[8] L. G. Roberts, "Multiple computer networks and inter-computer communications," presented at Ass. Comput. Mach. Operating Systems Principles, Gatlinburg, TN, Oct. 1967.

[9] J. W. Smith, "Determination of path lengths in a distributed network," RAND Corp., memo. RM-3578-PR, Aug. 1964.

[10] E. C. Wolf, "An advanced computer communication network," in *Proc. AIAA Comput. Network Syst. Conf.*, 1973.

Raymond L. Pickholtz (S'54-A'55-M'60) was born in New York, NY, on April 12, 1932. He received the B.E.E. and M.E.E. degrees from the City College of New York, New York, NY, in 1954 and 1958, respectively, and the Ph.D. degree from the Polytechnic Institute of Brooklyn, Brooklyn, NY, in 1966.

From 1954 to 1957, he was a Research Engineer at RCA Laboratories engaged in work on color television. From 1957 to 1961 he was a member of the technical staff of ITT Laboratories, Nutley, NJ, where he worked on guidance and communications problems. He was on the faculty of the Polytechnic Institute of Brooklyn from 1962 to 1972. He is presently Professor of Electrical Engineering and Computer Science at the George Washington University, Washington, DC. He is an Associate Editor of the *IEEE Transactions on Communications* and he has been a Consultant to IBM Research, Fairchild Industries, and Computer Sciences Corporation. He was chairman of the New York Chapter of the IEEE Information Theory Society in 1971-1972. He was financial chairman of the 1973 International Symposium on Information Theory. He is a member of the executive committee of the Washington, DC section of the IEEE. He was chairman of the Third Data Communications Symposium and his present research interests are in data communications and computer networks.

Dr. Pickholtz is a member of Eta Kappa Nu and Sigma Xi.

Caldwell McCoy, Jr. (M'65) received the B.S.E.E. degree from University of Connecticut, Storrs, in 1956. He received the M.S.E.E. and D.Sc.E.E. degrees from George Washington University, Washington, DC, in 1968 and 1975, respectively.

From 1956 to 1959 he was on active duty with the U.S. Air Force as a rated flyer in the Strategic Air Command. In 1959, he joined the Acoustics Division of the Naval Research Laboratory, Washington, DC. From 1959 to the present he has worked in design, testing, and evaluation of underwater detecting and processing systems for long range detection of submarines. The investigations deal with beam-forming, signal acquisition, tracking, and statistical analysis on underwater acoustic signals. Large scale data systems and computer networks are used extensively and resulted in Dr. McCoy doing research in the area of adaptive routing techniques for packet-switched networks. He authored and coauthored over 20 papers and publications.

Dr. McCoy is a member of the Acoustical Society of America, Toastmasters International, Sigma Xi, and Beta Sigma Gamma. He is an Advisor on Committee for Underwater Acoustic Communications, and Committee for Signal Processing Research at NRL. He was awarded a Thomas Edison Fellowship, and is listed in the Marquis Directory of "Who's Who in the East."

Timing Recovery in Digital Synchronous Data Receivers

KURT H. MUELLER AND MARKUS MÜLLER

Abstract—A new class of fast-converging timing recovery methods for synchronous digital data receivers is investigated. Starting with a worst-case timing offset, convergence with random binary data will typically occur within 10-20 symbols. The input signal is sampled at the baud rate; these samples are then processed to derive a suitable control signal to adjust the timing phase. A general method is outlined to obtain near-minimum-variance estimates of the timing offset with respect to a given steady-state sampling criterion. Although we make certain independence assumptions between successive samples and postulate ideal decisions to obtain convenient analytical results, our simulations with a decision-directed reference and baud-to-baud adjustments yield very similar results. Convergence is exponential, and for small loop gains the residual jitter is proportional and convergence time is inversely proportional to the loop gain. The proposed algorithms are simple and economic to implement. They apply to binary or multilevel PAM signals as well as to partial response signals.

I. INTRODUCTION

SYMBOL synchronization or timing recovery is one of the most critical receiver functions in synchronous communication systems. The receiver clock must be continuously adjusted in its frequency and phase to optimize the sampling instants of the received data signal and to compensate for frequency drifts between the oscillators used in the transmitter and receiver clock circuits. For binary or multilevel PAM signals, several timing recovery methods are known [1]-[9]. The timing information is usually derived from the data signal itself and based on some meaningful optimization criterion which determines the steady-state location of the timing instants. A crude

Paper approved by the Associate Editor for Data Communication Systems of the IEEE Communications Society for publication after presentation at the 1974 Zurich Seminar for Digital Communication, Zurich, Switzerland. Manuscript received October 9, 1975; revised December 11, 1975.

K. H. Mueller is with Bell Laboratories, Holmdel, NJ 07733.

M. Müller was with the Swiss Federal Institute of Technology, Zurich, Switzerland. He is now with the Overseas Department, General Radio Company, Zurich, Switzerland.

distinction can be made between three different kinds of methods.

Class A: The threshold crossings of the received baseband data signal (at zero if the signal is binary, or halfway between the reference levels if the signal is multilevel) are compared with the sampling phase. A correction of the sampling phase is initiated as a result of this comparison. The mean location of the crossings is estimated and the optimum sampling instant and maximum eye opening are assumed to be halfway between these crossings.

Class B: This method uses the signal derivative at the sampling instants. This derivative, or at least its sign, is usually correlated with the estimated data to produce the updating information required for the timing control loop. The resulting sampling phase is such that the mean square error between the signal and the appropriate reference levels is minimized, or, with slight changes, such that sampling will occur at the peak of the impulse response.

Systems of class A and class B have been investigated by Saltzberg [4]; timing recovery systems of class B have been described by Chang [5], Gitlin and Salz [6], and Kobayashi [7]. Both schemes can be used with a variety of algorithms within the control loop: adjustments can be made in increments that are error proportional or fixed, averaging may or may not be used prior to adjustments, dead zones can be introduced, and different parameters can be used during the initial training mode and during the subsequent tracking mode. Both types of systems operate on the baseband signal.

Class C: A spectral line at the clock frequency (or at a multiple of this frequency) is filtered out with a narrowband loop. Since such lines are not ordinarily encountered in bandwidth-efficient systems, some nonlinear processing of the signal is used to generate such lines. An early proposal of such a scheme is due to Bennett [8]. Square law devices have been investigated by Takasaki [9] and Franks and Bubrouski [10]. An advantage of these systems is their ability to work with either the baseband or the passband signal. However, performance with narrow-band near-Nyquist-limited systems is usually marginal since the recovered timing waveform amplitude and the SNR depend on the system's excess bandwidth. Timing recovery systems of the class C type are often used in PCM-repeaters because of their comparatively simple implementation.

The best timing phase for a given system will depend on the overall impulse response and thus on the characteristics of the communication channel. This is not only because of the unknown delay which is introduced by the channel. The main problems are caused by noise and linear distortion (intersymbol interference); these disturbances can severely limit the performance of a timing recovery loop. In some investigations [11], [12] an impulse response which is limited to one signaling interval is assumed; this seems unrealistic for bandlimited channels. In such channels, particularly if they are near-Nyquist-limited, the level transitions will be distributed over a large part of the signaling interval. Even in the absence of noise and distortion, timing information using threshold crossing methods (class A) can thus be obtained reliably only by averaging over a large number of transitions. This is not a serious drawback during steady-state tracking, but it tends to increase the initial training time. Similar considerations apply for methods based on classes B and C.

The mentioned timing recovery systems are usually implemented after several signal processing operations have taken place: the received signal is filtered, demodulated, filtered again, and probably passed through an automatic equalizer. For most of these operations, analog signal processing has been used and still is in use today. The signal that is needed to derive timing information is thus usually a continuous signal in both time and amplitude. Threshold crossing information or the derivative are easily obtained to realize a particular timing loop.

In the current trend towards fully digital receivers using medium- and large-scale integration (MSI and LSI) technology, we are confronted with a somewhat different situation. The signal in such a receiver is sampled and A/D converted at the input. It is available only at discrete time intervals for further processing. Basically, sampling could be performed at a high enough rate to allow a complete reconstruction of the signal. Analog timing recovery schemes could then be "digitized" and would still perform in a functionally equivalent way. However, such an approach is often a complex and expensive solution which leaves much to be desired. The high A/D conversion rate that is needed with such a scheme may be another serious drawback. Furthermore, it is desirable to sample in synchronism with the baud rate and many systems actually use only one sample per baud interval for signal processing; for example in digital automatic equalizers [13] or in digital demodulators [14]. Such a low sampling rate is justified because the final decisions at the output are also based on samples taken at the baud rate and the behavior of the data signal between the sampling instants is immaterial. Note that baud sampling will not permit an exact signal reconstruction by interpolation techniques, except in the case of a baseband signal which is strictly limited to the Nyquist frequency (i.e., half the signaling rate). The information required by all previously mentioned timing recovery schemes will just not be available with baud sampling. The use of higher sampling rates or additonal sampling of the signal derivative for timing recovery reasons does not seem to be an appealing solution. Obviously, a new approach is needed.

Another difference between analog and digital processing is depicted in Fig. 1 for the simple case of a synchronous baseband data receiver. In the analog version the signal is sampled after processing and conditioning, whereas in the digital receiver the unconditioned and probably distorted signal is sampled. Timing recovery from the digitized signal must always be achieved with a feedback loop, in contrast to analog processing where nonfeedback schemes are possible.

In this paper, we investigate some new methods suitable for timing recovery in digital synchronous data receivers. Sampling is assumed at the baud rate. Updating information for the control loop is derived from these samples and the estimated data values in a simple and straightforward way without the

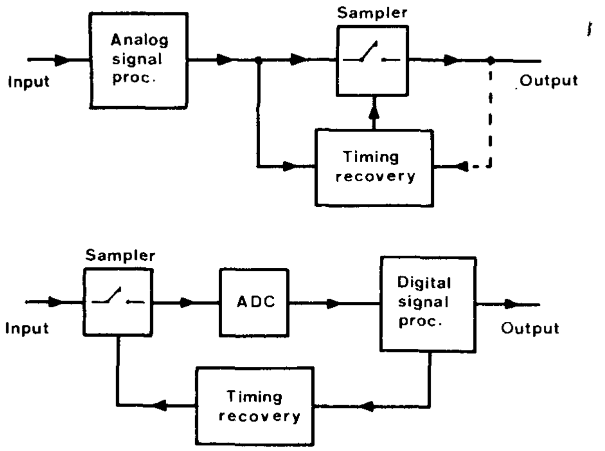


Fig. 1. Block diagram of timing recovery loop for baseband data receiver. (a) Analog signal processing. (b) Digital signal processing.

need of any further signal information. Our discussion will be limited to baseband signaling. This is justified because the more important linear modulation methods allow the concept of an equivalent baseband channel for system modeling [15]. We realize, however, that in a general data receiver, the adaptive loops for timing recovery, carrier phase control, and automatic equalization do not work independently of each other when jointly operated, and interaction must be carefully investigated. Such aspects have been studied for analog timing loops using signal differentiation [5], [7]. In the present paper, we will concentrate on timing recovery alone, but joint operation with other receiver parameters is under investigation and will be reported in a future publication.

After a short review of the timing problem we will outline how timing information can be derived from the samples of the impulse response. Since these samples are not available during transmission, a technique will be presented to obtain estimates directly from the signal samples. We will derive a bound for the minimum variance of these estimates and show how suboptimum estimates, suitable for simple implementation, can be obtained, which are close to this minimum. The method will be illustrated with some practical examples. As a next step, the convergence behavior of a timing control loop that uses these estimates in a stochastic adjustment algorithm will be studied. Finally, we will present several computer simulations that confirm the fast convergence properties, even with a decision-directed start-up.

II. REVIEW OF THE TIMING RECOVERY PROBLEM

Let us consider a synchronous baseband data transmission system with an overall impulse response $h(t)$; its output can then be described as

$$x(t) = \sum_k a_k h(t - kT) + n(t) \tag{1}$$

where $n(t)$ represents some additive Gaussian noise. The a_k's are data symbols chosen with equal probability and independently from previous symbols from a set of L equidistant values. Assume now, that the signal is sampled at instants $t = \tau + mT$; then

$$x(\tau + mT) = h(\tau)\left[a_m + \frac{1}{h(\tau)} {\sum_i}' a_{m-i} h(\tau + iT) + \frac{n(\tau + mT)}{h(\tau)}\right]. \tag{2}$$

The term $h(\tau)$ represents a gain factor which depends both on the overall system attenuation and the sampling phase τ. Within the brackets, two terms appear in addition to the desired data value a_m: The first one[1] is caused by intersymbol interference; it disappears if $h(\tau + iT) = \delta_{io}$, i.e., if the impulse response satisfies the Nyquist criterion. Since all echoes $h(\tau + iT)$ are functions of τ, it is clear that the intersymbol interference is heavily influenced by the choice of the sampling phase. The remaining term is due to additive noise which is assumed to be a stationary, zero mean random process. The sampling phase should ideally be chosen in such a way as to minimize error probability; but for practical implementation more convenient suboptimal criteria are preferable, such as sampling at the maximum eye opening (minimum intersymbol interference), or minimizing the mean square error. It is well known that, for these two objectives, the peak distortion D and the mean square distortion ϵ are appropriate quality measures [15] defined by

$$D(\tau) = \frac{1}{h_0} {\sum_i}' |h_i| \tag{3}$$

$$\epsilon(\tau) = \frac{1}{{h_0}^2} {\sum_i}' {h_i}^2 \tag{4}$$

where we have used the short notation $h_i = h(\tau + iT)$ for convenience. A channel is distortion free if a particular phase τ_0 exists such that $D(\tau_0) = \epsilon(\tau_0) = 0$. Whether the channel is distortion free or not, the usual objective is to find a phase τ that minimizes one of the performance measures (3) or (4). One obvious approach is to compute the partial derivative of the performance measure with respect to τ and make proportional timing updates in the opposite direction. Such steepest descent gradient algorithms will stop adjusting once the desired optimum phase is reached. Note that, instead of the mentioned derivative, any other (monotonic) function of τ could be used, provided it has the same root, or at least one that is close. This fact is used in the threshold crossing schemes discussed earlier. It also points the way for solving our problem at hand: all that is required is a timing function $f(\tau)$ that can be efficiently computed from baud-spaced signal samples and whose root is close to the minimum of a reasonably chosen performance measure. This will be done in two steps: we will first derive our timing function from the impulse response, and then, in a second step, show how this timing function (or estimates of it) can be derived from the signal samples [16].

[1] A prime on a summation indicates deletion of the zeroth term.

III. RELATING TIMING TO THE IMPULSE RESPONSE

In the present study we will limit our investigations to linear combinations of the samples of the impulse response,[2] i.e., to timing functions of the type

$$f(\tau) = \sum_i u_i h_i = u^T h. \tag{5}$$

The coefficients u_i are dimensionless, and for normalization, we will assume that the received signal power is unity. Note also that we will only be concerned with reasonably band-limited systems ($f_{\max} \leqslant 1/T$) whose impulse response oscillates over several signaling intervals.

The timing function (5) will determine the transfer characteristic of the control loop, and the resulting steady-state timing phase will be the one for which $f(\tau) = 0$. In the case of ideal Nyquist signaling this should of course be $\tau = 0$. For this example, since $h(t)$ is even, it is also easy to see that the constraint

$$u_0 = 0, \qquad u_i = -u_{-i}, \qquad \text{for } i \neq 0 \tag{6}$$

will define a class of transfer characteristics which have odd symmetry around the origin. This is preferable since it guarantees that offsets of both polarities are handled symmetrically. To achieve the same effect with an odd symmetry impulse response (e.g., Class IV or bipolar partial response [17]) the set of weights u_i would be chosen with even symmetry [18]. The combination of (5) and (6) requires that, for Nyquist signaling, $f(\tau)$ be of the form

$$f(\tau) = \sum_{i=1}^{L} u_i(h_i - h_{-i}) \tag{7}$$

in order to yield (in the absence of distortion) an odd symmetry detector characteristic. Note that, in the approach just outlined, timing information is derived from the symmetry error of the sampled impulse response.

From the large class of possible timing functions we will pick out two particular ones for a more detailed discussion, namely

Type A:

$$f(\tau) = \tfrac{1}{2}(h_1 - h_{-1}) = \tfrac{1}{2}\,[h(\tau + T) - h(\tau - T)]. \tag{8}$$

Type B:

$$f(\tau) = h_1 = h(\tau + T). \tag{9}$$

Here, type A refers obviously to a first-order symmetry error, i.e., $L = 1$ and $u_1 = 1/2$ in (7). A class of algorithms related to type A has first been proposed by Mueller and Spaulding [19]. The zero forcing of the first trailing echo h_1 was been mentioned by Lucky, Salz, and Weldon [15], and some preliminary investigations regarding both functions have been reported by the authors of this paper [16]. Although the above timing functions represent a somewhat subjective choice, they are nevertheless probably the most basic and simple functions satisfying all the requirements discussed earlier. Many of the problems that have to be studied with other timing functions will be highlighted in the discussion of the two examples (8) and (9).

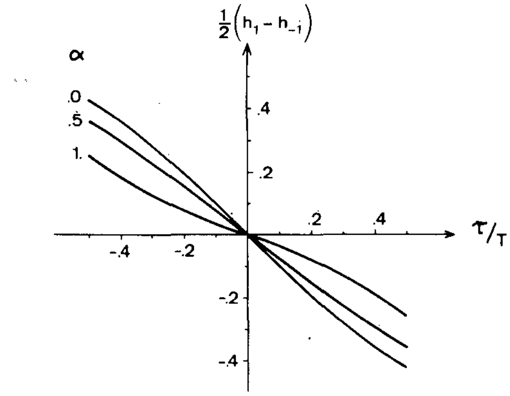


Fig. 2. Phase detector characteristics of a loop which forces $h_1 = h_{-1}$ (Nyquist channel with cosine rolloff α).

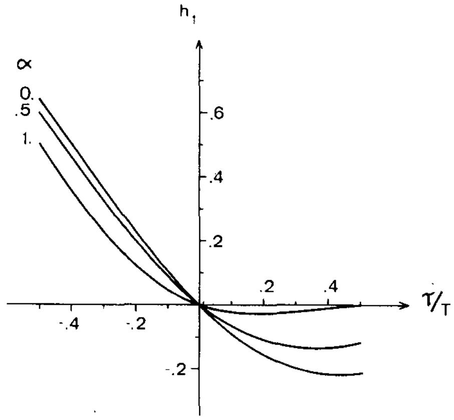


Fig. 3. Phase detector characteristics of a loop which forces $h_1 = 0$ (Nyquist channel with rolloff α).

The timing function type A is plotted in Fig. 2 for a Nyquist pulse with various rolloffs α. Note the excellent linearity around zero. The slope at zero which defines the phase detector gain constant is only mildly affected by the bandwidth (it decreases by a factor of two if α varies from 0 to 0.8). For α = 0 we have

$$f(\tau) = -\frac{1}{\pi}\frac{\sin(\pi\tau/T)}{(\tau/T)^2 - 1}, \tag{10}$$

and it can be shown that this holds also for first-order symmetry errors in duobinary or bipolar signaling formats. The inclusion of more distant echos ($L > 1$) would yield higher order error functions which we will, however, not discuss at this time.

The corresponding functions for scheme B are shown in Fig. 3; they do not look nearly as nice, and because of the strongly asymmetrical transfer behavior we would intuitively judge the stability of a control loop based on this timing information as rather poor, especially for larger amounts of

[2] The impulse response can be estimated from the sampled data signal; see Section IV.

excess bandwidth. However, we will find some advantages in a moment when we discuss operation in the presence of distortion. For zero excess bandwidth, both schemes yield the same slope at zero, i.e., for a small offset they will provide an identical correction. The resulting sampling instants for a distorted pulse are shown in Fig. 4. Algorithms based on A will choose their steady-state timing in such a way as to yield equal echoes h_1 and h_{-1}. Algorithms based on B will take the first zero crossing after the main pulse as their timing reference. Note that, for distorted signals, the two approaches will generally give different sampling phases. The sampling instants provided by scheme A will always be optimally located with even impulse responses; this means that both $D(\tau)$ and $\epsilon(\tau)$ are minimized in the presence of amplitude distortion alone. For channels where phase distortion is the main impairment, scheme B may be better. This is illustrated in Fig. 5 for a Nyquist system with cosine rolloff α that has been degraded by quadratic delay distortion.[3] For tight rolloff, both methods provide timing phases that result in a larger distortion than the minimum that could be achieved. For rolloffs above approximately $\alpha = 0.3$, scheme B coincides with the minimum; indeed for severe delay distortion it is always superior to the symmetrical scheme A. For small delays the difference is not so significant, particularly for rolloffs $\alpha \geqslant 0.4$. When the main channel impairment consists of rising delay distortion, the linearity of the transfer characteristic is improved in scheme B and degraded in scheme A.

During actual data transmission the sampled impulse response is not directly available to determine $f(\tau)$. In the next section we will therefore show how a low variance estimate z_k whose expected value equals $f(\tau)$ can be obtained directly from the signal samples.

IV. EXTRACTION OF TIMING INFORMATION

Due to the linear character of (1) and (5) it makes sense to assume a linear relationship for z_k in the form

$$z_k = g_k{}^T x_k \tag{11}$$

or alternatively

$$z_k = g_k{}^T e_k = g_k{}^T (x_k - h_0 a_k) \tag{12}$$

where $E\{z_k\} = f(\tau)$, and

$$x_k{}^T = (x_{k-m+1}, x_{k-m+2}, \cdots, x_k) \tag{13}$$

is the signal vector at $t = \tau + kT$ containing the last m input samples, a_k is the corresponding data vector

$$a_k{}^T = (a_{k-m+1}, a_{k-m+2}, \cdots, a_k), \tag{14}$$

and $e_k = x_k - a_k$ is the associated error vector. The objective is to obtain, through appropriate choice of the weighting vector g_k, a good estimate of $f(\tau)$. The elements of g_k are (yet undefined) functions of the data symbols

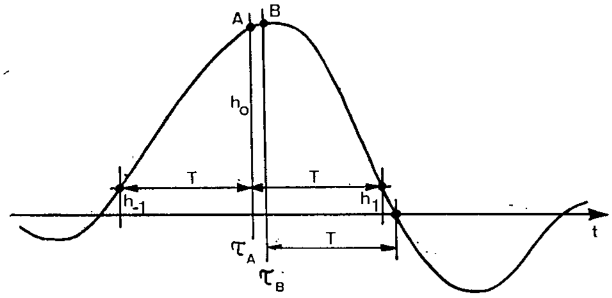


Fig. 4. Resulting sampling instants with two different timing functions. (a) $2f(\tau) = h(\tau + T) - h(\tau - T)$. (b) $f(\tau) = h(\tau + T)$.

$$g_k = \begin{pmatrix} g_1(a_{k-m+1}, \cdots, a_k) \\ g_2(a_{k-m+1}, \cdots, a_k) \\ \cdot \\ \vdots \\ g_m(a_{k-m+1}, \cdots, a_k) \end{pmatrix}. \tag{15}$$

Obviously these elements cannot be constants since this would make z_k the output of a nonrecursive transversal filter and such an arrangement would not provide any timing information. Note that g_k is assumed to depend only on data values contained within a_k, but this is no serious restriction since some function g_i can always be set zero to make z_k dependent on "outside data."

For our further analysis, we will need both the mean and the variance of z_k. We will first discuss the expected value of (11), which is unbiased by additive zero mean noise. The formation of the expected value can be split up into two operations [20],

$$E\{z_k\} = E\{g_k{}^T x_k\} = E\{E[g_k{}^T x_k / a_k]\}. \tag{16}$$

Because g_k depends only on data symbols that are contained within the vector a_k, the inner conditional expected value can be written as

$$E\{g_k{}^T x_k / a_k\} = g_k{}^T E\{x_k / a_k\} = g_k{}^T \nu_k \tag{17}$$

where we have introduced a new vector $\nu_k = E\{x_k / a_k\}$ whose components are given by

$$[\nu_k]_i = E\{x_{k-m+i} / a_k\} = \sum_{j=1}^{m} a_{k-m+j} h_{i-j}. \tag{18}$$

The conditional expected value of x_k is a linear function of the $2m - 1$ samples $h_{1-m}, \cdots, h_0, \cdots, h_{m-1}$ of the system impulse response. This can be stated more clearly in the form

$$\nu_k = E\{x_k / a_k\} = A_k{}^T h \tag{19}$$

where h contains the samples of the above-mentioned truncated impulse response and A_k is a $(2m - 1) \cdot m$ matrix

[3] Delay βT at Nyquist frequency.

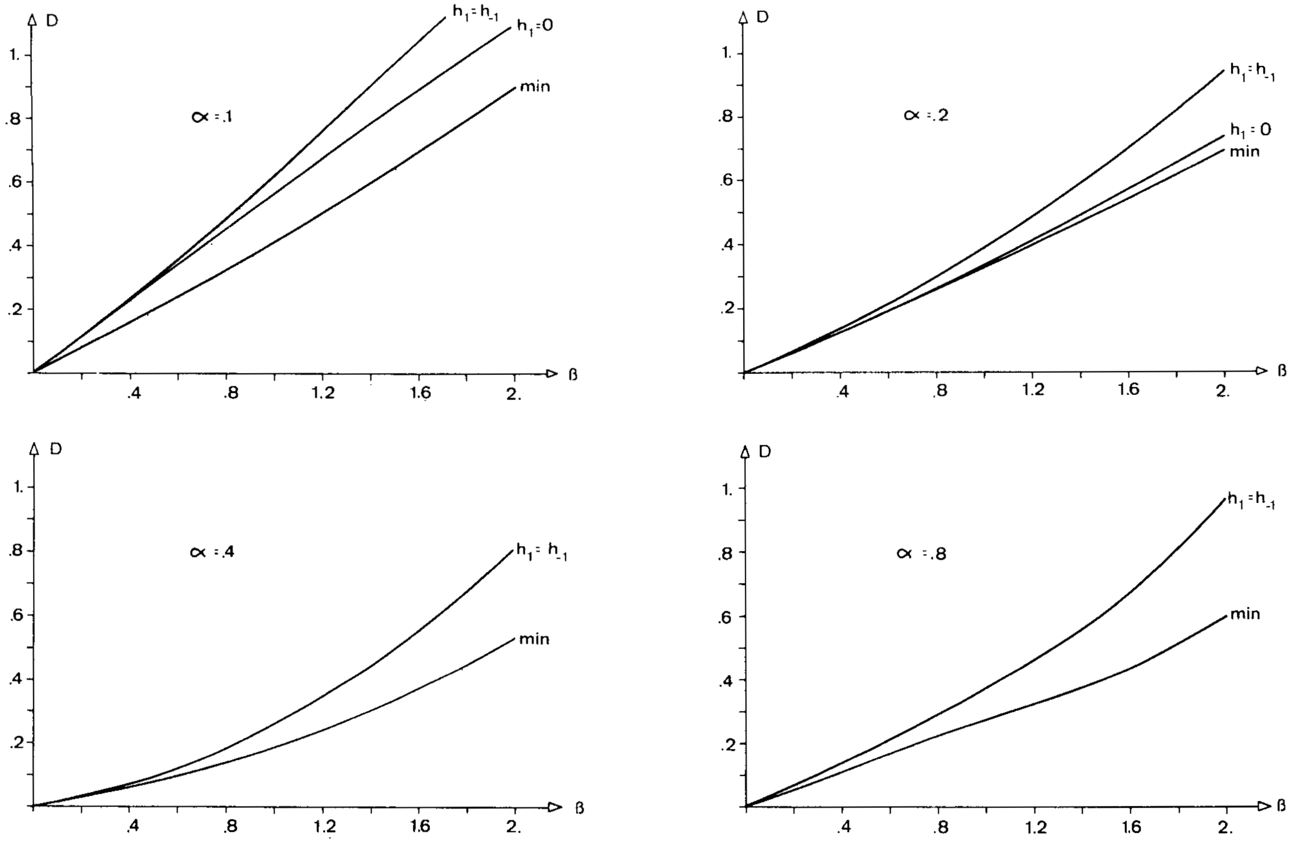


Fig. 5. Peak distortion for channels with cosine rolloff α and quadratic delay distortion.

$$A_k = \begin{bmatrix} a_k & 0 & 0 & \cdots & 0 \\ a_{k-1} & a_k & 0 & \cdots & 0 \\ a_{k-2} & a_{k-1} & a_k & \cdots & 0 \\ \vdots & & & & \vdots \\ a_{k-m+1} & & & \cdots & a_k \\ 0 & a_{k-m+1} & & \cdots & a_{k-1} \\ 0 & 0 & a_{k-m+1} & \cdots & a_{k-2} \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \cdots & a_{k-m+1} \end{bmatrix}. \tag{20}$$

By combining (16), (17), and (19) we obtain

$$E\{z_k\} = h^T E\{A_k g_k\}. \tag{21}$$

The expectation of the product of a signal vector and a data dependent weighting vector is thus a linear function of the samples of $h(t)$, precisely as we have specified for our timing function (5). The mean of (12) is obtained in an equivalent way,

$$E\{g_k{}^T e_k\} = E\{g_k{}^T x_k\} - h_0 E\{g_k{}^T a_k\} \tag{22}$$

where the second term yields a constant that can either be made zero or used to offset some bias in the first term (e.g., dependence on h_0).

A block diagram of a subsystem using these principles to extract timing information is depicted in Fig. 6. Data or error samples are entered into an m tap transversal filter-like structure; but it is important to note that the weighting coefficients are functions of the data symbols and are thus changing at the symbol rate. Such a function can be linear or nonlinear and involve one, several, or all data values of a_k. The resulting coefficients can be digital numbers requiring representation with one or several bits. With the exception of the most simple examples (e.g., a memory of only two symbols), the generation of g_k is most efficiently accomplished with a read-only memory (ROM) that contains the appropriate truth table.

So far we have not discussed the computation of g_k. Before this is done we will determine the variance of z_k because this will be a measure for the mean square error involved in the estimate of $f(\tau)$. First we evaluate

$$E\{z_k{}^2\} = E\{g_k{}^T x_k x_k{}^T g_k\} = E\{g_k{}^T E[x_k x_k{}^T/a_k] g_k\}. \tag{23}$$

The elements m_{ij} of the $m \times m$ matrix $M_k = E\{x_k x_k{}^T/a_k\}$ are given by

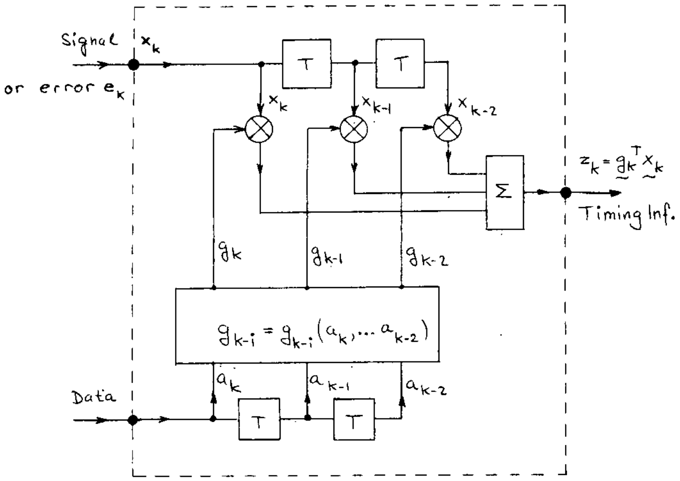


Fig. 6. Generalized block diagram of the proposed timing recovery scheme (shown for $m = 3$).

$$\begin{aligned} m_{ij} &= E\{x_{k-m+i}x_{k-m+j}/a_k\} \\ &= \sum_{\mu}\sum_{\nu} h_{k-m+i-\mu}h_{k-m+j-\nu}E\{a_\mu a_\nu/a_k\}. \end{aligned} \tag{24}$$

Under the usual assumption of statistical independence of data symbols we obtain

$$E\{a_\mu a_\nu/a_k\} = \begin{cases} a_\mu a_\nu, & \text{if } \mu, \nu \quad \in(k-m+1,k) \\ \overline{a^2}, & \text{if } \mu = \nu \quad \notin(k-m+1,k) \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

and thus

$$\begin{aligned} m_{ij} &= \sum_{p=1}^{m}\sum_{r=1}^{m} a_{k-m+p}a_{k-m+r}h_{i-p}h_{j-r} \\ &\quad + \overline{a^2}\sum_{p\notin(1-m)} h_{i-p}h_{j-p}. \end{aligned} \tag{26}$$

The matrix (24) can now be more conveniently expressed as

$$M_k = E\{x_k{}^T x_k/a_k\} = v_k v_k{}^T + Q \tag{27}$$

where the elements of Q are defined by the second term in (26). The second moment of z_k can be written as

$$E\{z_k{}^2\} = E\{g_k{}^T(v_k v_k{}^T + Q)g_k\} = E\{(g_k{}^T v_k)^2 + g_k{}^T Q g_k\} \tag{28}$$

and the variance of z_k becomes

$$\begin{aligned} S &= E\{z_k{}^2\} - E^2\{z_k\} \\ &= E\{(g_k{}^T v_k)^2\} - E^2\{g_k{}^T v_k\} + E\{g_k{}^T Q g_k\}. \end{aligned} \tag{29}$$

The results show that the variance of z_k depends very strongly on g_k. It is interesting to note that the matrix Q does not contain the main sample h_0; thus, if the channel is ideal and if the correct timing phase is used, we conclude that $Q = 0$ and $v_k = h_0 a_k$. The variance will then be a function of the inner product $g_k{}^T a_k$ and will be zero if g_k is chosen orthogonal to the data vector a_k; but we will say more about that in the next section. Finally, we would like to point out that only minor modifications occur if additive white Gaussian noise with variance σ^2 is included in our analysis. The matrix Q must then be replaced by $Q + \sigma^2 I$; the mean of z_k remains unchanged.

V. CHOOSING THE WEIGHTING VECTOR

The weighting vector g_k has to be selected according to

$$E\{z_k\} = f(\tau) = h^T u \tag{30}$$

in order to yield an unbiased estimate of $f(\tau)$. From (11), (21), and (30) we see that this requires

$$E\{A_k g_k\} = u. \tag{31}$$

In general, there is no unique solution for g_k; however, this freedom can favorably be used to select the particular weighting which provides the smallest variance for z_k. Since the advantages of such a solution are obvious, we will consider this optimization in somewhat more detail. Our goal is to minimize the variance (29) while simultaneously satisfying the vector constraint (31). Using a Lagrangian multiplier λ we define the functional

$$\phi = E\{g_k{}^T M_k g_k\} - E^2\{z_k\} - \lambda^T[u - E\{A_k g_k\}]. \tag{32}$$

Recall that $E^2\{z_k\}$ is a constant and that g_k, M_k, and A_k are depending only on the m data values contained in a_k. The expected values in (32) are thus sums of individual functions $g(a_k)$, each weighted with the probability of its a_k, and the extremum is found by setting

$$\frac{\partial\phi}{\partial g_k} = 0, \qquad \frac{\partial\phi}{\partial\lambda} = 0 \tag{33}$$

for each a_k. This yields

$$2M_k g_k - A_k{}^T\lambda = 0. \tag{34}$$

The same result would have been obtained by introducing a variation $g_k + \gamma\delta g_k$ and then requiring

$$\left.\frac{\partial\phi}{\partial\gamma}\right|_{\gamma=0} = 0. \tag{35}$$

If we assume for the moment that M_k is nonsingular, the optimum g_k may be expressed as

$$g_k = \tfrac{1}{2}M_k{}^{-1}A_k{}^T\lambda. \tag{36}$$

The second equation of (33) requires that

$$\lambda = 2E^{-1}\{A_k M_k{}^{-1}A_k{}^T\}u \tag{37}$$

so that finally

$$g_{k\,\text{opt}} = M_k{}^{-1}A_k{}^T E^{-1}\{A_k M_k{}^{-1}A_k{}^T\}u. \tag{38}$$

This formal solution can be inserted into (22) and will then yield a minimum variance

$$S_{\min} = u^T E^{-1}\{A_k M_k^{-1} A_k^T\} u - \bar{z}_k^2. \tag{39}$$

A few comments are in order to the above results. First we recall that the matrix Q can become very small in the vicinity of the optimum timing instant since it does not depend on the main pulse h_0. M_k will then be ill-conditioned since it is mainly determined by the singular matrix $v_k v_k^T$. Singularity of M_k can of course always be avoided if a noise term $\sigma^2 I$ is added to Q. Nevertheless, the evaluation of (38) can become quite involved. Furthermore, since $g_{k\text{opt}}$ depends on M_k and thus on Q, the optimum weighting vector is a function of the impulse response and is therefore influenced by the channel characteristics and, most important, by the timing offset itself. For this reason, any fixed weighting vector g_k can only be optimum for one special situation. We can thus interpret (39) as a lower bound. The variance associated with a fixed g_k can then be compared with this lower bound for a variety of channel parameters. This will be done later on when we have developed some particularly simple examples for g_k. Instead of evaluating (38) for some specific channels, we will in the following propose a simple, suboptimum, channel-independent approach to the problem which will lead us to a number of interesting g_k's.

Condition (31) may be expressed as

$$A_k g_k = u + d_k \tag{40}$$

where the components of d_k are zero mean random variables. The choice of the random vector d_k will affect the variance of z_k according to

$$S = E\{(h^T d_k)^2 + g_k^T Q g_k\} \tag{41}$$

which suggests that we keep d_k as small as possible; this applies in particular to the center components that are weighted with the usually large center samples of h. Of course, d_k cannot be selected arbitrarily, since the system (40) defines $2m - 1$ equations for the m components of g_k and the $2m - 1$ elements of d_k with the already mentioned zero mean constraint for d_k. The weighting vector g_k would of course be uniquely specified by choosing m independent equations of the system (40),

$$A_k^m g_k = u^m + d_k^m. \tag{42}$$

Which equations should be chosen? Certainly, all equations having a nonzero element in u must be considered. Furthermore, we require the variance S to be independent of the main sample h_0 to guarantee zero variance with a Nyquist channel operating at its proper timing phase. This implies that the d_k component associated with h_0 is zero, which in turn means that the center equation of (40), namely

$$a_k^T g_k = 0, \tag{43}$$

must belong to the reduced system (42). Note that the center element of u must also be zero to avoid any appearance of h_0 in $E\{z_k\}$ in order to allow for proper operation of the control loop as has been mentioned in Section III while discussing the choice of u.

In practice, it seems to be a logical start to select the m equations (42) symmetrically around the center of the original system, i.e., blocking out an $m \times m$ square from the rectangular matrix A_k. In addition, we would try to set all components of d_k^m equal to zero in a first approach, thereby avoiding in the main term of S all contributions of the usually largest samples in the vicinity of h_0. Thus we obtain a tentative solution

$$g_k = [A_k^m]^{-1} u^m \tag{44}$$

which needs to be checked against the remaining equations of the set (40). If those are not satisfied, we can try again, this time allowing nonzero values for at least some of the noncenter d_k^m elements or probably choosing different equations. On the other hand, a slight deviation from the specified u may be entirely tolerable. Although this method may sound somewhat heuristic, it nevertheless proved to be quite efficient and convenient in practice. Those who may prefer to determine the optimum weighting vector directly from (38) must bear in mind that the components of g_k are rational functions of the elements of a. Such a solution would also depend on the channel and on timing offset itself. For computer evaluation, special formula-manipulation programs like MATHLAB, ALPAK, or SYMBAL may thus be required. At many computer sites such compilers do not exist.

Finally we mention that the alternative approach (12) based on the error signal will of course always yield a low variance timing estimate since the components of e_k are given by

$$e_{k-m+i} = \sum_j{}' a_{k-m+i-j} h_j \tag{45}$$

and do not contain the main sample h_0. The variance S will thus also be independent of h_0. Note, however, that h_0 must be known to apply this technique. In practice, this is not a disadvantage since some kind of automatic gain control (AGC) will be used anyway to ensure a constant signal level for efficient operation of the A/D converter. For more accuracy, h_0 can be learned from the signal itself after an initial estimate has been used to start the process.

VI. PRACTICAL EXAMPLES

The procedure of determining a weighting vector in accordance with a given u, (i.e., timing function) and a specified memory length m will be illustrated with some simple examples. For scheme A, based on (8) and $m = 2$, (40) reads

$$\begin{pmatrix} a_k & 0 \\ a_{k-1} & a_k \\ 0 & a_{k-1} \end{pmatrix} \begin{pmatrix} g_{k-1} \\ g_k \end{pmatrix} = \begin{pmatrix} -1/2 \\ 0 \\ 1/2 \end{pmatrix} + \begin{pmatrix} d_{k-2} \\ 0 \\ d_k \end{pmatrix} \tag{46}$$

where we have already satisfied (43). One solution is

$$g_k = \frac{1}{2E\{a_k{}^2\}}\begin{pmatrix} -a_k \\ a_{k-1} \end{pmatrix} \tag{47}$$

$$d_k = \frac{1}{2E\{a_k{}^2\}}\begin{pmatrix} E\{a_k{}^2\} - a_k{}^2 \\ 0 \\ a_{k-1}{}^2 - E\{a_k{}^2\} \end{pmatrix}, \tag{48}$$

yielding the estimate

$$z_k = \tfrac{1}{2}(x_k a_{k-1} - x_{k-1} a_k)/E\{a_k{}^2\}. \tag{49}$$

Note that

$$E\{z_k\} = \tfrac{1}{2}(h_1 - h_{-1}), \qquad E\{d_k\} = 0 \tag{50}$$

and that $d_k = 0$ for binary transmission. The variance of (49) can either be evaluated from (41) and (48) or via direct calculation. After some manipulations we obtain

$$S = \tfrac{1}{2}\sum_k{}' h_k{}^2 - \left[2 - \frac{E\{a_k{}^4\}}{E^2\{a_k{}^2\}}\right]\frac{h_1{}^2 + h_{-1}{}^2}{4} + \frac{\sigma^2}{2E\{a_k{}^2\}} \tag{51}$$

where σ^2 is the variance of some added zero mean channel noise. The term in the brackets is unity for binary signaling and decreases with multilevel data. Note that, except for this term, the mean square error of the estimate (49) is simply related to the sum of the mean square distortion and the SNR.

Two possible implementations of (49) are shown in Fig. 7. The realization which uses only one quantizer seems preferable. Further, it should be noted that for binary signaling with $a_k = \pm 1$ the multiplier and summing arrangement is really reduced to a controlled adder-subtractor and is thus extremely simple.

We will now proceed to scheme B (zero forcing of h_1). Trying again $m = 2$ yields the trivial estimate

$$z_k = x_k a_{k-1}/E\{a_k{}^2\} \tag{52}$$

which satisfies (9) and (30), but not (43); i.e., the variance still depends on h_0. This can be avoided if we use the error signal e_k instead of x_k as defined in (12) and (45). We obtain the new estimate

$$z_k = a_{k-1}(x_k - a_k h_0)/E\{a_k{}^2\} \tag{53}$$

with a variance

$$S = \sum_k{}' h_k{}^2 - \left[2 - \frac{E\{a_k{}^4\}}{E^2\{a_k{}^2\}}\right] h_1{}^2 + \frac{\sigma^2}{E\{a_k{}^2\}} \tag{54}$$

which is very similar to (51). Since $h_1 = 0$ in the steady-state timing position, we conclude that the variance (54) will then be equal to the mean square distortion plus the SNR. The variance is roughly reduced by a factor of two for the two-sample estimate (49) when compared to the one-sample estimate (53), as we would expect due to the effect of "averaging." Observe, that for this comparison, we have normalized $f(\tau) = E\{z_k\}$to yield identical slope at the origin. An improved estimate could be obtained by averaging over several symbols, in general,

$$z_k(M > m) = \frac{1}{M - m + 1}\sum_{i=0}^{M-m} z_{k-i}(m). \tag{55}$$

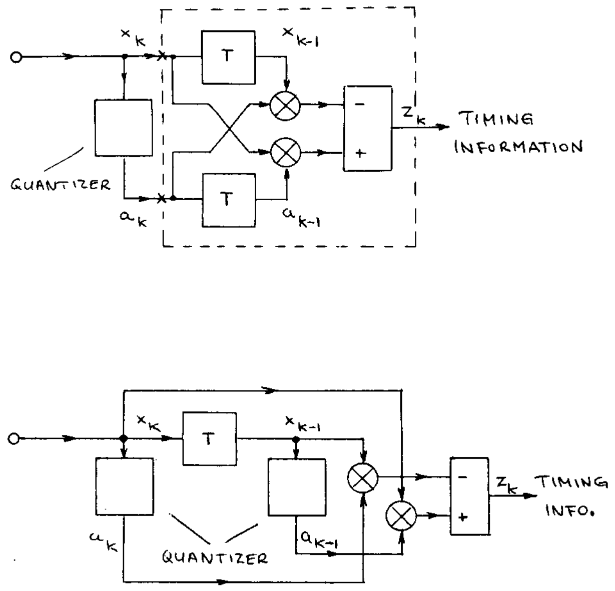


Fig. 7. Two implementations of a type A system.

An alternate solution, which does not contain h_0, can be derived from (40) if we allow $m > 2$. As an example, for binary data and $m = 3$ we get a weighting vector

$$g_k = \frac{1}{3}\begin{pmatrix} -a_{k-1} - 2a_k a_{k-1} a_{k-2} \\ a_k + 2a_{k-2} \\ a_{k-1} - a_k a_{k-1} a_{k-2} \end{pmatrix}. \tag{56}$$

Although a single estimate becomes more accurate for larger m's, the correlation between succeeding estimates (if taken at the symbol rate as discussed in the next section) increases and decision errors will propagate over several estimates. If scheme B is used, the calculations required for each estimate become quite involved for $m > 3$. The simple weighting vectors discussed in the previous examples will be economic to implement and will give satisfactory results so that there seems to be little reason to give detailed consideration to more complex schemes.

To illustrate this point, the variance (51) of the estimate (49) and the variance based on the weights (56) have been calculated as a function of the timing offset τ for a noiseless Nyquist channel with cosine rolloff $\alpha = 0.2$ and random binary signaling. The results are depicted in Fig. 8 (dashed curves). Simultaneously we have evaluated the lower bound (39) associated with the appropriate timing functions (solid

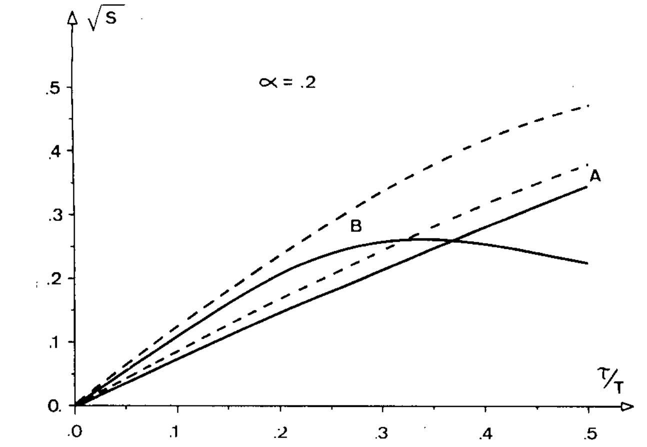


Fig. 8. Variance of estimate z_k (dashed curves) and theoretical minimum (solid curves). Curve A is estimate (49). Curve B is an estimate based on (56) for $\tau > 0$.

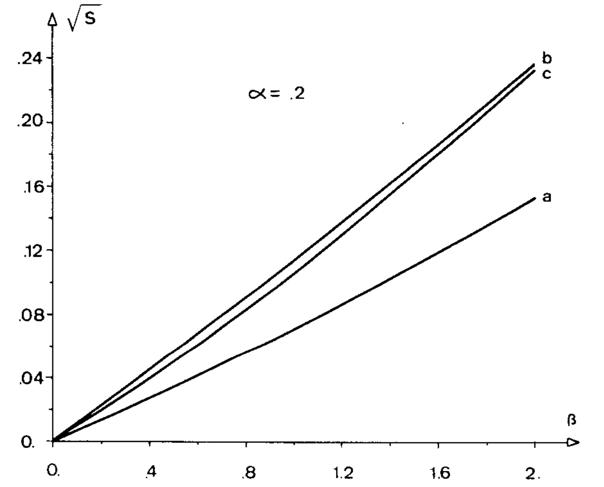


Fig. 9. Variance of estimate z_k in the presence of quadratic delay distortion. Curve a is estimate (49). Curve b is an estimate based on (56). Curve c is estimate (53).

curves). For very small offsets, the actual mean square error coincides with the lower bound, but even for larger offsets the estimate based on the symmetry error (curve A) yields a difference of only 1 or 2 dB. This small difference is due to the fact that $g_{k\text{opt}}$ depends on the impulse response and is thus not identical to (47) except near the origin. However, for practical purposes the difference is negligible and the fixed weighting can be considered near optimum.

The estimates, in general, are of course particularly reliable near the origin. For larger timing deviations the symbol estimation, which is needed if a decision directed algorithm is used, becomes unreliable anyway, so that this region is of limited practical interest. On the other hand, the mean square error for small offset will, for many channels, be primarily determined by noise and residual intersymbol interference. This is shown in Fig. 9 where the variance at the steady-state timing phase is plotted for different amounts of parabolic delay distortion. Included are estimates based on (49), (53), and (56). It is seen that, among this limited selection, the scheme based on the symmetry error performs best. By increasing the memory m, estimates with still smaller variance can of course be obtained. But since noise and quantization effects contribute in a dominant way to the steady-state variance, such further increase in complexity will usually give only marginal improvements in timing jitter.

VII. LOOP BEHAVIOR AND CONVERGENCE

The estimates discussed so far will now be used for adaptive timing recovery. Our timing control schemes will be such that the $(k + 1)$th and the kth adjustment of the sampling phase are related by the recursion

$$\tau_{k+1} = \tau_k + \gamma_k z_k(\tau_k). \tag{57}$$

An algorithm of the above form will in general not exactly "turn itself off" (unless γ decreases to zero) since z_k is a stochastic variable depending on the timing phase, the impulse response, the data symbols, and additive noise. This will cause some random fluctuations (jitter) around the steady-state timing phase. The mean of the corrections will be zero for a timing phase τ_0 such that

$$E\{z_k(\tau_0)\} = f(\tau_0) = 0 \tag{58}$$

and this determines of course the steady-state sampling instant. For our further discussion we will assume that this is at $\tau = \tau_0 = 0$, which is no restriction since we could always define a shift

$$\theta = (\tau - \tau_0)/T. \tag{59}$$

We will also assume that

$$\gamma_k = cT = \text{constant}. \tag{60}$$

This is in contrast to the usual (but unrealistic) stochastic approximation procedure [21], [22] where a gain constant is used that decreases to zero. We thus obtain the recursion

$$\theta_{k+1} = \theta_k - cz_k. \tag{61}$$

For the further analysis we express the estimate z_k as

$$z_k = f(\theta_k) + r_k = \theta_k s(\theta_k) + r_k. \tag{62}$$

Here r_k is a zero mean random variable with variance S and

$$-s_2 \leqslant s(\theta) = \frac{f(\theta)}{\theta} \leqslant -s_1 \tag{63}$$

i.e., the timing function is bounded by two straight lines with slopes s_1 and s_2. These bounds can be quite tight, particularly with timing functions of type A which can exhibit a very linear transfer characteristic with suitable chosen coefficients. The recursion (61) can now be expressed as

$$\theta_{k+1} = \theta_k[1 - cs(\theta_k)] - cr_k. \tag{64}$$

We now define

$$q_k = E\{\theta_k^2\}. \tag{65}$$

Squaring both sides of (64) and taking expected values yields

$$\begin{aligned} q_{k+1} &= E\{\theta_k^2[1 - cs(\theta_k)]^2\} + c^2 S(\theta_k) \\ &\quad - 2cE\{\theta_k r_k [1 - cs(\theta_k)]\}. \end{aligned} \tag{66}$$

The expected value of the crossterm is zero if θ_k and r_k are assumed to be uncorrelated. A closer analysis shows that $E\{\theta_k r_k\}$ depends on the autocorrelation products $cE\{r_k r_{k-i}\}$ with $i > 0$. Thus some interaction can be expected with timing estimates that use $m > 1$, and also due to intersymbol interference. However, if adjustments are made at intervals NT rather than T, and if N is chosen large enough, our assumption will be justified.[4]

$S(\theta)$ is composed of a constant noise term plus an offset dependent intersymbol interference term. The variance can thus be bounded in a similar way as the slope, i.e.,

$$v_1^2\theta^2 + S_\infty \leqslant S(\theta) \leqslant v_2^2\theta^2 + S_\infty \tag{67}$$

where S_∞ is the residual steady-state noise (including some intersymbol interference if the channel is not ideal) and v_1 and v_2 are two constants defining the segment of the rms error. Again, as in (63), these bounds can be quite tight since for many estimates the rms error is a relatively linear function of θ (see Fig. 8). We now obtain from (66)

$$q_k A_{\min} + c^2 S_\infty \leqslant q_{k+1} \leqslant q_k A_{\max} + c^2 S_\infty \tag{68}$$

where we have set for convenience

$$A_{\min} = (1 - cs_2)^2 + c^2 v_1^2 \tag{69}$$

$$A_{\max} = (1 - cs_1)^2 + c^2 v_2^2. \tag{70}$$

The bounds are thus defined by a first-order difference equation. To express its solution in a more compact form we define

$$q_{\infty\,\max} = \frac{c^2 S_\infty}{1 - A_{\max}} = \frac{cS_\infty}{2s_1 - c(s_1^2 + v_2^2)} \tag{71}$$

$$q_{\infty\,\min} = \frac{c^2 S_\infty}{1 - A_{\min}} = \frac{cS_\infty}{2s_2 - c(s_2^2 + v_1^2)}. \tag{72}$$

The bounds can now be written as

$$\begin{aligned} q_{\infty\,\min} &+ (q_0 - q_{\infty\,\min})A_{\min}^k \leqslant q_k \\ &\leqslant q_{\infty\,\max} + (q_0 - q_{\infty\,\max})A_{\max}^k. \end{aligned} \tag{73}$$

Note that the quantities $q_{\infty\min}$ and $q_{\infty\max}$ are lower and upper bounds for the steady-state MS jitter. In practice, this jitter can be precisely determined without the need for bounds. This is because the jitter amplitudes can be considered small and thus the behavior of s and v around the origin is given by the slopes

$$s_0 = \frac{d}{d\theta} f(\theta), \qquad \theta = 0 \tag{74}$$

$$v_0 = \frac{d}{d\theta} \sqrt{S(\theta) - S_\infty}, \qquad \theta = 0 \tag{75}$$

rather than by the bounds used in the preceding approach. Therefore, the resulting jitter is

$$q_\infty = \frac{cS_\infty}{2s_0 - c(s_0^2 + v_0^2)}. \tag{76}$$

System stability in the steady state is determined by the condition

$$A_0 = (1 - cs_0)^2 + c^2 v_0^2 < 1, \tag{77}$$

and thus requires

$$c < \frac{2s_0}{s_0^2 + v_0^2}. \tag{78}$$

The minimal A_0 is achieved with

$$c_0 = \frac{s_0}{s_0^2 + v_0^2} \tag{79}$$

and becomes

$$A_{0\,\min} = \frac{v_0^2}{s_0^2 + v_0^2} = 1 - s_0 c_0 \tag{80}$$

with a resulting jitter

$$q_{\infty\,0} = \frac{S_\infty}{s_0^2 + v_0^2} = \frac{c_0}{s_0} S_\infty. \tag{81}$$

In many cases the bounds on s and v are close together and the convergence behavior can, with sufficient accuracy, be expressed by s_0, v_0, and A_0.

We will now discuss the convergence time. Note, from (73), that the steady-state jitter is reached in an exponential way. We will, somewhat arbitrarily, define the convergence time as the number of symbols needed to reduce the average rms jitter to one percent of a signaling interval. Such a small standard deviation cannot always be obtained if c is initially selected according to (79), but basically q_∞ can be made arbitrarily small by gear shifting to a smaller c after initial training. We will neglect such considerations for the moment and assume $q_\infty = 0$. The convergence time is then bounded by

$$-\frac{\log(10^4 q_0)}{\log A_{\min}} < k < -\frac{\log(10^4 q_0)}{\log A_{\max}}. \tag{82}$$

For the worst case offset where $\theta = 0.5$, a noiseless system with the optimized parameters (79) and (80) would have a

[4] In actual simulations both methods yielded about identical results. For most practical estimates z_k and parameters c, the expected value of the crossterm (on a baud-to-baud basis) is indeed very small.

convergence time

$$k \approx \frac{7.82}{\ln\left[1+\left(\frac{s_0}{v_0}\right)^2\right]}. \tag{83}$$

In practice, the initial offset is random and will tend to be uniformally distributed over $-0.5 \leqslant \theta \leqslant 0.5$. If this is taken into account, it can be shown that, in the mean, a convergence time about 25 percent shorter than (83) can be expected. However, this figure is not very useful for practical system design where enough start-up time must be allocated for worst case behavior.

Since timing recovery schemes of the type described here are most likely to be used in digital receivers, some remarks should be made about quantization effects. Assume that θ can be adjusted in increments p; then adjustments will stop as soon as the correction term is smaller than the quantization interval, i.e., when

$$|cz_k| < p. \tag{84}$$

This dead zone effect can be avoided as long as

$$\theta^2 s_0^2 + 2s_0\theta r + r^2 > \frac{p^2}{c^2}. \tag{85}$$

Whether or not, at any specified offset θ, any further adjustment will occur at the next step will depend on both θ itself and the random variable r. Taking expected values and assuming "separated" adjustments we obtain the condition

$$E\{\theta^2\}(s_0^2 + v_0^2) + S_\infty > \frac{p^2}{c^2} \tag{86}$$

for the *MS* jitter. Since this must hold even for the limit q_∞, we can insert (76) and obtain after a few manipulations

$$\frac{2S_\infty c_0^2}{p^2}\left(\frac{c}{c_0}\right)^2 + \left(\frac{c}{c_0}\right) - 2 > 0. \tag{87}$$

The minimum value c/c_0 that satisfies (87), and thus avoids dead zone effects within the loop, is shown in Fig. 10 as a function of the parameter

$$R = \frac{c_0}{p}\sqrt{2S_\infty}. \tag{88}$$

For $R \to 0$ the curve would asymptotically reach $c \to 2c_0$. Although this is still within the stability range, it is generally undersirable to make $c > c_0$ and one would preferably set $c = c_0$ and tolerate some dead zone jitter in systems with $R < 1$.

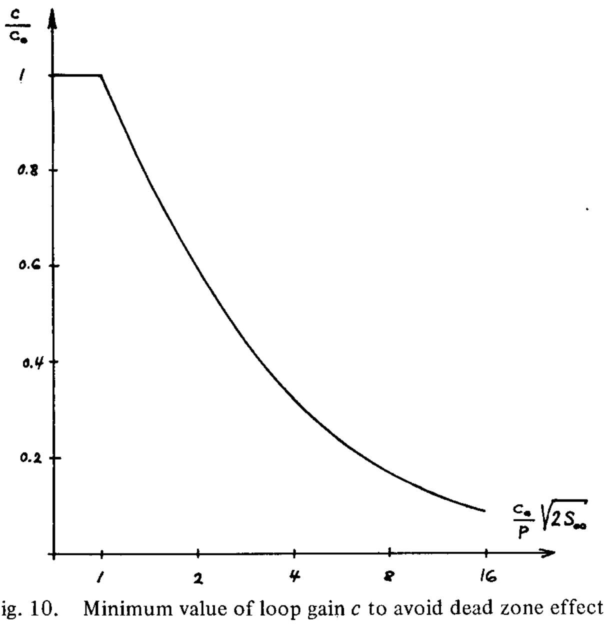


Fig. 10. Minimum value of loop gain c to avoid dead zone effects.

TABLE I
PARAMETERS s AND v FOR BINARY NYQUIST SIGNALING

α	s_0	s_1	s_2	v_0	v_1	v_2
0	1.000	0.849	1.000	1.070	0.860	1.070
0.2	0.963	0.826	0.963	0.851	0.754	0.851
0.4	0.858	0.762	0.858	0.662	0.662	0.671
0.6	0.702	0.675	0.702	0.504	0.504	0.605
0.8	0.519	0.519	0.582	0.367	0.367	0.550
1.0	0.333	0.333	0.500	0.239	0.239	0.500

VIII. SIMULATION RESULTS

Numerical evaluation of the results obtained in the previous section shows that timing can indeed be recovered with only a small number of adjustments, often less than 20. This has been confirmed by detailed simulation studies which have also shown that our formulas tend to be quite accurate, and with certain restrictions, are even usable when adjustments are performed at the symbol rate.

As an example, we will investigate the estimate (49). Table I shows the values of the parameters s and v for binary Nyquist signaling. For all rolloffs, the bounds of s and v occur either at $\theta = 0$ or $\theta = 0.5$. We select a 20 percent cosine rolloff channel with a 26-dB SNR and a quantization of $p = 1/256$. For $\alpha = 0.2$, we obtain $c_0 = 0.58$, but it is easy to verify that such a large value would yield unacceptable jitter amplitudes in the presence of channel impairments. Dead zone requirements would dictate $c > 0.1$. As a compromise between these extremes, we choose $c = 0.2$. This will yield acceptable jitter and avoid dead zone effects. The results of such a simulation are depicted in Fig. 11. Adaptation starts with a worst case timing offset $\theta = 0.5$ ($q = -6$ dB) and settling occurs with a final rms jitter $\sqrt{q} = 0.0125$ (-38.1 dB) which agrees perfectly with (76). Lower and upper bounds for the convergence have been plotted from (73) as a comparison basis. With the particular parameters used in this simulation, settling time occurs within 20 adjustments, which again, agrees very closely with the theoretical results of Section VII.

The simulation results in Fig. 11 have been obtained by averaging over a large number of adaptations, each with a

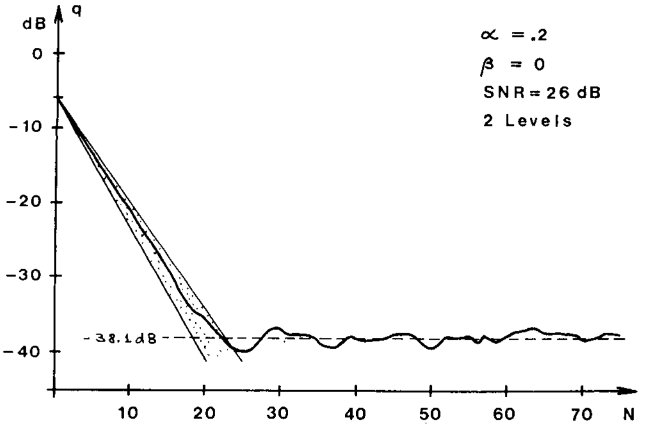


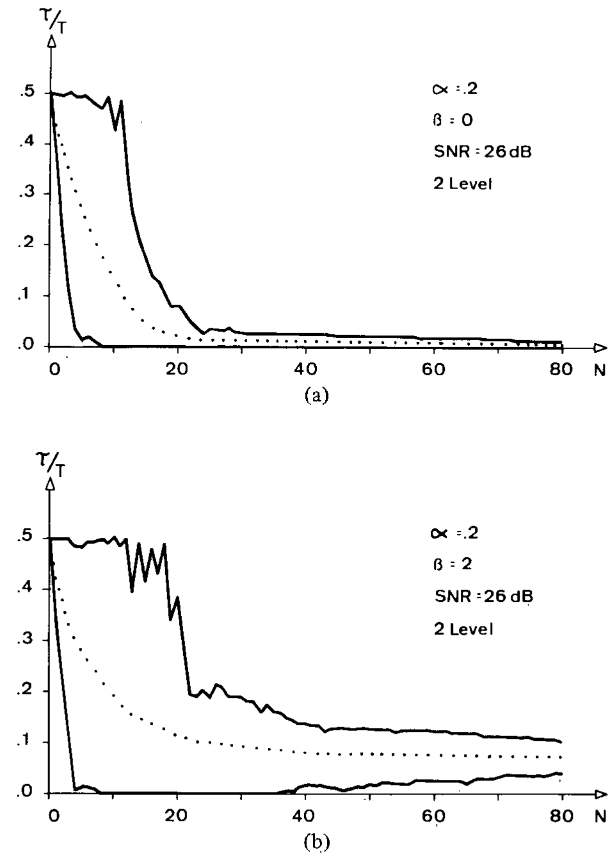
Fig. 11. Convergence with algorithm based on estimate (49).

different data sequence. To bring some practical realism into this procedure, a 63-bit maximum length sequence has been used and starting was assumed to be at random; i.e., averaging was performed over the results obtained with each of the 63 possible shifted data patterns. Such a sequence would also be suitable for initial training of an adaptive equalizer.

Similar simulations for other estimates have also shown good agreement with the theoretically predicted performance. Instead of discussing these results in detail, we will devote some space to related topics that are of more practical significance to the system designer, and we will modify our subsequent simulations accordingly.

First a few words will be said regarding the data symbols a_k that are required to compute the estimates z_k. With an initial offset of $T/2$ the first few decisions are practically uncorrelated with the actual data symbols. In theory, a synchronized reference could be provided at the receiver, but this is not an attractive solution since the synchronization process would probably require more time than the timing recovery. The use of a decision directed reference, even during start-up, is a much more appealing scheme; i.e., the data symbols are estimated by suitable quantization of the signal samples. Because of the large initial error rate such a loop is hard to analyze, but it can conveniently be investigated via simulation. Second, we will make adjustments at each baud to obtain fast convergence in *real time.* A somewhat larger value for c will be used initially and then reduced by a factor of four after 30 adjustments to obtain a small steady-state timing jitter. Finally, we will also demonstrate the variability of convergence as caused by the training sequence itself, i.e., the choice of the starting point. This can most conveniently be done by depicting the envelopes of all 63 curves. As previously, a rolloff $\alpha = 0.2$, a 26-dB SNR, and a quantization $p = 1/256$ will be used.

Fig. 12(a) shows the performance results of this modified loop, again with the estimate (49). The solid lines give upper and lower bounds for θ, while the dotted curve shows the rms value of the average jitter q as in Fig. 11, only that we are now using a linear scale. Convergence takes 5-15 symbols depending on the sequence start, which demonstrates that careful optimization of the training pattern is very important for rapid acquisition. If the channel is noise free, a slight improvement is obtained, particularly in the upper bound, but for SNR's as they occur at voice grade channels (25-35 dB), convergence is almost unaffected by noise.

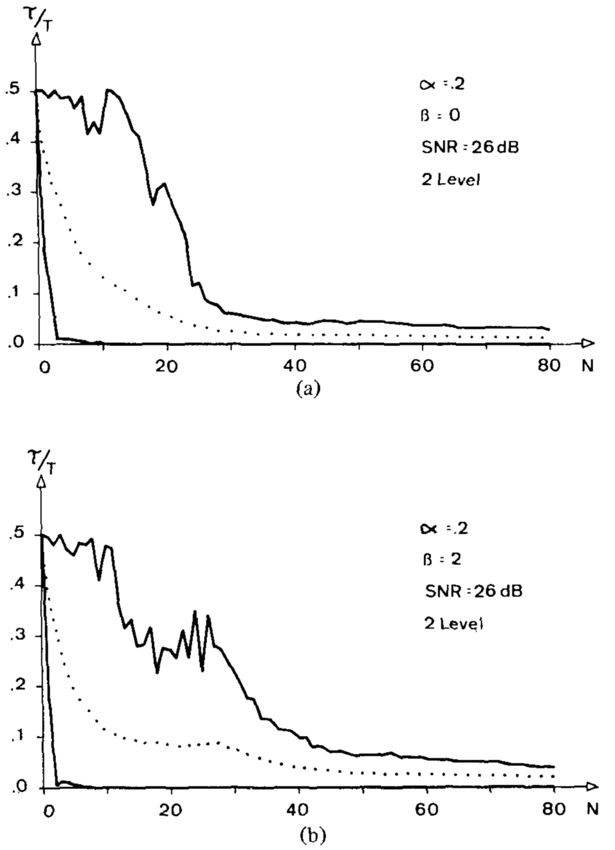


(a)

(b)

Fig. 12. Convergence behavior with estimate (49). (a) No distortion. (b) Quadratic delay distortion, $\beta = 2$.

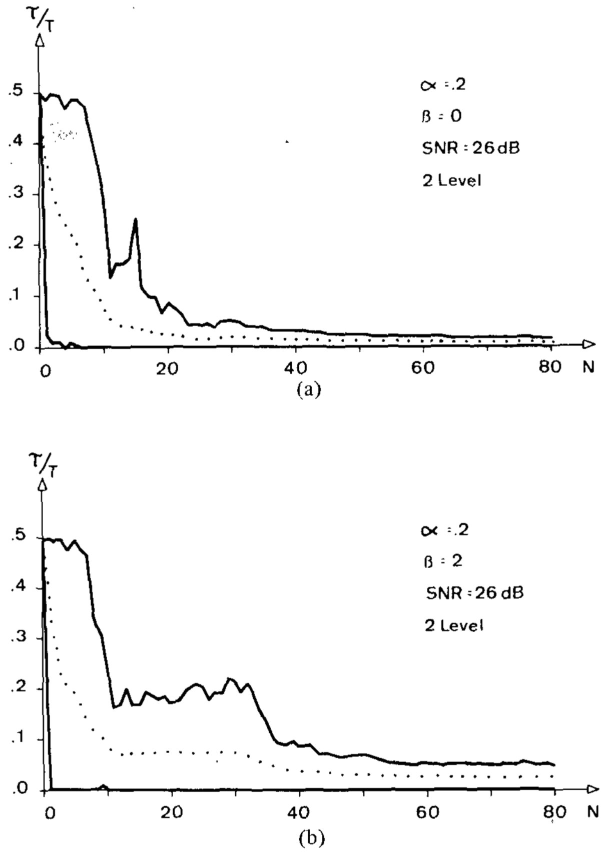
Fig. 12(b) demonstrates the behavior of the same loop with a channel that exhibits quadratic delay distortion ($\beta = 2$). The phase which minimizes peak distortion has been chosen as the reference, e.g., $\tau = 0$. The variance of the convergence characteristics has somewhat increased. The larger jitter, of course, is due to the remaining intersymbol interference which, in turn causes a high S_∞. This degradation is even enhanced by the timing functions inability to "find" exactly the maximum eye opening (see Section II and Fig. 5). In the case under study, settling occurs at a timing phase which yields a peak distortion of 0.9 whereas a value of 0.7 could be achieved with a more optimum sampling instant. Decreasing c will thus only help to a minor degree since a steady-state bias will remain. An algorithm which forces $h_1 = 0$ (scheme B) should provide superior steady-state behavior, at least for channels with delay distortion similar to that used in our example. This is depicted in Fig. 13 where the same simulations have been repeated with the weighting vector (56). The average settling time remains about the same, the upper bound is increased (particularly in the undistorted channel), but the bias in the presence of delay distortion is significantly reduced. A similar behavior can be expected with estimate (53) and is depicted in Fig. 14. Here



(a)

(b)

Fig. 13. Convergence behavior with weighting vector (56). (a) No distortion. (b) Quadratic delay distortion, $\beta = 2$.



(a)

(b)

Fig. 14. Convergence behavior with estimate (53). (a) No distortion. (b) Quadratic delay distortion, $\beta = 2$.

h_0 is learned simultaneously with the timing phase by using the linear recursion

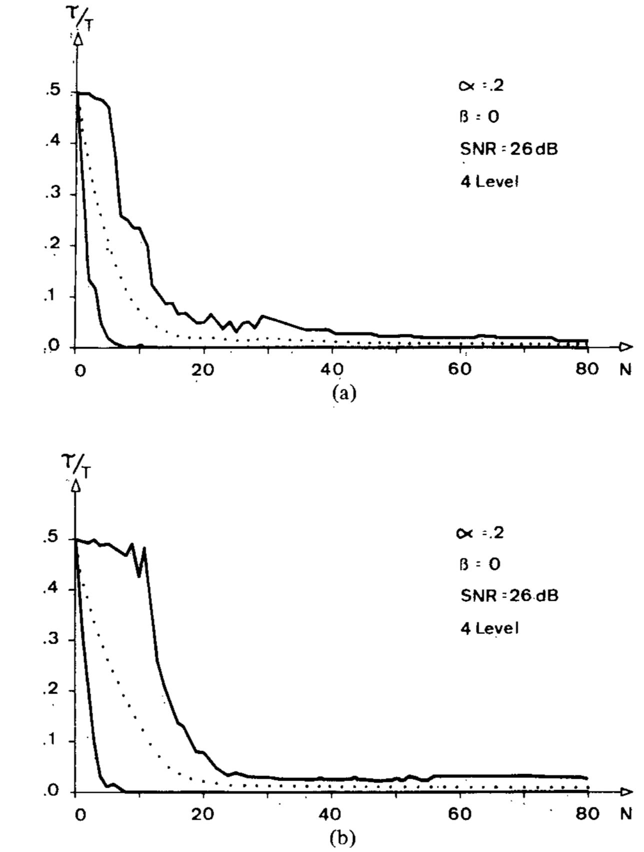
$$h_0^{(m+1)} = (1 - \epsilon) h_0^{(m)} + \epsilon x_m / a_m \; , \tag{89}$$

where we have set $\epsilon = 1/8$.

Remember that a decision-directed reference has been used in the preceding simulations. The use of an ideal reference was found to yield only a minor improvement in mean convergence time; even the decrease of the worst case bound is not that significant. The reason for this may be that the large corrections occurring within the first few adjustments will rapidly shift the phase away from its initial $\theta = 0.5$ position. It is unimportant in which direction this shift takes place; the reliability of the decisions will always improve, and the loop will be able to "lock in." Our experience indicates that this is not necessarily true with multilevel signaling. Such signals have shown convergence problems with decision-directed start-up, even under ideal channel conditions. The use of an ideal reference is very advantageous in such situations. As an example, Fig 15(a) shows the simulation results of a four-level system using estimate (49). Note that the 26-dB SNR is now a more serious degradation than with binary signaling, but even so, the timing loop settles very rapidly. The ideal reference is not needed if the system is started up with a binary signal first; the number of levels is then increased after a specified time. The behavior of such a system is demonstrated in Fig. 15(b), and one can conclude that this approach provides an efficient and highly practical scheme for multilevel signaling. Similar results have been obtained in this case with algorithms based on other estimates.

IX. CONCLUSIONS AND SUMMARY

We have presented a new class of timing recovery schemes for synchronous data receivers. All information is derived from the Nyquist spaced signal samples alone; no signal derivatives, zero crossings, square law devices, or narrow-band filters are required. Timing corrections are based on estimates which are products of the sampled signal vector (or error vector) and a weighting vector whose components are functions of the data symbols. The expected value of this estimate defines a timing function $f(\tau)$ which is of crucial importance for both the transfer characteristic of the control loop and the resulting steady-state sampling phase. Examples for suitable choices of $f(\tau)$ have been presented, and a procedure has been outlined to obtain an appropriate weighting vector that will yield a low variance estimate. A bound for the minimum variance has been given, and it was demonstrated that results close to this bound can be practically achieved. Due to this small variance very rapid convergence is obtained when these estimates are used in a timing loop with a stochastic adjustment algorithm. The theoretical results regarding lower and



(a)

(b)

Fig. 15. Convergence behavior of four-level signal. (a) Ideal reference. (b) Estimated reference with two-level start-up.

upper bounds on convergence as well as residual jitter have been very well confirmed by simulations. Our analysis was based on certain independence assumptions and on the availability of an ideal reference signal, but these do not seem to be restricting requirements. In fact, a large number of simulations using a decision-directed reference and baud-to-baud adjustment (where some dependence between adjacent estimates exists) have shown very similar results. For multilevel systems with decision-directed reference, it is necessary to start up with two levels first, and then switch to more levels after a certain number of symbols. For binary data, convergence can be obtained in less than 20 symbols. For channels with negligible or small distortion, estimates based on the symmetry error of the sampled impulse response seem to be preferable to all others since they yield timing functions of odd symmetry (detector characteristic), and can provide very small variance over a wide range of offsets. However, if distortion is severe, the "symmetric" estimates may produce more steady-state offset than is acceptable, and other schemes, e.g., forcing $h_1 = 0$, can give superior results in this respect. Noise levels as they are likely to occur on voice grade telephone channels have little influence on the rapid convergence. Steady-state jitter decreases with decreasing loop gain, but the loop gain cannot be reduced arbitrarily because of dead zone effects due to finite timing resolution. A good compromise between these requirements can, however, be rather easily established. To achieve both fast convergence and low jitter, a gearshifting arrangement for the loop gain may be used. Finally, it should be mentioned that the timing recovery schemes presented here are extremely suitable for receivers based on digital processing. The necessary computations are very simple; in some estimates a single addition or subtraction is all that is required.

ACKNOWLEDGMENT

The authors would like to thank M. Duenki, who initially proposed the concept of using the error vector e_k, as formulated in (12). They would further like to thank R. D. Gitlin and D. D. Falconer for their valuable comments, and D. A. Spaulding for earlier contributions in regard to estimate (49).

REFERENCES

[1] W. R. Bennett and J. R. Davey, *Data Transmission.* New York: McGraw-Hill, 1965.

[2] J. J. Stiffler, *Theory of Synchronous Communication.* Englewood Cliffs, NJ: Prentice-Hall, 1971.

[3] W. C. Lindsey and M. K. Simon, *Telecommunication Systems Engineering.* Englewood Cliffs, NJ: Prentice-Hall, 1973.

[4] B. R. Saltzberg, "Timing recovery for synchronous binary data transmission," *Bell Syst. Tech. J.*, vol. 46, pp. 593-622, March 1967.

[5] R. W. Chang, "Joint equalization, carrier acquisition, and timing recovery for data communication," in *Conf. Rec., Int. Conf. Commun.*, 1970.

[6] R. D. Gitlin and J. Salz, "Timing recovery in PAM-systems," *Bell Syst. Tech. J.*, vol. 50, pp. 1645-1669, May-June 1971.

[7] H. Kobayashi, "Simultaneous adaptive estimation and decision algorithm for carrier modulated data transmission systems," *IEEE Trans. Commun.*, vol. COM-19, pp. 268-280, June 1971.

[8] W. R. Bennett, "Statistics of regenerative digital transmission," *Bell Syst. Tech. J.*, vol. 37, pp. 1501-1542, Nov. 1958.

[9] Y. Takasaki, "Timing extraction in baseband pulse transmission," *IEEE Trans. Commun.*, vol. COM-20, pp. 877-884, Oct. 1972.

[10] L. E. Franks and J. P. Bubrouski, "Statistical properties of timing jitter in a PAM timing recovery system," *IEEE Trans. Commun.*, vol. COM-22, pp. 913-920, July 1974.

[11] A. L. McBride and A. P. Sage, "Optimum estimation of bit synchronization," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-5, pp. 525-536, May 1969.

[12] U. Mengali, "A self bit synchronizer matched to the signal shape," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-7, pp. 686-693, July 1971.

[13] A. Lender, "Decision directed adaptive equalization technique for high speed data transmission," *IEEE Trans. Commun. Technol.*, vol. COM-18, pp. 625-632, Oct. 1970.

[14] D. A. Spaulding, "A new digital coherent demodulator," *IEEE Trans. Commun.*, vol. COM-21, pp. 237-238, Mar. 1973.

[15] R. W. Lucky, J. Salz, and E. J. Weldon, *Principles of Data Communication.* New York: McGraw-Hill, 1968.

[16] K. H. Mueller and M. Mueller, "Adaptive timing recovery in digital synchronous data receivers," in *Conf. Rec., 1974 Int. Zurich Seminar on Digital Communications*, 1974.

[17] E. R. Kretzmer, "Generalization of a technique for binary data communication," *IEEE Trans. Commun. Technol.*, vol. COM-14, pp. 67-68, Feb. 1966.

[18] G. Schollmeier and N. Schatz, "The design of nonlinear phase tracking loops by simulation," *IEEE Trans. Commun.*, vol. COM-23, pp. 296-299, Feb. 1975.

[19] K. H. Mueller and D. A. Spaulding, unpublished memorandum, Apr. 24, 1973.

[20] A. Papoulis, *Probability, Random Variables and Stochastic Processes.* New York: McGraw-Hill, 1965.

[21] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, pp. 400-407, Sept. 1951.

[22] D. J. Sakrison, "Stochastic approximation: A recursive method for solving regression problems," in *Advances in Communication Theory,* vol. 2. New York: Academic, 1966.

Kurt H. Mueller received the Diploma in electrical engineering and the Ph.D. degree from the Swiss Federal Institute of Technology, Zurich, Switzerland, in 1961 and 1967, respectively.

From 1962 to 1969 he worked at various research, teaching, and supervisory positions at the Swiss Federal Institute of Technology, where he gave courses in signal theory and information theory. In 1969 he joined Bell Laboratories, Holmdel, NJ, where he was involved in a variety of problems in high-speed data communication. During 1972–1973 he was on leave of absence back at the Swiss Federal Institute of Technology. During 1973 he was also a member of the Executive Body of the European Informatics Network. His present work at Bell Laboratories is mostly in digital signal processing for data transmission systems.

Markus Müller was born in Zurich, Switzerland, on November 24, 1947. He received the Diploma in electrical engineering from the Swiss Federal Institute of Technology (SFIT), Zurich, Switzerland, in 1970.

From 1971 to 1973 he was a Teaching Assistant and from 1973 to 1975 he was a Research Assistant with the Telecommunication Department of the SFIT, where he was concerned with problems of data transmission on the switched telephone network. His main interest was in the investigation of synchronization problems, modulation selection, and equalizer structures in digital receivers for synchronous data transmission. Since 1976 he has been with the Overseas Department, General Radio Company, Zurich, Switzerland.

The Capture Effect in FM Receivers

KRIJN LEENTVAAR AND JAN H. FLINT

Abstract—**In this paper a theoretical explanation of the capture effect is given by calculating the instantaneous frequency of the output signal of a limiter when two frequency modulated (FM) signals are present at the limiter input. When this signal is applied to a demodulator with unlimited bandwidth, the output signal of the demodulator proves to have an extreme capture effect. When however the demodulator bandwidth is limited, the capture effect is shown not be be extreme. This phenomenon is explained and possibilities are given to minimize the capture effect.**

Some of the results of measurements on limiters and demodulators are given in this paper; they prove that a weak capture effect can be obtained. A method of calculating the degree of capturing is included.

INTRODUCTION

WHEN a frequency modulated (FM) receiver has two different FM signals with unequal amplitudes falling within the passband at the same time, the modulation of the weaker signal no longer exists at the demodulator output or at least is attenuated to a very high degree. This also appears when the stronger signal is unmodulated. This phenomenon is known as the capture effect.

In this paper, first the phasor diagram will be considered by which it is possible to calculate the output signal of a limiter and its instantaneous frequency when two FM signals are present at the limiter input. To illustrate the problem the frequency spectrum of the output signal is calculated. A function is given to express the mean frequency of the limiter output signal.

It is possible to explain the reduction of the capture effect by limiting the bandwidth of the demodulator.

A method of calculating these effects is given for a Foster-Seely demodulator.

I. THE PHASOR DIAGRAM

Suppose the two different signals at the input of the limiter are a_1 and a_2. These signals are shown in Fig. 1. The signals may be expressed as

$$a_1 = A_1 \cos \phi_1 = \mathrm{Re}\ [A_1 e^{j\phi_1}] \qquad (1)$$

$$a_2 = A_2 \cos \phi_2 = \mathrm{Re}\ [A_2 e^{j\phi_2}] \qquad (2)$$

where

Paper approved by the Associate Editor for Radio Communication of the IEEE Communications Society for publication without oral presentation. Manuscript received December 10, 1973; revised February 14, 1975.

The authors are with the Physics Laboratory of the National Defense Research Organisation TNO, The Hague, The Netherlands.