating convolved signals and its relation with cepstrum analysis has been treated by Oppenheim.[7]

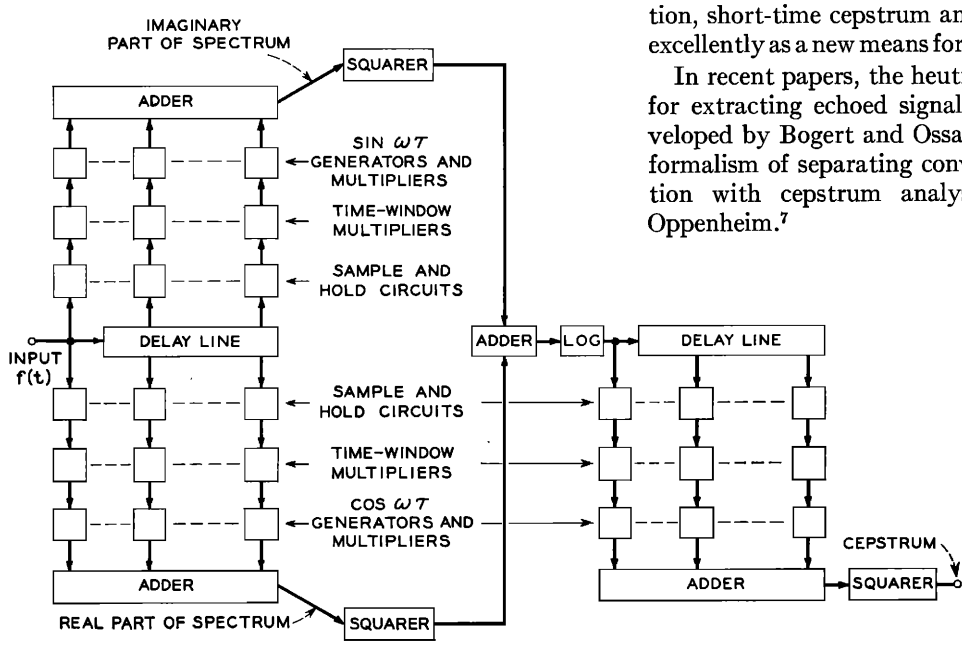


FIG. 1. Block diagram of sampled-data device for performing short-time cepstrum analysis.

[3] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, in *Proceedings of the Symposium on Time Series Analysis*, by M. Rosenblatt, Ed. (John Wiley & Sons, Inc., New York, 1963), Chap. 15, pp. 209–243.

[4] M. R. Schroeder and B. S. Atal, "Generalized Short-Time Power Spectra and Autocorrelation Functions," J. Acoust. Soc. Am. 34, 1679–1683 (1962).

[5] J. L. Kelly, Jr., Carol Lochbaum, and V. A. Vyssotsky, "A Block Diagram Compiler," Bell System Tech. J. 40, 669–677 (1961).

[6] B. P. Bogert and J. F. Ossanna, "The Heuristics of a Stationary Complex Echoed Gaussian Signal in Stationary Gaussian Noise," IEEE Trans. Information Theory **IT-12**, No. 3, 343 (1966).

[7] A. V. Oppenheim, "Nonlinear Filtering of Convolved Signals," Mass. Inst. Technol. Res. Lab. Electron. Quart. Progr. Rept. No. 80, 168–175 (January 1966).

II. CESTRUM-PITCH DETERMINATION

In its most basic form, the system for producing voiced speech sounds consists only of the vocal source and the vocal tract as shown in Fig. 2. The source signal $s(t)$ is the periodic puffs of air admitted through the vocal cords. The effect of the vocal tract is completely specified by its impulse response $h(t)$ such that the output speech signal $f(t)$ equals the convolution of $s(t)$ and $h(t)$. Alternatively, if $S(\omega)$ is the spectrum of the vocal source and $H(\omega)$ is the transfer function or spectrum of the vocal tract, then the spectrum of the speech signal equals the product of $S(\omega)$ and $H(\omega)$. Expressed algebraically,

$$f(t)=s(t)*h(t), \tag{1}$$

$$F(\omega)=S(\omega)\cdot H(\omega), \tag{2}$$

with

$$F(\omega)=\mathfrak{F}[f(t)], \tag{3}$$

$$S(\omega)=\mathfrak{F}[s(t)], \tag{4}$$

$$H(\omega)=\mathfrak{F}[h(t)], \tag{5}$$

where $*$ denotes convolution, $\mathfrak{F}$ denotes Fourier transformation, and the Fourier transforms of $s(t)$ and $h(t)$ are assumed to exist.

The source signal and, therefore, the speech signal, are quasiperiodic for voiced-speech sounds. If the period is T seconds, then the power spectrum $|F(\omega)|^2$ of the speech signal consists of harmonics spaced T^{-1} Hz. Thus, the power spectrum of a voiced speech signal is "periodic" along the frequency axis with "period" equal to the reciprocal of the period of the time signal being analyzed. The obvious way to measure this "period" in the power spectrum is to take the Fourier transform of the spectrum that will have a peak corresponding to the "period." This spectrum of the power spectrum is more commonly known as the autocorrelation function of the original time signal. Mathematically, the autocorrelation function $r(\tau)$ is defined as

$$r(\tau)\equiv\mathfrak{F}[|F(\omega)|^2]. \tag{6}$$

The speech power spectrum equals the product of the spectra of the vocal source and the vocal tract. But the Fourier transform of a product equals the convolution of the Fourier transforms of the two multiplicands. Thus,

$$r(\tau)=\mathfrak{F}[|S(\omega)|^2|H(\omega)|^2] \tag{7}$$

$$=\mathfrak{F}[|S(\omega)|^2]*\mathfrak{F}[|H(\omega)|^2] \tag{8}$$

$$=r_s(\tau)*r_h(\tau), \tag{9}$$

where $r_s(\tau)$ and $r_h(\tau)$ are the autocorrelation functions of $s(t)$ and $h(t)$, respectively. The effects of the vocal source and vocal tract are therefore convolved with each other in the autocorrelation functions. This results in broad peaks and in some cases multiple peaks in the autocorrelation function; thus, an autocorrelation approach to pitch determination is, in general, unsatisfactory.[8]

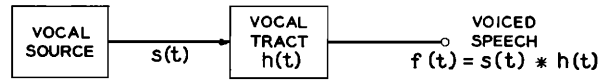


FIG. 2. Basic system for the production of voiced speech sounds. $h(t)$ is the impulse response of the vocal tract.

The solution is to devise a new function in which the effects of the vocal source and vocal tract are nearly independent or easily identifiable and separable. The Fourier transform of the *logarithm* of the power spectrum is such a new function and, indeed, separates the effects of the vocal source and tract. The reason for this is that the logarithm of a product equals the sum of the logarithms of the multiplicands:

$$\log|F(\omega)|^2=\log[|S(\omega)|^2\cdot|H(\omega)|^2] \tag{10}$$

$$=\log|S(\omega)|^2+\log|H(\omega)|^2. \tag{11}$$

The Fourier transform of the logarithm power spectrum preserves the additive property and is

$$\mathfrak{F}[\log|F(\omega)|^2]=\mathfrak{F}[\log|S(\omega)|^2]+\mathfrak{F}[\log|H(\omega)|^2]. \tag{12}$$

The source and tract effects are now additive rather than convolved as in the autocorrelation. The importance of this can be intuitively explained with the assistance of Fig. 3. The effect of the vocal tract is to produce a "low-frequency" ripple in the logarithm spectrum, while the periodicity of the vocal source manifests itself as a "high-frequency" ripple in the logarithm spectrum. Therefore, the spectrum of the logarithm power spectrum has a sharp peak corresponding to the high-frequency source ripples in the logarithm spectrum and a broader peak corresponding to the low-frequency formant structure in the logarithm spectrum. The peak corresponding to the source periodicity can be made more pronounced by squaring the second spectrum. This function, the square of the Fourier transform of the logarithm power spectrum, is called the "cepstrum," borrowing Tukey's terminology.

To prevent confusion between the usual frequency components of a time function and the "frequency" ripples in the logarithm spectrum, Tukey has used the paraphrased word *quefrency* in describing the "frequency" of the spectral ripples. Quefrencies have the units of cycles per hertz or, simply, seconds. Adopting this terminology, the cepstrum consists of a peak occurring at a high quefrency equal to the pitch period in seconds and low-quefrency information corresponding to the formant structure in the logarithm spectrum.

Thus far, no mention has been made about the time length of the signal under analysis. As mentioned before, for seismic signals, a single cepstrum analysis is performed for the whole seismic event. But speech param-

[8] M. R. Schroeder, "Vocoders: Analysis and Synthesis of Speech," Proc. IEEE 54, No. 5, 720–734 (1966).

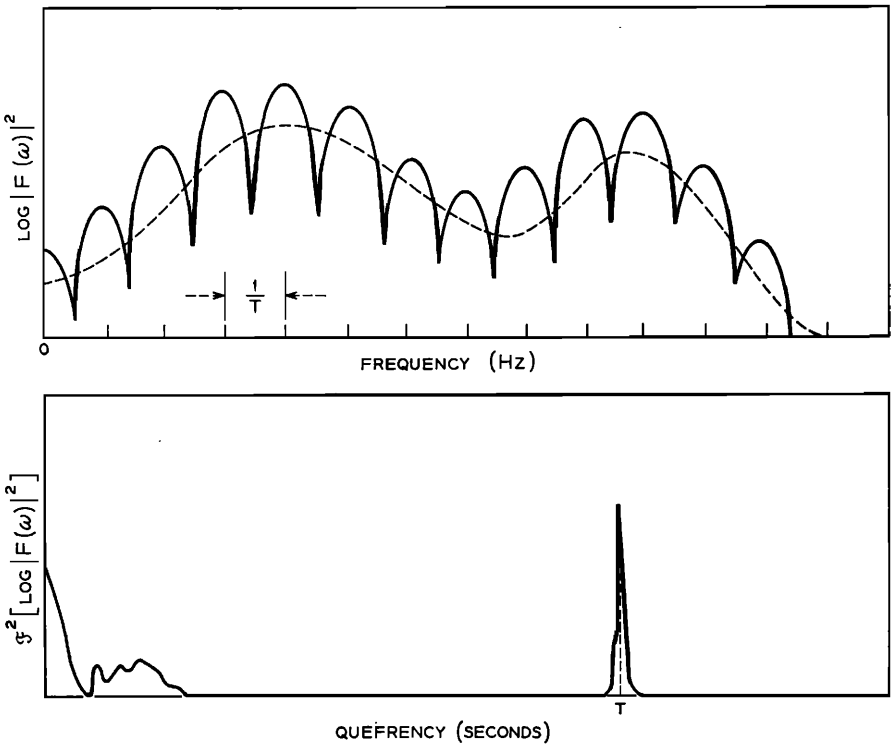


FIG. 3. Logarithm power spectrum (top) of a voiced speech segment showing a spectral periodicity resulting from the pitch periodicity of the speech. The power spectrum of the logarithm spectrum, or cepstrum (bottom), therefore has a sharp peak corresponding to this spectral periodicity.

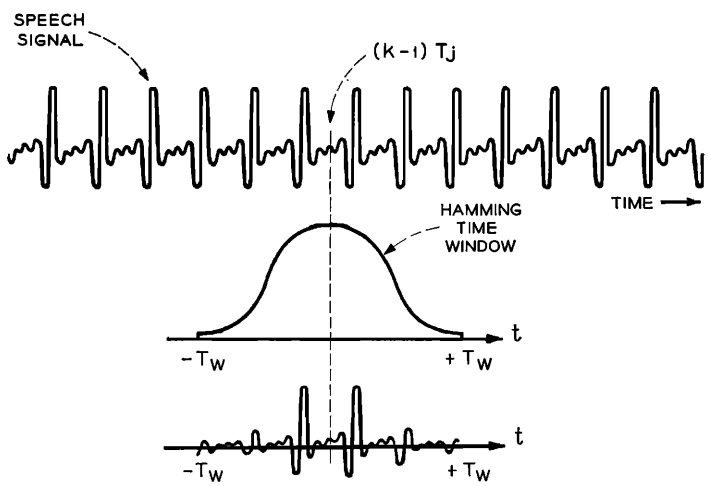


FIG. 4. Basic operations required for obtaining the short-time cepstrum of a speech signal. The hamming time window of length T_w sec moves in jumps of T_J sec.

K^{TH} SHORT-TIME SPECTRUM:

$$F_K(\omega) \equiv \int_{-T_W}^{T_W} [\text{speech}]\, e^{-j\omega t}\, dt$$

K^{TH} LOG POWER SPECTRUM:

$$\text{LOG}\left|F_K(\omega)\right|^2 = \text{LOG}\left[\mathcal{R}e^2 F_K(\omega) + \mathcal{I}m^2 F_K(\omega)\right]$$

K^{TH} SHORT-TIME CEPSTRUM:

$$C_K(q) \equiv \left| \int_0^{\omega_c} \text{LOG}\left|F_K(\omega)\right|^2 \cos\, \omega q\, dq \right|^2$$

eters—and, in particular, pitch—change with time; therefore a series of cepstra for short time segments of the signal are required. This is accomplished by multiplying the time signal by a function that is zero outside some finite time interval. The function performs something like a window through which the time signal is viewed, and its effects are discussed later in more detail. As shown in Fig. 4, the time-limited signal is spectrum analyzed once to obtain the log spectrum and then again to produce the cepstrum. A new portion of the time signal then enters the window and is similarly analyzed to produce another cepstrum. This process, when performed repetitively, results in a series of short-time cepstra. The time window, if desired, could also look at overlapping portions of the signal.

The resultant cepstra are automatically examined to determine the maximum peaks corresponding to voiced speech intervals and the frequency of these peaks. This information is used to decide if the speech segment is voiced or unvoiced and, if voiced, to determine the pitch period.

Both the effects of the time window and a mathematical justification for the spectral ripples were neglected in the preceding discussion and are now taken up. The time-limited signal to be analyzed is

$$g(t)=[s(t)*h(t)]\cdot w(t) \tag{13}$$

from Eq. 1, where $w(t)$ is the time window, defined to be zero for $|t|>T_W$. But, the periodic source signal $s(t)$ can be represented as the superposition of an infinite series of identical signals $s_0(t)$ repeated every T seconds:

$$s(t)=\sum_{n=-\infty}^{\infty} s_0(t-nT) \tag{14}$$

$$=s_0(t)*\sum_{n=-\infty}^{\infty}\delta(t-nT). \tag{15}$$

Substitution into Eq. 13 gives

$$g(t)=\{[s_0(t)*\sum_{n=-\infty}^{\infty}\delta(t-nT)]*h(t)\}w(t). \tag{16}$$

The Fourier transform or complex spectrum $G(\omega)$ of $g(t)$ is

$$G(\omega)=\left\{\left[s_0(\omega)\sum_{n=-\infty}^{\infty}\delta\left(\omega-n\frac{2\pi}{T}\right)\right]H(\omega)\right\}*W(\omega) \tag{17}$$

$$=\left[\sum_{n=-\infty}^{\infty}H(\omega)S_0(\omega)\delta\left(\omega-n\frac{2\pi}{T}\right)\right]*W(\omega) \tag{18}$$

$$=\left[\sum_{n=-\infty}^{\infty}H\left(n\frac{2\pi}{T}\right)S_0\left(n\frac{2\pi}{T}\right)\delta\left(\omega-n\frac{2\pi}{T}\right)\right]*W(\omega), \tag{19}$$

where $S_0(\omega)$, $H(\omega)$, and $W(\omega)$ are the Fourier transforms of $s_0(t)$, $h(t)$, and $w(t)$, respectively.

The results of the preceding show that if the original speech signal $s(t)*h(t)$ is not time-limited, then the complex spectrum consists of an infinite series of impulses spaced T^{-1} Hz and with amplitude $H(n2\pi/T)\times S_0(n2\pi/T)$. If the non-time-limited signal is band-limited, the complex spectrum would be frequency limited or zero for $|\omega|>\omega_{\max}$. The effect of time limiting the speech signal with a multiplicative time window $w(t)$ is a convolution of the corresponding spectral window $W(\omega)$ with the spectral impulses of the non-time-limited complex spectrum. Thus, the impulses are broadened and assume the shape of $W(\omega)$. The complex spectrum is now no longer frequency-limited, since $W(\omega)$ is the transform of a time-limited function and, therefore, cannot be zero over any finite frequency interval. Hence, the complex spectrum is not strictly frequency-limited, but can be described as being approximately frequency-limited if $W(\omega)$ has very small side lobes. Also, the main lobe of $W(\omega)$ determines the spectral resolution, and therefore a $W(\omega)$ with low-amplitude side lobes and a narrow main lobe is required. Although these requirements are mutually exclusive, a good compromise is the hamming time window,[9]

$$w(t)=0.54+0.46\cos(\pi t/T_W);\quad |t|\leq T_W \tag{20}$$
$$=0;\quad |t|>T_W.$$

The hamming spectral window has a maximum side lobe 44 dB below its peak response.

III. NUMERICAL COMPUTATION OF CEPSTRA

The Fourier transform $F(\omega)$ of some function of time $f(t)$ is defined as

$$F(\omega)\equiv\int_{-\infty}^{\infty}f(t)e^{-j\omega t}dt. \tag{21}$$

If $f(t)$ is time limited by some multiplicative time window $w(t)$ such that $w(t)=0$ for $|t|>T_W$ and if complex exponentiation is separated into real and imaginary parts, Eq. 21 becomes

$$F(\omega)=\int_{-T_W}^{T_W}w(t)f(t)\cos(\omega t)dt$$
$$-j\int_{-T_W}^{T_W}w(t)f(t)\sin(\omega t)dt. \tag{22}$$

Furthermore, since $F(\omega)$ has a time-limited transform, namely, $w(-t)f(-t)$, then by Nyquist's sampling theorem applied to the frequency domain, ω can be represented as $\omega=m\Delta\omega$, where $\Delta\omega\leq 2\pi/(2T_W)$. Also, since $f(t)$ is band-limited to 0 to $\omega_c/(2\pi)$ Hz, t can be represented as $t=l\Delta t$, where $\Delta t=2\pi/(2\omega_c)$. Thus, the integrations in Eq. 22 can be replaced by summations,

[9] R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra* (Dover Publications, Inc., New York, 1959).

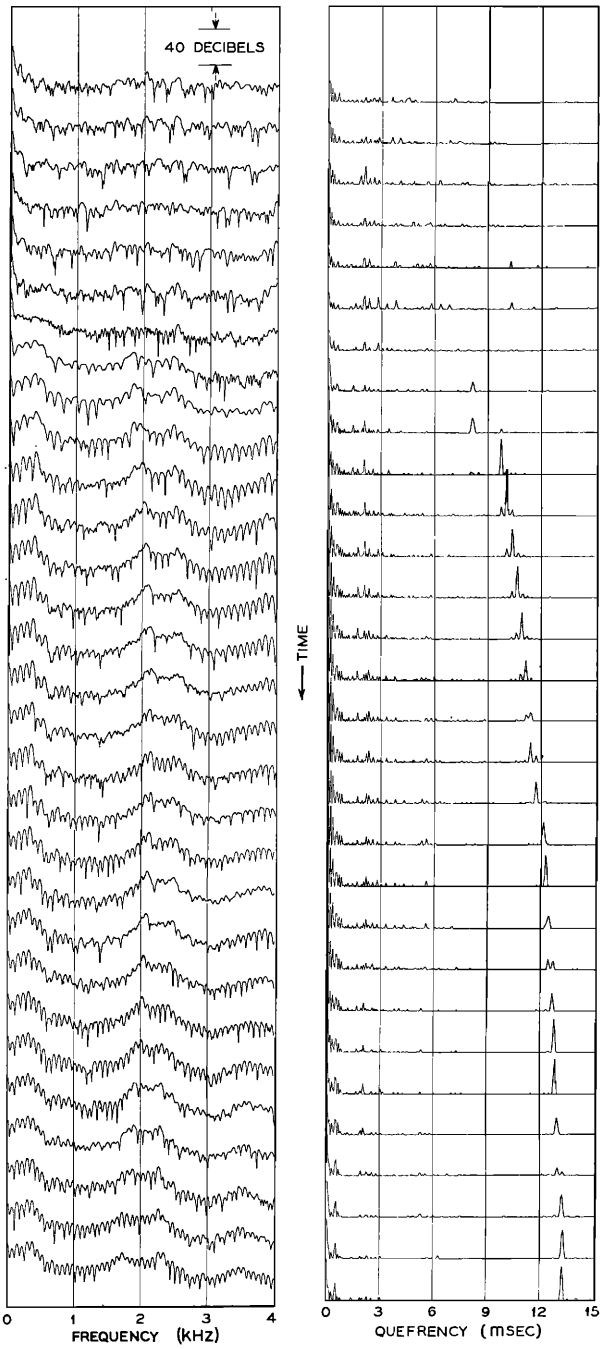


FIG. 5. Short-time logarithm spectra (left) and short-time cepstra (right) for a male talker (L.G.) recorded with a condenser microphone. The 40 msec-long hamming time moved in jumps of 10 msec.

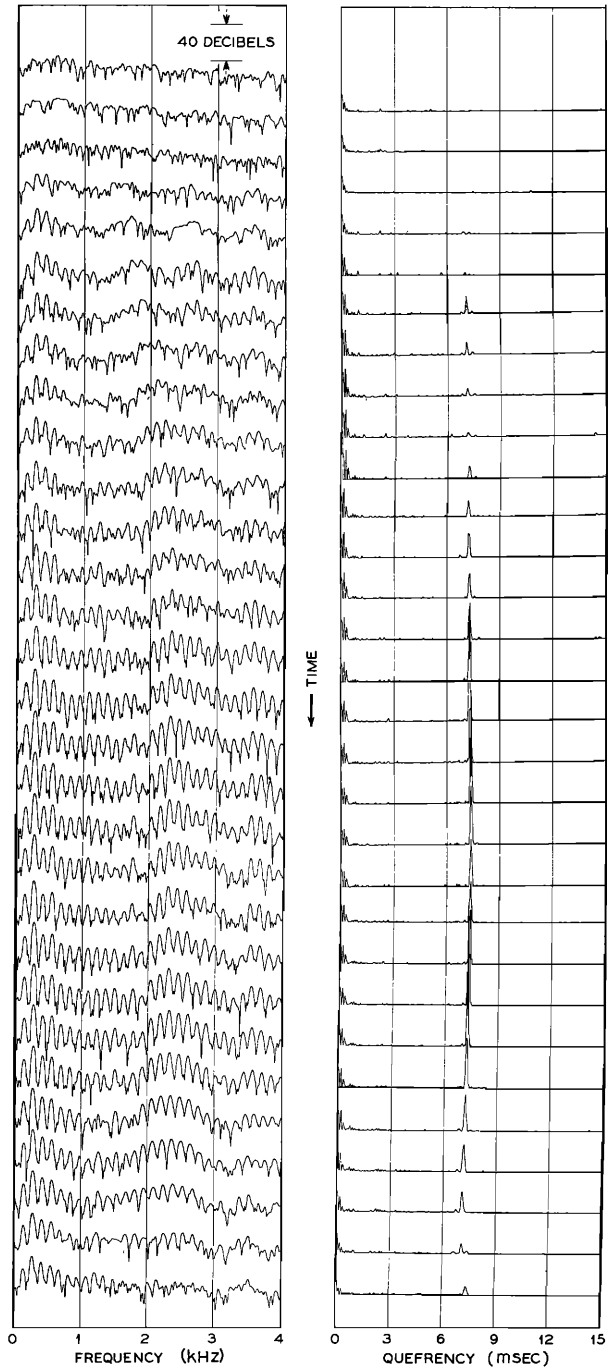


FIG. 6. Short-time logarithm spectra (left) and short-time cepstra (right) for a male talker (F.L.C.) recorded from a 500-type telephone set with carbon microphone.

so that $F(\omega)$ becomes

$$F(m\Delta\omega)=\Delta t \sum_{l=-L}^{L} w(l\Delta t) f(l\Delta t) \cos(lm\Delta t\Delta\omega) - j\Delta t \sum_{l=-L}^{L} w(l\Delta t) f(l\Delta t) \sin(lm\Delta t\Delta\omega), \quad (23)$$

where $L=T_W\Delta t$.

This equation led to the concept of a delay line for storing $2L+1$ samples of the input signal (sample and hold circuits at the taps of the delay line) so that the signal being analyzed remains constant during the analysis (window multipliers, function generators for cosine and sine, and adders as shown in Fig. 1). The real and imaginary parts of the spectrum produced by

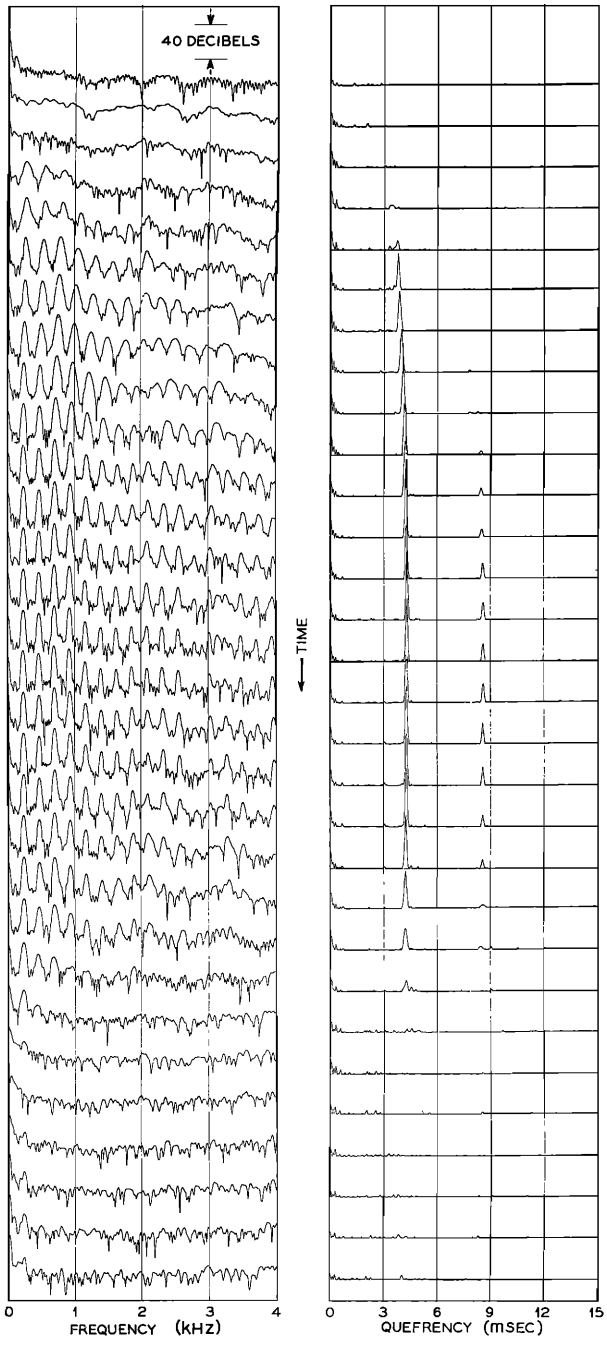


FIG. 7. Short-time logarithm spectra (left) and short-time cepstra (right) for a female talker (S.S.) recorded with a condenser microphone.

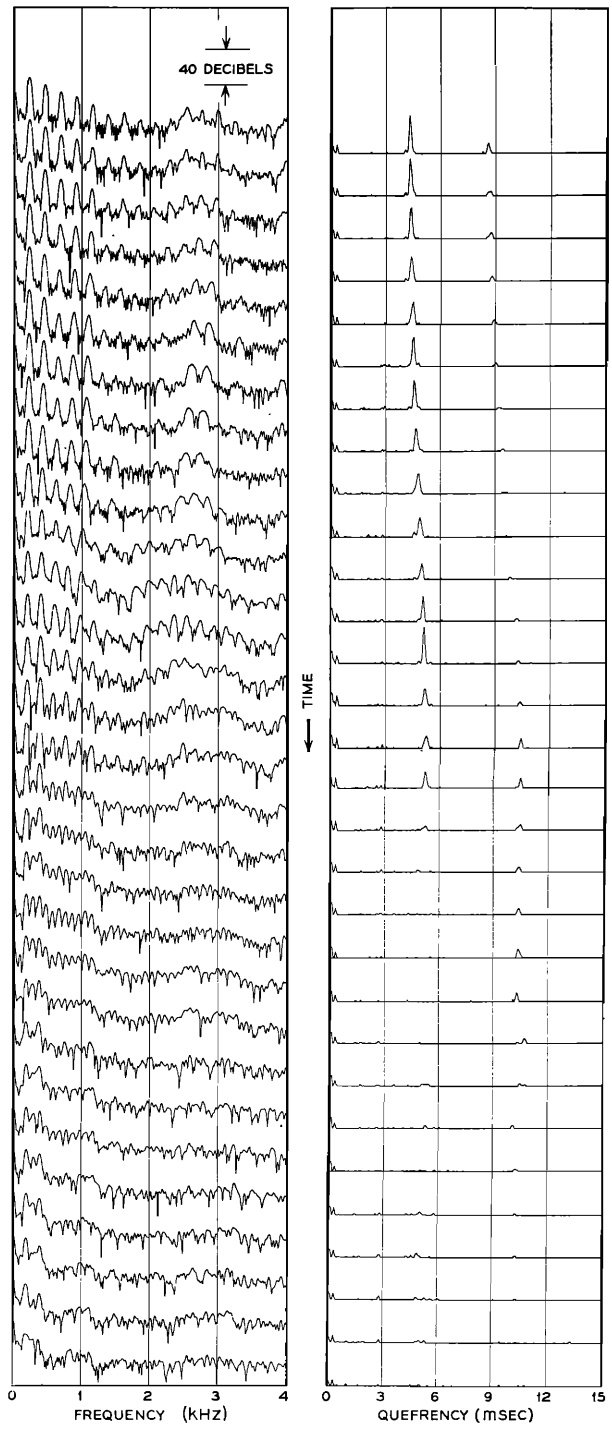


FIG. 8. Short-time logarithm spectra (left) and short-time cepstra (right) of "(scr)eaming," spoken by a female talker (S.S.) and recorded with a condenser microphone. A doubling in pitch period occurs at the end of the utterance.

this sampled-data spectrum analyzer are squared and added to generate the power spectrum. The logarithm of the power spectrum is used as the input to a similar power-spectrum analyzer whose output is the cepstrum.

This sampled-data analyzer was simulated on an IBM–7094 digital computer by using the BLODI compiler. The input speech to the computer was band-limited to 4 kHz, sampled every 10^{-4} secs, and digitized; the time window extended from -15 to $+15$ msec. The results reported in the previous paper were obtained with this computer simulation, which consumed nearly 2 h of computer time to analyze only 2 sec of speech.

The program was extremely unwieldy and changes in any parameters were difficult. Obviously, some streamlining of the program was required if further progress in cepstrum-pitch detection were to be accomplished.

Only a single spectrum in a series of spectra is defined by Eq. 22. If the time window moves in jumps of T_J sec, then the kth short-time spectrum $F_k(m)$ is defined as

$$F_k(m)=\sum_{l=-L}^{L} w(l)f[(k-1)K+l]\cos(lm\Delta t\Delta\omega) - j\sum_{l=-L}^{L} w(l)f[(k-1)K+l]\sin(lm\Delta t\Delta\omega), \quad (24)$$

where $L=T_W/\Delta t$, $K=T_J/\Delta t$ and $m=0,1,\cdots,\omega_c/\Delta\omega$.

The kth short-time power spectrum is the magnitude squared of the kth short-time spectrum:

$$|F_k(m)|^2=\{\sum_{l=-L}^{L} w(l)f[(k-1)K+l]\cos(lm\Delta t\Delta\omega)\}^2 + \{\sum_{l=-L}^{L} w(l)f[(k-1)K+l]\sin(lm\Delta t\Delta\omega)\}^2. \quad (25)$$

Although the complex spectrum may be sampled at $\Delta\omega\leq 2\pi/(2T_W)$, the power spectrum should be sampled at $\Delta\omega\leq 2\pi/(4T_W)$. This is because the Fourier transform of the power spectrum is the autocorrelation function that for a signal time limited to $\pm T_W$ sec is itself time-limited to $\pm 2T_W$ sec. By Nyquist's sampling theorem, the power spectrum therefore must be sampled at $\Delta\omega\leq 2\pi/(4T_W)$. Strictly speaking, if the Fourier transform of the power spectrum is time-limited, then the Fourier transform of the logarithm power spectrum is generally not time-limited. But from experience that the aliasing is negligible, the log power spectrum is sampled at the same interval as the power spectrum. Since the computer is used in taking logarithms, the logarithm of zero is forced to be noninfinite.

The cepstrum $C(\tau)$ is now formally defined as the power spectrum of the logarithm power spectrum. Since the log power spectrum is an even function, this definition is equivalent to the square of the cosine transform of the log power spectrum, or

$$C(\tau)\equiv\left\{\int_0^\infty \log|F(\omega)|^2\cos(\omega\tau)d\omega\right\}^2. \quad (26)$$

For $C(\tau)$ to be sampled, the Fourier transform of $C(\tau)$ must be band-limited. However, $C(\tau)$ is the product of two cosine transforms, and therefore the Fourier transform of $C(\tau)$ is the convolution of the Fourier transforms of the individual cosine transforms. But, since the cosine transform of $\log|F(\omega)|^2$ is also an even function, the Fourier transform of the cosine transform of $\log|F(\omega)|^2$ simply gives $\log|F(\omega)|^2$. Thus, the Fourier transform of $C(\tau)$ equals the convolution of $\log|F(\omega)|^2$ with itself. Since $\log|F(\omega)|^2$ is very small for $|\omega|>\omega_c$, the convolution is very nearly limited to the interval $|\omega|\leq 2\omega_c$. Nyquist's theorem can therefore be applied, and the cepstrum can be sampled so that $\tau=n\Delta\tau$ with $\Delta\tau\leq 2\pi/(4\omega_c)$. Thus, the kth short-time cepstrum $C_k(n)$ can be calculated as

$$C_k(n)=\sum_{m=0}^{M}\log|F_k(m)|^2\cos(mn\Delta\tau\Delta\omega), \quad (27)$$

where $\Delta\omega\leq 2\pi/(4T_W)$, $M=\omega_c/\Delta\omega$, $n=0,1,\cdots,N$ with N some arbitrary upper limit on the desired quefrencies in the cepstrum.

The numerical operations indicated by Eqs. 25 and 27 were programmed in the FORTRAN language. To conserve execution time, all sine and cosine operations were performed as table lookups from calculated sine and cosine tables. Also, the computation of the sine and cosine transforms utilized even and odd symmetry in the input signal to reduce further the number of calculations. Nevertheless, the program was still very lengthy and required about 0.8 h to compute the cepstra for about 2 sec of speech. Recently, an algorithm has been developed by Cooley and Tukey for performing fast numerical Fourier transformations.[10] This algorithm has been incorporated into the cepstrum program and has resulted in a program about eight times faster than the previous one.

A very important factor in the computer calculation of short-time cepstra has been facilities for the automatic plotting of the spectra and cepstra. These facilities consist of a cathode-ray tube and camera, both under the direct control of the digital computer.

IV. EXAMPLES OF SPEECH SPECTRA AND CEPSTRA

The computer technique described in the preceding portions of this paper was used to analyze a few selected sentences and words. The speech was low-pass filtered to 4 kHz and sampled every 10^{-4} sec. The hamming time window was 40 msec long and moved in jumps of 10 msec. The spectral components were calculated at frequency intervals of 12.5 Hz up to a maximum frequency of 4 kHz; the cepstral components were calculated at invervals of 0.0625 msec up to a maximum quefrency of 15 msec. The results of the calculations were automatically plotted on microfilm by the computer with corresponding spectra and cepstra shown adjacent to each other. Time progresses downwards in jumps of 10 msec.

Figure 5 shows the spectra and cepstra of a male talker (L.G.) recorded with a condenser microphone; Fig. 6 is for a different male talker (F.L.C.) recorded

[10] J. W. Cooley and J. W. Tukey, "An Algorithm for the Machine Calculation of Complex Fourier Series," Math. of Computation **19**, 297–301 (1965).

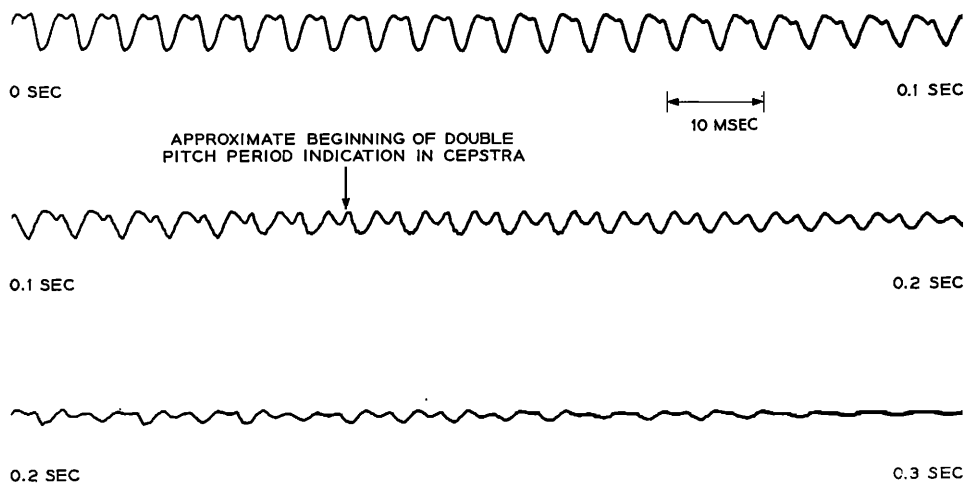


FIG. 9. Speech waveform of the "ing" portion of "(scr)eaming," showing the approximate location of the switch to double pitch period indicated by the cepstra.

from a 500-type telephone set with carbon transmitter; Fig. 7 is female speech (S.S.) recorded from a condenser microphone. In all three examples, the voiced-speech intervals are clearly indicated by the sharp peaks in the cepstra. The cepstral peaks in Fig. 5 for the voiced-speech intervals of Curves 11–15 are particularly interesting since they consist of a major peak with two smaller peaks on either side. This occurred because the pitch was changing rapidly such that each 40-msec analysis interval contained different pitch periods. Actually, the 40-msec hamming window looks mostly at only the center 20 msec since the tails of the window are strongly weighted down in amplitude. Thus, very little smoothing is actually present, and the largest cepstral peak corresponds to the dominant pitch period mostly within the 20-msec center interval.

Figure 8 shows the spectra and cepstra of the utterance *(scr)eaming* spoken by a female talker (S.S.) into a condenser microphone. At about the 12th cepstrum, a second "rahmonic" appears and gradually grows in amplitude until, at about the 17th cepstrum, its amplitude exceeds the fundamental peak at about 5.2 msec. The fundamental peak then disappears, leaving only the cepstral peak at 10.4 msec. This would imply a doubling of pitch period at the end of the "··· ing" sound, and, indeed, speech synthesized with the doubled excitation sounds natural and compares better with the original than excitation that does not double in period at the "··· ing" portion. The spectra corresponding to this transition show the alternate harmonics gradually growing in amplitude until they fill in the gaps between the harmonics corresponding to the lower pitch period. The actual speech waveform is shown in Fig. 9, and the point of transition is indicated. Although the doubling is discernible towards the end of the signal, the cepstrum gives an indication of doubling earlier than would be determined by visual inspection of the waveform.

The spectra and cepstra of the word *chase* spoken by a famale talker (B.M.) and recorded with a condenser microphone is shown in Fig. 10. The 12th, 13th, and 14th cepstra have small second rahmonics at about 8.8 msec that are smaller in amplitude than the fundamental cepstral peak at about 4.4 msec. However, the 19th through 21st cepstra have second rahmonics with amplitudes exceeding the fundamental. This type of doubling of pitch period imbedded in voiced speech sounds wrong when used as excitation for a vocoder and is therefore considered as undesirable. The spectra for the double pitch consist of harmonics corresponding to the 4.4-msec pitch period with interlaced harmonics that fade in and out across the spectrum. This type of spectrum is caused by minute jitter in the pitch-pulse timing.[11] If the vocal source signal $s(t)$ is assumed to consist of air puffs at $\cdots 0,\ T+\epsilon,\ 2T,\ 3T+\epsilon,\ \cdots$, then

$$s(t)=\sum_{n=-\infty}^{\infty}[s_0(t-2nT)+s_0(t-\epsilon-2nT-T)]$$
$$=s_0(t)*\sum_{n=-\infty}^{\infty}[\delta(t-2nT)+\delta(t-2nT-T-\epsilon)]. \tag{28}$$

The Fourier transform of the summation portion corresponding to the jittered pulses is

$$J(\omega)=[1+e^{-j\omega(T+\epsilon)}]\sum_{n=-\infty}^{\infty}\delta\left(\omega-n\frac{2\pi}{2T}\right). \tag{29}$$

But,

$$|1+e^{-j\omega(T+\epsilon)}|^2=2[1+\cos\omega(T+\epsilon)], \tag{30}$$

so that the power spectrum consists of impulses every $1/2T$ Hz with an amplitude fluctuation of $[1+\cos\omega(T+\epsilon)]$. If there is not jitter, then $\epsilon=0$; and, since $[1+\cos\omega T]=0$ for $\omega=(\pi/T)n$ (where $n=1,3,5,\ \cdots$),

[11] B. Gold and J. Tierney, "Pitch-Induced Spectral Distortion in Channel Vocoders," J. Acoust. Soc. Am. **35**, 730–731 (1963).

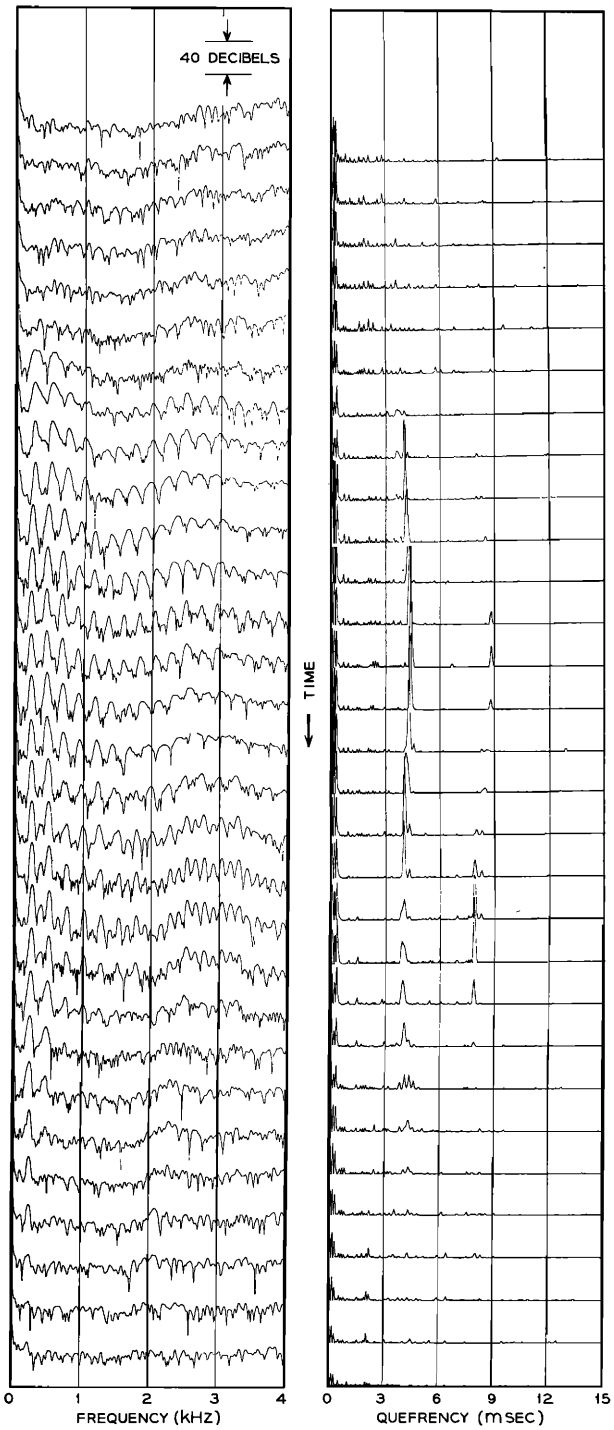


FIG. 10. Short-time logarithm spectra (left) and short-time cepstra (right) of "chase," spoken by a female talker (B.M.) and recorded with a condenser microphone. The 19th through 21st cepstra have second rahmonics that exceed the fundamental and that would result in an undesired indication of pitch-period doubling.

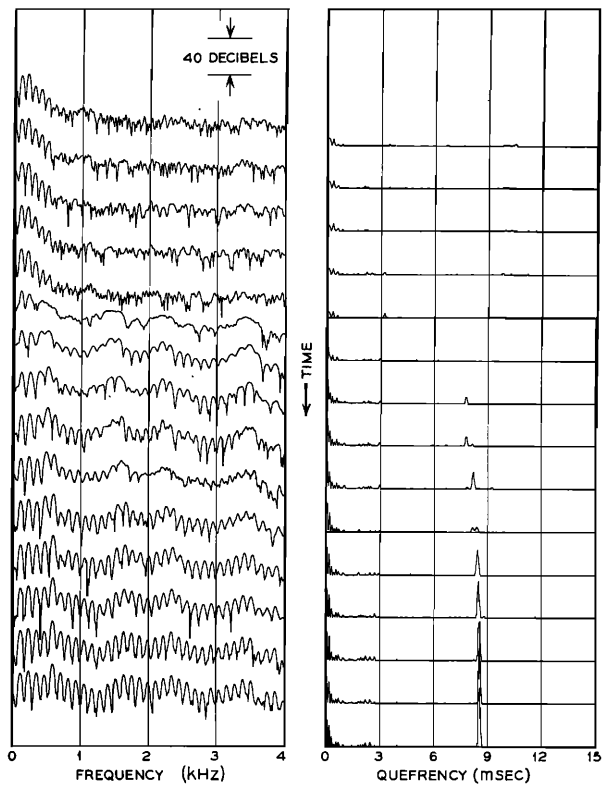


FIG. 11. Short-time logarithm spectra (left) and short-time cepstra (right) of "(o)be(y)," spoken by a male talker (R.C.L.) and recorded with a condenser microphone. The explosion occurs at the sixth spectrum and cepstrum.

the odd harmonics disappear, thereby leaving impulses every $1/T$ Hz. However, if ϵ is not zero, the spectrum starts with impulses spaced $1/T$ Hz, but gradually impulses appear at $1/2T$-Hz intervals and then periodically fade in and out across the spectrum. The jitter can be calculated from the frequency in the spectrum at which the amplitude of the impulses are first equal, since at this frequency the Nth cosine wave with period $1/(T+\epsilon)$ Hz has a maximum situated exactly between two adjacent impulses. For the spoken word *chase*, this occurred at 3 kHz corresponding to an $\epsilon \approx 0.08$ msec, which is smaller than the accuracy of one previous measurement of pitch perturbations.[12]

The spectra and cepstra shown in Figs. 11–13 are for a male speaker (R.C.L.) recorded with a condenser microphone. These speech utterances were chosen by O. Fujimura in his investigations at Bell Telephone Laboratories of speech sounds. The first set of spectra and cepstra show the explosion in the word *obey* (occurring at the sixth line of Fig. 11) as exemplified by a completely ripple-free spectrum. Figure 13 shows the spectra and cepstra for the voiced fricative portion of the word *razor* at the sixth through ninth lines.

V. AUTOMATIC TRACKING OF CEPSTRAL PEAKS

The cepstral peaks corresponding to voiced speech intervals can easily be picked visually. However, these

[12] P. Lieberman, "Perturbations in Vocal Pitch," J. Acoust.Soc. Am. 33, 597–603 (1961).

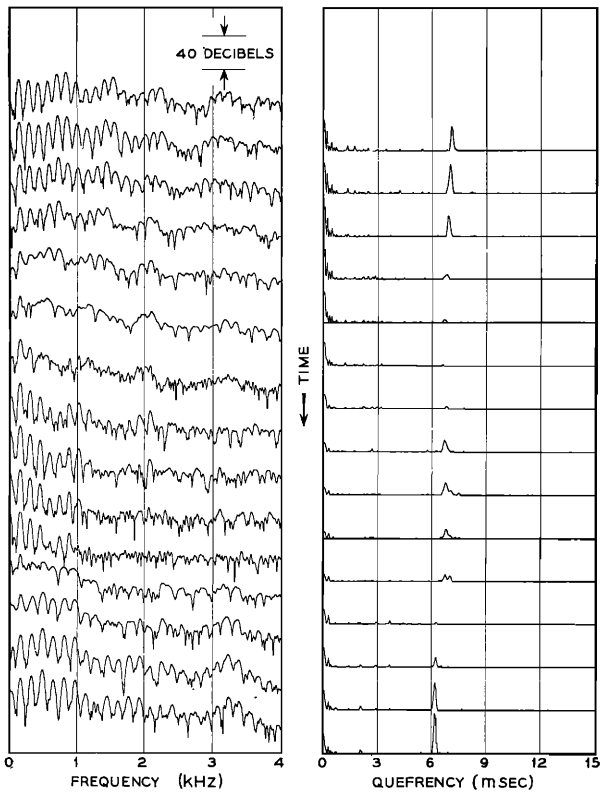


FIG. 12. Short-time logarithm spectra (left) and short-time cepstra (right) of "(b)abbl(ed)," spoken by a male talker (R.C.L.) and recorded with a condenser microphone. The explosion occurs at the sixth spectrum and cepstrum.

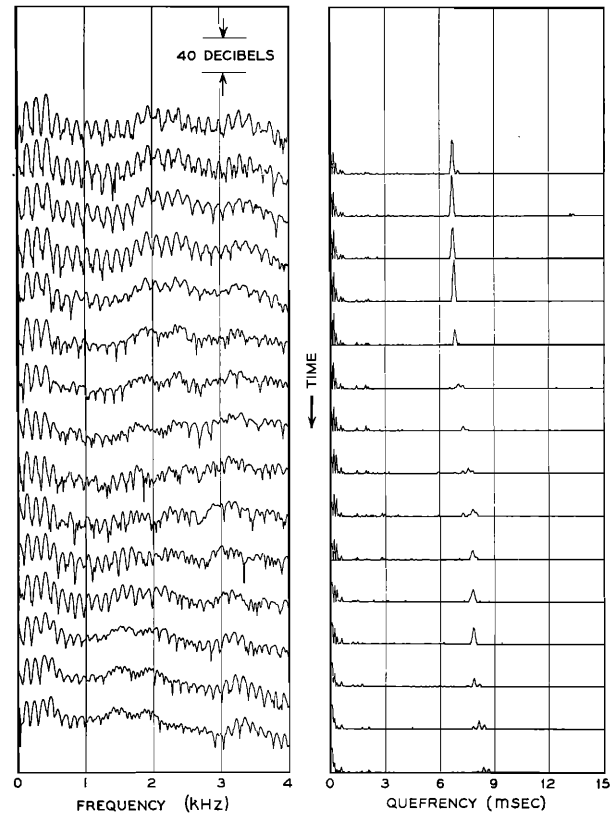


FIG. 13. Short-time logarithm spectra (left) and short-time cepstra (right) of "(r)azor," spoken by a male talker (R.C.L.) and recorded with a condenser microphone. The voiced fricative occurs at the sixth through ninth spectra and cepstra.

peaks must be picked automatically if cepstrum techniques are to be used in a pitch detection scheme. This section of the paper describes the heuristic development of an algorithm for picking the cepstral peak that best describes the pitch of the speech for that time interval. The criterion of "best" was evaluated by using the pitch data as excitation of a computer-simulated vocoder and then comparing the vocoded speech with the original speech.

The examples of cepstra indicate that the cepstral peaks are clearly defined and are quite sharp. Hence, the peak-picking scheme is to determine the maximum value in the cepstrum exceeding some specified threshold. Since pitch periods of less than 1 msec are not usually encountered, the interval searched for the peak in the cepstrum is 1–15 msec.

Since the cepstral peaks decrease in amplitude with increasing quefrency, a linear multiplicative weighting was applied over the 1–15-msec range. The weighting was 1 at 1 msec and 5 at 15 msec. The Fourier transform of the power spectrum of the time window equals the convolution of the time window with itself,

$$\mathfrak{F}[|W(\omega)|^2]=\mathfrak{F}[W(\omega)W(-\omega)] \qquad (31)$$
$$=w(t)*w(-t).$$

Thus, the higher-quefrency components in the power spectrum decrease as the time window convolved with itself. Although the mathematics becomes unwieldy for an exact solution, it is reasonable to expect the higher-quefrency components in the logarithm of the power spectrum to decrease similarly, thereby explaining the need of weighting of the higher quefrencies in the cepstrum. The linear weighting with range of 1–5 was chosen empirically by using periodic pulse trains with varying periods as input to the cepstrum program.

The cepstral peaks at the end of a voiced-speech segment usually decrease in amplitude and would fall below the peak threshold. The solution is to decrease the threshold by some factor (2) over a quefrency range of ± 1 msec of the immediately preceding pitch period when tracking the pitch in a series of voiced-speech segments. The threshold reverts to its normal value over the whole cepstrum range after the end of the series of voiced segments.

There is also the possibility that an isolated cepstral peak might exceed the threshold, thereby resulting in a false indication of a voiced speech segment. In fact, some isolated flaps of the vocal cords have been observed as the cause of such an isolated cepstral peak. In any event, such peaks should not be considered as

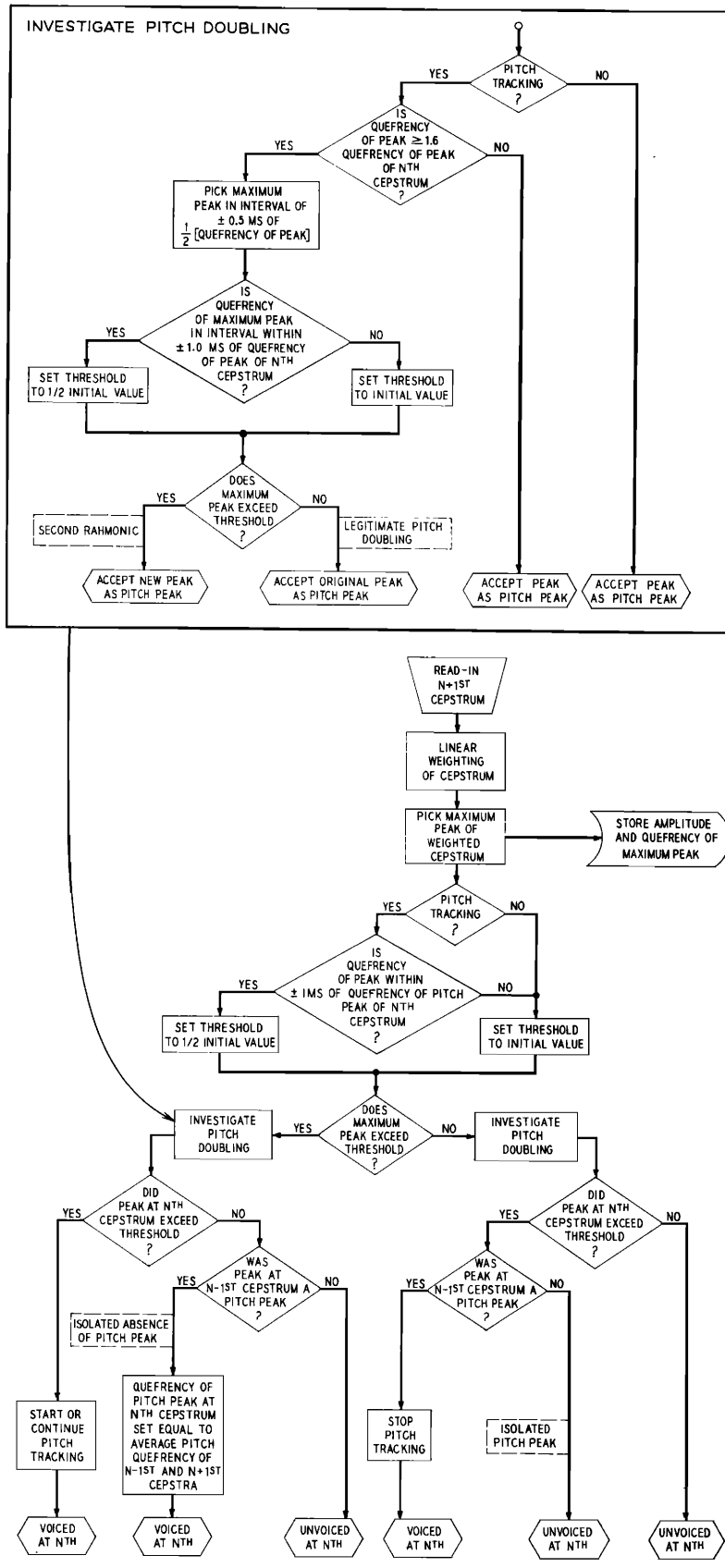


FIG. 14. Flow chart of the algorithm used to decide if the Nth cepstrum represents a voiced speech interval.

voiced, and this is accomplished by disregarding any cepstral peaks exceeding the threshold if the immediately preceding cepstrum and immediately following cepstrum indicate unvoiced speech. This means that the immediately following cepstrum must be peak searched before a decision can be made about the present cepstrum. Hence, a delay of one cepstrum must be introduced to eliminate this requirement of knowledge about the future. Before deciding about the "present" cepstrum, however, knowledge about the preceding and following cepstrum is also required for the algorithm used to eliminate another problem, namely, pitch doubling.

An example of legitimate pitch doubling occurred at the end of the word *screaming*, as shown in Fig. 8. However, the second rahmonic of a cepstral peak sometimes exceeds the fundamental, and the second rahmonic should not be chosen as representing the pitch period. Thus, the peak picking algorithm should eliminate false pitch doubling caused by a second rahmonic but should also allow legitimate pitch doubling. For legitimate doubling, there is no cepstral peak at a one-half quefrency, but for erroneous doubling, there is such a peak at one-half quefrency since this is the fundamental. The algorithm capitalizes upon this observation by looking for a cepstral peak exceeding the threshold in an interval of ± 0.5 msec of one-half the quefrency of the double-pitch peak. If such a peak is found, then it is assumed that it represents the fundamental, and the double-pitch indication is wrong. The threshold is reduced by a factor of 2 if the maximum peak in the ± 0.5-msec interval falls within ± 1.0 msec of the immediately preceding pitch period. Pitch doubling has occurred whenever the cepstral peak exceeding the threshold is at a quefrency of ≥ 1.6 times the immediately preceding pitch period.

A flow chart of the peak-picking algorithm is shown in Fig. 14. The algorithm determines whether the cepstral peak of the Nth cepstrum represents a voiced speech segment. Information about the $N-1$th cepstrum is stored, and the $N+1$th cepstrum is peak picked before deciding about the Nth cepstrum. The $N+1$th cepstrum is read in, linear weighting is applied, and the maximum peak is picked. If the preceding two cepstra represented voiced-speech segments, then pitch tracking is in effect, and the threshold is reduced to $\frac{1}{2}$ its initial value if the quefrency of the peak is within ± 1.0 msec of the quefrency of the pitch peak of the Nth cepstrum. The previously determined peak in the $N+1$th cepstrum is now compared with the threshold. Pitch doubling is investigated whether the peak exceeds or does not exceed the threshold. Both cases are checked since the peak might represent pitch doubling and yet not exceed the initial value of the threshold. But, the fundamental peak could still exceed the $\frac{1}{2}$ initial value threshold. If the maximum peak exceeds the threshold, it is tentatively chosen as a pitch peak representing a voiced-speech segment at the $N+1$th cepstrum. The

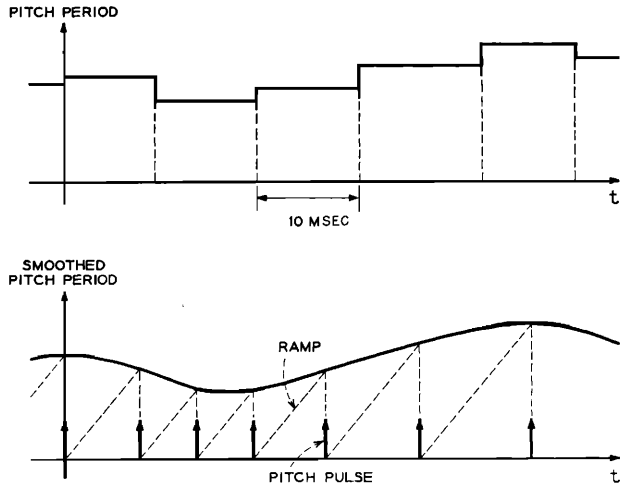


FIG. 15. Method for deriving pitch pulses from pitch period data supplied by cepstral peak picker.

information about the $N+1$th cepstrum and $N-1$th cepstrum is then used to decide if the Nth cepstral peak represents an isolated voiced segment or an isolated absence of voicing in a series of voiced-speech segments. The final result is an indication of whether the Nth cepstrum represents a voiced or an unvoiced speech segment. If the segment is voiced, the pitch period is also given.

A computer program was written to perform the operations required by the algorithm. The voicing and pitch-period information were both printed on paper and written on magnetic tapes for later processing.

VI. VOCODER EXCITATION

The final judge of any vocal-pitch detection scheme is its ability to perform satisfactorily in determining the excitation for a vocoder. Vocodor excitation in the form of pitch pulses during voicing and white noise during nonvoicing thus had to be derived from the results of the cepstral peak picking.

The cepstral peak picker produced two outputs on digital magnetic tape. The first tape contained voicing information as two dc levels corresponding to a voiced or unvoiced speech interval. The levels were constant for the 10-msec corresponding to the speech time jumps. The second tape contained the pitch period as dc level signals that also were constant for 10 msec. These two tapes formed the input to the excitation generator.

The voicing and pitch-period signals are first each smoothed by a pair of 33-Hz low-pass filters. The pitch pulses are derived from the smoothed pitch signal as shown in Fig. 15 by running a counter up until it equals the smoothed pitch signal. An impulse is then emitted, and the counter is reset to zero before again starting its count. If the smoothed pitch signal is measured in tenths of a millisecond and the counter counts in tenths of a millisecond, then the timing between the emitted

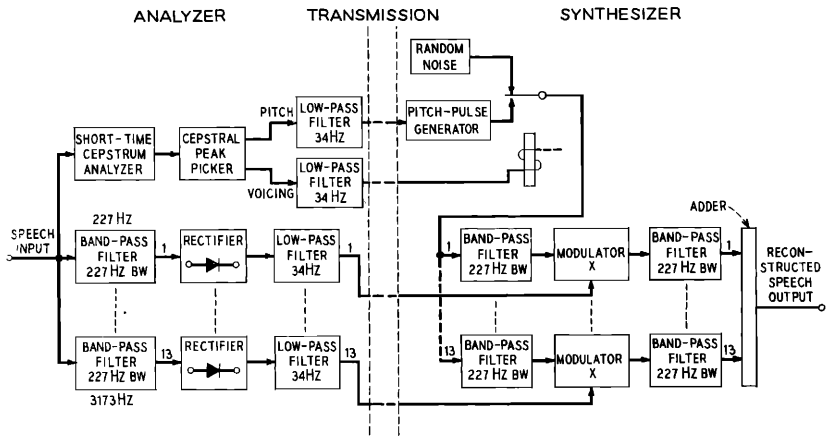


FIG. 16. Block diagram of 13-spectrum channel vocoder with excitation derived from a cepstrum pitch detector.

impulses equals the pitch period. The smoothed voicing signal is used to control a double-throw switch for choosing either pitch pulses or white noise as a final excitation output.

This technique was devised and simulated on the computer by M. M. Sondhi using the BLODI programming language. The output of the program was still another digital magnetic tape, which was then used as the excitation input to a 13-spectrum channel vocoder designed by Golden.[13] The vocoder was also simulated on the computer using the BLODI programming language. The spectrum channel information was derived from a computer-simulated vocoder analyzer and, together with the excitation, they formed the input to the computer-simulated synthesizer. The whole operation from speech signal to simulated vocoder output is shown in Fig. 16. The digital computer generated numerous visual outputs on microfilm including the short-time spectra and cepstra, the voicing and pitch-period variations, the original speech signal, and the vocoded speech signal. These visual outputs were extremely valuable in devising the final versions of all the different portions of the chain making up the complete pitch-detection scheme.

The complete scheme, including the vocoder, was used to modify and improve all portions of the chain by comparing the vocoded speech with the original speech. In particular, the pitch-period doubling at the end of the word *screaming* was determined to be aurally correct by such a comparison of original with vocoded speech.

The synthetic speech from the computer-simulated cepstrum-excitation channel vocoder was compared both with the original speech and with the synthetic speech from a computer-simulated voice-excited vocoder and the same computer-simulated channel vocoder, but with the full-band speech as excitation. Although only a few sentences spoken by four talkers were used in these informal paired-comparison tests, the pitch quality of the channel vocoder with cepstrum pitch detection was judged to be excellent by experienced vocoder critics. This optimism was sufficient to initialize construction of a real-time cepstrum pitch detector.[14]

VII. IMPLEMENTATION OF CEPSTRUM ANALYZERS

In its most basic form, cepstrum-pitch detection requires two spectrum analyses with logic circuitry for picking the cepstral peak corresponding to the pitch period of a voiced-speech segment. Thus, a means for performing two spectrum analyses in real time is required for a hardware implementation of a cepstrum-pitch detector. The requirements of real-time operation and good frequency resolution in the spectrum analyzers are somewhat difficult to satisfy and have therefore resulted in the correct opinion that a hardware cepstrum analyzer would be difficult to construct.

However, techniques are available for performing real-time spectrum analyses that could be adapted to cepstrum analysis. One such method performs the spectrum analysis by a circulating delay line with a time-variable phase shifter operating upon a heterodyned version of the time signal. This method, described by Bickel and Bernstein[15] has been successfully used by Weiss, Vogel, and Harris in an implementation of a cepstrum analyzer.[16,17] Still another method, similar to a spectrum analyzer described by Gill, uses a heterodyne filter operating on a time-swept version of the input signal.[18] Kelly and Kennedy have utilized this

[13] R. M. Golden, "Digital Computer Simulation of a Sampled-Data Voice-Excited Vocoder," J. Acoust. Soc. Am. **35**, 1358–1366 (1963).

[14] J. M. Kelly and R. N. Kennedy, "An Experimental Cepstrum Pitch Detector for Use in a 2400-bit/sec Channel Vocoder," presented at the 72*nd* meeting of Acoustical Society of America (Nov. 1966), Paper 1H3.

[15] H. J. Bickel and R. I. Bernstein, U. S. Patent No. 3,013,209.

[16] H. J. Bickel, "Spectrum Analysis with Delay-Line Filter," IRE WESCON Conv. Rec. **1959** (Part 8), 59–67 (1959).

[17] M. R. Weiss, R. P. Vogel, and C. M. Harris, "Implementation of a Pitch Extractor of the Double-Spectrum-Analysis Type," J. Acoust. Soc. Am. **40**, 657–662 (1966).

[18] J. S. Gill, "A Versatile Method for Short-Term Spectrum Analysis in 'Real-Time,'" Nature **189**, No. 4759, 117–119 (14 Jan. 1961).

method in yet another successful implementation also including logic circuitry to track the cepstral peak.[14] They have also derived vocoder excitation from their cepstra and have produced excellent-quality vocoded speech utilizing a complete hardware system of cepstrum analyzer and vocoder.

Both methods utilize analog-circuit techniques during all or part of the spectrum analysis. Digital techniques, however, have progressed to the state where a completely digital implementation should be possible. The Cooley–Tukey algorithm greatly reduces the number of multiplications and additions, and might be of practical use in such a completely digital cepstrum analyzer.

Another promising method utilizes the spectrum analyzing properties of a lens.[19,20] A lens forms at its focal plane an image that is the Fourier transform of the image at the object plane. Since this is a *spatial* Fourier transform, the signal must be frozen in time with light intensity made proportional to signal amplitude. A coherent light source is required to illuminate the spatial representation of the signal, and there are some questions concerning the most efficient way to convert the time signal into such a spatial signal. But, the technique seems particularly promising (since parallel processing is very convenient), so that thousands of signals could be analyzed almost simultaneously.

VIII. PSEUDO-AUTOCOVARIANCE OR CEPSTRUM?

In their article in Rosenblatt's book, Bogert *et al.*,[3] define the cepstrum as "autocovariance and Fourier transformation . . . [of] the log spectrum of the original process." Since the Fourier transform of the autocovariance of some function is identical with the power spectrum of the same function, the cepstrum should be equivalent to the power spectrum of the log power spectrum of the original process. Furthermore, since the log power spectrum is an even function of frequency, the cepstrum should equal the square of the cosine transform of the log power spectrum.

Later in the article, Bogert *et al.* define a pseudo-autocovariance as "the Fourier transform of [the] log· · · power spectrum." The "pseudo" prefix is logically used since the Fourier transform of the non-logged power spectrum is the usual autocovariance. Thus, the cepstrum should equal the square of the pseudo-autocovariance. But, in their definition of the cepstrum, Bogert *et al.* had meant to assume that the log spectrum existed for all positive frequencies (private communication). As a result, their cepstrum equals the sum of the squares of the sine transform and the cosine transform of the log power spectrum. Stated mathematically, their definition of the cepstrum is

$$C_{\text{Bogert}}(\tau)=\{\mathfrak{F}_{\sin}[\log|F(\omega)|^2]\}^2+\{\mathfrak{F}_{\cos}[\log|F(\omega)|^2]\}^2, \quad (32)$$

where $\mathfrak{F}_{\sin}$ and $\mathfrak{F}_{\cos}$ denote Fourier sine transformation and Fourier cosine transformation, respectively; $F(\omega)$ is the complex Fourier transform of the original process; and $F(\omega)\equiv 0$ for $\omega<0$. The pseudo-autocovariance is

$$R_c(\tau)=\mathfrak{F}_{\cos}[\log|F(\omega)|^2] \quad (33)$$

and its square is identical with the definition of the cepstrum used in this paper. A pseudoquadrature autocovariance can be defined as

$$R_s(\tau)=\mathfrak{F}_{\sin}[\log|F(\omega)|^2], \quad (34)$$

so that

$$C_{\text{Bogert}}(\tau)=[R_c(\tau)]^2+[R_s(\tau)]^2. \quad (35)$$

Two different definitions of the cepstrum can certainly lead to some confusion, but in this paper the cepstrum has consistently been defined as the square of the cosine transform of the log power spectrum.

The digital computer was programmed to calculate the following short-time functions: the square of the cosine transform of the one-sided log power spectrum (pseudo-autocovariance squared), the square of the sine transform of the one-sided log power spectrum (pseudoquadrature autocovariance squared), and the sum of the squares of the cosine and sine transorms of the one-sided log power spectrum (Bogert's cepstrum). The input signal was a male talker recorded from a 500-type telephone handset with additive white noise (signal-to-noise ratio approximately 12 dB). The three short-time functions with the corresponding log power spectrum are shown in Fig. 17 for a voiced speech segment. The pseudoquadrature autocovariance is very noisy, so that Bogert's cepstrum does not have peaks as sharp as the pseudo-autocovariance alone. Clearly, in retrospect, these results are good justification for using only the pseudo-autocovariance for speech pitch detection.

IX. CONCLUSION

Some of the advantages claimed for cepstrum pitch detection and confirmed by computer simulation are, first, that the fundamental frequency component need not be present in the time signal, since the spectral ripples or fine structure caused by the harmonics give rise to the cepstral peak. For this reason, cepstrum pitch detection is particularly well suited to such bandpass-filtered signals as telephone speech. Since only the power spectrum is used, phase is completely ignored. Additive white noise is not too degrading if it does not destroy the spectral ripples. Actually, a clearly defined cepstral peak has been obtained for speech signals with a 6-dB signal-to-noise ratio over the 40-msec analysis interval.

[19] L. J. Cutrona, E. N. Leith, C. J. Palermo, and L. J. Porcello, "Optical Data Processing and Filtering Systems," IRE Trans. Information Theory **IT-6**, 386–400 (1960).

[20] B. Julesz, A. M. Noll, and M. R. Schroeder, "Optical Cepstrum Analysis" (unpublished morandum).

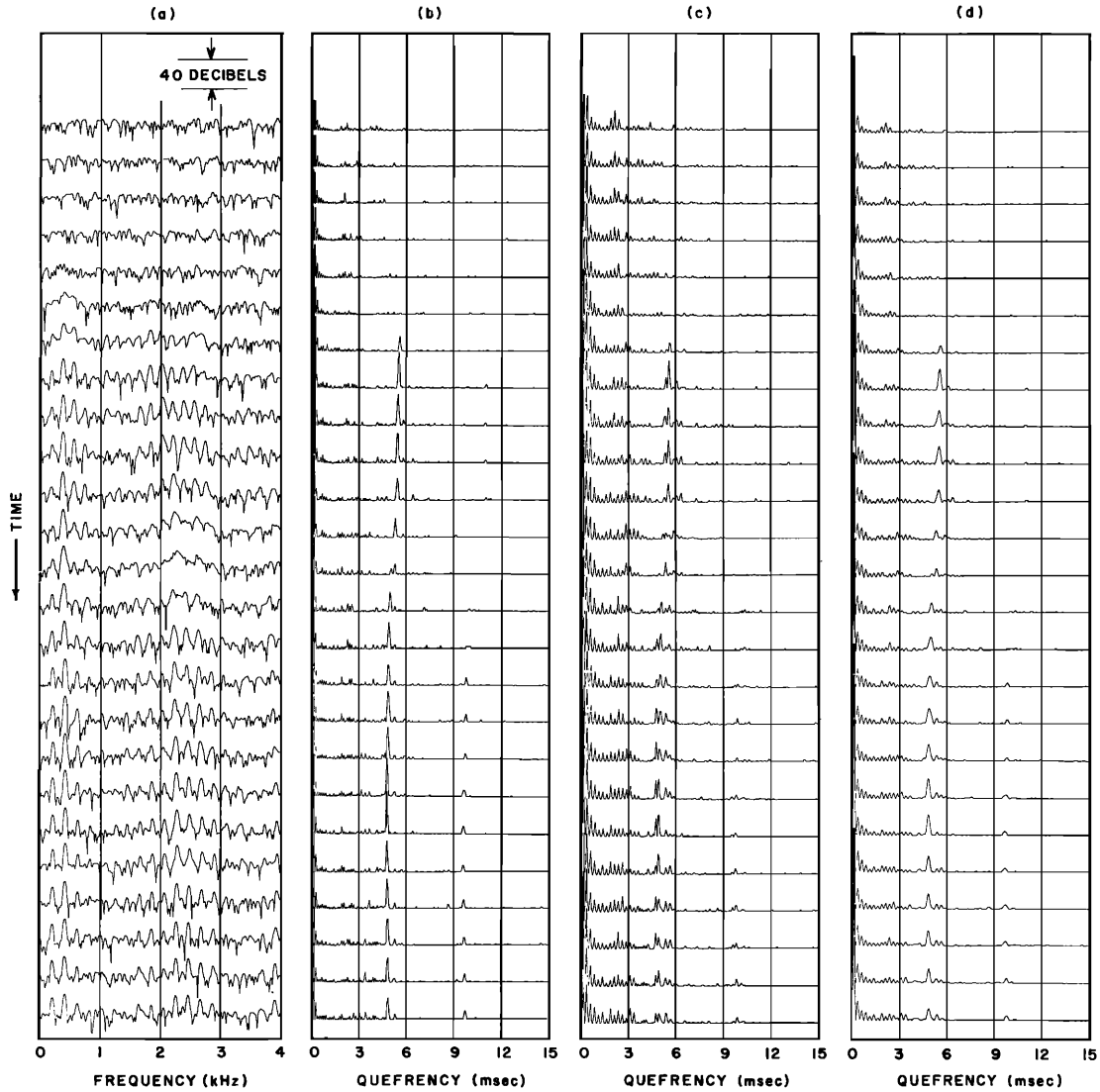


FIG. 17. (a) Short-time logarithm spectra, (b) pseudo-autocovariance squared, (c) pseudoquadrature autocovariance squared, and (d) Bogert's cepstra (defined as the sum of the squares of the cosine and sine transforms of the logarithm spectra) for a male talker (F.L.C.) recorded from a 500-type telephone handset and with additive white noise (signal-to-noise ratio $\approx$12 dB).

Of course, cepstrum pitch detection is insensitive to narrow-band white noise, since such noise would at most obscure only a few spectral ripples.

Cepstrum pitch detection has to some extent changed our over-all concept of a vocoder. Previously, most diagrams of a channel vocoder showed considerable detail about the channel filters while the pitch detector was usually shown as a small block at the bottom, although the pitch detector itself was sometimes quite elaborate. However, the spectrum-channel information is obtained as an intermediate step during the cepstrum-analysis process. Thus, our new concept of a vocoder analyzer shows an involved diagram of a pitch detector with the spectrum-channel information obtained as a by-product! (See Fig. 18.) Perhaps this is more realistic, because it has long been recognized that accurate pitch information is the most challenging aspect of vocoder design. The spectrum-channel information has perhaps been reduced to its true relative importance.

But where does all this effort lead us?. It seems that vocoder design is becoming conceptually more complicated with asymptotic, though not necessarily insignificant, improvements in quality. The vocoder schemes and pitch detectors are becoming increasingly exotic, as exemplified by cepstrum pitch detection. Also, such new speech transmission methods as microwave, satellites, and the promise of light communication over laser beams might someday change the present restrictions on available bandwidth. The future of vocoders for speech bandwidth compression might seem bleak. Why

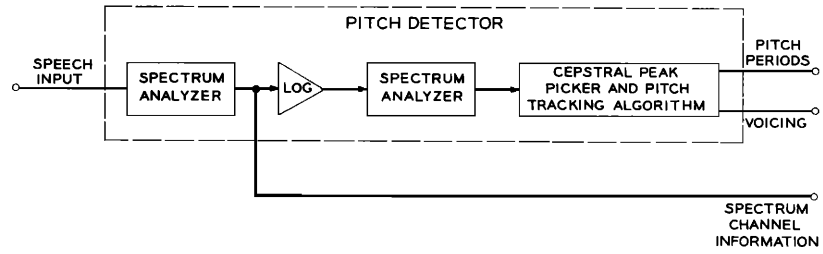


FIG. 18. New concept of spectrum channel vocoder in which the spectrum channel information is obtained as a by-product of the cepstrum pitch detector.

continue, then, with vocoder development, and—in particular—why be concerned with pitch detectors?

Special-purpose vocoders can be useful in removing certain types of speech distortion. For example, the "Donald Duck" quality of speech spoken in the helium environment used in certain underwater quarters such as Sealab can be eliminated by frequency shifting of the vocoder channel signals.[21] The transmission of speech can be made private or secure by the use of vocoders.[8] An accurate pitch-detection scheme would become a very important tool in speech research by fostering research in pitch fluctuations and patterns. Thus, further research and development of pitch detectors is warranted not only to produce speech bandwidth-compression vocoders but also as a fundamental tool for speech research and for special-purpose vocoders.

As mentioned previously, cepstrum analysis performs remarkably well as a vocal-pitch detector. However, a more general conclusion has evolved from the concept of cepstrum analysis: that the spectrum itself can be regarded as a signal and can be processed by standard signal-analysis techniques. With such a viewpoint, cepstrum analysis and other signal processing of the spectrum do not seem quite so exotic.

[21] R. M. Golden, "Improving Naturalness and Intelligibility of Helium–Oxygen Speech Using Vocoder Techniques," J. Acoust. Soc. Am. 40, 621–624 (1966).