Epoch Extraction from Linear Prediction Residual for Identification of Closed Glottis Interval

TIRUPATTUR V. ANANTHAPADMANABHA AND B. YEGNANARAYANA, MEMBER, IEEE

Abstract—**In voiced speech analysis epochal information is useful in accurate estimation of pitch periods and the frequency response of the vocal tract system. Ideally, linear prediction (LP) residual should give impulses at epochs. However, there are often ambiguities in the direct use of LP residual since samples of either polarity occur around epochs. Further, since the digital inverse filter does not compensate the phase response of the vocal tract system exactly, there is an uncertainty in the estimated epoch position. In this paper we present an interpretation of LP residual by considering the effect of the following factors: 1) the shape of glottal pulses, 2) inaccurate estimation of formants and bandwidths, 3) phase angles of formants at the instants of excitation, and 4) zeros in the vocal tract system. A method for the unambiguous identification of epochs from LP residual is then presented. The accuracy of the method is tested by comparing the results with the epochs obtained from the estimated glottal pulse shapes for several vowel segments. The method is used to identify the closed glottis interval for the estimation of the true frequency response of the vocal tract system.**

I. Introduction

VOICED speech analysis consists of determining the frequency response of the vocal tract system and the glottal pulses representing the voice source. Although the source of excitation for voiced speech is a sequence of glottal pulses, the significant excitation of the vocal tract system can be considered, to a first approximation, to be at discrete instants of time, called epochs. There can be more than one epoch within a pitch period but the major excitation usually coincides with the glottal closure. Determination of epochs solves, to a large extent, the basic problem of defining the pitch periods. This is because pitch is a subjective attribute of voicing and a precise definition for pitch is generally difficult due to variations in the periodicity, shape, and amplitude of the glottal pulses [1]. Many times the glottal flow is zero beyond the instant of glottal closure, and speech signal in the closed glottis interval represents the force-free response of an all-pole system. Analysis of speech signal over such an interval provides an accurate estimate of the frequency response of the vocal tract system [2]-[4]. Epochs can be effectively used in identifying the closed glottis interval.

Despite its importance in speech analysis, there is as yet no satisfactory method for extracting epochal information from the speech signal. The ideal solution to this problem is to extract glottal pulses. But extraction of glottal pulses from the speech wave is computationally tedious and the results require careful interpretation [5]. Because of the difficulty in identifying epochs from continuous speech, empirical rules are often proposed for estimation of pitch periods [6]. A large value in the error signal (LP residual) obtained by linear prediction (LP) analysis is supposed to indicate the epoch location [8], [9]. Speech samples following the estimated epoch or those belonging to an interval with low values of LP residual are assumed to belong to the closed glottis interval. However, there are some difficulties in the direct use of LP residual for epoch identification which will be discussed in later sections. Sobakin proposed a measure for the linear predictability of a signal over a given interval [3]. This measure, with some modification, was used by Strube to identify the instants of glottal closure [10]. For each frame of data, a $(p+1)\times(p+1)$ covariance matrix is formed where p is the order of the predictor. The Gram determinant of the covariance matrix will be small if the samples within the frame are linearly related. Sobakin suggested that the center of the frame for which the Gram determinant is a minimum should be considered as the instant of glottal closure. Strube showed that the beginning of the frame for which the logarithm of the Gram determinant reached a maximum was the instant of glottal closure. Compared to the direct use of LP residual, the log determinant method was found to be less ambiguous. But Strube's method is computationally so complex that it is suitable only for special investigations. The authors have proposed an epoch filtering technique for epoch extraction from voiced speech [11]. The technique involves the choice of a suitable frequency domain window and it was found to be effective for analyzing clean data.

In this paper a more general method [12] for extracting epochal information is described. The technique involves the application of the epoch filter theory to LP residual. We shall show that the technique can be effectively used in identifying the presence of multiple excitations within a pitch period and hence, the closed glottis interval. In Section II a brief review of LP analysis is presented to show the limitations of LP residual for extracting epochal information. A detailed analysis of LP residual is presented in Section III to show the effect of source and system parameters and analysis variables on LP residual. A method for unambiguous identification of epochs from LP residual is then presented in Section IV. In Section V the relation between epochs and glottal pulses is studied. Determination of the closed glottis interval and its significance are also discussed in Section V.

Manuscript received January 24, 1978; revised November 28, 1978.

The authors were with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India. They are now with the Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA 15213.

0096-3518/79/0800-0309$00.75

II. Linear Prediction Analysis

In linear prediction analysis [7]-[9], [13] voiced speech is assumed to be the output of an all-pole digital filter excited by a sequence of impulses. According to this model the nth sample of speech signal can be approximated by a linear weighted sum of p previous samples. The difference between the actual value and the approximated value is called the prediction error signal or the LP residual. Energy in the prediction error signal is minimized to determine the weights called the LP coefficients (LPC's). Thus, for speech signal $\{s(nT)\}$, the predicted value of the nth sample is given by

$$\hat{s}(nT) = -\sum_{k=1}^{p} a_k s(nT - kT) \tag{1}$$

where a_k, $k = 1, 2, \cdots, p$ are the LPC's. The total squared error is given by

$$E = \sum_{n} [s(nT) - \hat{s}(nT)]^2. \tag{2}$$

Minimization of E with respect to a_ks leads to the following linear simultaneous equations, called normal equations.

$$\sum_{k=1}^{p} a_k R_{i-k} = -R_i; \quad i = 1, 2, \cdots, p \tag{3}$$

where R_k is the kth autocorrelation coefficient given by

$$R_k = \sum_{n=0}^{(N+1-|k|)} s'(nT)\, s'(nT + kT) \tag{4}$$

where $s'(nT)$ is the nth sample of the windowed speech signal and N is the number of samples in the frame. In this study a Hanning window centered at half the frame length is used. The prediction error is given by the equation

$$t(nT) = s(nT) - \hat{s}(nT) \tag{5}$$

$$= \sum_{k=0}^{p} a_k s(nT - kT) \tag{6}$$

where a_0 is unity. Thus, $t(nT)$ is a convolution of the sequence $\{a_k\}$ with the speech signal $\{s(nT)\}$. In other words, $t(nT)$ can be considered to be the output of a digital filter whose (finite) impulse response is the sequence $\{a_k\}$. If the speech signal were to be truly the response of an all-pole model, then the predictability will be exact at all instants except at the excitation instants. Then, the LP residual will be a sequence of impulses. This reasoning is put forward to explain that a large value of LP residual occurs at the excitation instant. The filter $A(z)$, given by the z-transform

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}, \tag{7}$$

is called a digital inverse filter. Since this filter flattens the short-time spectrum of the speech signal being analyzed, the reciprocal of the frequency response of $A(z)$ gives an estimate of the envelope of the short-time spectrum.

Although the LP residual contains information pertaining to the excitation, epoch identification directly from LP residual is not recommended due to the following problems. LP analysis assumes an all-pole model for representing the combined effect of the impulse response of the vocal tract system and the glottal pulse shape. The all-pole model implicitly assumes a minimum phase characteristic for the speech signal. If this is not valid, the phase response of the vocal tract system is not compensated exactly by the digital inverse filter. Phase compensation will also be affected when formants and their bandwidths are inaccurately estimated. Effect of uncompensated phase on LP residual is not known. Moreover, the inverse filter does not compensate for zeros which may be introduced due to the finite duration of a glottal pulse or the nasal coupling. These factors cause multiple peaks of either polarity to occur around the epochs in the LP residual and make unambiguous estimation of epochs from the LP residual difficult [10]. LP residuals of two vowel segments are shown in Fig. 1 to illustrate the above remarks.

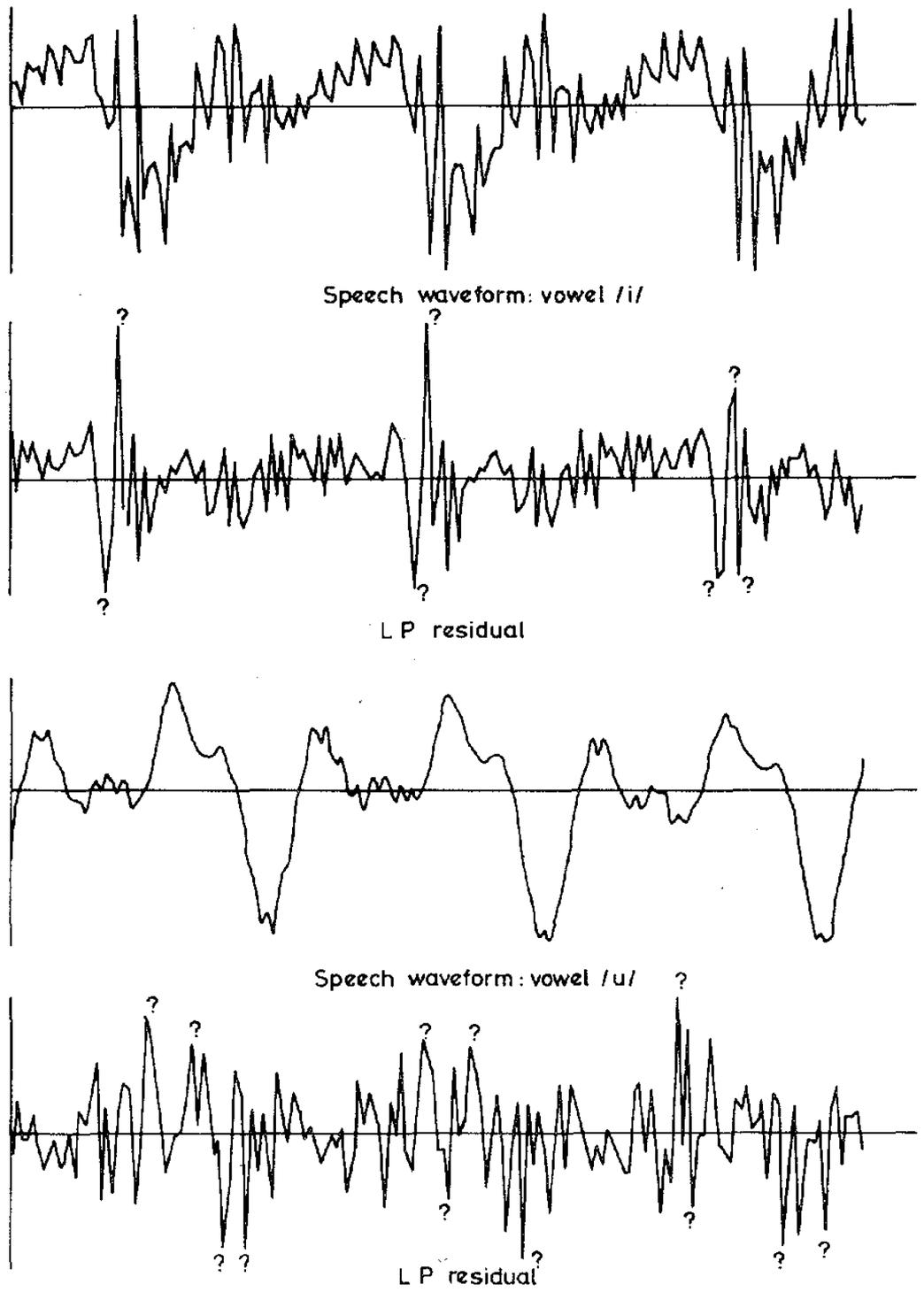

Fig. 1. Ambiguities in the use of the LP residual for epoch identification.

III. Interpretation of LP Residual

Ideally, the output of the digital inverse filter for voiced speech should consist of impulses separated by pitch periods. However, such an output is seldom observed except for vowel sounds synthesized using impulses as the excitation function. The purpose of this section is to explain the observed deviations in the LP residual from the expected impulse sequence. We shall also discuss the accuracy of estimating the epoch from the LP residual.

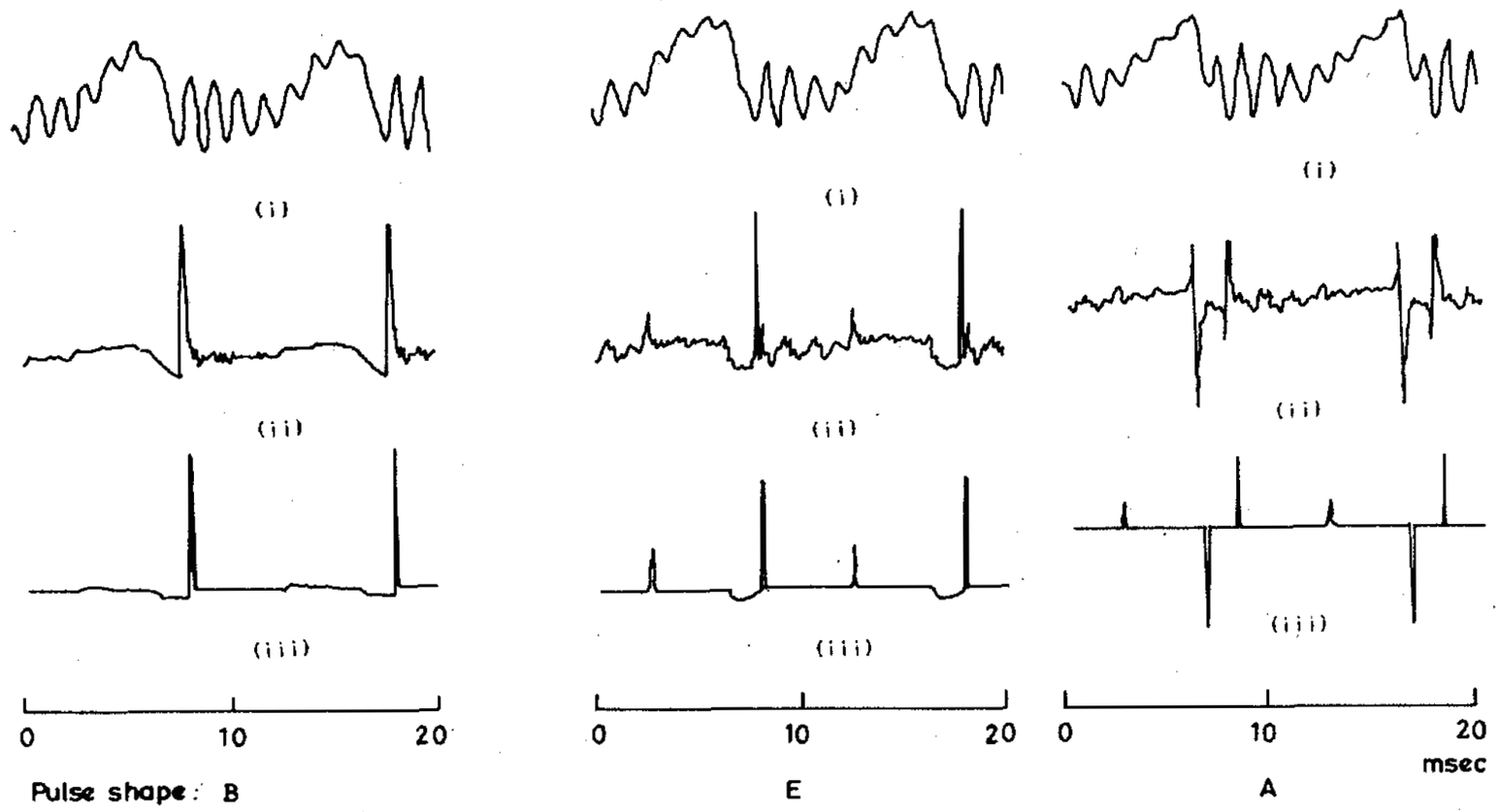

Fig. 2. Relation between glottal pulses and LP residual: (i) vowel waveform, (ii) LP residual, and (iii) second derivative of glottal pulses.

A. Glottal Pulse and LP Residual

Let $V(z)$ be the z-transform of the vocal-tract impulse response inclusive of radiation characteristics. Let $g(nT)$ be the quasiperiodic sequence of glottal pulses exciting the system $V(z)$ to produce the voiced speech signal $s(nT)$. According to the source-system model for voiced speech [14], the z-transform of $s(nT)$ can be written as

$$S(z) = G(z)\, V(z). \tag{8}$$

It will be useful for further discussion if we use an alternative, but mathematically equivalent, model for voiced speech. In this model $g(nT)$ is considered to be the output of a double integrator excited by the second derivative $g^{(2)}(nT)$ of $g(nT)$. The cascade of the double integrator and the system $V(z)$ is represented as the system $\hat{V}(z)$. Then,

$$S(z) = G^{(2)}(z)\, \hat{V}(z) \tag{9}$$

where $G^{(2)}(z)$ is the z-transform of $g^{(2)}(nT)$. In general, the digital inverse filter in LP analysis is an optimum spectral whitening filter [15]. Since $g^{(2)}(nT)$ is known to possess usually flat spectral characteristics [5], it is reasonable to assume that the digital inverse filter $A(z)$ flattens the spectrum of $\hat{V}(z)$ only; i.e.,

$$\hat{V}(z)\, A(z) \simeq 1. \tag{10}$$

The LP residual $t(nT)$, which is the output of $A(z)$, is therefore given by

$$t(nT) \simeq g^{(2)}(nT). \tag{11}$$

This simple analysis brings out the relationship between the LP residual and the glottal pulses. LP analysis of synthetic vowels was performed to illustrate this relationship. The synthetic vowel waveform, LP residual, and the second derivative of the synthetic glottal pulses are shown in Fig. 2 for the three typical pulse shapes described in [16].

B. Effect of Inaccurate Estimation of Vocal Tract Response

Inaccurate estimation of formants and their bandwidths introduces undesirable signal components in the LP residual. We shall explain such an effect by considering the case of a single resonator.

Consider the impulse response of a single resonator given by

$$f(nT) = \begin{cases} 0; & n < 0 \\ 1; & n = 0 \\ -\hat{a}_1 f(nT - T) - \hat{a}_2 f(nT - 2T) & n > 0. \end{cases} \tag{12}$$

The reciprocal of the z-transform of $f(nT)$ is given by

$$F^{-1}(z) = 1 + \hat{a}_1 z^{-1} + \hat{a}_2 z^{-2}. \tag{13}$$

Let the estimated LPC's for the signal $f(nT)$ be

$$a_1 = \hat{a}_1 + \tilde{a}_1, a_2 = \hat{a}_2 + \tilde{a}_2 \tag{14}$$

where $\tilde{a}_1$ and $\tilde{a}_2$ are the errors in $\hat{a}_1$ and $\hat{a}_2$, respectively. Then, $A(z)$ is given by

$$A(z) = 1 + \hat{a}_1 z^{-1} + \hat{a}_2 z^{-2} + \tilde{a}_1 z^{-1} + \tilde{a}_2 z^{-2}. \tag{15}$$

The z-transform of LP residual $t(nT)$ for the signal $f(nT)$ can be written as

$$T(z) = F(z)\, A(z) \tag{16}$$

$$T(z) = 1 + \tilde{a}_1 z^{-1} F(z) + \tilde{a}_2 z^{-2} F(z). \tag{17}$$

Hence, we get

$$t(nT) = \delta(nT) + \tilde{a}_1 f(nT - T) + \tilde{a}_2 f(nT - 2T). \tag{18}$$

Thus, we notice that in addition to the impulse, scaled and delayed versions of the original signal appear in the LP residual. Errors in the coefficients can be related to errors in formant frequency and bandwidth in the frequency response of $f(nT)$. For example, an error of 5.7 percent in $\hat{a}_2$ corresponds to an error of 100 Hz in the bandwidth of a resonator with formant at 1250 Hz and bandwidth 100 Hz (10 kHz sampling). It may be noted at this stage that large errors in the LPC's occur in the analysis of high-pitched sounds [17]. Hence, LP residual for high-pitched sounds does not contain clearly distinguishable peaks.

C. Effect of the Phase Angles of Formants

Consider the following representation for a speech signal which includes the amplitudes and phase angles of the formants

$$h(nT) = \sum_i C_i \exp(-a_i nT) \sin(b_i nT + \phi_i) \tag{19}$$

where C_i, ϕ_i, a_i/π, $b_i/2\pi$ represent, respectively, the amplitude, phase angle, bandwidth, and formant frequency of the ith formant. In general, the z-transform of $h(nT)$ possesses both poles and zeros. The effect of zeros on the LP residual will be discussed in Section III-D. The amplitude of the LP residual around the excitation instant, i.e., $nT = 0$, depends on the phase angles in a complex manner. But the effect of the phase angles can be explicitly seen for the case of a single resonator, with $C_1 = 1$, for which the LP residual $t(nT)$ at $nT = 0$ is $h(0) = \sin(\phi_1)$. Thus, for $\phi_1 = 0$, the LP residual at the instant of excitation will be zero instead of an impulse of unit magnitude as in the ideal case. The LP residual $t_H(nT)$ for the quadrature component of $h(nT)$ will be $\cos(\phi_1)$. Thus, for $\phi_1 = 0$, the LP residual $t_H(nT)$ at $nT = 0$ for the quadrature component will be an impulse of unit magnitude. However, by computing the Hilbert envelope $t_0(nT)$, given by

$$t_0(nT) = [t^2(nT) + t_H^2(nT)]^{1/2}, \tag{20}$$

the ambiguity caused by the phase angle in locating the epoch may be overcome. The signal $t_H(nT)$ can be computed as the Hilbert transform of $t(nT)$.

The LP residual $t(nT)$ and its Hilbert transform $t_H(nT)$ are shown in Fig. 3 for two vowel sounds. Although several other factors could have influenced the LP residual, it can be said that the effect of phase angles $\phi_i s$ is to introduce bipolar swings near the excitation instant. In the LP residual $t(nT)$ for vowel /i/ there appears to be a zero crossing at the excitation instant. This contention is supported by the fact that there is an unambiguous peak in $t_H(nT)$ at the same location. For arbitrary phase angles $\phi_i s$, bipolar swings occur both in $t(nT)$ and $t_H(nT)$, as can be seen for the case of vowel /e/.

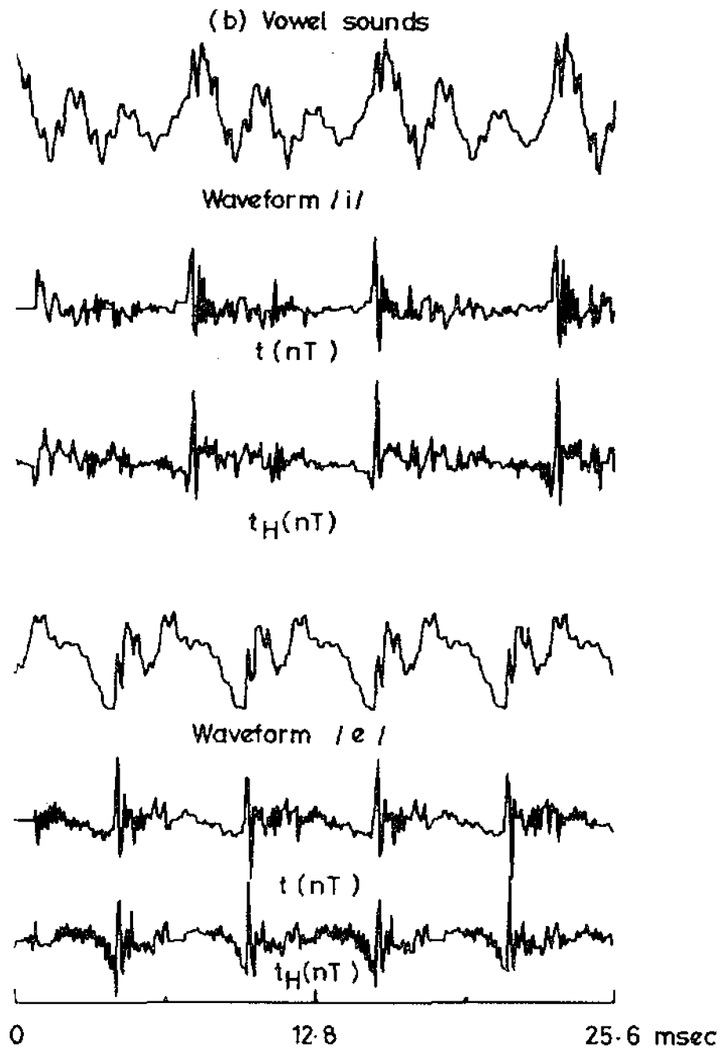

Fig. 3. Effect of the phase angles of formants on the LP residual.

D. Effect of Zeros

The presence of glottal and nasal zeros may affect the accuracy of the estimation of formants and bandwidths and thus cause undesirable signal components to be introduced into the LP residual. The inverse filter $A(z)$ will compensate, as best it can, for the total combination of poles and zeros present in the vocal tract system. When the vocal tract system violates the assumption of an all-pole model, the inverse filter will be a net poorer match to the frequency response of the system. Assuming the effect of other factors to be small, in the presence of zeros, instead of a single impulse, a sequence of impulses appears in the LP residual. The length of the sequence depends on the number of zeros present in the transfer function. An antiresonance caused by a complex conjugate zero pair can be represented in the time domain by a sequence of the form $(1, b_1, b_2)$. The zeros of the z-transform of the sequence $(1, b_1, b_2)$ lie outside the unit circle in the z plane when the magnitude of b_2 is greater than unity. Hence, in the presence of such a zero-pair in the impulse response of the vocal tract system, the largest value of the LP residual occurs delayed by two samples. In general, if there are L zeros outside the unit circle in the z-transform $T(z)$ of $t(nT)$, the largest value in the LP residual occurs delayed by L samples. Thus, the estimated epoch location would be correct only if $T(z)$ has no zeros outside the unit circle.

We shall consider the effect of glottal zeros at this stage. A sequence of the form $(1, b)$ introduces a single zero into the spectrum. On the other hand, in the case of two impulses separated in the time domain by k samples, periodic zeros separated by $2\pi/kT$ radians appear in the spectrum. It is known that a typical glottal pulse possesses slope discontinuities at certain specific instants [5], [1], [18]. Hence, the second derivative $g^{(2)}(nT)$ of a glottal pulse contains impulses of specific strengths at those instants [1, ch. 6, p. 241], [18]. Let us denote such an impulse sequence by $e(nT)$. The signal $g^{(2)}(nT)$ contains, besides $e(nT)$, a residual signal which we shall denote as $r(nT)$. Since the impulses in $e(nT)$ are well separated, a typical glottal pulse produces distributed zeros in the spectrum. The effect of glottal zeros can be studied in terms of $e(nT)$. The magnitude spectrum of $r(nT)$ should asymptotically fall off at least as fast as 6 dB per octave [1, ch. 6, p. 241]. The energy in $r(nT)$ is usually on the order of 8 percent of the energy in $e(nT)$ for typical pulse shapes [12]. Hence, $r(nT)$ has most of its energy concentrated in the low-frequency region of the spectrum. The components $e(nT)$ and $r(nT)$ for the typical pulse shapes may be identified in Fig. 2. The impulses in $e(nT)$ excite the resonances of the vocal tract system. Hence, the instants of occurrence of impulses in $e(nT)$ may be considered to be the excitation instants or epochs. The problem is to determine these instants and amplitudes of the excitation impulses from the speech signal.

Ideally, it should be possible to identify the epochs from the LP residual as per (11). However, due to various factors discussed above, direct identification of epochs from LP residual is difficult. In the next section we describe a method based on envelope computation as mentioned in Section III-C, which

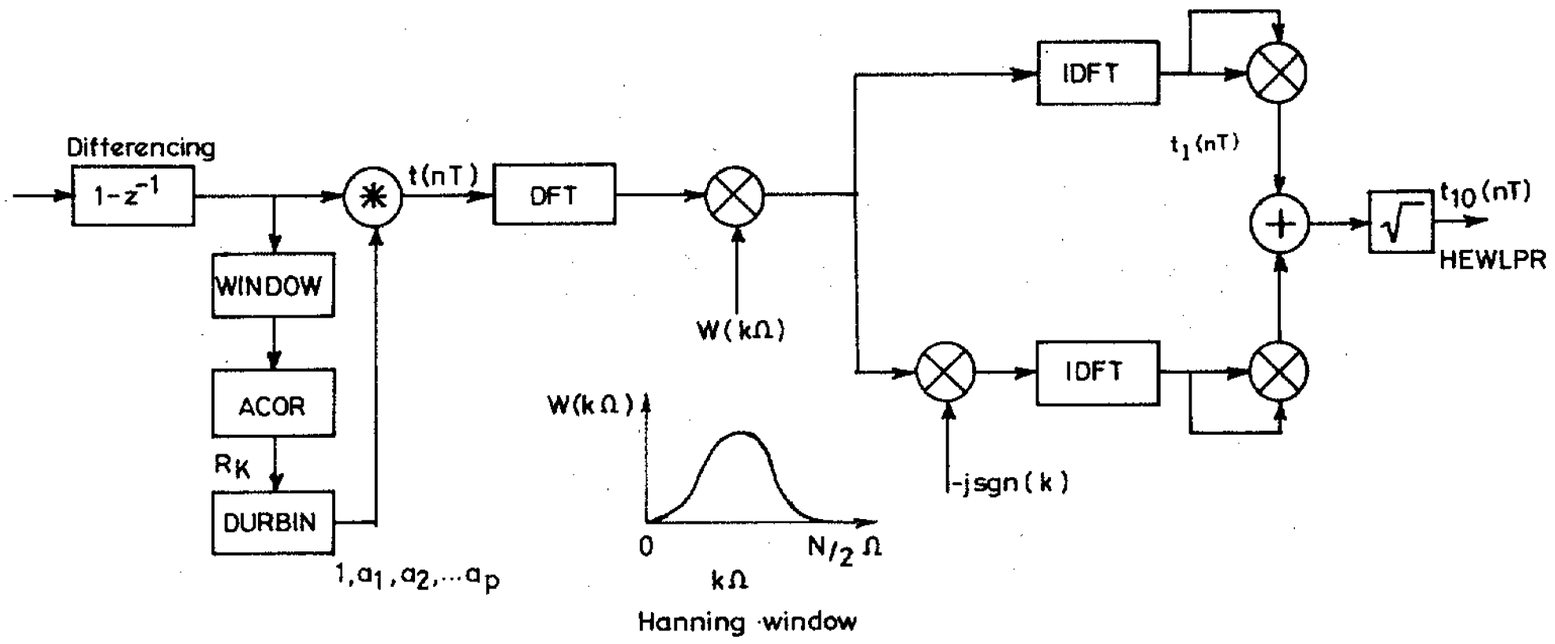

Fig. 4. Computational scheme for the EFLPR method.

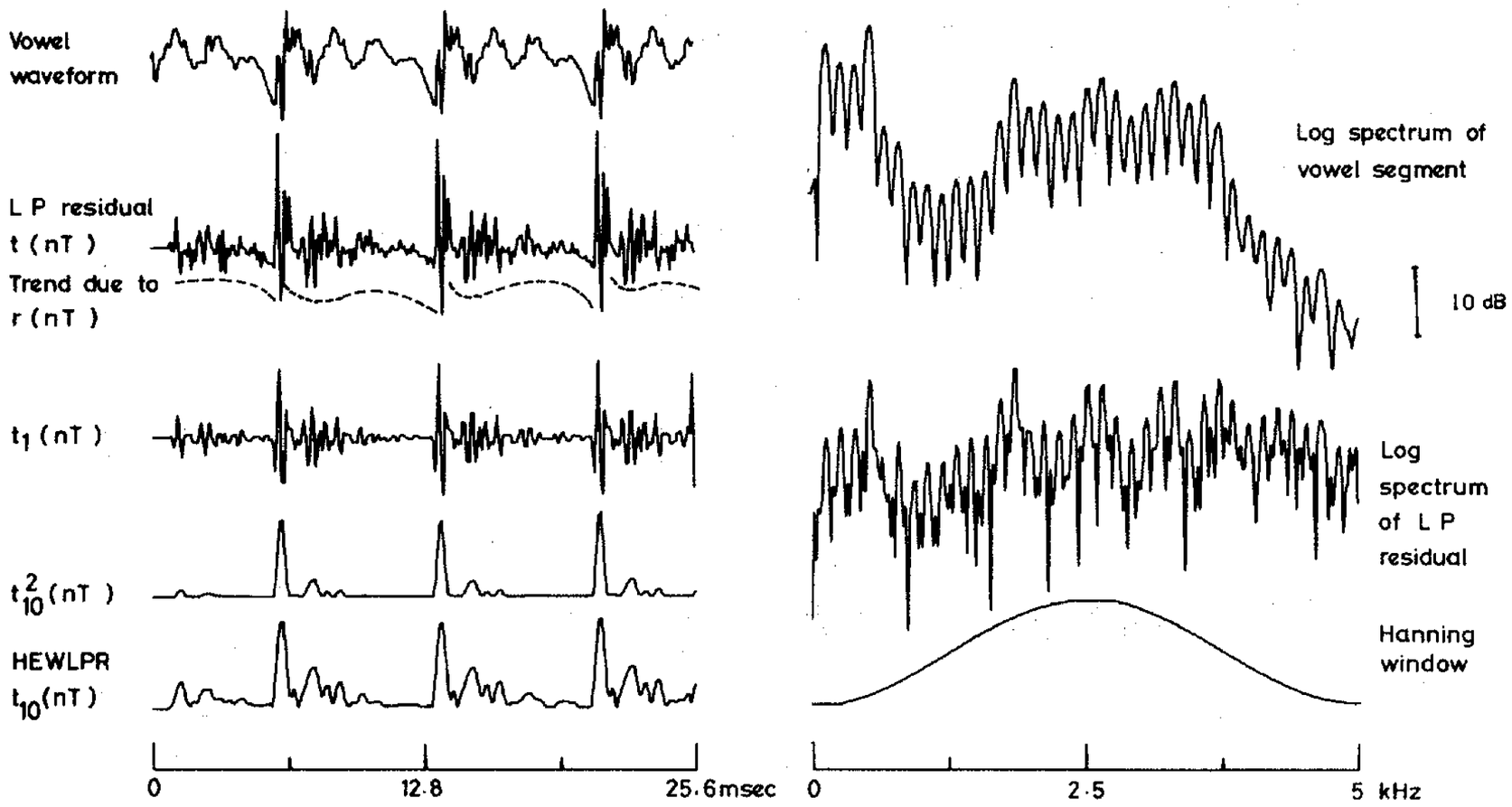

Fig. 5. Illustration of the EFLPR method for a vowel segment.

partially overcomes some of the limitations discussed above. The method is not intended to improve the basic accuracy of LP analysis itself but rather it is aimed at reducing the ambiguities involved in the direct use of the LP residual for epoch identification.

IV. Epoch Filtering of LP Residual

The epoch filtering technique has been proposed for epoch extraction from voiced speech [11]. The output of an epoch filter has a limited resolution and the technique is suitable for use on clean data. By performing the epoch filtering of the LP residual, the output resolution is improved compared to the output of the epoch filter, and the ambiguities in interpreting the LP residual are also overcome. An epoch filter performs a bandpass operation of the speech signal over a region where the signal spectrum is nearly flat. Bandpass filtering is realized by using a frequency domain window function and FFT algorithms. The effect of the phase spectrum of the speech signal over the passband on the bandpass filter output has been discussed in the theory of the epoch filter [11]. Epoch filter theory suggests the computation of the Hilbert envelope of a bandpass filter output to resolve the ambiguities in epoch identification. Since the spectrum of the LP residual possesses flat spectral characteristics, the entire frequency range from zero to folding frequency may be used in the epoch filter. The LP residual consists of two components $e(nT)$ and $r(nT)$ as discussed before. We are interested only in extracting $e(nT)$. As most of the energy of $r(nT)$ is concentrated in the low-frequency region, a frequency domain window function which attenuates the low-frequency components of the LP residual is used in the epoch filter. Also, from considerations of noise, the frequency components near the folding frequency are given a lower weightage. We have used a Hanning window centered at half the folding frequency in the epoch filter (Fig. 4). The computation of the Hilbert envelope overcomes the effect of phase upon the LP residual discussed in Section III-C. The computational steps involved in this method of epoch filtering of LP residual (EFLPR) are shown in Fig. 4. The LP residual obtained after multiplying the residual transform with a Hanning window function in the frequency domain is denoted by $t_1(nT)$. The output, denoted by $t_{10}(nT)$, will be referred to as HEWLPR (Hilbert envelope of windowed LP residual). The EFLPR method is illustrated in Fig. 5 for a vowel segment. The normal LP residual $t(nT)$ contains significant samples of either polarity within every pitch period. The low-frequency trend in $t(nT)$ due to $r(nT)$ is distinctly absent in the signal $t_1(nT)$. The epoch filtered LP residual $t_{10}(nT)$ shows unambiguous peaks at epoch locations. A secondary peak corresponding to the glottal opening can also be seen in $t_{10}(nT)$.

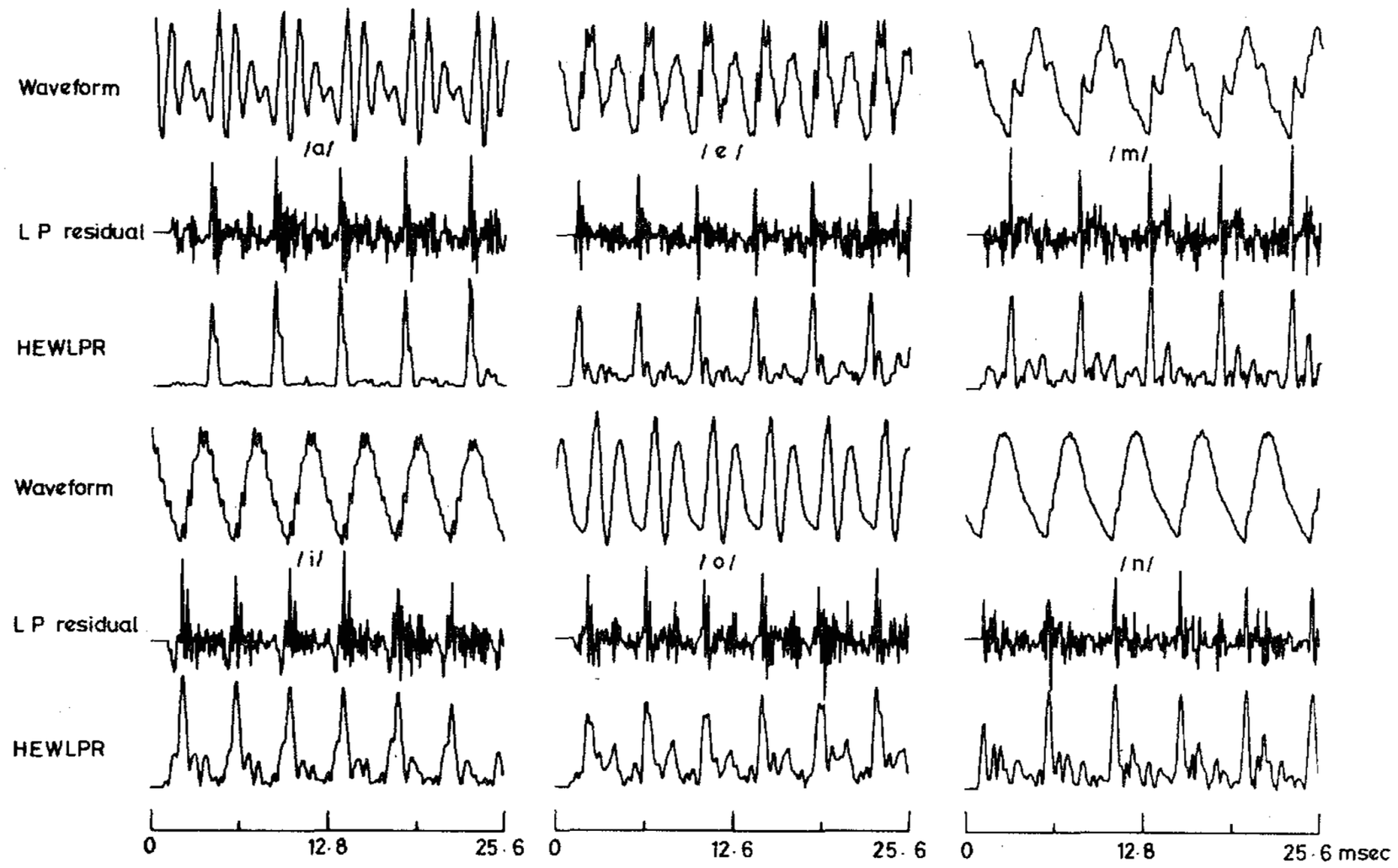

Fig. 6. Epoch filtered LP residual for different voiced segments of speech.

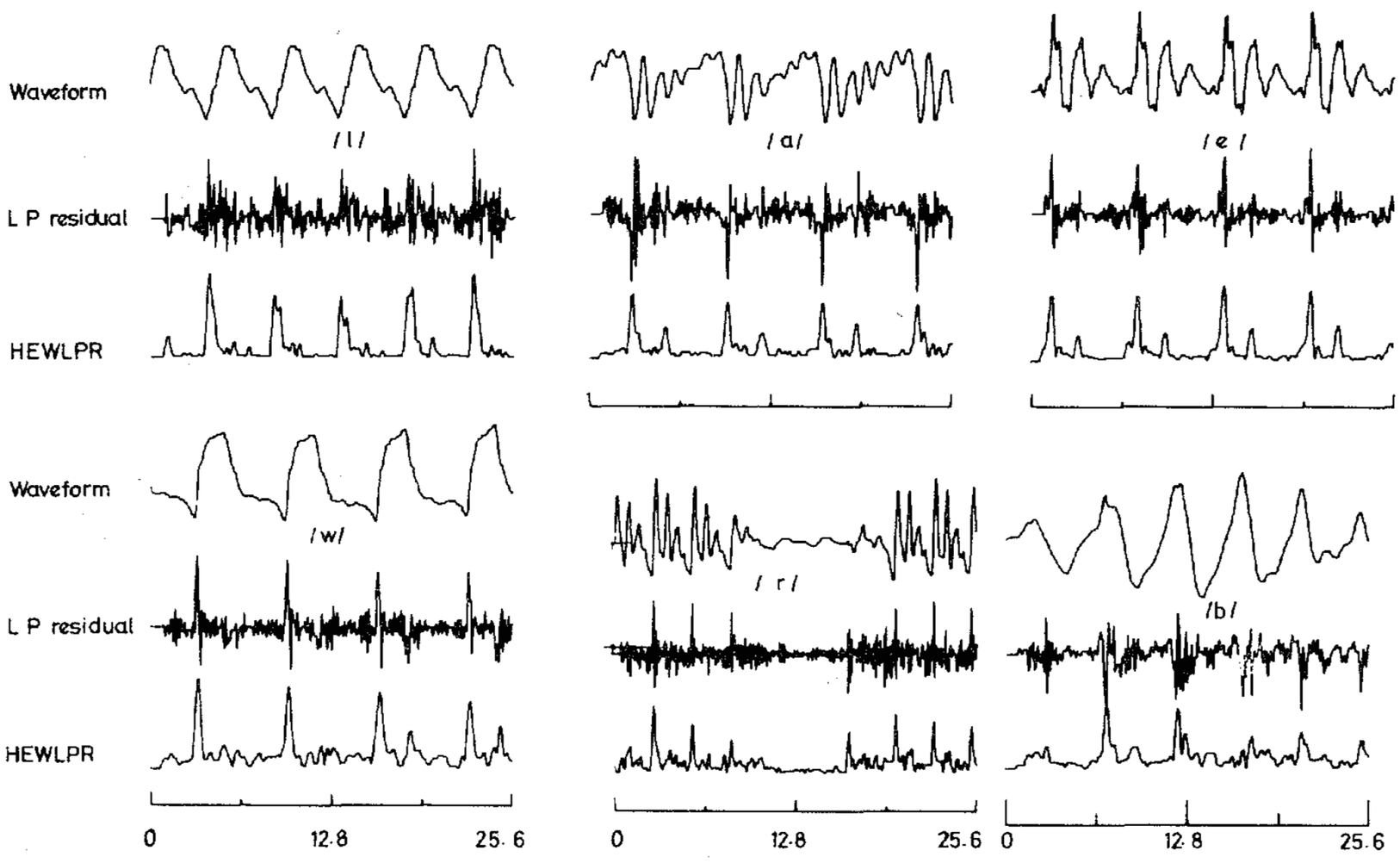

Fig. 7. Epoch filtered LP residual for different voiced segments of speech.

The results obtained by the EFLPR method for several segments of speech sounds are shown in Figs. 6 and 7. These results illustrate the effectiveness of the EFLPR method in resolving the ambiguities present in the use of LP residual for epoch identification.

The EFLPR method relies on the accuracy of LP analysis for satisfactory performance. For sounds like vowel /u/ [10], high-pitched voiced sounds [17], and voiced fricatives, it is noticed that the LP residual cannot be used satisfactorily for the extraction of epochs. Large errors in the estimation of formants and bandwidths occur in the LP analysis of these sounds. In order to produce frication for a voiced fricative, the glottis never closes completely. Moreover, there are two sources simultaneously exciting the vocal tract system. Further, the transfer function of the system contains zeros for these sounds. Hence, it is difficult to analyze fricatives for extracting epochs. LP residual and HEWLPR for some difficult cases are shown in Fig. 8. It may be noted that it is difficult to identify the epochs from LP residual or HEWLPR. In such cases one may use the zero-phase inverse filtering technique described in [20]. Here the LP residual is obtained by passing the speech signal through a zero-phase inverse filter $A_0(z)$. The filter $A_0(z)$ has the same magnitude response as that of $A(z)$ but possesses a zero phase. The output of $A_0(z)$ is then

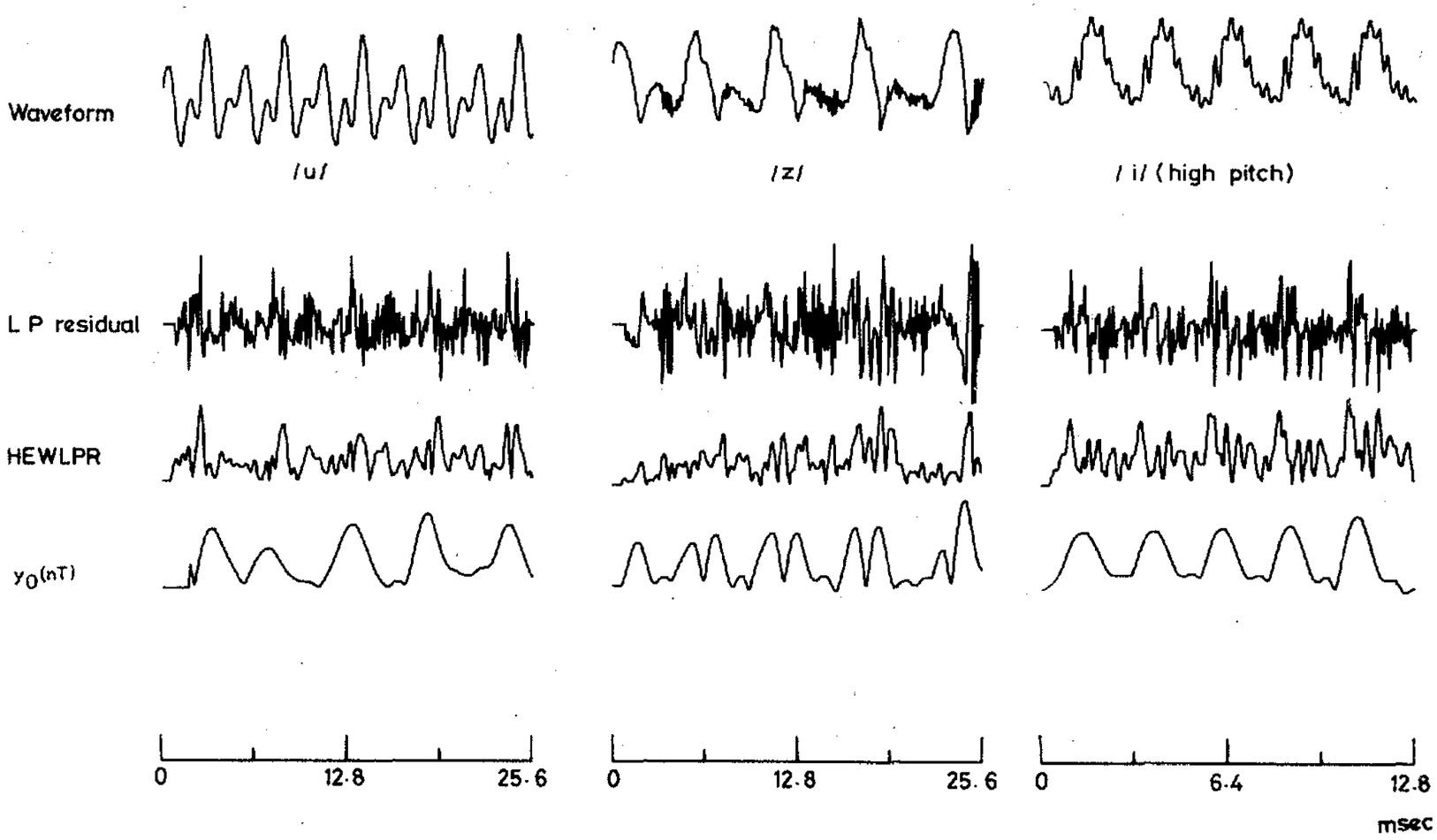

Fig. 8. HEWLPR for some difficult cases.

processed by an epoch filter. A frequency domain window (of bell cosine form) of 1250 Hz width is used. The location of the window function has to be chosen adaptively in order to obtain an unambiguous output. This procedure is, however, time consuming, and also resolution obtainable at the output is limited.

V. Analysis over the Closed Glottis Interval

An important application of EFLPR method is in the identification of the closed glottis interval. It is known that epochs usually occur at glottal opening and closure [1, ch. 6, p. 241], [18]. In order to test the accuracy of the EFLPR method the approximate shape of glottal pulses for vowel sounds is estimated, and then a comparison of pulse shape and HEWLPR is made. Also, the speech signal in the closed glottis interval is analyzed by the covariance method of linear prediction [8] to estimate the frequency response of the vocal tract system.

A. Extraction of Glottal Pulses

The following steps outline the method used for estimating the approximate shape of glottal pulses.

1) Linear predictor coefficients a_i, $i = 1, 2, \cdots, 8$ for a differenced vowel segment of about four pitch periods are computed.

2) Formant frequencies (F_k) are obtained from the derivative of the phase spectrum of the digital inverse filter $A(z)$ [19].

3) Let $b_{1k} = -2\exp(-\pi B_k T)\cos(2\pi F_k T)$ and $b_{2k} = \exp(-2\pi B_k T)$.

A filter with impulse response $(1, b_{1k}, b_{2k})$ has an antiresonance at F_k with bandwidth B_k. Sequences $(1, b_{1k}, b_{2k})$, for each k, $k = 1, 2, \cdots, M$, where M is the number of resonances, are computed. For each formant, the knowledge of F_k as found in Step 2) is used. The sequences are calculated for several trial bandwidth values (B_k) for each formant. The speech segment is convolved with the sequences $(1, b_{1k}, b_{2k})$ for different choices of bandwidths of the antiresonant filters until the formant ripples in the output are reduced to a minimum.

4) The output of the cascade of antiresonant filters is integrated to obtain the approximate shape of glottal pulses. Integration is performed to account for the radiation effect.

The output of the antiresonant filters at various stages obtained by the above method are illustrated in Fig. 9. It is interesting to note that the high-frequency ripples in the output obtained after removing the first formant indicates the presence of multiple excitations within a pitch period (marked by arrows). This result is confirmed by the epoch filter output $t_{10}(nT)$ which has two epochal peaks per pitch period.

Speech waveform, epoch filtered LP residual (HEWLPR), and the estimated glottal pulse shapes for different vowel sounds are shown in Figs. 10 and 11. Different vowel sounds uttered by the same speaker and a given vowel sound, viz., /e/, uttered by different speakers are considered for analysis. These sounds were produced in a computer room into a microphone held about four inches away from the speaker and were directly digitized. Care was taken to avoid nasalization. From the figures we notice that the major excitation occurs at the instant of glottal closure. Further, in several cases we observe a peak in HEWLPR at the instant corresponding to the opening of the glottis. Speech samples following the principal peak in HEWLPR can be considered to belong to the closed glottis interval. In cases where an epoch occurs at the glottal opening also, then the closed glottis interval can be clearly identified.

B. Analysis in the Closed Glottis Interval

The accuracy of epoch identification was tested by performing LP analysis of the speech signal in the closed glottis interval. Since this interval contains speech data corresponding to the force-free response of an all-pole system, the frequency response of the system can be accurately estimated. It is well known that the covariance formulation of LP analysis is suitable for estimating LPC's from short segments of

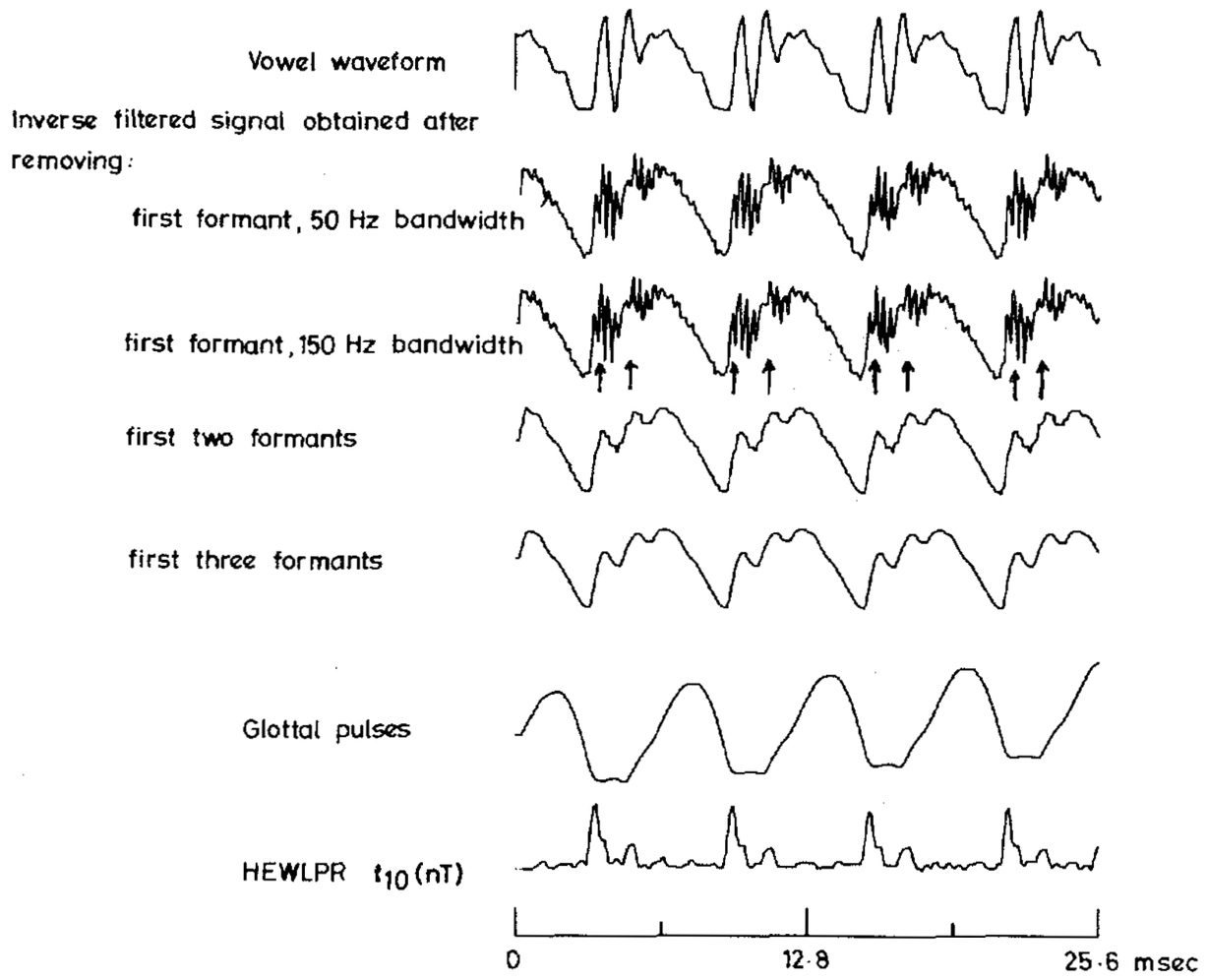

Fig. 9. Extraction of glottal pulses.

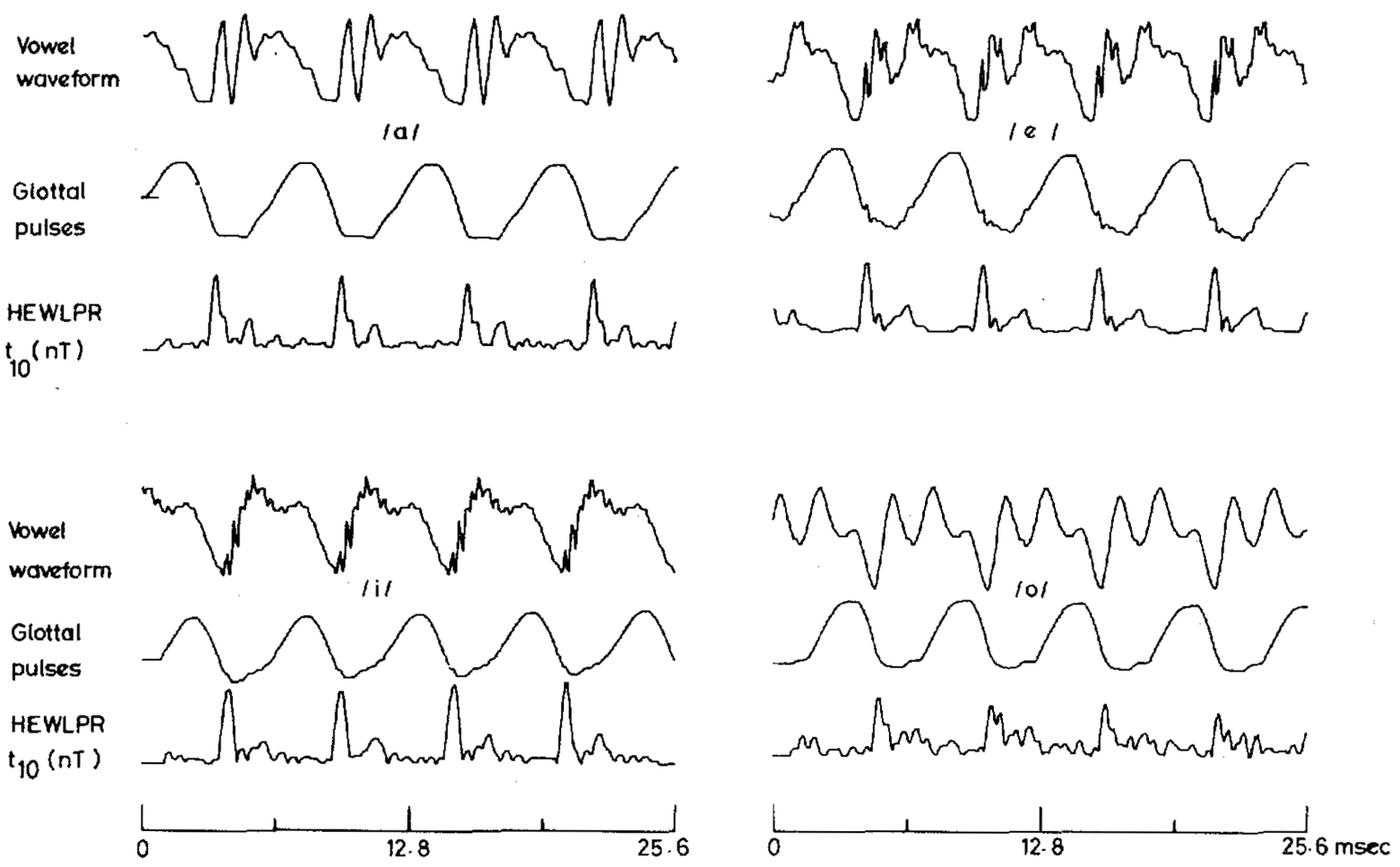

Fig. 10. Glottal pulses and HEWLPR for different vowels spoken by the same speaker.

speech data [8].. The covariance method consists in solving the following set of linear simultaneous equations:

$$\sum_{i=1}^{p} K_i Q_{ik} = -Q_{0k}, \qquad 1 \leqslant k \leqslant p \tag{21}$$

where

$$Q_{ik} = \sum_{n=0}^{N-1} s(nT - iT)\, s(nT - kT) \tag{22}$$

is the covariance of $\{s(nT)\}$ in the analysis interval, N is the number of samples in the interval, p is the order of the predictor, and K_i, $i = 1, 2, \cdots\cdots, p$ are the LPC's. It may be noted that p past samples are required for computing the covariance. In the closed glottis interval analysis all the $(N+p)$ samples should be within this interval.

Synthetic and natural sounds were analyzed by the covariance method of linear prediction analysis to estimate the frequency response of the vocal tract system. Speech samples belonging to various intervals within a pitch period were considered for analysis. These intervals relative to glottal pulse are as follows:

(i) the interval inclusive of the glottal opening;
(ii) the mid-portion of the glottal pulse;
(iii) the interval inclusive of closure; and
(iv) the closed glottis interval.

Estimated LPC's for the above intervals are denoted by K_i^j,

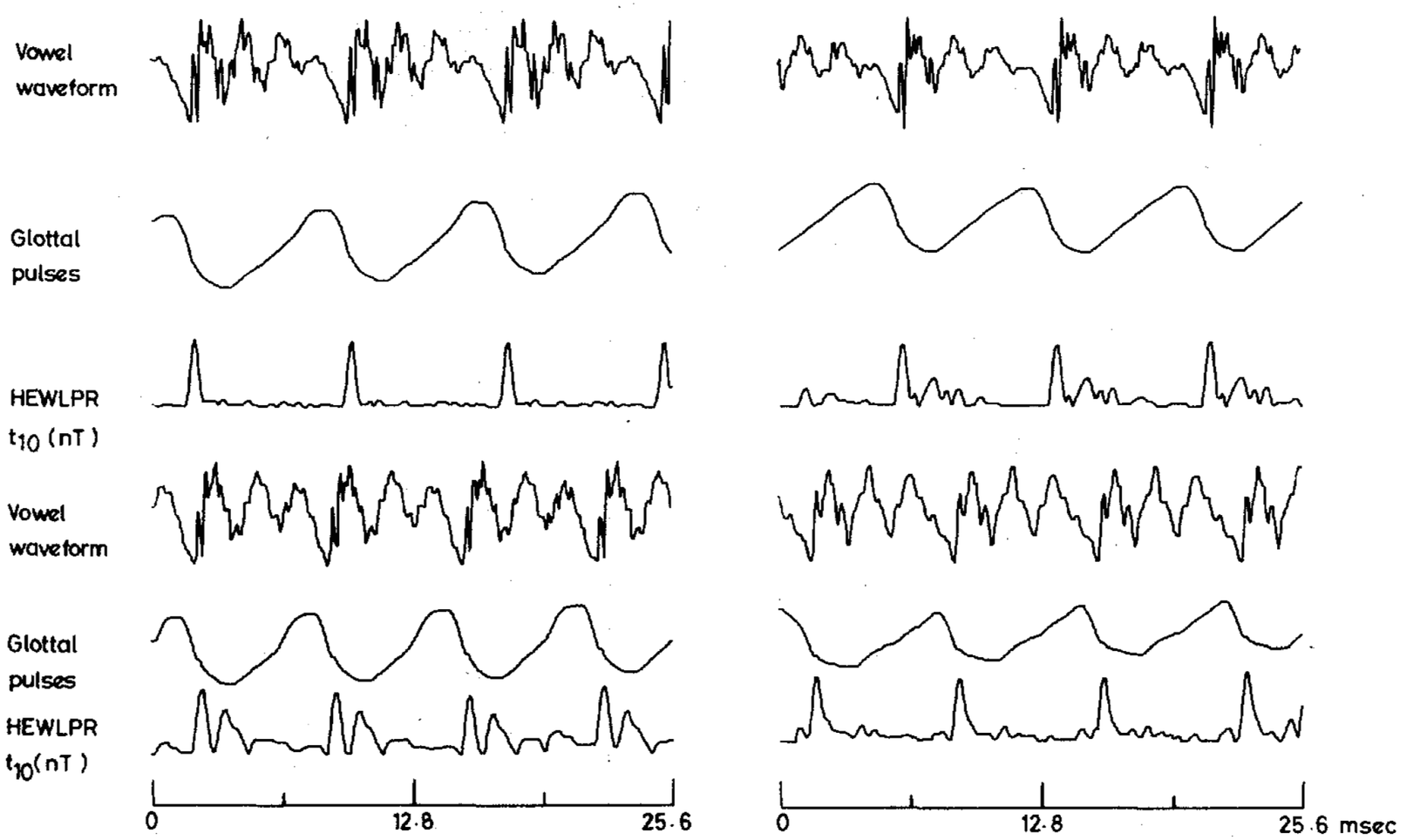

Fig. 11. Glottal pulses and HEWLPR for the vowel /e/ spoken by different speakers.

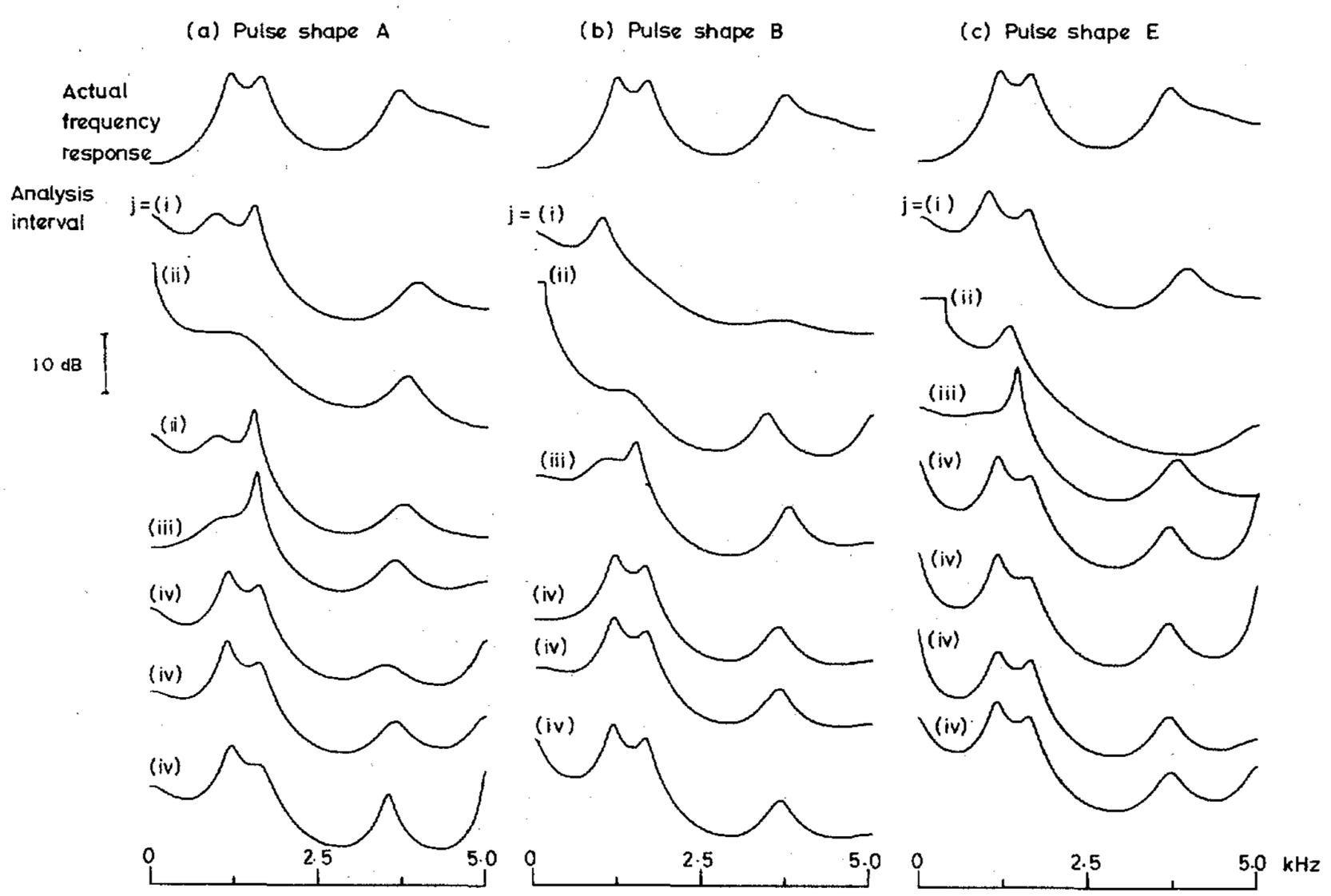

Fig. 12. Estimated frequency response of $V(z)$ for different analysis intervals: synthetic vowels.

where j = (i)-(iv) represents the analysis interval. For synthetic vowels, the closed glottis interval was 4 ms. Hence, p and N were chosen to be eight and twenty, respectively. Three different pulse shapes (A, B, and E described in [16]) were considered. For the natural vowels shown in Fig. 10, the estimated closed glottis interval was about 2 ms and hence, a value of eight was chosen for both p and N.

The results of analysis for synthetic and natural vowels are shown in Figs. 12 and 13. These figures show the frequency response of the vocal tract system determined by the LPC's for the different analysis intervals. In the case of synthetic vowels, the estimated frequency response compares well with the actual frequency response only when the analysis is performed over the closed glottis interval (iv). The three curves marked as (iv) correspond to different intervals in the closed glottis region. Bandwidths of formants are also estimated correctly. For other choices of the analysis interval, viz., (i, ii, and iii), the estimated frequency response is strongly influenced by the choice of the analysis interval. It is interesting to note that there is a marked change in the spectral shape

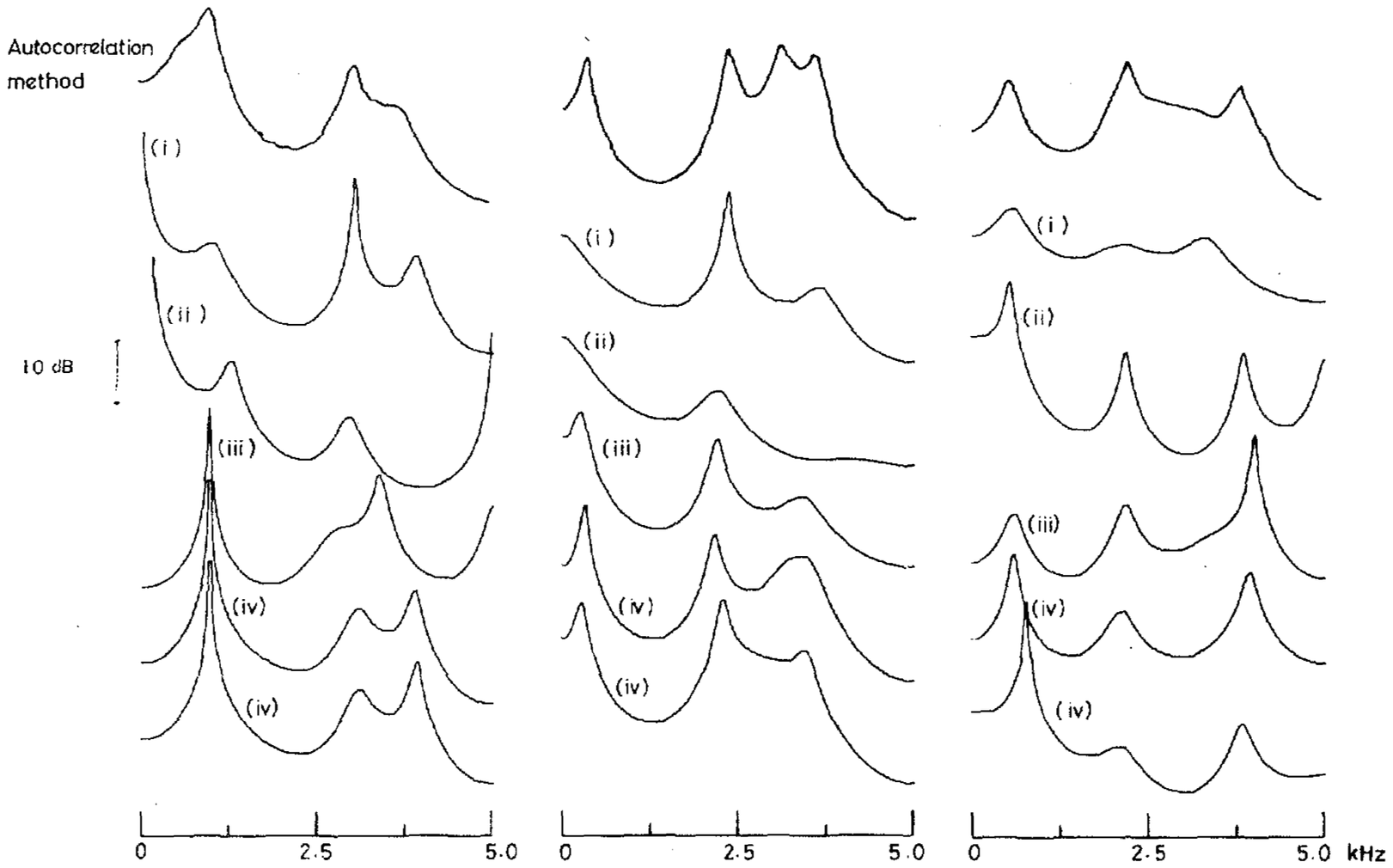

Fig. 13. Estimated frequency response of $V(z)$ for different analysis intervals: natural vowels.

as the analysis interval just crosses the instant of glottal closure. The accuracy of epoch identification for natural vowels can be tested by observing similar changes in the estimated frequency responses for different analysis intervals. Fig. 13 illustrates this effect. For the case of natural vowels the estimated frequency responses [curves marked (iv)] for different intervals do not always match because the excitation is not strictly zero over the closed glottis region.

VI. Conclusion

Direct use of the LP residual for extracting epochal information is not very effective owing to the occurrence of samples of either polarity of large values around the instant of significant excitation. A systematic analysis of LP residual has shown that the ambiguities arise mainly due to zeros in the vocal tract system and the phase angles of formants at the instant of excitation. A method for unambiguous identification of epochs from the LP residual consists of computing the Hilbert envelope of the error signal. The accuracy of epoch identification was tested by extracting glottal pulses and comparing the instants of slope discontinuities in the glottal pulses with the epoch locations. An important application of epoch extraction methods is in the accurate identification of the closed glottis interval. Such an interval provides a means for computing the true frequency response of the supraglottal vocal cavity. The importance of excluding the instant of glottal closure from the analysis interval and the accuracy of identification of the closure instant from the Hilbert envelope of LP residual were clearly demonstrated through the LP analysis performed over several intervals within a pitch period. The presence of several excitation instants within a pitch period suggests that the conventional method of characterizing the voice source by a single impulse per pitch period may not be adequate. Epoch characterization of the voice source effectively represents the pitch variations and the important features of the glottal pulse.

References

[1] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, 2nd ed. New York: Springer-Verlag, 1972, ch. 5, pp. 184–186.

[2] E. N. Pinson, "Pitch synchronous time-domain estimation of formant frequencies and bandwidths," *J. Acoust. Soc. Amer.*, vol. 35, Aug. 1963, pp. 1264–1273.

[3] A. N. Sobakin, "Digital computer determination of formant parameters of the vocal tract from a speech signal," *Soviet Phys.–Acoust.* (Transl.: Amer. Inst. Phys.), vol. 18, July–Sept. 1972.

[4] K. Steiglitz and B. Dickinson, "The use of time-domain selection for improved linear prediction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 34–39, Feb. 1977.

[5] J. N. Holmes, "Formant excitation before and after glottal closure," in *Conf. Rec., IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1975, pp. 39–42.

[6] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 399–418, Oct. 1976.

[7] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[8] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, Aug. 1971.

[9] J. Makhoul and J. I. Wolf, "Linear prediction and the spectral analysis of speech," Bolt Beranek and Newman Inc., Cambridge, MA, BBN Rep. 2304, 1972.

[10] H. W. Strube, "Determination of the instant of glottal closure from the speech wave," *J. Acoust. Soc. Amer.*, vol. 56, pp. 1625–1629, Nov. 1974.

[11] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 562–570, Dec. 1975.

[12] T. V. Ananthapadmanabha, "Epoch extraction and its application to voiced speech analysis," Ph.D. dissertation, Dep. Elec. Commun. Eng., Indian Institute of Science, Bangalore, India, Sept. 1977.

[13] J. D. Markel and A. H. Gray, *Linear Prediction of Speech.* New York: Springer-Verlag, 1976.

[14] G. Fant, *Acoustic Theory of Speech Production,* 2nd ed. The Hague, The Netherlands: Mouton, 1970, ch. 1, pp. 15–21.

[15] A. H. Gray and J. D. Markel, "A spectral flatness measure for studying the autocorrelation method of linear prediction analysis of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 207–217, June 1974.
[16] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Amer.*, vol. 49, pp. 583–590, Feb. 1971.
[17] B. S. Atal and M. R. Schroeder, "Recent advances in predictive coding–Applications to voiced speech synthesis," Speech Commun. Seminar, Stockholm, 1974.
[18] M. V. Mathews, J. E. Miller, and E. E. David, "Pitch synchronous analysis of voiced sounds," *J. Acoust. Soc. Amer.*, vol. 33, pp. 178–186, 1961.
[19] B. Yegnanarayana, "Formant extraction from linear prediction phase spectra," *J. Acoust. Soc. Amer.*, vol. 63, p. 1638, May 1978.
[20] T. V. Ananthapadmanabha and B. Yegnanarayana, "Zero-phase inverse filtering for extraction of source characteristics," in *Conf. Rec., IEEE Int. Conf. on Acoust., Speech, Signal Processing*, 1977, pp. 336–339.

An Approach to Segmenting Speech into Vowel- and Nonvowel-Like Intervals

HIDEKI KASUYA, MEMBER, IEEE, AND HISASHI WAKITA, MEMBER, IEEE

Abstract—A speaker-independent algorithm is given for segmenting continuous speech in English into vowel-like (*V*) and nonvowel-like (*NV*) intervals. The algorithm has three stages: *measurements* (parameter extraction), *phonetic feature detection*, and *V/NV decision*. In *measurements*, the broad-band rms energy, the back-to-total cavity volume ratio (BTR), the signed front-to-back maximum area ratio (SFBR), and the normalized high-to-low frequency energy ratio (HLR) are computed. The BTR and SFBR are new parameters derived from linear prediction area functions and are interpreted in terms of the speech spectrum. The BTR is useful for distinguishing nasal segments from *V* segments, while the SFBR is effective for detecting the bursts of voiced plosives. In *phonetic feature detection*, three independent types of intervals are detected on the basis of the parameters: silence, preliminary *V/NV*, and turbulence noise. The *V/NV decision* stage accomplishes the final *V/NV* interval decision.

Interspeaker differences are handled by normalizing the frequency scale on the basis of an estimated average vocal-tract length.

Ten sentences spoken by each of two males and two females resulted in 93.3 percent correct *V/NV* segment-detection decisions (92.9 percent for design speakers, and 93.7 percent for test speakers).

I. Introduction

At an early stage of automatic segmentation and phonetic labeling of continuous speech, it is appropriate to classify sound segments into two different classes: 1) sound segments corresponding to vowels, diphthongs, semivowels, and liquids (vowel-like segments), and 2) sound segments corresponding to the remaining consonants (nonvowel-like segments).

Vowels, diphthongs, semivowels, and liquids (perhaps except /l/) are all primarily characterized by the resonances of the vocal tract, whereas most consonants are produced with anti-resonances as well. Because of this basic difference in spectral structure, efficient algorithms for automatic processing can be expected to be different for the two categories. In particular, if the vowel-like/nonvowel-like decision can be made accurately on the basis of the *gross characteristics* of the speech wave, the use of formant frequencies, which typically require elaborate extraction algorithms [1], can be avoided for nonvowel-like segments which constitute a considerable portion of the material. On the other hand, analysis and perception experiments indicate that within the vowel-like class the vowels, diphthongs, semivowels, and liquids can be distinguished by taking the temporal characteristics of formant frequencies into account [2]–[6].

This paper presents an algorithm which segments continuous speech produced by arbitrarily selected speakers into vowel-like (*V*) and nonvowel-like (*NV*) intervals.

Our approach to finding effective cues for the *V/NV* interval decision is to utilize the rms energy as a primary parameter, and then to employ several other parameters to compensate for failures of the rms energy. Various kinds of rms energy functions have been used in previous studies of segmentation and phonetic labeling. Stowe [7], Mermelstein [8], and Lea *et al.* [9] have reported that the large energy dips in frequency-weighted rms energy functions provide significant cues for locating syllable boundaries. The rms dips have also been used to detect some consonant segments [6], [10], [11]. However, even if a well-determined frequency band has been selected to measure the rms energy, it still seems to be difficult to use only the rms energy functions to classify vowels, diphthongs, semivowels, and liquids as one class, and the nasals as the other class [6]. In addition, the rms energy function is not sufficient for detecting nonintervocalic nasals.

Manuscript received March 1, 1977; revised August 16, 1977, December 19, 1977, June 19, 1978, and February 16, 1979.

H. Kasuya is with the Department of Electronics, Utsunomiya University, Utsunomiya, Japan.
H. Wakita is with the Speech Communications Research Laboratory, Inc., Los Angeles, CA 90007.

0096-3518/79/0800-0319$00.75