[9] S. P. Chan, *Introductory Topological Analysis of Electrical Networks*. New York: Holt, Rinehart, and Winston, 1969.
[10] S. Seshu and M. B. Reed, *Linear Graphs and Electrical Networks*. Reading, Mass.: Addison-Wesley, 1961.
[11] J. F. Kaiser, "Some practical considerations in the realization of linear digital filters," in *Proc. 3rd Allerton Conf. Circuit and System Theory*, 1965, pp. 621–633.
[12] R. Golden, "Digital filters," in *Modern Filter Theory and Design*. New York: Wiley-Interscience, 1973, ch. 12.
[13] L. B. Jackson, J. F. Kaiser, and J. F. McDonald, "An approach to the implementation of digital filters," *IEEE Trans. Audio Electroacoust. (Special Issue on Digital Filters: The Promise of LSI Applied to Signal Processing)*, vol. AU-16, pp. 413–421, Sept. 1968.

Epoch Extraction of Voiced Speech

T. V. ANANTHAPADMANABHA AND B. YEGNANARAYANA

Abstract—**A general theory of epoch extraction of overlapping nonidentical waveforms is presented. The theory is applied to outputs of models of voiced speech production mechanism and to actual speech data. Some typical glottal waveshapes are considered to explain their effect on the speech output. It is shown that the points of excitation of the vocal tract can be precisely identified for continuous speech. It is possible to obtain accurate pitch information by this method even for high-pitched sounds. The epoch extraction has wide applications in speech analysis, speaker verification, speech synthesis, and pitch perception studies.**

I. Introduction

SPEECH can be considered as the output of a linear system for which neither the excitation nor the system is known. In particular, voiced speech is formed by exciting a quasi-stationary vocal tract system with quasi-periodic puffs of air [1]. The responses due to successive excitations overlap forming a superposition of nearly identical waveforms. Such a superposition of multiple signals is referred to as a composite signal [2]. The decomposition of composite signals is of considerable importance in speech analysis and synthesis. The formation of composite signals can also be seen in other areas like seismology [3], radar [4], electrophysiology [2], etc.

The main difficulty in the analysis of voiced speech is the overlapping of successive impulse responses in time domain and the lack of knowledge of the excitation waveform. Linear filters can be designed to separate out superposed waveforms [2] provided of course the individual waveforms have energy in different frequency bands. In voiced speech the successive responses are nearly identical and hence cannot be separated in the frequency domain also. A convenient alternative viewpoint is to regard voiced speech as a convolution of impulses at the pitch epochs and the impulse response of the vocal tract system. The system response now incorporates the glottal waveform also [5]. Voiced speech analysis is then essentially a deconvolution problem aimed at obtaining the resonances of the vocal tract and the characteristics of the excitation. The characteristics of the excitation are the glottal waveform, the duty ratio, the pitch period, and the excitation epochs. In this paper we are mainly concerned with the epoch extraction.

Inverse filtering techniques [6], [7] have been proposed for finding the glottal waveform. In these techniques speech signal is passed through an inverse network which removes the first or the first two resonances of the vocal tract. The output of the inverse network is then passed through a suitable low-pass filter to obtain the glottal waveforms. The parameters of the inverse filter need to be adjusted depending upon the formant structure of the vocal tract. This requires a priori estimate of the formants and their bandwidths and hence the inverse filtering technique is suited for sustained vowels only.

In pitch-synchronous analysis [8] of voiced sounds the spectrum of a single pitch period of voiced speech is fitted with the spectrum of an assumed pole-zero model in some optimal sense. The poles are ascribed to the vocal tract system and the zeros to the excitation. The impulse response of the all-zero transfer function gives an estimate of glottal waveform.

Both the above approaches are unsuitable for the analysis of high-pitched sounds. The duration available for analysis will not adequately represent the impulse response of the vocal tract system.

Some of the methods [9]–[13] for the determination of the pitch period are based on the assumption that voiced speech is a result of a periodic impulse train exciting a time-invariant linear system. The spectrum of the impulse train will be periodic in the frequency domain with a quefrency of the pitch period. The gross resonant structure of the vocal tract is removed from the spectrum of voiced speech to obtain a flat spectrum containing the periodicities due to pitch. It is possible to get only the average pitch over the chosen duration by means of this technique because it uses only the spectral information. The exact pitch initiation points can be obtained only when the actual phase information is also used. The accuracy of pitch estimation is therefore limited due to the simplified model assumed. Moreover, these methods fail to extract the pitch information from high-pitched sounds. This is due to the fact that there is considerable temporal overlap of the impulse responses which causes the spectral difference between the system and the excitation less marked.

Manuscript received December 4, 1974; revised February 2, 1975.
The authors are with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012, India.

At present there is no reliable technique for obtaining epochs or the instants of excitation of the vocal tract [1]. Although the closure of the glottis is considered as the point of excitation, it could occur at other points as well [7], [14]. Usually the epoch is regarded as the point of maximum discontinuity in the derivative of the glottal waveform [1]. It is generally difficult to extract the epochs from speech waveform due to the varying shapes of the glottal pulses and the nonstationary nature of the vocal tract.

Through our study of composite signal decomposition, we have arrived at a technique for epoch determination of identical wavelets overlapping in time [15]. We shall present here the generalization of the technique for epoch extraction of voiced segments of speech.

II. Theory of Epoch Extraction

A. Epochs

An epoch can be defined as the occurrence of a reference point on a given waveform. The reference point can be selected so as to have some significance as related to the waveform. For example, the starting point of a wavelet can be chosen to be an epoch [16], [17]. In the present work epochs are defined as follows.

Let $f(t)$ be a function defined over an interval (a, b) and zero outside the interval. Also let $f(t)$ possess continuously differentiable derivatives in the interval (a, b). Then the point of discontinuity of the lowest ordered derivative will be regarded as an epoch. The epoch therefore can occur either at a or at b or at both a and b. Here by the term "discontinuity of the lowest ordered (n) derivative" we mean that at $t = t_i$, $f^{(n+1)}(t)$ is discontinuous, but $f^{(n)}(t)$ is continuous. Some examples are given in Table I. The value of $(n + 1)$ will be referred to as the order of the epoch.

In the above definition we have restricted the function to possess continuously differentiable derivatives in the interval (a, b). If the function is piecewise continuous, then the interval (a, b) can be further subdivided so that the restriction is satisfied in each of the subintervals. Then the number of epochs could be more than two for a given function. Thus, for example, for a triangular function there will be epochs of order one at the ends as well as at the apex.

B. The Asymptotic Expansion of Fourier Transform and the Epochs

For the type of function $f(t)$ defined in Section II-A, it has been shown [18] that the Fourier transform $F(\omega)$ can be written as

$$F(\omega) = \frac{f(a)e^{-j\omega a} - f(b)e^{-j\omega b}}{j\omega} + \cdots + \frac{f^{(n)}(a)e^{-j\omega a} - f^{(n)}(b)e^{-j\omega b}}{(j\omega)^n} + \cdots. \quad (1)$$

Let us assume that the function $f(t)$ has an epoch of order k at $t = t_i$ where $t_i = a$ or b. Then $f^{(n)}(t) = 0$ outside the interval (a, b) for all n. Since $f^{(n)}(t)$ is continuous at $t = t_i$ for $n = 0, 1, \cdots, k - 1$, it follows that

TABLE I
Examples of some Functions and Their Epochs

FUNCTION	EPOCHS	ORDER
$f(t) = u(t-a) - u(t-b)$	$t_i = a, b$	0 $f(t_i^-) \neq f(t_i^+)$
$f(t) = t\exp(-kt)$	$t_i = 0$	1 $f(t_i^-) = f(t_i^+)$ $f^{(1)}(t_i^-) \neq f^{(1)}(t_i^+)$
$f(t) = \begin{cases} .5 - .5\cos[2\pi(t-a)/(b-a)] & \text{for } a \leq t \leq b \\ 0 & \text{otherwise} \end{cases}$	$t_i = a, b$	2 $f(t_i^-) = f(t_i^+)$ $f^{(1)}(t_i^-) = f^{(1)}(t_i^+)$ $f^{(2)}(t_i^-) \neq f^{(2)}(t_i^+)$
$f(t) = \begin{cases} k_1(t-a) & a \leq t \leq c \\ A - k_2(t-c) & c \leq t \leq b \\ 0 & \text{otherwise} \end{cases}$	$t_i = a, b, c$	1 $f(t_i^-) = f(t_i^+)$ $f^{(1)}(t_i^-) \neq f^{(1)}(t_i^+)$

$$f^{(n)}(t_i^-) = f^{(n)}(t_i^+) = 0 \text{ for } n = 0, 1, \cdots, k - 1. \quad (2)$$

For large ω we can regard

$$\frac{1}{\omega^n} << \frac{1}{\omega^k} \text{ for } n > k. \quad (3)$$

Combining (1)–(3) we get

$$F(\omega)\,|_{\text{large }\omega} \approx R(\omega)e^{-j\omega t_i} \quad (4)$$

where

$$R(\omega) = f^{(k)}(t_i)/(j\omega)^k. \quad (5)$$

C. Theory of Epoch Filter

In this section we shall use the above definition of epochs and the asymptotic theorem to develop the theory of epoch filter for the extraction of epochs. Let the frequencies ω_c and $\omega_c - B$ be sufficiently large so that (3) is satisfied. We have

$$\frac{1}{(\omega_c - B)^k} = \frac{1}{\omega_c^k}\left[1 - \frac{B}{\omega_c}\right]^{-k}. \quad (6)$$

If $kB/\omega_c << 1$, i.e., if B is very small compared to ω_c, then

$$\frac{1}{(\omega_c - B)^k} \approx \frac{1}{\omega_c^k}. \quad (7)$$

From (5) and (7) we get

$$R(\omega) \approx \frac{f^{(k)}(t_i)}{[j\,\text{sgn}(\omega)]^k\,\omega_c^k} \quad \text{for} \quad \omega_c - B \leq |\omega| \leq \omega_c. \quad (8)$$

For even values of k $(= 2l)$

$$R(\omega) \approx (-1)^l \frac{f^{(2l)}(t_i)}{\omega_c^{2l}} = P_i \text{ (an even real constant)}. \quad (9)$$

For odd values of $k(= 2l - 1)$

$$R(\omega) \approx j \operatorname{sgn}(\omega) \frac{f^{(2l-1)}(t_i)}{\omega_c^{2l-1}}$$

$$\approx j \operatorname{sgn}(\omega)\, Q_i \text{ (an odd imaginary constant)}. \quad (10)$$

In a more general situation like for example the complex echoed signal [19], $R(\omega)$ will have both real and imaginary parts. Then

$$R(\omega) \approx P_i + j\, Q_i \operatorname{sgn}(\omega). \quad (11)$$

We define a window function $G(\omega)$ as

$$G(\omega) = \begin{cases} 1 \text{ for } \omega_c - B \leqslant |\omega| \leqslant \omega_c \\ 0 \text{ otherwise}. \end{cases} \quad (12)$$

The signal $f(t)$ is passed through a filter whose transfer function is $G(\omega)$. Then the output $\hat{f}_1(t)$ of the filter is given by

$$\hat{f}_1(t) = \mathcal{F}^{-1}\{G(\omega)R(\omega)e^{-j\omega t_i}\}$$

$$= P_i\, g(t - t_i) - Q_i g_H(t - t_i) \quad (13)$$

where

$$g(t) = \frac{\sin(Bt/2)}{\pi t} \cos \omega_c t + \frac{1 - \cos(Bt/2)}{\pi t} \sin \omega_c t \quad (14)$$

and

$$g_H(t) = -\frac{1 - \cos(Bt/2)}{\pi t} \cos \omega_c t + \frac{\sin(Bt/2)}{\pi t} \sin \omega_c t. \quad (15)$$

For sampled signals

$$t = nT = 2\pi n/2\omega_c \quad (16)$$

where T is the sampling interval. Then

$$\sin \omega_c t = \sin n\pi = 0$$

and

$$\cos \omega_c t = \cos n\pi = (-1)^n.$$

Therefore, we get

$$g(nT) = (-1)^n \frac{\sin(BnT/2)}{\pi nT} \quad (17)$$

and

$$g_H(nT) = -(-1)^n \frac{1 - \cos(BnT/2)}{\pi nT}. \quad (18)$$

The alternate points of $\hat{f}_1(nT)$ have opposite signs. This effect can be removed by changing the sign of the alternate samples. The output $\hat{f}(nT)$ of the epoch filter now becomes

$$\hat{f}(nT) = P_i \frac{\sin[B(n - n_i)T/2]}{\pi(n - n_i)T} + Q_i \frac{1 - \cos[B(n - n_i)T/2]}{\pi(n - n_i)T} \quad (19)$$

where

$$n_i T = t_i.$$

The phase of $F(\omega)$ at large ω decides the values of P_i and Q_i. When $Q_i = 0$ we get a sampling function centered at the epoch. On the other hand if $P_i = 0$, the function $\hat{f}(nT)$ crosses zero at the epoch and will possess an odd symmetry with respect to the epoch position. If a peak is desired at the epoch in the latter case, the Hilbert transform $\hat{f}_H(nT)$ of $\hat{f}(nT)$ is computed.

$$\hat{f}_H(nT) = P_i \frac{1 - \cos[B(n - n_i)T/2]}{\pi(n - n_i)T} - Q_i \frac{\sin[B(n - n_i)T/2]}{\pi(n - n_i)T}. \quad (20)$$

In a general situation since both P_i and Q_i can exist, the epoch can be obtained by computing the sum of the squares of (19) and (20).

$$\hat{f}_a^2(nT) = \hat{f}^2(nT) + \hat{f}_H^2(nT). \quad (21)$$

Therefore,

$$\hat{f}_o(nT) = (P_i^2 + Q_i^2)^{1/2} \frac{2 \sin[B(n - n_i)T/4]}{\pi(n - n_i)T}. \quad (22)$$

This gives a sampling function around the epoch irrespective of the phase of $F(\omega)$ for large ω.

Since all steps up to (20) are linear operations, for multiple epochs we get

$$\hat{f}(nT) = \sum_i \left[P_i \frac{\sin[B(n - n_i)T/2]}{\pi(n - n_i)T} + Q_i \frac{1 - \cos[B(n - n_i)T/2]}{\pi(n - n_i)T} \right]. \quad (23)$$

However, if more than one epoch is present the nonlinear operation in step (21) may produce peaks slightly displaced from the true epochs due to crossterms. By choosing a proper frequency window function the sidelobe leakage of adjacent epochs can be minimized. Then (22) also can be generalized, i.e.,

$$\hat{f}_o(nT) = \sum_i (P_i^2 + Q_i^2)^{1/2} \frac{2 \sin[B(n - n_i)T/4]}{\pi(n - n_i)T}. \quad (24)$$

III. Computational Considerations

The resolution of epochs at the output of the epoch filter depends upon the choice of the window width (B). A nearly flat spectrum over a large width can be obtained by sampling the signal at a higher rate [15]. The sidelobes associated with the uniform window function $G(\omega)$ affect the epoch estimation. Hence a proper window function [20] must be chosen. We have used a discrete Hanning window centered around the folding frequency in the discrete Fourier transform (DFT) domain.

Generally, the two endpoints of an input frame act as impulse discontinuities. The contribution of these discontinuities in the frequency domain is large compared to that of the epochs. This results in large peaks at the endpoints of the

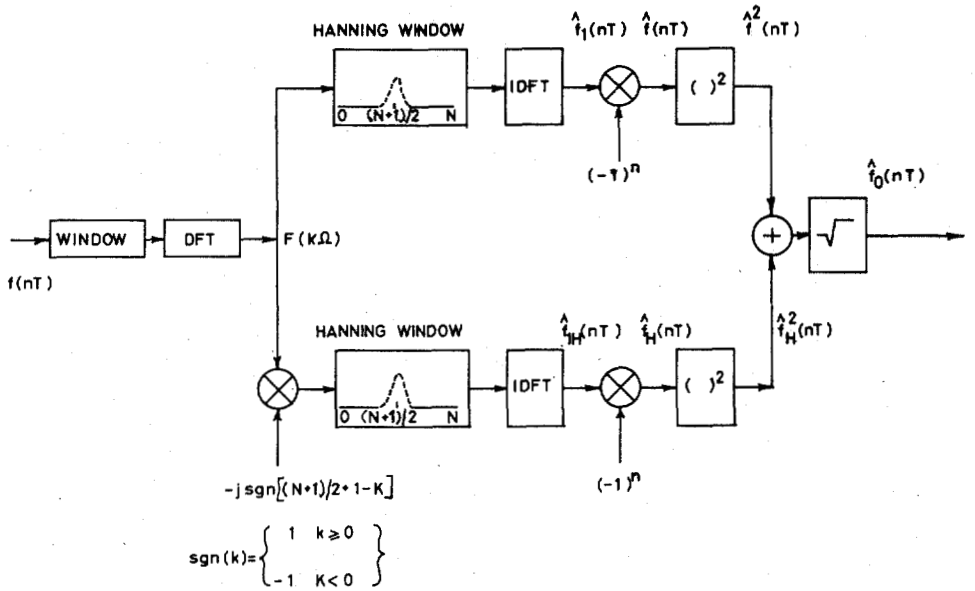

Fig. 1. Epoch filter.

output of the epoch filter. To suppress these undesired peaks a suitable window function must be used.

The computational procedure adopted for epoch extraction is illustrated schematically in Fig. 1. The epochs are indicated by peaks in the output $\hat{f}_o(nT)$ of the epoch filter.

IV. Application to Voiced Speech

In this section the epoch filter theory is first applied to the output of models of voiced speech production mechanism and then to the actual speech data.

A. Model 1: Impulse Excitation of a Two-Resonator System

A two-resonator model with resonant frequencies at 500 and 1000 Hz and bandwidths 50 and 100 Hz, respectively is considered for the system. The system is excited by four impulses spaced 8 ms apart. The response of the system is shown in Fig. 2(a). This is a superposition of successive impulse responses. Each impulse response has an epoch at the excitation point. The output $\hat{f}(nT)$ of an epoch filter is shown in Fig. 2(b). It is seen that the epochs obtained correspond to the excitation instants.

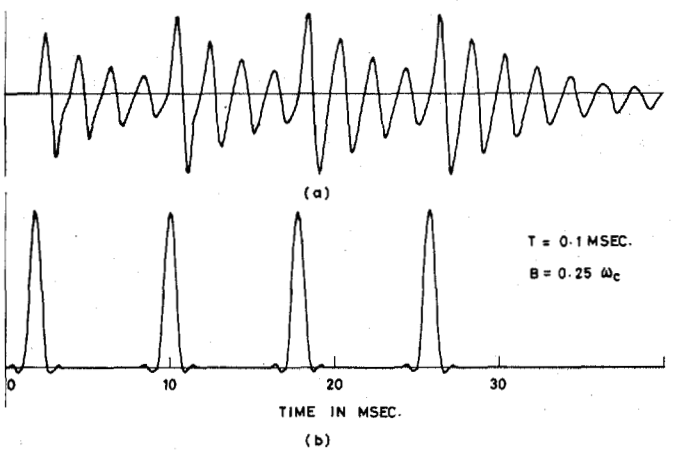

Fig. 2. Impulse excitation of a two-resonator model: (a) Response due to impulse sequence and (b) output $\hat{f}(nT)$ of the epoch filter.

B. Model 2: Glottal Wave Excitation of a Two-Resonator System

Glottal pulses are triangular-like waveforms with marked points of discontinuity in their derivatives and a definite duty ratio [1]. A set of five representative glottal pulses [21] are shown in Fig. 3(a). These waves have a positive slope during 40 percent and negative slope during 16 percent of one pitch period. The spacing of these waves was chosen to be 10 ms. These waves have specific points of discontinuities in their first derivatives. The waveform A has three such points, two at the ends and one at the apex, B and C have one such point each at the closure of the glottis while D has none and E has two such points both at the open and closure of the glottis. These points correspond to the epochs of order one as defined earlier. The output $\hat{f}(nT)$ shown in Fig. 3(b) is obtained when these waves are passed through the epoch filter. It can be seen from the figure that the epochs are obtained correctly. The relative amplitudes and polarity of the epochs are also maintained. It should particularly be noted that peaks corresponding to waveforms D do not appear significantly as there are no discontinuities in its first derivative.

The response of the two resonator vocal tract model excited by these five glottal waves is shown in Fig. 4(a). It is difficult to know the exact points of epochs and their relative amplitudes from the response shown in the figure. The Fourier transform of the response is the product of the Fourier transforms of the impulse response of the system and glottal waves. At high frequencies the Fourier transform of the impulse response of the system gives a rational decay. Therefore the epoch theory can be applied to the response. The output of the response when passed through the epoch filter is shown in Fig. 4(b). This figure can be compared with Fig. 3(b). The peaks correctly give the points of discontinuity of the first

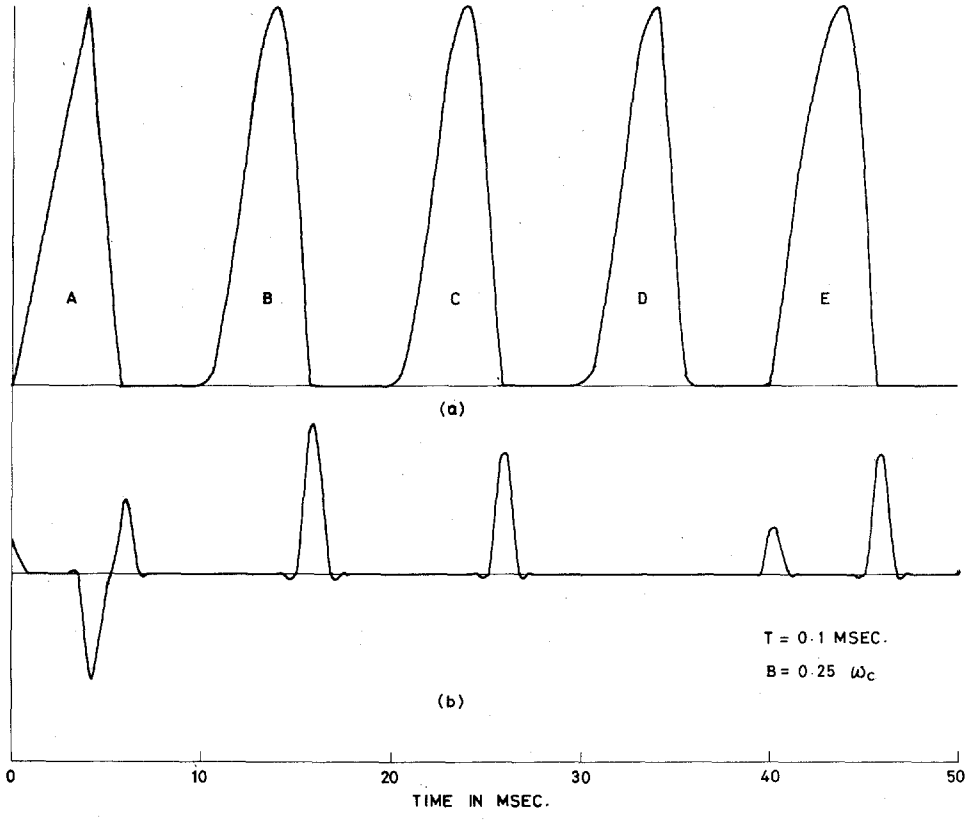

Fig. 3. Glottal waveforms: (a) Typical glottal waveforms and (b) output $\hat{f}(nT)$ of the epoch filter.

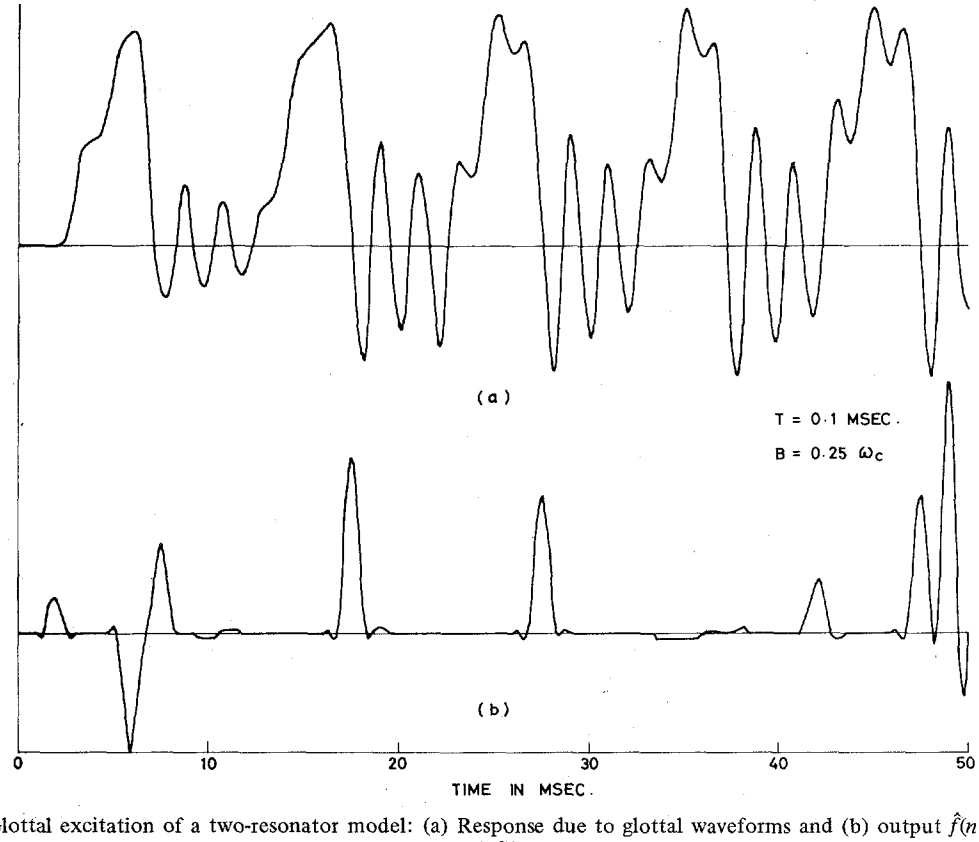

Fig. 4. Glottal excitation of a two-resonator model: (a) Response due to glottal waveforms and (b) output $\hat{f}(nT)$ of the epoch filter.

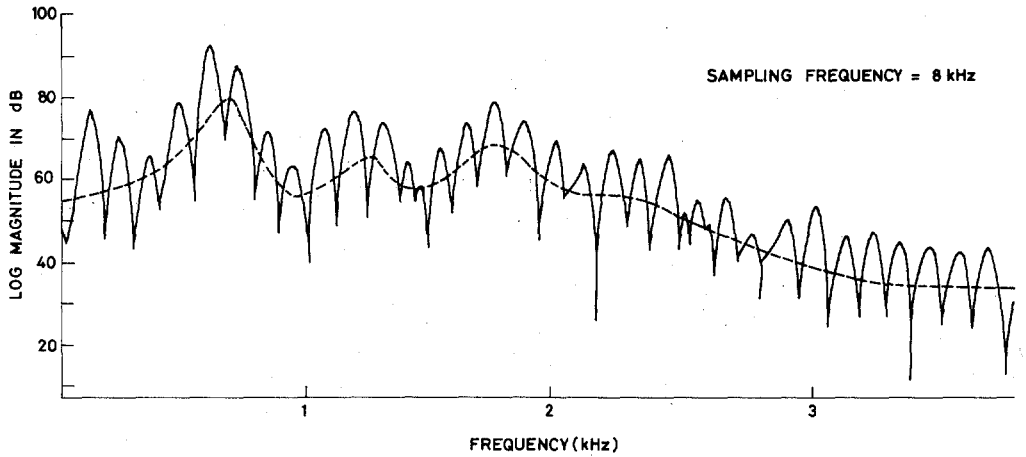

Fig. 5. Log spectrum of a voiced speech segment.

derivative of the glottal waveform. From this it can be concluded that the system gets excited at all points of discontinuity though the major excitation coincides with the greatest discontinuity in the derivative which usually occurs at the closure of the glottis.

The response in Fig. 4(a) for the glottal waveform D corresponds mostly to the first resonance, i.e., 500 Hz. This waveform does not contain significant high-frequency components to excite higher formants. Only a point of discontinuity can excite all the resonances of the vocal tract and hence can be regarded as the point of excitation. This may be the reason for perceptually lower ranking of waveform D [21]. It can be noted that the epoch corresponding to B is larger compared to C. This may possibly be the reason for higher ranking of B in perceptual tests [21].

C. Speech Data

The determination of variations in the pitch period for band-limited speech is generally difficult. The usual method of pitch extraction assumes the excitation to be periodic. This is not required if the epochs of the excitation of the vocal tract can be obtained. In this section, epoch extraction from speech data is discussed.

Speech data from a sentence "cats and dogs each hate the other" spoken by a male speaker representative of general American English (GAE) are considered for analysis. The log spectrum of a portion of voiced speech [Fig. 6(a)] is shown in Fig. 5. The gross spectrum (shown by dotted line) corresponds to the system transfer function whereas the superposed fluctuations are due to the periodicity of the excitation [22]. It can be seen that the tail end spectrum is nearly flat. A Hanning window over the nearly flat portion is used in the epoch filter. For the voiced speech segment shown in Fig. 6(a) the outputs $\hat{f}(nT)$ and $\hat{f}^2_o(nT)$ of the epoch filter are given in Fig. 6(c) and (d), respectively. The output of a digital inverse filter [23] is also shown in Fig. 6(b) for comparison. It is clear that the positions of the epochs and hence the pitch interval can be unambiguously obtained. The average pitch period for each

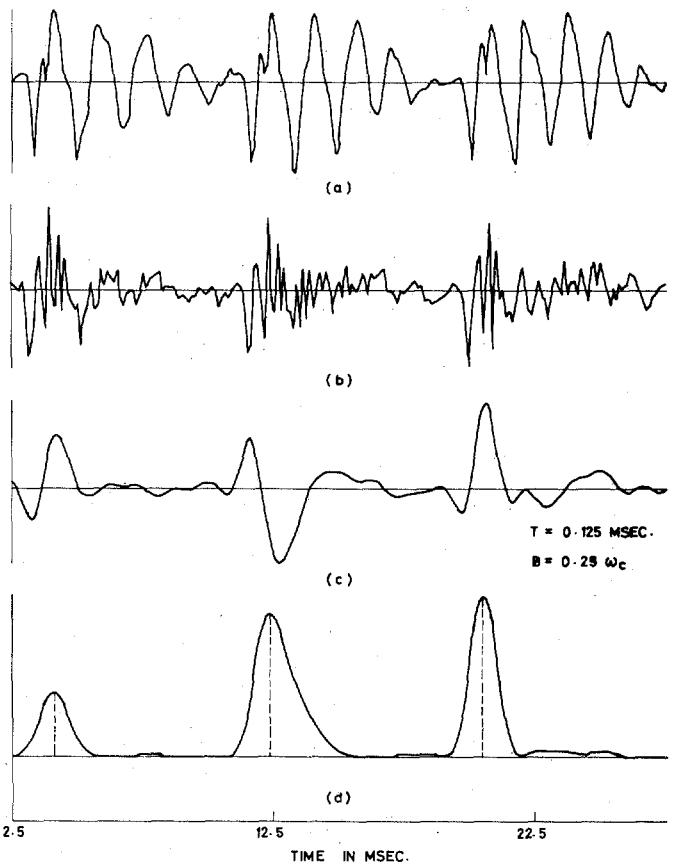

Fig. 6. Epoch extraction of voiced speech: (a) Voiced speech segment, (b) output of a digital inverse filter with 12 coefficients, (c) output $\hat{f}(nT)$ of the epoch filter, and (d) output $\hat{f}^2_o(nT)$ of the epoch filter.

frame is calculated from the output of the epoch filter. This is compared with the value obtained using cepstrum technique for several frames of voiced speech in Table II.

The outputs of the epoch filter for four successive frames is shown in Fig. 7. It is seen that the epoch positions as well as the shape of the output waveforms are maintained in spite of the shift. This also shows that the vocal tract need not be strictly stationary.

TABLE II
PITCH PERIOD ESTIMATE USING EPOCH FLITER AND CEPSTRUM TECHNIQUE

Frame No.	Average Pitch Period Obtained using Epoch Filter (Milliseconds)	Average Pitch Period Obtained using Cepstrum (Milliseconds)
1	8.500	8.457
2	8.500	8.750
3	8.500	8.825
4	8.375	8.312
5	8.125	8.194
6	7.875	7.825
7	7.375	7.250
8	7.250	7.000
9	7.000	6.940
10	6.875	7.000
11	7.875	8.156
12	8.125	8.460
13	8.250	8.540
14	8.375	8.250
15	8.500	8.500

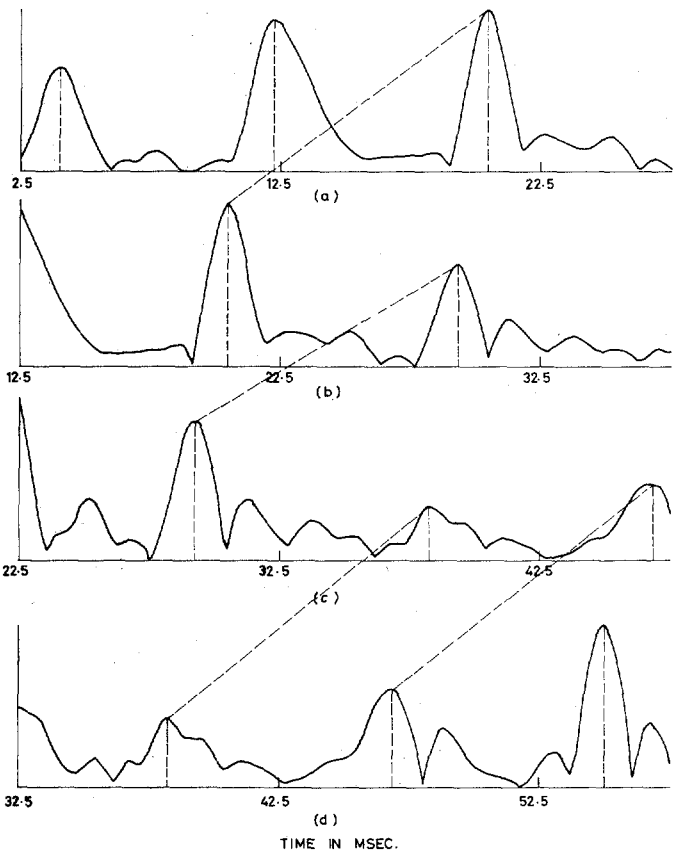

Fig. 7. Output $\hat{f}_O(nT)$ of the epoch filter for four successive overlapping frames.

The application of the epoch filter technique for sounds other than steady-state vowels is illustrated in Figs. 8–10. A portion of the speech waveform from the consonant-vowel transition in the word "cats" is shown in Fig. 8(a). The output of the epoch filter for this segment is shown in Fig. 8(b). The onset of voicing is clearly evident from the figure. A portion of speech waveform from the vowel-nasal transition in the word "and" is shown in Fig. 9(a). The output [Fig. 9(b)]

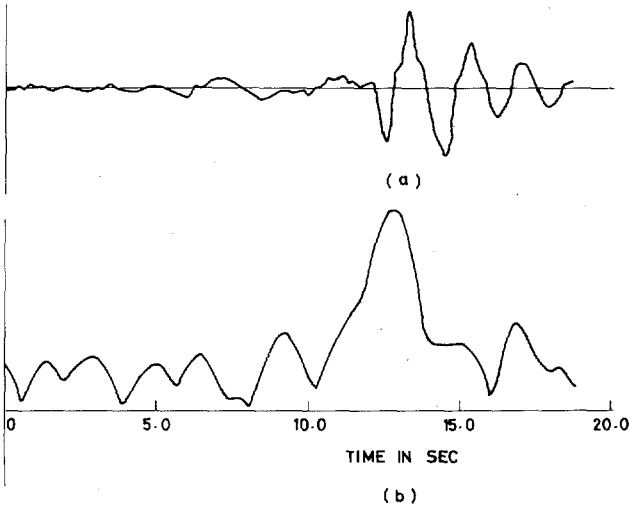

Fig. 8. (a) Speech segment of the consonant-vowel transition in the word "cats" and (b) output $\hat{f}_O(nT)$ of the epoch filter.

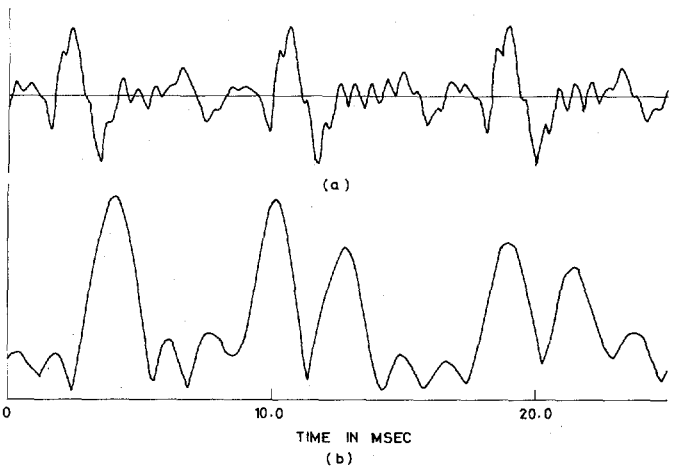

Fig. 9. (a) Speech segment of the vowel-nasal transition in the word "and" and (b) output $\hat{f}_O(nT)$ of the epoch filter.

of the epoch filter for this segment indicates double excitation of the vocal tract within a single pitch period. The output of the epoch filter for the fricative / s / in the word "cats" is shown in Fig. 10. As this sound is unvoiced, the peaks in the output are randomly spaced compared to the output for a voiced sound.

D. Experimental Work

Since an epoch filter performs mainly a bandpass operation, speech signal was passed through one-third octave bandpass filter to obtain the epochs. A portion of the speech waveform and the output of the bandpass filter centered around 10 and 20 kHz are shown in Figs. 11 and 12. It is seen that the epochs can easily be located from the sharp pulses at the output. It is therefore clear that the rational decay of the lower order resonances of the vocal tract dominates over the higher order resonances even at 20 kHz. The experiment was conducted in an ordinary room with normal background noise and reverberation. For glottal waves having more than one discontinu-

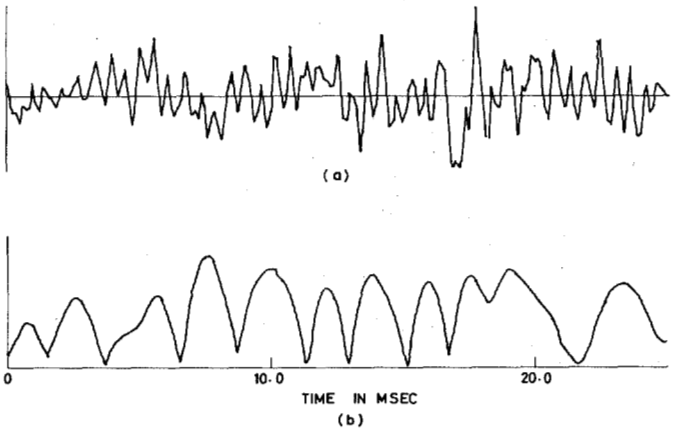

Fig. 10. (a) Speech segment of unvoiced fricative / s / and (b) output $\hat{f}_O(nT)$ of the epoch filter.

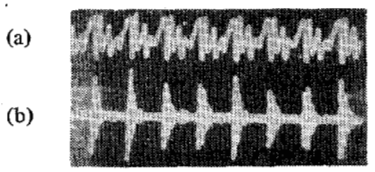

Fig. 11. (a) Voiced speech segment and (b) output of one-third octave bandpass filter centered at 10 kHz.

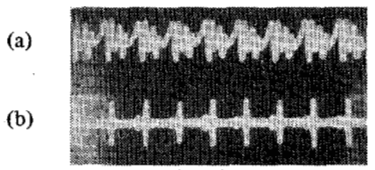

Fig. 12. (a) Voiced speech segment and (b) output of one-third octave bandpass filter centered at 20 kHz.

ity, multiple peaks were seen within a single pitch period at the output of the filter.

V. Conclusions

Besides obtaining the epochs and the pitch period, the present theory reveals several other points. The relative amplitudes of the excitation are preserved in the output. In some cases there are distinctly two peaks in a single pitch period. Possibly these peaks could be the other points of excitation which can be made use of in a synthesizer. An approximate shape of the glottal wave can be estimated from the amplitudes and positions of these peaks.

An interesting observation is that pitch information, a low-frequency datum, is obtained by processing high-frequency portion of speech. In fact speech was prefiltered to 3.4 kHz before sampling at 8 kHz. This is close to telephone quality speech. Pitch extraction for telephone quality speech has been regarded as one of the challenging problems in speech area [11]. Another feature is that since this technique does not rely on the spectral differences between the vocal tract transfer function and the excitation spectrum, even high-pitched speech can be analyzed to obtain the pitch. Because of the linearity of the process, even the transition regions will not pose any problem. This technique could possibly be used for voiced-unvoiced decision.

It is well known that the glottal wave is highly characteristic of the speaker whereas the vocal tract parameters are mainly characteristic of the speech [24]. Although the present technique is not aimed at obtaining the glottal waveshape, the important characteristics of the glottal wave and the variations in pitch period are still obtained. These characteristics are extracted irrespective of the formant structure of the vocal tract. This may prove useful for speaker verification and in pitch perception studies of connected speech.

Acknowledgment

The authors wish to thank Prof. B. S. Ramakrishna for his interest and encouragement and Dr. V. V. S. Sarma for his valuable suggestions. The authors also wish to thank Prof. M. D. Srinath of the Information and Control Sciences Center, Institute of Technology, Southern Methodist University, Dallas, Tex. for providing us with the speech data. The first author wishes to thank the Council for Scientific and Industrial Research (CSIR) for the award of Senior Research Fellowship.

References

[1] J. L. Flanagan, *Speech Analysis Synthesis and Perception.* New York: Springer-Verlag, 1965; second ed., 1972.
[2] D. G. Childers and M. T. Pao, "Complex demodulation of transient wavelet detection and extraction," *IEEE Trans. Audio Electroacoust.* (*Special Issue on Digital Filtering*), vol. AU-20, pp. 295–308, Oct. 1972.
[3] B. P. Bogert, M. J. Healy, and J. W. Tukey, "The quefrency analysis of time series for echoes: Cepstrum, pseudoautocovariance, cross cepstrum, and saphe cracking," in *Proc. Symp. Time Series Analysis*, M. Rosenblatt, Ed. New York: Wiley, 1963, pp. 209–243.
[4] D. K. Barton, "Radar measurement accuracy in log-normal clutter," in *Conf. Rec. 1971 EASCON*, New York, N.Y., pp. 246–251.
[5] B. Gold and C. M. Rader, *Digital Processing of Signals.* New York: McGraw-Hill, 1969, ch. 8.
[6] M. Rothenberg, "A new inverse filtering technique for deriving glottal air flow waveform during voicing," *J. Acoust. Soc. Amer.*, vol. 53, pp. 1632–1645, 1973.
[7] R. L. Miller, "Nature of the vocal cord wave," *J. Acoust. Soc. Amer.*, vol. 31, pp. 667–677, 1959.
[8] M. V. Mathews, J. E. Miller, and E. E. David, "Pitch synchronous analysis of voiced sounds," *J. Acoust. Soc. Amer.*, vol. 33, pp. 179–186, 1961.
[9] M. M. Sondhi, "New methods of pitch extraction," *IEEE Trans. Audio Electroacoust.* (*Special Issue on Speech Communication and Processing–Part II*), vol. AU-16, pp. 262–266, June 1968.
[10] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293–309, Feb. 1967.
[11] —, "Cepstrum and some close relatives," in *Signal Processing*, J. W. R. Griffiths, P. L. Stocklin, and C. van Schooneveld, Eds., NATO Advanced Study Inst. New York: Academic, 1973.
[12] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367–377, Dec. 1972.
[13] B. S. Atal and B. L. Hanaver, "Speech analysis and synthesis by linear prediction," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, 1971.
[14] J. N. Holmes, "An investigation of the volume velocity wave-

form at larynx during speech by means of an inverse filter," in *Proc. Stockholm Speech Commun. Seminar*, Royal Inst. Technol., Stockholm, Sweden, Sept. 1962.
[15] T. V. Ananthapadmanabha and B. Yegnanarayana, "A decomposition technique for composite signals," to be published.
[16] E. A. Robinson, *Statistical Communication and Detection.* London, England: Griffin, 1967, ch. 9, p. 253.
[17] T. Y. Young, "Epoch detection–A new method for resolving overlapping signals," *Bell Syst. Tech. J.*, vol. 44, pp. 401–425, Mar. 1965.
[18] A. Papoulis, *Systems and Transforms with Application in Optics.* New York: McGraw-Hill, 1968, ch. 7.
[19] B. F. Cron, "Phase distortion of a pulse caused by bottom reflection," *J. Acoust. Soc. Amer.*, vol. 37, pp. 486–492, 1965.
[20] R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra.* New York: Dover, 1968.
[21] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Amer.*, vol. 49, pp. 583–590, 1971.
[22] R. W. Schafer, "A survey of digital speech processing techniques," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 28–35, Mar. 1972.
[23] J. D. Markel, "Digital inverse filtering–A new tool for formant trajectory estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 129–137, June 1972.
[24] J. I. Makhoul and J. J. Wolf, "Linear prediction and the spectral analysis of speech," Bolt Beranek and Newman, Inc., Cambridge, Mass., BBN Rep. 2304, Aug. 1972.

A Semiautomatic Pitch Detector (SAPD)

CAROL A. McGONEGAL, LAWRENCE R. RABINER, SENIOR MEMBER, IEEE, AND AARON E. ROSENBERG, MEMBER, IEEE

Abstract–**The purpose of this paper is to describe a technique for semiautomatically determining the pitch contour of an utterance. The method is significantly more sophisticated than the standard technique of hand tracking of pitch periods from a waveform display of the utterance and leads to a fairly robust measurement of the pitch period. This technique utilizes a simultaneous display (on a 10 ms section-by-section basis) of the low-pass filtered waveform, the autocorrelation of a 400-point segment of the low-pass filtered waveform, and the cepstrum of the same 400-point segment of the wideband recording. For each of the separate displays (i.e., waveform, autocorrelation, and cepstrum) an independent estimate of the pitch period is made on an interactive basis with the computer, and the final pitch period decision is made by the user based on results of each of the measurements. The technique has been tested on a large number of utterances spoken by a variety of speakers with very good results. Formal tests of the method were made in which four people were asked to use the method on three different utterances, and their results were then compared. During voiced regions, the standard deviation in the value of the pitch period was about 0.5 samples across the four people. The standard deviation of the location of the time at which voiced regions became unvoiced, and vice versa was on the order of half a section duration, or 5 ms. The major limitation of the proposed method is that it requires about 30 min to analyze 1 s of speech. However, the increased accuracy and robustness of the results indicate that the tradeoff of time for accuracy is a good one for many applications.**

I. Introduction

FOR SOME applications, an extremely accurate and reliable measurement of the pitch contour of an utterance is required. One such application is a comparison and evaluation study of a variety of pitch detection algorithms which has recently been performed at Bell Laboratories [1]. Another application was a study of the inter and intra speaker similarities in pitch contours for several utterances [2]. Other applications include studies for determination of linguistic rules for pitch generation for use in speech synthesis applications [3], [4]. Although a large number of pitch detection algorithms have been proposed in the literature, none of them is able to achieve the performance of a human who is knowledgeable in the area of speech communications with a fairly sophisticated interactive display of the speech waveform.

The usual method of manual pitch detection is for the user to mark pitch periods on a period-by-period basis, directly on a display of the speech waveform. Although such a technique is often quite good, there are segments of some speech sounds during which the waveform periodicity is not clearly visible in the waveform due to rapid spectral changes in the sound [5]. During such intervals a rough indication of the pitch period can be obtained from the waveform, but due to the changing spectrum, the pitch period estimate can be off by several samples. It is the purpose of the paper to describe a semiautomatic pitch detection (SAPD) technique which is significantly more sophisticated than the standard manual pitch tracking method described above, and which has been found to yield reliable, and repeatable estimates of the pitch period across a variety of speakers and utterances.

Manuscript received May 19, 1975.
The authors are with Bell Laboratories, Murray Hill, N.J. 07974.

II. The Analysis System

Fig. 1 shows a block diagram of the SAPD processing. The speech signal $s(n)$ sampled at a 10 kHz rate is processed to give three simultaneous displays for each section of speech. For two of the displays the speech is low-pass filtered by an $N = 99$ linear phase, finite impulse response (FIR), low-pass digital filter with a passband cutoff frequency of 900 Hz, and a stopband cutoff frequency of 1100 Hz. Fig. 2 shows a plot of the log magnitude frequency response of the filter. The passband ripple of the filter is about 0.03 and the stopband ripple is down about 50 dB. The low-pass filtered speech waveform