Automatic and Reliable Estimation of Glottal Closure Instant and Period

YAN MING CHENG, MEMBER, IEEE, AND DOUGLAS O'SHAUGHNESSY, MEMBER, IEEE

Abstract—**The automatic and reliable determination of the Glottal Closure Instant (GCI) has recently become a fundamental requirement for high-quality speech synthesis and speech coding. Moreover, pitch estimation based on GCI determination can exhibit changes period-by-period; thus, nonstationary pitch variation can be detected. We use maximum-likelihood epoch determination as the basis to locate GCI's, and discuss the application of the Hilbert transformation to improve performance and reliability. A description of the system and the voiced/unvoiced/mixed (V/UV/M) decision procedure follows. The results of this study show that the algorithm works in noise-free, noisy, and very-noisy signals for vowels as well as for voiced consonents. Our pitch estimation only needs a very short frame length, and gives accurate results at voice onset.**

I. Introduction

CONSIDER the pitch period as a unit in a voiced speech signal. Within this unit, the behavior of the speech-production organs can be clearly specified and the system is very close to being time invariant. Some modern speech-synthesis and speech-coding strategies tend to model the signal period-by-period [1] and require epoch (period boundary) determination. The period variation carries phonemic, linguistic, and speaker information. It is also helpful and sometimes necessary for speech recognition; for instance, the micro-melody (short-time and intrinsic period variation) can aid some phoneme recognition, e.g., for plosive constants. The macro-melody (long-time period variation) specifies the stress, the notion of question versus statement, speaker identity, etc. Unfortunately, there has been little work devoted to epoch determination, and most period estimation methods require a long signal duration [2], yielding an average period over a long time span.

In the past several years, a few methods of epoch determination have been presented. One locates the glottal closure instant (GCI) using predictive error, but the presence of a pulse at the input of the vocal tract is not always predictable. Algorithms with more complete consideration based on this assumption fall under the heading of autocovariance matrix determinant evaluation [3]. This method, however, does not work with all vowel signals; indeed, some vowels cause great difficulty in determining GCI's. The problem is due to many significant residual pulses occurring around the GCI from both the input pulse and the large predictive error. Otherwise, this method is adequate, but computationally expensive. An alternative method uses the occurrence of discontinuities in derivatives of the glottal airflow [4], [5]. This algorithm works well for most clean vowel signals. However, for vowels which do not possess manifest discontinuities in the derivatives of glottal airflow, e.g., strong deceleration of the airflow of front and high vowels during the closing glottis, they fail. A drawback of this kind of method is the confusion the discontinuities introduce by the derivatives or by the noise excitation and contaminant noise. Thus, restrictions of clean data and of certain vowel signals are imposed on its application. Therefore, until now, there has not been an epoch determination algorithm whose performance covers all speech signals, and virtually no period estimators are based on epoch determination.

In contrast to progress in epoch determination, direct period estimation has been further developed. However, one must compromise between the reliability of performance and the analysis frame length. For instance, some methods (e.g., Maximum Likelihood [6], AMDF [7], Center Clipping Autocorrelation [8]) are able to estimate the period with short time frames, about 20 ms. However, there is ambiguity in finding the true period peak in a set of comparable peaks. Other methods (e.g., Cepstrum [9], SIFT [10], Harmonic Matching [11]) reliably locate the period, but the analysis frame is very long, more than 40 ms. The drawback of a long frame is the smoothing of the micro-melody. None of the latter methods can sense period-to-period variations or nonstationary period variations within the long frame.

The increase in demand of epoch determination and for period estimation with speech synthesis, coding, and recognition development encourages creation of a more sophisticated algorithm. Ideally, one should find epochs reliably and accurately, even at voice onsets, and then obtain directly the period variation. In this paper, we describe an algorithm which approaches the ideal and pays much attention to its applicability, e.g., to various types of signals, contaminant noise, etc. In the next section, we describe the nucleus of the algorithm, maximum-likelihood epoch determination (MLED). In Section III, using the Hilbert transformation to enhance the reliability of epoch determination is discussed. Section IV describes the actual system and the use of its byproducts to make the

Manuscript received February 17, 1988; revised February 18, 1989. This work was supported in part by NSERC-Canada and FCAR-Quebec.

The authors are with INRS-Telecommunications, 3 Place du Commerce, Nuns' Island, Verdun, P.Q., Canada H3E 1H6.

IEEE Log Number 8931338.

0096-3518/89/1200-1805$01.00 © 1989 IEEE

voiced/unvoiced/mixed (V/UV/M) excitation decision. Section V discusses the results, and Section VI concludes this work.

II. Maximum-Likelihood Epoch Determination (MLED)

The maximum-likelihood theory for epoch detection in radar applications has been established for some time [12], [13]. Here we adapt this theory for application to speech signals. First, we assume the speech signal within a pitch period is induced by a pulse at one epoch (or by an event). Generally, we define this epoch as a representation of the glottal closure instant (GCI), because the GCI induces the sound vibration and introduces most of the energy within each period. Assuming that speech production can be modeled as an all-pole linear system, the wavelet due to an epoch can be expressed as

$$\hat{s}(n) = \begin{cases} \sum_{i=1}^{p} a_i \hat{s}(n-i) & 0 < n \leq \infty \\ G & n = 0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where G is a positive arbitrary constant and p is the order of the polynomial. This is an autoregression of outputs with null initial conditions except at the epoch, where the initial condition is specified by a Kronecker delta pulse. It is evident that there is a unique discontinuity in the wavelet at the epoch because it is easy to prove that the autoregression process is continuous throughout without input. The importance of using the wavelet induced by the epoch instead of direct use of the discontinuities is to guarantee that the discovered discontinuity is that evoking the vibration in the vocal tract.

We suppose that the difference between the observed signal, $s(n + n_0)$, $n \in [0, N-1]$ (n_0 is a sequence alignment delay), and the wavelet is a Gaussian process and that the N observations with corresponding wavelet values construct a Gaussian process with N independent dimensions and with uniform variance σ.

$$x(n) = s(n + n_0) - \hat{s}(n)$$

$$\boldsymbol{X} = \boldsymbol{S} - \hat{\boldsymbol{S}}$$

where

$$\boldsymbol{X} = [x(0), x(1), x(2), \cdots, x(N-1)]$$

$$\boldsymbol{S} = [s(n_0), s(1 + n_0), s(2 + n_0), \cdots, s(N - 1 + n_0)]$$

$$\hat{\boldsymbol{S}} = [\hat{s}(0), \hat{s}(1), \hat{s}(2), \cdots, \hat{s}(N-1)].$$

This supposition is relevant for speech signals [6], [14]. Thus, the conditional probability density, or likelihood function, of the epoch appearance as a function of the parameters can be written as

$$p(\boldsymbol{X}/\Omega) = \frac{1}{(2\pi\sigma^2)^{N/2}} \cdot \exp\left\{ -\sum_{n=0}^{N-1} [s(n + n_0) - \hat{s}(n)]^2 / 2\sigma^2 \right\}, \tag{2}$$

where the Ω is the parameter space, $\Omega = [\sigma, a_1, \cdots, a_p, n_0]$. The maximum-likelihood estimation means that the epoch occurs when the parameter values maximize the likelihood function [15]. The equivalence of maximizing this estimator is to maximize the logarithm likelihood estimator

$$\ln [p(\boldsymbol{X}/\Omega)] = -\frac{N}{2} \ln (2\pi\sigma^2) - \frac{\sum_{n=0}^{N-1} [s(n + n_0) - \hat{s}(n)]^2}{2\sigma^2}. \tag{3}$$

Using $\partial \ln (p(\boldsymbol{X}/\Omega))/\partial\sigma = 0$, we obtain the optimal variance of the Gaussian process, which yields a partially maximum-likelihood function

$$\sigma_{opt} = \sqrt{\frac{2\pi}{N} \sum_{n=1}^{N-1} (s(n + n_0) - \hat{s}(n))^2}. \tag{4}$$

The parameter n_0 is a nonlinear function of the logarithmic likelihood function. An explicit expression for the optimal value of n_0 is not available. Thus, the optimal value can be obtained by enumerating all the possibilities and searching for the maximum. We can write (3) as

$$\begin{aligned} \ln [p(\boldsymbol{X}/\Omega)] &= -\frac{N}{2} \ln (2\pi\sigma^2) - \frac{\sum_{n=0}^{N-1} [s^2(n + n_0) + \hat{s}^2(n) - 2s(n + n_0)\hat{s}(n)]}{2\sigma^2} \\ &= -\frac{N}{2} \ln (2\pi\sigma^2) - \sum_{n=0}^{N-1} \frac{s^2(n + n_0)}{2\sigma^2} - \sum_{n=0}^{N-1} \frac{\hat{s}^2(n)}{2\sigma^2} + \sum_{n=0}^{N-1} \frac{s(n + n_0)\hat{s}(n)}{2\sigma^2}. \end{aligned} \tag{5}$$

Viewing $\ln [p(\boldsymbol{X}/\Omega)]$ as a function of n_0 and assuming that the observed signal energy is constant as a function of n_0 (if N is large enough), the term $\Sigma_{n=0}^{N-1} s(n + n_0)\hat{s}(n)$, hereafter the "MLED signal," alone influences the logarithmic likelihood function. In fact, the MLED signal is a cross-correlation function between the speech signal and the wavelet due to an epoch. Thus, the partial maxi-

mum-likelihood function can be obtained by

$$\max_{n_0 = n_{0opt}} \sum_{n=0}^{N-1} s(n + n_0)\, \hat{s}(n). \tag{6}$$

Combining (1) and (3) and setting $\partial \ln (p(X/\Omega))/\partial a_k = 0$, we obtain a set of equations

$$\frac{1}{\sigma^2} \sum_{n=k}^{N-1} \hat{s}(n-k) \left[s(n + n_0) - \sum_{i=1}^{p} a_i \hat{s}(n-i) \right] = 0, \tag{7}$$

for $k = 1, \cdots, p$.

An assumption, relevant for GCI detection, is that when the appropriate n_0 is achieved ($n_0 = n_{0opt}$) equality in the sense of the maximum likelihood, ignoring carry-over energy from prior periods,

$$s(n + n_{0opt}) = \hat{s}(n), \qquad 0 < n < N \tag{8}$$

holds with the constraint of N less than a period (note that both the assumption and the constraint are temporary and will be removed). Thus, (7) becomes a set of correlation equations, referred to as the autocorrelation method [16], substituting $s(n + n_{0opt})$ for $\hat{s}(n)$:

$$\sum_{i=1}^{p} a_i \Phi(i-k) = \Phi(k) \tag{9}$$

where

$$\Phi(k) = \sum_{n=k}^{N-1} s(n + n_{0opt})\, s(n + n_{0opt} - k).$$

In other words, the autoregressive coefficients of the wavelet, which produce a maximum-likelihood function, are that of the speech linear-prediction coefficients with the autocorrelation method. In this sense, (1) is also denoted a "matched filter." As is well known, such coefficients identify the impulse response of the vocal tract filter, if p is less than the period. Considering the stationarity of a speech signal within a frame, we can approximate $\Phi(k)$ without the above assumption and constraints

$$\Phi(k) \approx \sum_{n=k}^{N_f - 1} s(n)\, s(n-k),$$

where N_f is the frame size.

To investigate if the method noted above is affected by noise, we consider two types of noise: a) additive noise independent of the wavelet process, $\epsilon_a(n)$; b) noise correlated with the wavelet process, which is the predictive error noise, excitation noise, etc., $\epsilon_c(n)$ with the first $q + 1$ nonnull cross correlations, $\rho_0, \rho_1, \cdots, \rho_q$. Replacing the observed clean signal by the noisy signal, we have from (6):

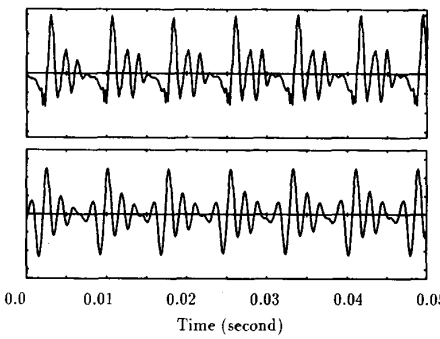


Fig. 1. Typical signals in Maximum-Likelihood Epoch Determination (MLED) for vowel /a/; top: speech signal, bottom: MLED signal. The abscissa is time and the ordinate is amplitude.

$$\begin{aligned}
&\max_{n_0 = n_{0opt}} \sum_{n=0}^{N-1} \left[s(n + n_0) + \epsilon_a(n) + \epsilon_c(n) \right] \hat{s}(n) \\
&\quad = \max_{n_0 = n_{0opt}} \sum_{n=0}^{N-1} s(n + n_0)\, \hat{s}(n) + \sum_{n=0}^{N-1} \epsilon_a(n)\, \hat{s}(n) \\
&\qquad + \sum_{n=0}^{N-1} \epsilon_c(n)\, \hat{s}(n) \\
&\quad \overset{N \to \infty}{\Rightarrow} \max_{n_0 = n_{0opt}} \sum_{n=0}^{N-1} s(n + n_0)\, \hat{s}(n) + \rho_0.
\end{aligned} \tag{10}$$

This equation means the method approaches asymptotically to a no-noise influence model, as N increases. Fig. 1 shows a typical result for the application of the maximum-likelihood epoch determination (MLED) for an actual speech signal. The strongest positive pulse indicates the defined GCI within a period. Based on examination of all the results, it is best to use the 50 percent amplitude (from zero to the positive maximum of a pulse) point on the rising slope (or left slope) of the strongest positive pulse within a period as the mark for the GCI. This criterion cannot be currently proven; it is empirical. This introduces a difference of about 0–8 samples corresponding to that using the maximum point of the strongest pulse for all signals. Experiments show that the difference is more important for high vowels than low vowels.

III. Improvement of the Performance by a Hilbert Transformation

As shown in Fig. 1, the MLED creates not only a strong and sharp epoch pulse, but also a set of weaker pulses which represent the suboptimal epoch candidates within a period. The strength ratio between the proper epoch pulse and the subpulses varies greatly and depends on the utterance condition, signal properties, etc. The variety creates ambiguity for the decision. The conventional solution is to employ a complex logical system (or expert system) to deal with the different cases to achieve a reliable decision. This type of solution appears expensive and inconvenient for practical applications. On the other hand, there is an important aspect intrinsically related to the

GCI—periodicity—which is not used in the MLED procedure. We believe this information can be used to reduce the subpulses. Our object is to find, via the MLED signal, a "selection signal" to emphasize the contrast between the epoch pulse and the subpulses while not greatly increasing computation.

We define the selection signal $g_{\sqcap}(n)$ as a periodic pulse signal with a certain pulse-width Δ, superimposed on the GCI's. The size of Δ can tolerate the period variation due to nonstationarity within the analysis frame. The shape of the pulse might be rectangular, triangular, Gaussian, etc. (or one similar to these but not closed-form, except for the Kronecker delta pulse, which has null pulse-width making it insignificant for the selection signal, as well as difficult to be generated by the MLED signal because of its requirement of spectral flatness). To find the general properties of the selection signal, we note expressions for the rectangular, triangular, and Gaussian versions, based on a Fourier series expansion neglecting an initial delay:

Rectangular

$$g_{\sqcap R}(n) = \sum_{m=0}^{(M-1)/2} \frac{\sin\left(\frac{m\pi\Delta}{P}\right)}{\pi m} \cos\left(\frac{2\pi mn}{M}\right)$$

Triangular

$$g_{\sqcap T}(n) = \sum_{m=0}^{(M-1)/2} \frac{P\left[1 - \cos\left(\frac{m\pi\Delta}{P}\right)\right]}{\Delta(m\pi)^2} \cos\left(\frac{2\pi mn}{M}\right)$$

Gaussian

$$g_{\sqcap G}(n) = \sum_{m=0}^{(M-1)/2} \frac{\sqrt{\Delta/\pi}}{P} e^{-(2\pi\Delta m/P)^2} \cos\left(\frac{2\pi mn}{M}\right),$$

where P is the period in time and $M = \lfloor f_s \times P \rfloor$ ($\lfloor u \rfloor$ denotes the largest integer less or equal to u) is the period in samples. For the expression of the Gaussian pulse signal, we assume $P >> \Delta$.

Examining the three typical expressions, two general properties for the selection signal can be obtained: 1) the cosinusoidal basis alone is used, i.e., the spectrum is symmetric and real; 2) the spectrum has its maximum at the origin then gradually decreases its amplitude (for some pulse shapes, it may possess the spectral "sidelobes"). Consider the MLED signal

$$\hat{f}(n_0) = \sum_{n=0}^{N-1} s(n + n_0)\, \hat{s}(n). \tag{11}$$

According to (8), the z-transform of both $s(n)$ and $\hat{s}(n)$ with a normalization of gain is

$$\hat{S}(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = S(z) z^{+n_{0opt}}.$$

The z-transform of $\hat{f}(n_0)$ is

$$\hat{F}(z) = \hat{S}(z) \cdot S(z^{-1}) = \hat{S}(z) \cdot \hat{S}(z^{-1}) z^{+n_{0opt}}$$

and its Fourier transform is

$$\begin{aligned}\hat{F}(e^{j\omega}) &= \hat{S}(e^{j\omega}) \cdot \hat{S}(e^{-j\omega}) e^{j\omega n_{0opt}} \\ &= \hat{S}(e^{j\omega})\, \hat{S}^*(e^{j\omega}) e^{j\omega n_{0opt}} = \left|\hat{S}(e^{j\omega})\right|^2 e^{j\omega n_{0opt}}.\end{aligned} \tag{12}$$

Ignoring the initial delay, (12) means that $\hat{f}(n_0)$ has a symmetric and real spectrum, but has no maximum at the origin. We now demonstrate that using the "Hilbert envelope" of $\hat{f}(n_0)$ can reconstruct the required selection signal. We express a periodic and no-initial-delay MLED signal as

$$f(n_0) = \sum_{k=0}^{(M-1)/2} \mu_k \cos\left(\frac{2\pi kn_0}{M}\right), \tag{13a}$$

and its Hilbert transform as

$$f_H(n_0) = \sum_{k=0}^{(M-1)/2} \mu_k \sin\left(\frac{2\pi kn_0}{M}\right) \tag{13b}$$

for $k \in [0, (M-1)/2]$. The "Hilbert envelope" is

$$g_{He}(n_0) = \left[f^2(n_0) + f_H^2(n_0)\right]^{1/2}. \tag{14}$$

$$\begin{aligned}g_{He}^2(n_0) &= \left[\sum_{k=0}^{(M-1)/2} \mu_k \cos\left(\frac{2\pi kn_0}{M}\right)\right]^2 \\ &\quad + \left[\sum_{k=0}^{(M-1)/2} \mu_k \sin\left(\frac{2\pi kn_0}{M}\right)\right]^2 \\ &= \sum_{j=0}^{(M-1)/2} \sum_{i=0}^{(M-1)/2} \mu_i \mu_j \left[\cos\left(\frac{2\pi in_0}{M}\right)\right. \\ &\quad \left. \cdot \cos\left(\frac{2\pi jn_0}{M}\right) + \sin\left(\frac{2\pi in_0}{M}\right) \sin\left(\frac{2\pi jn_0}{M}\right)\right] \\ &= \sum_{j=0}^{(M-1)/2} \sum_{i=0}^{(M-1)/2} \mu_i \mu_j \cos\left(\frac{2\pi(i-j)n_0}{M}\right).\end{aligned}$$

Extending the definition of μ_k for all k:

$$\mu_k = \mu_{k \oplus M/2} \qquad k \in (-\infty, \infty)$$

where $k \oplus M/2$ denotes the values of k modulo $M/2$

$$\begin{aligned}g_{He}^2(n_0) &= \sum_{i=0}^{(M-1)/2} \sum_{j=i-(M-1)/2}^{i} \mu_i \mu_j \cos\left(\frac{2\pi(i-j)n_0}{M}\right) \\ &= \sum_{m=0}^{(M-1)/2} \left[\sum_{i=0}^{(M-1)/2} \mu_i \mu_{i-m}\right] \cos\left(\frac{2\pi mn_0}{M}\right) \\ &= \sum_{m=0}^{(M-1)/2} c_m \cos\left(\frac{2\pi mn_0}{M}\right).\end{aligned} \tag{15}$$

This equation means the signal $g_{He}(n_0)$ satisfies the two general properties of a selection signal mentioned above because c_m is an autocorrelation sequence. Thus, taking into account the initial delay, the selection signal should be

$$g_{\sqcap}(n_0) = \left[\hat{f}^2(n_0) + \hat{f}_H^2(n_0)\right]^{1/2}. \tag{16}$$

As is well known, the Hilbert transform can be identified as a filter [17] with the transfer function

$$\mathcal{H}(\omega) = \begin{cases} -j & 0 < \omega < \pi \\ 0 & \omega = 0, \pi \\ j & -\pi < \omega < 0 \end{cases} \tag{17a}$$

or discrete-time impulse response

$$h(n) = \begin{cases} \frac{2}{\pi} \frac{\sin^2(\pi n/2)}{n} & n \neq 0 \\ 0 & n = 0. \end{cases} \tag{17b}$$

Therefore, $f_H(n)$ can be obtained from $f(n)$ in the frequency domain by a simple operation or in the time domain by convolution. Because the Hilbert transform can be realized with little computation, this method is efficient. We note that the contrast between the main pulse and secondary pulse in a typical selection signal (Fig. 2) is emphasized compared to the MLED signal. One way to make the selection signal more pulse-like is to use average-value subtraction, which allows obtaining the null signal between pulses:

$$\hat{g}_{\sqcap}(n_0) = \begin{cases} g_{\sqcap}(n_0) - \overline{g_{\sqcap}(n_0)} & \text{if } g_{\sqcap}(n_0) \geq \overline{g_{\sqcap}(n_0)} \\ 0 & \text{if } g_{\sqcap}(n_0) < \overline{g_{\sqcap}(n_0)}, \end{cases}$$

$$\overline{g_{\sqcap}(n_0)} = \frac{\sum_{n_0=0}^{N_f-1} g_{\sqcap}(n_0)}{N_f}, \tag{18}$$

where N_f is the frame size.

Finally, the GCI Determination Signal (GCIDS) is the MLED signal $\hat{f}(n_0)$ multiplied by the selection signal $\hat{g}_{\sqcap}(n_0)$

$$\theta(n_0) = \hat{f}(n_0) \cdot \hat{g}_{\sqcap}(n_0). \tag{19}$$

There are other ways to further improve the performance, e.g., center-clipping the MLED signal before executing the Hilbert transform. However, the present method provides a sufficiently high performance for practical applications.

IV. Description of the System

This section describes the current system of GCI determination as well as the algorithms for period estimation and V/UV/M excitation decisions. Fig. 3 shows the block diagram of the system. The input data comprise a frame of the sampled speech signal. At the beginning, there is a no-delay low-pass filter (NDLPF) to decrease the high-frequency noise. The NDLPF has gradual spectral falloff; in practice, we use a cutoff frequency equal to one-quarter of the sampling frequency, $f_c = f_s/4$ samples/s. The conventional autocorrelation method of linear prediction is applied to get the LP spectral coefficients a_i, using all samples of the input frame. The AR filter box is that of (1), which generates 4 ms of an impulse-response sequence. The cross-correlation box generates a corresponding frame of the MLED signal. In the current system, we use a frequency-domain operation for the Hilbert transform, i.e., take an FFT of the frame of MLED signal, multiply the resulting spectrum by $\mathcal{H}(\omega)$ in (17a) and then take the inverse FFT. The average-value subtraction box executes (18).

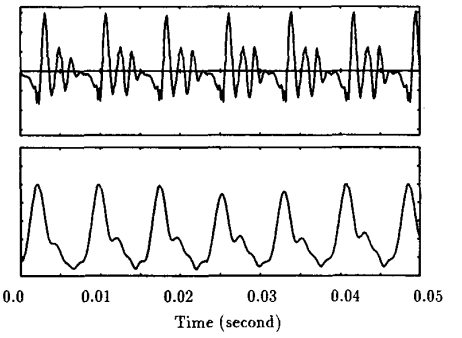


Fig. 2. Typical signals for vowel /ɑ/; top: speech signal, bottom: selection signal. The abscissa is time and the ordinate is amplitude.

Despite the comments of Section II, we prefer, at the output, a GCIDS using the maximum value to point out the GCI's. Thus, we introduce a box of time-shifting adaptation to compensate for the time-difference between the 50 percent amplitude point and the maximum amplitude point. The normalization box adjusts the maximum amplitude within an analysis frame. The GCI decision box is a set of logical relations to find the most probable GCI's. The period-estimation box is a simple logical system to measure the distance between two consecutive pulses in the selection signal because the selection signal is very reliable. If the measurement fails (e.g., an unrealistic pitch period is found), the analysis frame is very likely an unvoiced or silent segment. Experiments show that mixed excitation has a larger pulse width, and a flatter or more sinusoidal wave-like shape in the selection signal than voiced excitation, but the selection signal is always periodic. This fact can be understood in that the spectrum of the selection signal is more concentrated in the low-frequency region, because the harmonic spectrum in a mixed-excitation speech signal occupies only a part of the frequency range.

The flatness-measurement box notes, within a period of the selection signal, the ratio of the number of samples with amplitude above the average value to that with amplitude under the average value, because the next section shows that speech signals of mixed excitation or of high vowels tend to have a flatter selection signal. If this ratio approaches or exceeds 1, the analysis frame is possibly a mixed-excitation signal; if the ratio is close to 0, it is possibly a pure voiced-excitation signal. For unvoiced signals, the selection signal is very weak or null compared to the original signal energy. The zero-crossing rate is useful and efficient in speech signals to distinguish voiced and unvoiced (including mixed-excitation) signals, but it is not efficient to discriminate unvoiced and mixed excitations. The statistical properties of this factor have been

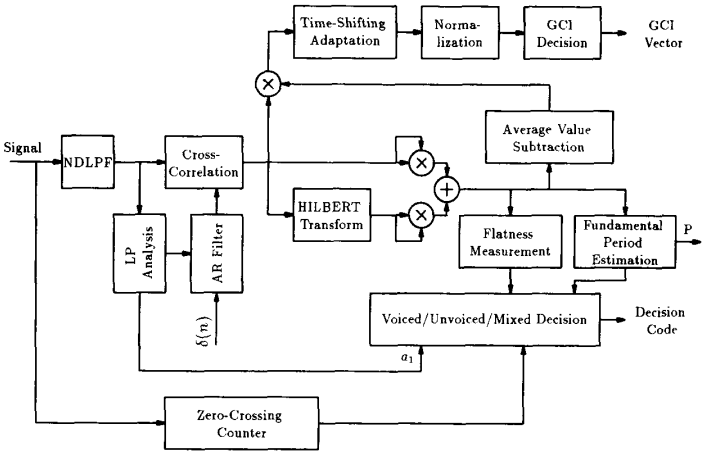


Fig. 3. Block diagram of the GCI determination system.

shown in [18]. This factor is included in our system in the zero-crossing counter box. Another factor which has similar characteristics to the zero-crossing rate is a byproduct of our system—the first LPC coefficient, a_1. Generally, it is positive when the signal is purely voiced, and it is negative when unvoiced or mixed [18].

Finally, a V/UV/M decision box assembles the flatness measurement, the first LPC coefficient, the zero crossing, and the success/failure switch of period estimation to make the final decision through a set of logical relations. Using binary opposition, low/high or success/failure, we express very briefly the relations as follows: a) for a pure-voiced frame, a low zero-crossing rate (note that the first LPC coefficient is only used to emphasize the zero-crossing binary opposition) and a successful period estimation have to be satisfied; b) for a mixed-excitation frame, a high flatness ratio, a high zero-crossing rate, and a successful period estimation have to be satisfied (note that high vowels cannot simultaneously have a high flatness and a high zero-crossing rate; thus they cannot be confused with mixed excitation signals); and c) for an unvoiced frame, a high zero-crossing rate and a period estimation failure have to be satisfied.

V. Performance Study and Results

In this section, we exhibit the results according to different signal types, including noisy signals. Utterances from male speakers were sampled at 10 000 samples/s. An antialiasing filter with a 4.3-kHz cutoff frequency was applied before sampling and the analysis frame length was 256 samples ($N_f = 256$) (this length provided adequate performance). The overlap between successive frames was 56 samples. Twelve coefficients and a rectangular window were used in the LP analysis of the system. The summation interval of the cross correlation of the observed data was 40 samples ($N = 40$).

Fig. 4 shows the results of GCI determination for a synthetic signal produced by exciting a set of formants with an artificial glottal airflow. The GCIDS indicates well the GCI's of the artificial glottal airflow. Thus, we can believe the reliability of our GCI definition.

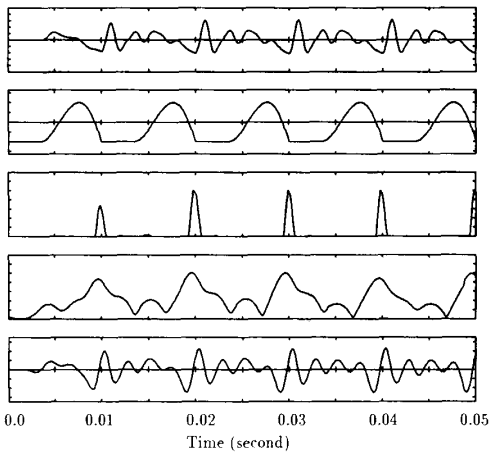


Fig. 4. The results of GCI determination for the synthetic vowel /o/; from top to bottom: the synthetic signal, the artificial glottal airflow, the GCI determination signal (GCIDS), the selection signal, and the MLED signal. The abscissa is time and the ordinate is amplitude.

Fig. 5 shows a segment of the vowel /a/, its GCIDS, the selection signal (without average subtraction), and the MLED signal. The MLED signal has a positive main pulse and a comparable positive subpulse per period. Its waveform is similar to the speech signal, and has a positive main pulse indicating the minimum of the speech signal, where we think a GCI occurs. The selection signal is positive and has a larger pulse riding on the GCI per period. The strength of the subpulse is very weak compared to that of the main pulse. After the average subtraction, the selection signal has one pulse per period. Using this signal to weight the MLED signal gives the GCIDS, which reliably locates the GCI's. According to (13) and (15), the MLED and selection signals should be symmetric about their maxima, but this is not the case in the figure. The reason is that the spectra of the MLED and se-

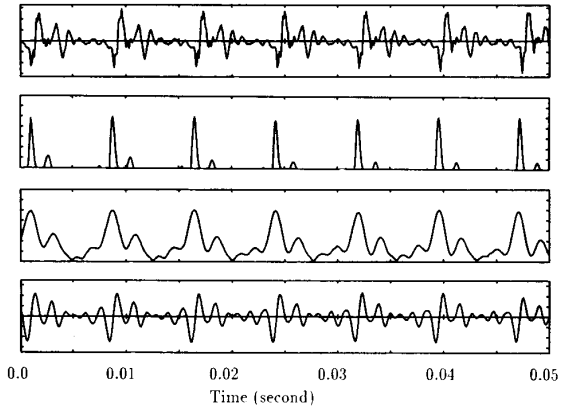


Fig. 5. The results of GCI determination for the synthetic vowel /ɑ/; from top to bottom: speech, the GCI determination signal (GCIDS), the selection signal, and the MLED signal. The abscissa is time and the ordinate is amplitude.

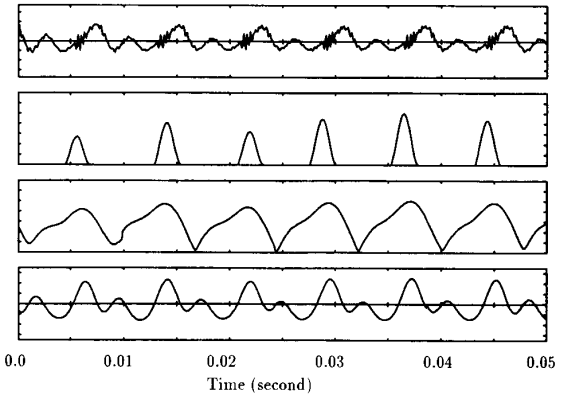


Fig. 6. The result of GCI determination for vowel /i/. The order of panels and the axes are the same as in Fig. 5.

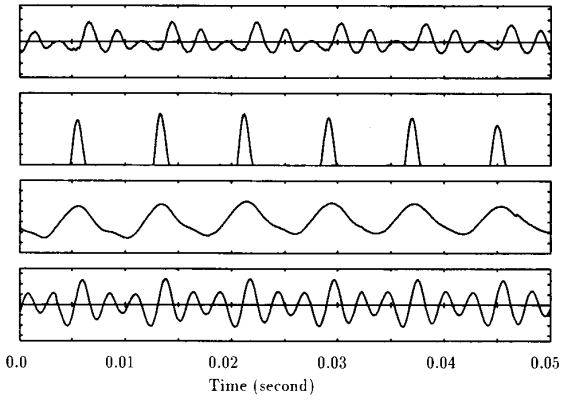


Fig. 7. The result of GCI determination for vowel /u/. The order of panels and the axes are the same as in Fig. 5.

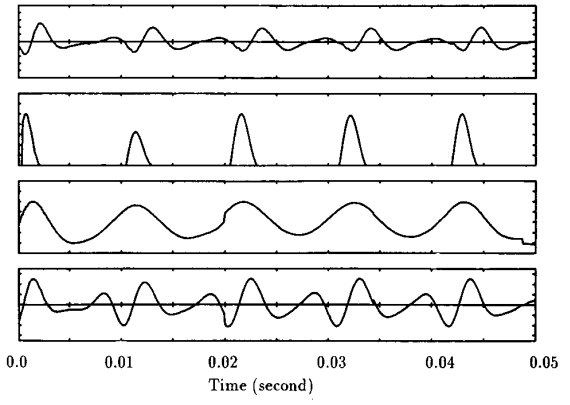


Fig. 8. The result of GCI determination for vowel /m/. The order of panels and the axes are the same as in Fig. 5.

lection signals are not real and symmetric because speech does not exactly follow an all-pole model; thus, the assumption in (8) is, in fact, an approximation. However, the introduced error has little influence on the GCI detection.

The performance for the vowel /i/ is noted in Fig. 6. As above, the GCI's are accurately and reliably found. One difference compared to Fig. 5 is that the width of all of the pulses is larger, especially for the selection signal. This phenomenon is found for all high vowels. We believe that it is due to the low first-formant frequency for these vowels. In Fig. 6, one can see a slight time-shifting between the maximum of the GCIDS and that of the MLED signal. As noted earlier, this time-shifting is to adapt to the 50 percent amplitude criterion as a GCI cue.

For the vowel /u/ (Fig. 7), almost all previous algorithms based on speech-signal prediction error in the past several years [3], [5] fail. The reason can be attributed to a high concentration of energy at low frequency, which results in most of the prediction error at low frequency. In contrast, the efficient epoch information is at high frequency [4]. The performance of the present algorithm does not suffer from this concentration of energy, and reliable results are obtained. From the performance for this vowel, the efficiency of the selection signal, which effectively eliminates the subpulses of the MLED signal, is well demonstrated. For the nasal consonant /m/ (Fig. 8), the performance is not reduced by nasal-tract coupling.

Voiced fricatives combine both voice and noise sources as excitation in human speech production. Because the voiced signal for this class may be masked by the noise, the theoretical basis for epoch detection, i.e., signal discontinuity, is invalid. Moreover, for these phonemes, the glottis may never close or only close incompletely. Thus, GCI's are not physically manifest for voiced fricatives, and their detection is more difficult. We found no algorithms in the literature which report success for GCI detection of this class. In our algorithm, the GCI definition can be interpreted as a time of waveform discontinuity and also as a time when most energy is injected within a period. This definition avoids the obstacles imposed by voiced fricatives. Figs. 9 and 10 show results for the voiced palato-alveolar fricative /ʒ/ and the voiced labiodental fricative /v/, respectively. It is evident that the presence of the noise source does not prevent reliable GCI detection. Like the high vowels, the voiced fricatives have

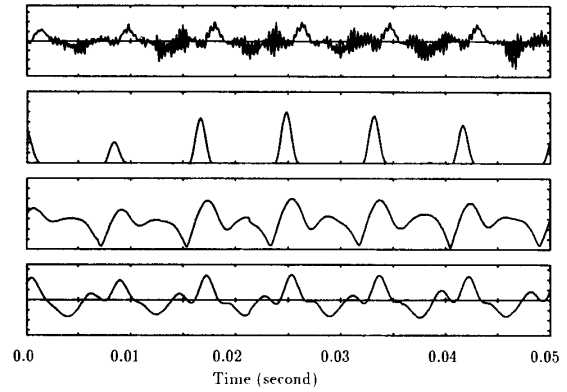


Fig. 9. The results of GCI determination for the voiced palato-alveolar fricative /ʒ/. The order of panels and the axes are the same as in Fig. 5.

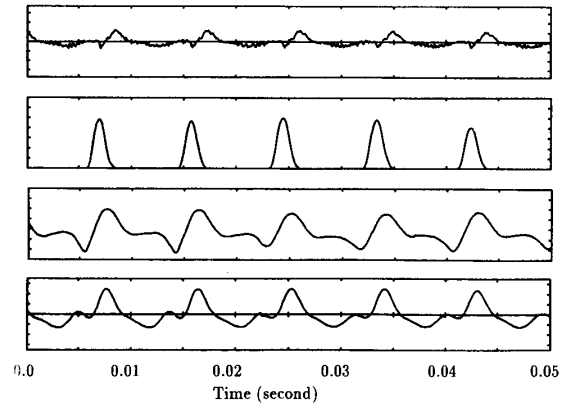


Fig. 10. The result of GCI determination for the voiced labiodental fricative /v/. The order of panels and the axes are the same as in Fig. 5.

larger-width pulses in both the MLED and selection signals. In the original speech signal, it is difficult to find the GCI's visually because of the incomplete glottal closure. This, however, seems to have no influence on the GCIDS.

Voiced plosives have a rapid onset of glottal vibration and incorporate a sudden change in vocal-tract properties. Theoretically, the abrupt change should degrade the performance of the MLED. Fig. 11 shows the performance on a syllable /dɔ/. The GCIDS finds, at the beginning, an irregular pulse with an abnormal period length (about 12.5 ms), then the more regular pulses and more uniform periods. Visually, it is difficult to estimate voicing at the beginning of the syllable. Experiments on computer glottis simulation, e.g., [19], have shown that the glottal-vibration onset always has a very long first period which is almost independent of the articulatory commands. This phenomenon is often denoted as part of the micro-melody. Thus, we can believe the reliability of the GCI detection. The performance degradation due to the abrupt change of vocal-tract properties can be understood as a reduction of the absolute amplitude of the GCIDS pulses. However, the local maximum of each pulse remains the same. After normalization, the absolute amplitude reduction is not visible.

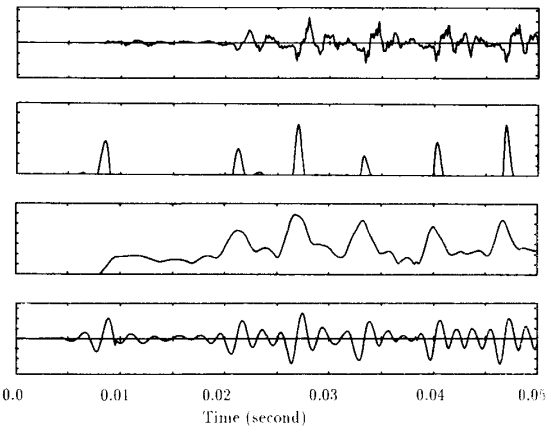


Fig. 11. The result of GCI determination for the syllable /dɔ/. The order of panels and the axes are the same as in Fig. 5.

Consider now the problem of a speech signal contaminated by strong white noise. We have already noted that, theoretically, this kind of noise has little influence on GCI detection when N is large enough. However, a few earlier assumptions may be invalid when the contamination noise increases significantly. Thus, it is worthwhile to determine practically the noise tolerance of the algorithm. In this experiment, we added white noise to a diphone /ɑj/. Four typical cases, where the ratio of signal-to-noise energy (SNR) is equal to ∞ (clean data), 30, 10, and 0 dB, are examined. Fig. 12 shows the performance for the noise-contaminated signals. For the clean, 30, and 10 dB SNR signals, the corresponding GCIDS have not been evidently degraded. For the 0 dB SNR signal, there are a few pulses in its GCIDS where amplitudes are reduced, which nonetheless hold the local maximum. Therefore, the ability to determine the GCI's is not reduced with strong noise.

In addition, this experiment tested the ability of our algorithm to follow nonstationary pitch. The diphone was spoken with a gradual decrease of fundamental frequency. For all of the signals, the gradual variation is tracked well.

Suppose that the input speech signal is distorted in amplitude or phase by a deterministic system during speech recording, transmission, etc. For amplitude distortion, normalizing the GCIDS amplitude can remove its influence. It is difficult to find a method that can generally remove phase distortion, because such distortion is highly variable and cannot be easily described by one model. However, one class of phase distortion, which can be described by an all-pass pole-zero model, is possibly diminished by our method. The phase modeling, which is successfully applied to speech-signal phase compensation [20], is

$$T(z) = \frac{B(1/z)}{B(z)},$$

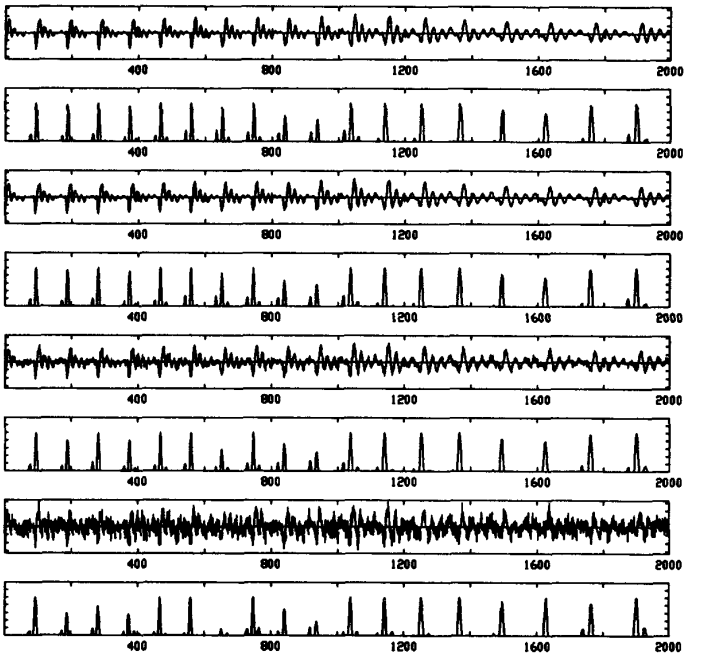


Fig. 12. The results of the contaminant-noise influence test. From top to bottom, the first and second panels are the clean speech and its GCIDS, respectively; the third and fourth ones are the 30-dB SNR speech and its GCIDS, respectively; the fifth and sixth ones are the 10-dB SNR speech and its GCIDs, respectively; the seventh and the eighth ones are the 0-dB speech and its GCIDS, respectively. The abscissa is time and the ordinate is amplitude.

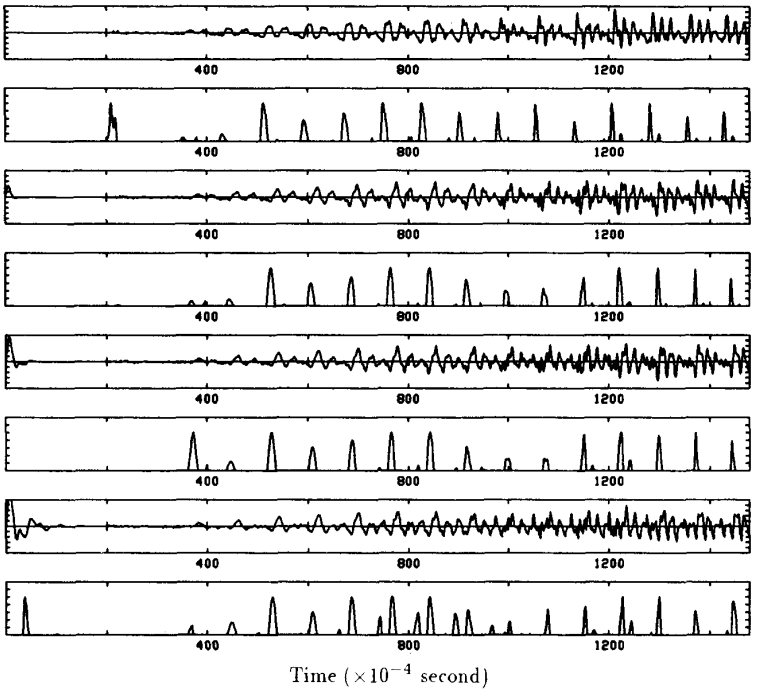


Fig. 13. The results for a phase-distortion signal. The speech segment is extracted from the beginning of the word "brought." From top to bottom, the first and second panels are the nondistorted speech and its GCIDS, respectively; the third and fourth ones are the distorted signal with BW = 400 Hz (identical bandwidth) in $T(z)$ and its GCIDS, respectively; the fifth and sixth ones are the distorted signal with BW = 200 Hz and its GCIDS, respectively; the seventh and the eighth ones are the distorted signal with BW = 100 Hz and its GCIDS, respectively. The abscissa is time and the ordinate is amplitude.

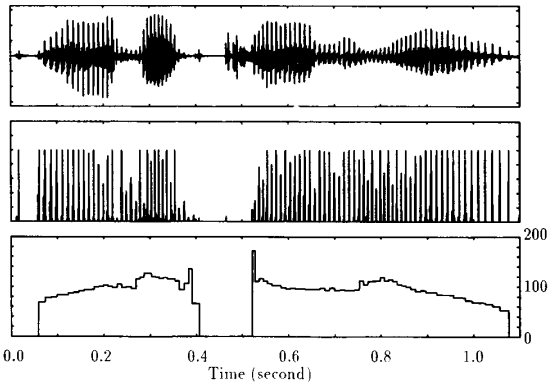


Fig. 14. The results of period estimation. From top to bottom: the speech signal (a sentence: "Robert came here"); the GCIDS, and the F0 contour. The abscissa is time. The ordinate for the first and second panels is amplitude; that for the third panel is frequency in hertz.

where $B(z)$ is a polynomial with all roots within the unit circle. We randomly placed five pairs of conjugate roots to construct $B(z)$ and $T(z)$. Fig. 13 shows a typical result with the five root pairs at frequencies 200, 500, 1000, 1700, and 2400 Hz, and with an identical bandwidth for all roots. The GCI's of the distorted signal are generally evident. However, when the bandwidth is below 100 Hz, the strength of secondary pulses for some vowels becomes important, and these pulses can be removed by tracking the period variation. That our GCI detection method can resist this class of phase distortion can be attributed to a number of supplementary poles in the matched filter, which roughly models the phase distortion.

Finally, Fig. 14 shows the result of period estimation based on GCI determination with a sentence ("Robert came here") from a male speaker. The pitch contour was not smoothed. The nonstationary pitch variation and even a sudden pitch jump can be demonstrated fairly well.

VI. Conclusion

In this paper, we present an automatic and reliable glottal closure instant determination algorithm. As a byproduct, nonstationary fundamental period estimation is achieved. The essential computation includes a twelve-pole speech linear-prediction analysis, a cross correlation, and both direct and inverse FFT's (or a convolution). This cost is not thought to be high. For normal speech, the high performance of the algorithm holds for a wide variety of speech signals. In addition, we find an important performance resistance to bad speech conditions or distorted signals. The algorithm may fail if the distortion is due to echos or reflections, because an echo or reflection contains other GCI's, which mask the GCI's of the true signal.

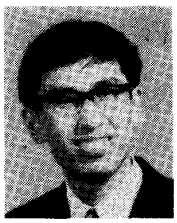
The importance of this work is to provide a powerful tool to support pitch-synchronized analysis, which has recently been largely accepted as a methodology to give more precision and reliability. Previously, pitch alignment was manually or semiautomatically executed. Even some auxiliary instruments (e.g., electroglottograph, optoglottograph, etc.) have been employed to aid the alignment. This inconvenience restricts pitch-synchronized analysis to laboratory applications. Our algorithm facilitates removing this inconvenience.

Acknowledgment

One of the authors wishes to thank INRS-Telecommunications for the award of a postdoctoral fellowship and for the remarkable material support to realize this work.

References

[1] P. Hedelin, "A glottal LPC-vocoder," in *Proc. Int. Conf. IEEE ASSP*, San Diego, CA, 1984.
[2] W. Hess, *Pitch Detemination of Speech Signals*. Berlin: Springer-Verlag, 1983.
[3] H. W. Strube, "Determination of the instant of glottal closure from the speech wave," *J. Acoust. Soc. Amer.*, vol. 56, no. 5, pp. 1625–1629, 1974.
[4] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 562–570, 1975.
[5] ——, "Epoch extraction from linear prediction residual for identification and closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 309–319, 1979.
[6] J. D. Wise, J. R. Caprio, and T. W. Parks, "Maximum likelihood pitch estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 418–423, 1976.
[7] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 353–362, 1974.
[8] M. M. Sondhi, "New methods of pitch extraction," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 262–266, June 1968.
[9] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, no. 2, pp. 293–309, 1970.
[10] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367–377, Dec. 1972.
[11] P. Martin, "Détection de F_0 par Intercorrelation avec un Fonction Peigne," *Journées d'Etude sur Parole*, pp. 221–232, 1981.
[12] C. W. Helstrom, *Statistical Theory of Signal Detection*. New York: McGraw-Hill, 1960.
[13] T. Y. Young, "Epoch detection—A method for resolving overlapping signals," *Bell Syst. Tech. J.*, vol. 44, pp. 401–426, 1965.
[14] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *Electron. Commun.*, vol. 53-A, pp. 36–43, 1970.
[15] R. S. Lipster and A. N. Shiryayev, *Statistics of Random Processes II—Applications*. New York: Springer-Verlag, 1978.
[16] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. Berlin, Germany: Springer-Verlag, 1976.
[17] R. Ansari, "IIR discrete-time Hilbert transformers," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 1116–1119, 1987.
[18] B. S. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with application to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 201–212, 1976.
[19] K. Ishizaka and J. Flanagan, "Synthesis of voiced sounds from a two-mass model of the voice cords," *Bell Syst. Tech. J.*, vol. 51, pp. 1233–1268, 1972.
[20] P. Hedelin, "Phase compensation in all-pole speech analysis," in *Proc. Int. Conf. IEEE ASSP*, New York, 1988.



Yan Ming Cheng (M'87) was born in Jiang Su, China, on July 1, 1957. He received the B.Sc. degree in electrical engineering in 1982 from the Nanjing Institute of Technology (now called Southeast University), Nanjing, China, and the D.E.A. and the Ph.D. degrees from the Institute of Speech Communication of E.N.S.E.R.G.-Institut National Polytechnique de Grenoble, France, in 1983 and 1986, respectively.

In 1987 he joined the Center of Telecommunications of the Institut National de la Recherche Scientifique (INRS), University of Quebec, Canada, where he was a Postdoctoral Researcher. He is currently a Research Associate at INRS-Telecommunications. His research interests are in the broad areas of speech signal processing, analysis, synthesis, recognition, and very-low-bit coding.

Dr. Cheng received the 1986 CNRS (Centre National de Recherche Scientifique) thesis award in France.

Douglas O'Shaughnessy (S'74-M'76) received the B.Sc. and M.Sc. degrees in 1972, and the Ph.D. degree in 1976 from the Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science. He also received the B.Sc. degree in mathematics from M.I.T. in 1972.

After a brief postdoctoral position at M.I.T., he joined INRS-Telecommunications, a research institute affiliated with the University of Quebec, where he is now a full Professor. As an auxiliary Professor, he also teaches courses in the Electrical Engineering Department at McGill University. He has worked on English and French synthesis-by-rule and modeling of intonation. His main interests lie in speech synthesis, coding, and recognition. He is the author of a text, *Speech Communications* (Reading, MA: Addison-Wesley, June 1987).