Carrier and Bit Synchronization in Data Communication—A Tutorial Review

L. E. FRANKS, FELLOW, IEEE

Abstract—**This paper examines the problems of carrier phase estimation and symbol timing estimation for carrier-type synchronous digital data signals, with tutorial objectives foremost. Carrier phase recovery for suppressed-carrier versions of double sideband (DSB), vestigial sideband (VSB), and quadrature amplitude modulation (QAM) signal formats is considered first. Then the problem of symbol timing recovery for a baseband pulse-amplitude modulation (PAM) signal is examined. Timing recovery circuits based on elementary statistical properties are discussed as well as timing recovery based on maximum-likelihood estimation theory. A relatively simple approach to evaluation of timing recovery circuit performance in terms of rms jitter of the timing parameters is presented.**

I. INTRODUCTION

IN digital data communication there is a hierarchy of synchronization problems to be considered. First, assuming that a carrier-type system is involved, there is the problem of *carrier synchronization* which concerns the generation of a reference carrier with a phase closely matching that of the data signal. This reference carrier is used at the data receiver to perform a coherent demodulation operation, creating a baseband data signal. Next comes the problem of synchronizing a receiver clock with the baseband data-symbol sequence. This is commonly called *bit synchronization*, even when the symbol alphabet happens not to be binary.

Depending on the type of system under consideration, problems of *word-*, *frame-*, and *packet-synchronization* will be encountered further down the hierarchy. A feature that distinguishes the latter problems from those of carrier and bit synchronization is that they are usually solved by means of special design of the message format, involving the repetitive insertion of bits or words into the data sequence solely for synchronization purposes. On the other hand, it is desirable that carrier and bit synchronization be effected without multiplexing special timing signals onto the data signal, which would use up a portion of the available channel capacity. Only timing recovery problems of this type are discussed in this paper. This excludes those systems wherein the transmitted signal contains an unmodulated component of sinusoidal carrier (such as with "on-off" keying). When an unmodulated component or pilot is present, the standard approach to carrier synchronization is to use a phase-locked loop (PLL) which locks onto the carrier component, and has a narrow enough loop bandwidth so as not to be excessively perturbed by the sideband components of the signal. There is a vast literature on the performance and design of the PLL and there are several textbooks dealing with synchronous communication systems which treat the PLL in great detail [1]-[5]. Although we consider only suppressed-carrier signal formats here, the PLL material is still relevant since these devices are often used as component parts of the overall phase recovery system.

For modulation formats which exhibit a high bandwidth efficiency, i.e., which have a large "bits per cycle" figure of merit, we find the accuracy requirements on carrier and bit synchronization increasingly severe. Unfortunately, it is also in these high-efficiency systems that we find it most difficult to extract accurate carrier phase and symbol timing information by means of simple operations performed on the received signal. The pressure to develop higher efficiency data transmission has led to a dramatically increased interest in timing recovery problems and, in particular, in the ultimate performance that can be achieved with optimal recovery schemes.

We begin our review of carrier synchronization problems with a brief discussion of the major types of modulation format. In each case (DSB, VSB, or QAM), we assume coherent demodulation whereby the received signal is multiplied by a locally generated reference carrier and the product is passed through a low-pass filter. We can get some idea of the phase accuracy, or degree of coherency, requirements for the various modulation formats by examining the expressions for the coherent detector output, assuming a noise-free input. Let us assume that the message signal, say, $a(t)$, is incorporated by the modulation scheme into the complex envelope $\beta(t)$ of the carrier signal.[1]

$$y(t) = \text{Re}\ [\beta(t) \exp(j\theta) \exp(j2\pi f_0 t)] \tag{1}$$

and the reference carrier $r(t)$ is characterized by a constant complex envelope

$$r(t) = \text{Re}\ [\exp(j\hat{\theta}) \exp(j2\pi f_0 t)]. \tag{2}$$

From (A-8), the output of the coherent detector is

$$z_1(t) = \tfrac{1}{2}\ \text{Re}\ [\beta(t) \exp(j\theta - j\hat{\theta})]. \tag{3}$$

For the case of DSB modulation, we have $\beta(t) = a(t) + j\theta$, so $z_1(t)$ is simply proportional to $a(t)$. The phase error $\theta - \hat{\theta}$ in the reference carrier has only a second-order effect

Manuscript received June 28, 1979; revised March 26, 1980.

The author is with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003.

[1] See the Appendix for definitions and basic relations concerning complex envelope representation of signals.

0090-6778/80/0800-1107$00.75

on detector performance. The only loss is that phase error causes a reduction, proportional to $\cos^2(\theta - \hat{\theta})$, in signal-to-noise ratio at the detector output when additive noise is present on the received signal.

For VSB modulation, however, phase error produces a more severe distortion. In this case $\beta(t) = a(t) + j\tilde{a}(t)$, where $\tilde{a}(t)$ is related to $a(t)$ by a time-invariant filtering operation which causes a cancellation of a major portion of one of the sidebands. In the limiting case of complete cancellation of a sideband (SSB), we have $\tilde{a}(t) = \hat{a}(t)$, the Hilbert transform of $a(t)$ [6]. The coherent detector output (3) for the VSB signal is

$$z_1(t) = \tfrac{1}{2}\, a(t)\cos(\theta - \hat{\theta}) - \tfrac{1}{2}\,\tilde{a}(t)\sin(\theta - \hat{\theta}) \tag{4}$$

and the second term in (4) introduces an interference called *quadrature distortion* when $\hat{\theta} \neq \theta$. As $\tilde{a}(t)$ has roughly the same power level as $a(t)$, a relatively small phase error must be maintained for low distortion, e.g., about 0.032 radian error for a 30 dB signal-to-distortion ratio.

In the QAM case, two superimposed DSB signals at the same carrier frequency are employed by making $\beta(t) = a(t) + jb(t)$, where $a(t)$ and $b(t)$ are two separate, possibly independent, message signals. A dual coherent detector, using a reference carrier and its $\pi/2$ phase-shifted version, separates the received signal into its in-phase (I) and quadrature (Q) components. Again considering only the noise-free case, these components are

$$\begin{aligned} c_I(t) &= \tfrac{1}{2}\, a(t)\cos(\theta - \hat{\theta}) - \tfrac{1}{2}\, b(t)\sin(\theta - \hat{\theta}) \\ c_Q(t) &= \tfrac{1}{2}\, b(t)\cos(\theta - \hat{\theta}) + \tfrac{1}{2}\, a(t)\sin(\theta - \hat{\theta}). \end{aligned} \tag{5}$$

From (5) it is clear that $\hat{\theta} \neq \theta$ introduces a *crosstalk* interference into the I and Q channels. As $a(t)$ and $b(t)$ can be expected to be at similar power levels, the phase accuracy requirements for QAM are high compared to straight DSB modulation.

From the previous discussion we see that the price for the approximate doubling of bandwidth efficiency in VSB or QAM, relative to DSB, is a greatly increased sensitivity to phase error. The problem is compounded by the fact that carrier phase recovery is much more difficult for VSB and QAM, compared to DSB.

II. CARRIER PHASE RECOVERY

Before examining specific carrier recovery circuits for the suppressed-carrier format, it is helpful to ask, "What properties must the carrier signal $y(t)$ possess in order that operations on $y(t)$ will produce a good estimate of the phase parameter θ?" A general answer to this question lies in the *cyclostationary* nature of the $y(t)$ process.[2] A cyclostationary process has statistical moments which are periodic in time, rather than constant as in the case of stationary processes [2], [6], [7]. To a large extent, synchronization capability can be characterized by the lowest-order moments of the process, such as the mean and autocorrelation. The $y(t)$ process is said to be cyclostationary in the wide sense if $E[y(t)]$ and $k_{yy}(t + \tau, t) = E[y(t + \tau)y(t)]$ are both periodic functions of t. A process modeled by (1) is typically cyclostationary with a period of $1/f_0$ or $1/2f_0$. The statistical moments of this process depend upon the value of the phase parameter θ and it is not surprising that efficient phase estimation procedures are similar to moment estimation procedures. It is important to note here that we are regarding θ as an unknown but nonrandom parameter. If instead we regarded θ as a random parameter uniformly distributed over a 2π interval, then the $y(t)$ process would typically be stationary, not cyclostationary.

A general property of cyclostationary processes is that there may be a correlation between components in different frequency bands, in contrast to the situation for stationary processes [8]. For carrier-type signals, the significance lies in the correlation between message components centered around the carrier frequency ($+f_0$) and the image components around ($-f_0$). This correlation is characterized by the cross-correlation function $k_{\beta\beta^*}(\tau) = E[\beta(t + \tau)\beta(t)]$ for a $y(t)$ process as in (1) when $\beta(t)$ is a stationary process.[3]

Considering first the DSB case with $\beta(t) = a(t) + j\theta$, and using (A-10) we have

$$\begin{aligned} k_{yy}(t + \tau, t) &= \tfrac{1}{2}\,\mathrm{Re}\,[k_{aa}(\tau)\exp(j2\pi f_0\tau)] \\ &\quad + \tfrac{1}{2}\,\mathrm{Re}\,[k_{aa}(\tau)\exp(j4\pi f_0 t + j2\pi f_0\tau + j2\theta)] \end{aligned} \tag{6}$$

where the second term in (6) exhibits the periodicity in t that makes $y(t)$ a cyclostationary process.

We are assuming that $y(t)$ contains no periodic components. Consider what happens, however, when $y(t)$ is passed through a square-law device. We see immediately from (6) that the output of the squarer has a periodic mean value, since

$$\begin{aligned} E[y^2(t)] &= k_{yy}(t, t) \\ &= \tfrac{1}{2}\, k_{aa}(0) + \tfrac{1}{2}\, k_{aa}(0)\,\mathrm{Re}\,[\exp(j2\theta + j4\pi f_0 t)]. \end{aligned} \tag{7}$$

If the squarer output is passed through a bandpass filter with transfer function $H(f)$ as shown in Fig. 1, and if $H(f)$ has a unity-gain passband in the vicinity of $f = 2f_0$, then the mean value of the filter output is a sinusoid with frequency $2f_0$, phase 2θ, and amplitude $\frac{1}{2}E[a^2(t)]$. In this sense, the squarer has produced a periodic component from the $y(t)$ signal.

It is often stated that the effect of the squarer is to produce a discrete component (a line at $2f_0$) in the spectrum of its output signal. This statement lacks precision and can lead to serious misinterpretations because $y^2(t)$ is not a stationary process, so the usual spectral density concept has no meaning. A stationary process can be derived from $y^2(t)$ by phase randomizing [6], but then the relevance to carrier phase recovery is lost because the discrete component has a completely indeterminate phase.

[2] In [2], these processes are called *periodic nonstationary*.

[3] Despite its appearance, this is not an autocorrelation function, due to the definition of autocorrelation for complex processes; see (A-11).

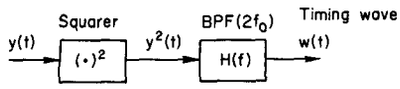

Fig. 1. Timing recovery circuit.

The output of the bandpass filter in Fig. 1 can be used directly to generate a reference carrier. Assuming that $H(f)$ completely suppresses the low-frequency terms [see (A-8)] the filter output is the *reference waveform*

$$w(t) = \tfrac{1}{2} \operatorname{Re} \{[\omega \otimes \beta^2](t) \exp(j2\theta) \exp(j4\pi f_0 t)\} \tag{8}$$

where the convolution product $[\omega \otimes \beta^2]$ represents the filtering action of $H(f)$ in terms of its low-pass equivalent $\Omega(f)$ in (A-5). For the DSB case, $\beta^2(t) = a^2(t)$ is real and $\omega(t)$ is real[4] if $H(f)$ has a symmetric response about $2f_0$. Then the phase of the reference waveform is 2θ and the amplitude of the reference waveform fluctuates slowly [depending on the bandwidth of $H(f)$]. The reference carrier can be obtained by passing $w(t)$ through an infinite-gain clipper which removes the amplitude fluctuations. The square wave from the clipper can drive a frequency divider circuit which halves the frequency and phase. Alternatively, the bandpass filter output can be tracked by a PLL and the PLL oscillator output passed through the frequency-divider circuit.

There is another tracking loop arrangement, called the Costas loop, where the voltage-controlled oscillator (VCO) operates directly at f_0. We digress momentarily to describe the Costas loop and to point out that it is equivalent to the squarer followed by a PLL [1]-[3]. The equivalence is established by noting that the inputs to the loop filters in the two configurations shown in Fig. 2 are identical. In the PLL quiescent lock condition, the VCO output is in quadrature with the input signal so we introduce a $\pi/2$ phase shift into the VCO in the configurations of Fig. 2. Then using (A-8) to get the output of the multiplier/low-pass filter combinations, we see that the input to the loop filter is

$$v(t) = \tfrac{1}{8} \operatorname{Re} [A^2 \beta^2(t) \exp(j2\theta - j2\hat{\theta} - j\pi/2)] \tag{9}$$

in both configurations if the amplitude of the VCO output is taken as $\frac{1}{2} A^2$ in the squarer/PLL configuration, and taken as A in the Costas loop.

Going back to (8), we see that phase recovery is perfect if $[\omega \otimes \beta^2]$ is real. Assuming $\omega(t)$ real, a phase error will result only if a quadrature component [relative to $\beta^2(t)$] appears at the output of the squarer. This points out the error, from a different viewpoint, of using the phase randomized spectrum of the squarer output to analyze the phase recovery performance because the spectrum approach obliterates the distinction between I and Q components. For the DSB case, a quadrature component will appear at the squarer output only if there is a quadrature component of interference added to the input signal $y(t)$. We can demonstrate this effect by considering the

[4] A real $\omega(t)$ corresponds to the case where the cross-coupling paths between input and output I and Q components in Fig. 10 are absent. If the bandpass function $H(f)$ does not exhibit the symmetrical amplitude response and antisymmetrical phase response about $2f_0$ for a real $\omega(t)$, then there simply is a fixed phase offset introduced by the bandpass filter.

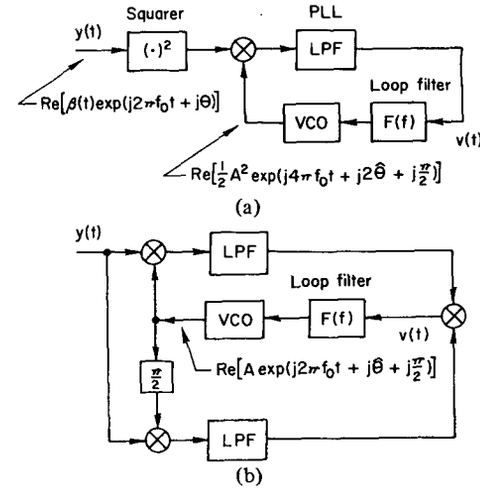

Fig. 2. Carrier phase tracking loops. (a) Squarer/PLL (b) Costas loop.

input signal to be $z(t) = y(t) + n(t)$ where $n(t)$ is white noise with a double-sided spectral density of N_0 W/Hz. We can represent $n(t)$ by the complex envelope, $[u_I(t) + ju_Q(t)] \exp(j\theta)$ where, from (A-15) the I and Q noise components relative to a phase θ are uncorrelated and have a spectral density of $2N_0$. The resulting phase of the reference waveform (8) is

$$2\hat{\theta} = 2\theta + \tan^{-1}\left[\frac{2\omega \otimes (\beta u_Q + u_I u_Q)}{\omega \otimes (\beta^2 + 2\beta u_I + u_I^2 - u_Q^2)}\right]. \tag{10}$$

We can approximate the phase error $\phi = \hat{\theta} - \theta$ (also called *phase jitter* because $\hat{\theta}$ is a quantity that fluctuates with time) by neglecting the noise $\times$ noise term in the numerator and both signal $\times$ noise and noise $\times$ noise terms in the denominator in (10). Furthermore, we replace $\omega \otimes \beta^2$ by its expected value (averaging over the message process) and use the $\tan^{-1} x \cong x$ approximation. With all these simplifications, which are valid at sufficiently high signal-to-noise ratio and with sufficiently narrow-band $H(f)$, it is easy to derive an expression for the variance of the phase jitter.

$$\operatorname{var} \phi = (2N_0 B) S^{-1} \tag{11a}$$

$$= \left(\frac{S}{N}\right)^{-1} \left(\frac{B}{W}\right) \tag{11b}$$

where

$$B \triangleq \int_{-\infty}^{\infty} |\Omega(f)|^2 \, df = \int_{0}^{\infty} |H(f)|^2 \, df$$

is the *noise bandwidth* of the bandpass filter, recalling that we have set $\Omega(0) = 1$. The message signal power is $S = E[a^2(t)]$ and for the second version of the jitter formula (11b) we have assumed a signal bandwidth of W Hz and have defined a noise power over this band of $N = 2N_0 W$. This allows the satisfying physical interpretation of jitter variance being inversely proportional to signal-to-noise ratio and directly proportional to the bandwidth ratio of the phase recovery circuit and the message signal. For the smaller signal-to-noise ratios, the accuracy and convenience of the expression can be maintained by incorporating a correction factor known as the *squaring loss* [3].

When the signal itself carries a significant quadrature component, as in the case of the VSB signal, there will be a quadrature component at the squarer output that interferes with the phase recovery operation even at high signal-to-noise ratios. Let us suppose that the VSB signal is obtained by filtering a DSB signal with a bandpass filter with a real transfer function (no phase shift) and with a cutoff in the vicinity of f_0. The resulting quadrature component for the VSB signal is $\tilde{a}(t) = [p_Q \otimes a](t)$ and $p_Q(t)$ is derived from the low-pass equivalent transfer function for the bandpass filter in accordance with (A-7). The real transfer function condition makes $p_Q(t)$ an odd function of time, which also makes the cross-correlation function for $a(t)$ and $\tilde{a}(t)$ an odd function.

The result is that, for $\beta(t) = a(t) + j\tilde{a}(t)$, the autocorrelation for the VSB signal is

$$k_{yy}(t+\tau, t) = \tfrac{1}{2} \operatorname{Re} [\{k_{aa}(\tau) + k_{\tilde{a}\tilde{a}}(\tau) + j2k_{\tilde{a}a}(\tau)\} \cdot \exp(j2\pi f_0\tau)] + \tfrac{1}{2} \operatorname{Re} [\{k_{aa}(\tau) - k_{\tilde{a}\tilde{a}}(\tau)\} \cdot \exp(j4\pi f_0 t + j2\pi f_0\tau + j2\theta)]. \quad (12)$$

Comparing (12) with (6), we see that the second, cyclostationary, term is much smaller for the VSB case than the DSB case since the autocorrelation functions for $a(t)$ and $\tilde{a}(t)$ differ only to the extent that some of the low-frequency components in $\tilde{a}(t)$ are missing because of the VSB rolloff characteristic. Although the jitter performance will be poorer, the phase recovery circuit in Fig. 1 can still be used since the mean value of the reference waveform is a sinusoid exhibiting the desired phase, but with an amplitude which is proportional to the difference in power levels in $a(t)$ and $\tilde{a}(t)$.

$$E[w(t)] = \tfrac{1}{2} [k_{aa}(0) - k_{\tilde{a}\tilde{a}}(0)] \operatorname{Re} [\exp(j4\pi f_0 t + j2\theta)]. \quad (13)$$

However, it is not possible to get a very simple formula for the variance of phase jitter, as in (11), because the power spectral density of the quadrature component of $\beta^2(t)$, which is proportional to $a(t)\,\tilde{a}(t)$, vanishes at $f = 0$, unlike in the additive noise case. An accurate variance expression must take into account the particular shape of the $\Omega(f)$ filtering function as well as the shape of the VSB rolloff characteristic.

Our examination of phase recovery for DSB (with additive noise) and VSB modulation formats has indicated that rms phase jitter can be made as small as desired by making the width of $\Omega(f)$ sufficiently small. The corresponding parameter in the case of the tracking loop configuration is called the loop bandwidth [3]. These results, however, are for *steady-state* phase jitter since the signals at the receiver input were presumed to extend into the remote past. The difficulty with a very narrow phase recovery bandwidth is that excessive time is taken to get to the steady-state condition when a new signal process begins. This time interval is referred to as the *acquisition time* of the recovery circuit and in switched communication networks or polling systems it is usually very important to keep this interval small, even at the expense of the larger steady-state phase jitter. One way to accommodate the conflicting objectives in designing a carrier recovery circuit is to specify a minimum phase-recovery bandwidth and then adjust other parameters of the system to minimize the steady-state phase jitter.

Another problem with a very narrow-band bandpass filter is in the inherent mistuning sensitivity, where mistuning is a result of inaccuracies in filter element values or a result of small inaccuracies or drift in the carrier frequency. This problem is avoided with tracking loop configurations since they lock onto the carrier frequency. One the other hand, tracking loops have some problems also, one of the more serious being the "hang-up" problem [9] whereby the nonlinear nature of the loop can produce some greatly prolonged acquisition times.

Although we have modeled the phase recovery problem in terms of a constant unknown carrier phase, it may be important in some situations to consider the presence of fairly rapid fluctuations in carrier phase (independent of the message process). Such fluctuations are often called *phase noise* and if the spectral density of these fluctuations has a greater bandwidth than that of the phase recovery circuits, there is a phase error due to the inability to track the carrier phase. Phase error of this type, even in steady state, becomes *larger* as the bandwidth of the recovery circuits decreases.

Another practical consideration is a π-radian phase ambiguity in the phase recovery circuits we have been discussing. The result is a polarity ambiguity in the coherently demodulated signal. In many cases this polarity ambiguity is unimportant, but otherwise some *a priori* knowledge about the message signal will have to be used to resolve the ambiguity.

For a QAM signal with $\beta(t) = a(t) + jb(t)$, where $a(t)$ and $b(t)$ are independent zero-mean stationary processes, we get

$$k_{yy}(t+\tau, t) = \tfrac{1}{2} \operatorname{Re} [\{k_{aa}(\tau) + k_{bb}(\tau)\} \exp(j2\pi f_0\tau)] + \tfrac{1}{2} \operatorname{Re} [\{k_{aa}(\tau) - k_{bb}(\tau)\} \cdot \exp(j4\pi f_0 t + j2\pi f_0\tau + j2\theta)] \quad (14)$$

and the situation is very similar to the VSB case (12). In this case where $a(t)$ and $b(t)$ are uncorrelated, the mean reference waveform has the correct phase, but the amplitude vanishes if the power levels in the I and Q channels are the same.

$$E[w(t)] = \tfrac{1}{2} [k_{aa}(0) - k_{bb}(0)] \operatorname{Re} [\exp(j4\pi f_0 t + j2\theta)]. \quad (15)$$

Hence, unless the QAM format is intentionally unbalanced, the squaring approach in Fig. 1 does not work. We briefly examine what happens when the squarer is replaced by a fourth-power device in the recovery schemes we have been considering. From (1), we can obtain

$$y^4(t) = \tfrac{1}{8} \operatorname{Re} [\beta^4(t) \exp(j8\pi f_0 t + j4\theta)] + \tfrac{1}{2} \operatorname{Re} [|\beta(t)|^2 \beta^2(t) \exp(j4\pi f_0 t + j2\theta)] + \tfrac{3}{8} |\beta(t)|^4. \quad (16)$$

Now if we use a bandpass filter tuned to $4f_0$ which passes only

the first term in (16), then the mean reference waveform at the filter output is

$$E[w(t)] = \tfrac{1}{4} \operatorname{Re}\left[\{\overline{a^4} - 3(\overline{a^2})^2\} \exp(j8\pi f_0 t + j4\theta)\right] \quad (17)$$

still assuming independent $a(t)$ and $b(t)$ and a balanced QAM format, i.e., $k_{aa}(0) = k_{bb}(0) = \overline{a^2}$. Hence, a mean reference waveform exists even in the balanced QAM case if a fourth-power device is used.[5]

One very popular QAM format is quadriphase-shift keying (QPSK) where the standard carrier recovery technique is to use a fourth-power device followed by a PLL or to use an equivalent "double" Costas loop configuration [3]. The QPSK format, with independent data symbols, can be regarded as two independent binary phase-shift-keyed (BPSK) signals in phase quadrature. In a nonbandlimited situation each BPSK signal can be regarded as DSB-AM where the message waveform has a rectangular shape characterized by $a(t) = \pm 1$. In this case, the complex envelope of the QPSK signal is characterized by $\beta(t) = (\pm 1 \pm j)/\sqrt{2}$ or $\beta(t) = \exp\ (j(\pi/4) + j(\pi/2)k)$ with $k = 0, 1, 2,$ or 3. The result is that $\beta^4\ (t) = -1$ and the $4f_0$ component in (16) is a pure sinusoid with no fluctuations in either phase or amplitude. For PSK systems with a larger alphabet of phase positions, the result of (17) cannot generally be used as the I and Q components are no longer independent. Analysis of the larger alphabet cases shows that higher-order nonlinearities are required for successful phase recovery [3], [10]. For any balanced QAM format, such as QPSK, the phase recovery circuits discussed here give a $\pi/2$-radian phase ambiguity. This problem is often handled by use of a differential PSK scheme, whereby the information is transmitted as a sequence of phase changes rather than absolute values of phase.

III. PAM TIMING RECOVERY

The receiver synchronization problem in baseband PAM transmission is to find the correct sampling instants for extracting a sequence of numerical values from the received signal. For a synchronous pulse sequence with a pulse rate of $1/T$, the sampler operates synchronously at the same rate and the problem is to determine the correct sampling phase within a T-second interval. The model for the baseband PAM signal is

$$x(t) = \sum_{k=-\infty}^{\infty} a_k g(t - kT - \tau) \quad (18)$$

where $\{a_k\}$ is the message sequence and $g(t)$ is the signaling pulse. We want to make an accurate determination of τ, from operations performed on $x(t)$. We assume that $g(t)$ is so defined that the best sampling instants are at $t = kT + \tau$; $k = 0, \pm 1, \pm 2, \cdots$. The objective is to recover a close replica of the message sequence $\{a_k\}$ in terms of the sequence $\{\hat{a}_k = x(kT + \hat{\tau})\}$, assuming a normalization of $g(0) = 1$. In the noise-free case, the difference between a_k and $\hat{a}_k$ is due to intersymbol interference which can be minimized by proper shaping of the data pulse $g(t)$. With perfect timing ($\hat{\tau} = \tau$), the intersymbol interference is

$$\hat{a}_k - a_k = \sum_{n \neq k} a_n g(kT - nT) \quad (19)$$

and this term can be made to vanish for pulses satisfying the Nyquist criterion, i.e., $g(nT) = 0$ for $n \neq 0$. For bandlimited Nyquist pulses, the intersymbol interference will not be zero when $\hat{\tau} \neq \tau$, and if the bandwidth is not significantly greater than the Nyquist bandwidth ($1/2T$) the intersymbol interference can be quite severe even for small values of timing error. The problem is especially acute for multilevel (nonbinary) data sequences where timing accuracy of only a few percent of the symbol period is often required.

Symbol timing recovery is remarkably similar in most respects to carrier phase recovery and we find that similar signal processing will yield suitable estimates of the parameter τ. In the discussion to follow, we assume that $\{a_k\}$ is a zero-mean stationary sequence with independent elements. The resulting PAM signal (18) is a zero-mean cyclostationary process, although there are no periodic components present [6]. The square of the PAM signal does, however, possess a periodic mean value.

$$E[x^2(t)] = \overline{a^2} \sum_k g^2(t - kT - \tau). \quad (20)$$

Using the Poisson Sum Formula [6], we can express (20) in the more convenient form of a Fourier series whose coefficients are given by the Fourier transform of $g^2(t)$.

$$E[x^2(t)] = \frac{\overline{a^2}}{T} \sum_{\ell} A_\ell \exp\left(\frac{j2\pi\ell}{T}(t - \tau)\right) \quad (21)$$

where

$$A_\ell \triangleq \int_{-\infty}^{\infty} G\left(\frac{\ell}{T} - f\right) G(f)\, df.$$

For high bandwidth efficiency, we are often concerned with data pulses whose bandwidth is at most equal to twice the Nyquist bandwidth. Then $|G(f)| = 0$ for $|f| > 1/T$ and there are only three nonzero terms ($\ell = 0, \pm 1$) in (21).

This result suggests the use of a timing recovery circuit of the same form as shown in Fig. 1, where now the bandpass filter is tuned to the symbol rate, $1/T$. Alternate zero crossings of $w(t)$, a *timing wave* analogous to the reference waveform in Section II, are used as indications of the correct sampling instants. Letting $H(1/T) = 1$, the mean timing wave is a sinusoid with a phase of $-2\pi\tau/T$, for a real $G(f)$.

$$E[w(t)] = \frac{\overline{a^2}}{T} \operatorname{Re}\left[A_1 \exp\left(j\frac{2\pi t}{T} - j\frac{2\pi\tau}{T}\right)\right]. \quad (22)$$

We see that the zero crossings of the mean timing wave are at a fixed time offset ($T/4$) relative to the desired sampling instants.

[5] Unless $a(t)$ and $b(t)$ are Gaussian processes, for then $\overline{a^4} = 3(\overline{a^2})^2$.

This timing offset can be handled by counting logic in the clock circuitry, or by designing $H(f)$ to incorporate a $\pi/2$ phase shift at $f = 1/T$.

The actual zero crossings of $w(t)$ fluctuate about the desired sampling instants because the timing wave depends on the actual realization of the entire data sequence. Different zero crossings result for different data sequences and for this reason the fluctuation in zero crossings is sometimes called *pattern-dependent* jitter to distinguish it from jitter produced by additive noise on the PAM signal. To evaluate the statistical nature of the pattern-dependent jitter, we need to calculate the variance of the timing wave. This is a fairly complicated expression in terms of $H(f)$ and $G(f)$ but it can be evaluated numerically to study the effects of a variety of parameters (bandwidth, mistuning, rolloff shape, etc.) relating to data pulse shape and the bandpass filter transfer function [11]. For a relatively narrow-band real $H(f)$ and real $G(f)$ bandlimited as mentioned previously, the variance expression has the form

$$\text{var } w(t) = C_0 + C_1 \cos \frac{4\pi}{T}(t - \tau) \tag{23}$$

where $C_0 \geqslant C_1 > 0$ are constants depending on $G(f)$ and $H(f)$. The cyclostationarity of the timing wave is apparent from this expression. As the bandwidth of $H(f)$ approaches zero, the value of C_1 approaches C_0 so that the variance has a great fluctuation over one symbol period. Note that the minimum variance occurs just at the instant of the mean zero crossings, hence the fluctuations in zero crossings are much less than would be expected from a consideration of the average variance of the timing wave over a symbol period. This again points out the error in disregarding the cyclostationary nature of the timing wave process as, for example, in using the power spectral density of the squarer output to analyze the jitter phenomenon.

The mean timing wave (22) can be regarded as a kind of *discriminator characteristic* or *S-curve* for measuring the parameter τ. For the bandlimited case we are discussing here, this S-curve is just a sinusoid, with a zero crossing at the true value of the parameter. Discrimination is enhanced by increasing the slope at the zero crossing. As this slope is proportional to A_1, we can see how the shape of the data pulse $g(t)$ affects timing recovery. From (21) we see that the value of A_1 depends on the amount of overlap of the functions $G(f)$ and $G(1/T - f)$, and hence it depends on the amount by which the bandwidth of $G(f)$ exceeds the $1/2T$ Nyquist bandwidth. With no excess bandwidth, $A_1 = 0$ and this method of timing recovery fails. The situation improves rapidly as the excess bandwidth factor increases from 0 to 100 percent. With very large increases in bandwidth there are more harmonic components in the mean timing wave, and its zero crossing slope can be further increased without increasing signal level by proper phasing of these components. On the other hand, there are systems where the spectral distribution is such that the fractional amount of energy above $1/2T$ is very small. An important case is that of (class IV) partial response signaling where the pulse shape is chosen to produce a spectral null at $1/2T$. This spectral null in combination with a sharp baseband rolloff characteristic can result in very small values of A_1. Although the class IV partial response format exhibits a relatively high tolerance to timing error [12], it is likely that some other recovery scheme may have to be used. Some of the proposed schemes [13], [14] closely resemble the data-aided approach discussed in Section IV.

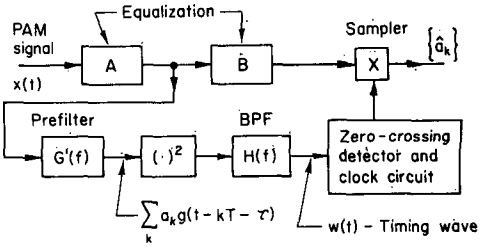

Fig. 3. Baseband PAM receiver with timing recovery.

Calculation of the statistical properties of the actual zero crossings of the timing wave is difficult. A useful approximation can be obtained by locating the zero crossings by linear extrapolation using the mean slope at the mean zero crossing. When this approach is used, the expression for timing jitter variance becomes [11]

$$\text{var } \hat{\tau}/T = (\overline{a^2})^{-1} \left(\frac{T}{2\pi A_1}\right)^2 (C_0 - C_1). \tag{24}$$

In order to reduce this pattern-dependent jitter, there is fortunately an attractive alternative to making the bandwidth of $H(f)$ very small, which increases acquisition time in the same manner as for carrier phase recovery circuits, or to making the bandwidth of $G(f)$ very large. There are symmetry conditions that can be imposed upon $H(f)$ and $G(f)$ that make $C_1 = C_0$ in (24), resulting in nonfluctuating zero crossings. These conditions are simply that $G(f)$ be a bandpass characteristic symmetric about $1/2T$, with a bandwidth not exceeding $1/2T$, and $H(f)$ be symmetric about $1/T$. The symmetry in $G(f)$ can be accomplished by prefiltering the PAM signal before it enters the squarer [11], [15]. Since the timing recovery path is distinct from the data signal path, the prefiltering can be performed without influencing the data signal equalization, as shown in the baseband receiver configuration of Fig. 3.

Although we are dealing with a baseband signal process, it is interesting to observe that the timing jitter problem can be studied by means of complex envelopes and decomposition into I and Q components, as in the carrier phase recovery case [16]. One way to do this is to let $\gamma(t)$ be the complex envelope of $g(t)$, relative to a frequency $f_0 = 1/2T$. This makes $\Gamma(f)$ bandlimited to $|f| < 1/2T$. Then, taking $\tau = 0$ for convenience, the output of the squarer is

$$x^2(t) = \tfrac{1}{2} \text{ Re } \left[\left\{ \sum_k (-1)^k a_k \gamma(t - kT) \right\}^2 \exp(j2\pi t/T) \right]$$

$$+ \tfrac{1}{2} \left| \sum_k (-1)^k a_k \gamma(t - kT) \right|^2. \tag{25}$$

The second term in (25) can be disregarded as not being passed by $H(f)$. The first term is expressed by a complex envelope relative to $f_0 = 1/T$. It is the quadrature component (imaginary part) of this complex envelope that produces timing jitter. This component is

$$b_Q(t) = \sum_k \sum_m a_k a_m (-1)^{k+m} c_I(t - kT) c_Q(t - mT) \tag{26}$$

where $c_I(t)$ and $c_Q(t)$ are the real and imaginary parts of $\gamma(t)$. The Fourier transform of (26), evaluated at $f = 0$, is

$$B_Q(0) = \int_{-\infty}^{\infty} \left| M\left(\nu - \frac{1}{2T}\right) \right|^2 C_I(\nu) C_Q(-\nu)\, d\nu = 0 \tag{27}$$

where $M(f) = \Sigma\, a_k \exp(-j2\pi kTf)$ is the transform of the data sequence. The integral (27) vanishes because, for a real $G(f)$ i.e., a real $\Gamma(f)$, the integrand is an odd function. The situation is similar to the VSB carrier signal case, where the spectrum of the quadrature component at the squarer output vanished at $f = 0$. We see here also that the particular shape of $H(f)$ will have a major influence when calculating the jitter variance because the spectrum of the jitter-producing component goes to zero just at the center of its passband.

IV. MAXIMUM-LIKELIHOOD PARAMETER ESTIMATION

The foregoing carrier- and bit-synchronization circuits were developed on a rather heuristic basis and a natural question arises as to how much improvement in parameter estimation could result from the choice of other circuit configurations or circuit parameters. It seems natural to regard θ and τ as unknown but nonrandom parameters which suggests the maximum-likelihood (ML) estimator as the preferred strategy [17]. Some authors have used the maximum *a posteriori* probability (MAP) receiver by modeling θ and τ as random parameters with specified *a priori* probability density functions. However, in most situations the *a priori* knowledge about θ is only accurate to within many carrier cycles, or in the case of τ, to within many symbol periods. As our concern is with estimation modulo 2π for θ or modulo T for τ, we would use a "folded" version of the *a priori* density functions, resulting in a nearly uniform distribution over the interval. In this case, the ML approach estimates and MAP estimates would be essentially identical. We find that the phase and timing-recovery circuits based on the ML approach may not be drastically different from the circuits already considered. In fact, under the proper conditions, the circuits we have examined can be close approximations to ML estimators. One of the main advantages of the ML approach, in addition to suggesting appropriate circuit configurations, is that simple lower bounds on jitter performance can be developed to serve as benchmarks for evaluating performance of the actual recovery circuits employed.

In this section we begin with discussion of ML carrier phase recovery with a rather general specification of the message signal process. We show that the Costas loop, or the equivalent squarer/BPF, can be designed to closely approximate the ML phase estimator. Then we present a similar development for ML estimation of symbol timing for a baseband PAM signal. We introduce the idea of using information about the data sequence to aid the timing recovery process and we later make comparisons to show the effectiveness of such data-aided schemes. Extension of the idea to joint recovery of both carrier phase and symbol timing parameters is discussed in Section V.

To formulate the problem in terms of ML estimation, we require that the receiver perform operations on a T_0-second record of the received signal, $z(t) = y(t, \theta) + n(t)$, to estimate the parameter θ, assumed essentially constant over the T_0 -second interval. This interval is called the *observation interval* and the T_0 parameter would be selected in accordance with acquisition time requirements. Estimation procedures based on data from a single observation interval will be referred to as *one-shot estimation.* We find that the one-shot ML estimators lead to the simplest methods for evaluating jitter performance. On the other hand, the preferred implementation of recovery circuits is usually in the form of tracking loops where the parameter estimates are being continuously updated. Fortunately, it is a relatively simple matter to relate the rms one-shot estimation error to the steady-state error of the tracking loop and the loop bandwidth is directly related to the T_0 parameter.

We shall assume that the additive noise $n(t)$ is Gaussian and white with a double-sided spectral density of N_0 W/Hz. Initially we consider the situation where $y(t, \theta)$ is completely known except for the parameter θ. The resulting likelihood function, with argument $\hat{\theta}$ which can be regarded as a trial estimate of the parameter, is given by

$$L(\hat{\theta}) = \exp\left\{ -\frac{1}{2N_0} \int_{T_0} [z(t) - y(t, \hat{\theta})]^2\, dt \right\}. \tag{28}$$

The ML estimate is the value of θ which minimizes the integral in (28). This integral expresses the signal space distance between the functions $z(t)$ and $y(t, \hat{\theta})$ defined on the interval T_0 [6], [17]. Expanding the binomial term in (28), we see that

$$\Lambda(\hat{\theta}) = \ln L(\hat{\theta}) = \frac{1}{N_0} \int_{T_0} z(t) y(t, \hat{\theta})\, dt + \text{constant} \tag{29}$$

since $z^2(t)$ is independent of $\hat{\theta}$, and if $\hat{\theta}$ is a time shift or phase shift parameter, then the integral of $y^2(t, \hat{\theta})$ over a relatively long T_0 interval would have only a small variation with $\hat{\theta}$. The first term in (29) is often called the correlation between the received signal $z(t)$ and the reference signal $y(t, \hat{\theta})$ so that in this "known-signal" case, the ML receiver is a correlator, and $\hat{\theta}$ is varied so as to maximize the correlation.

When $y(t, \hat{\theta})$ contains random message parameters, the appropriate likelihood function for estimating θ is obtained by averaging $L(\hat{\theta})$—not $\ln L(\hat{\theta})$—over these message parameters. We shall illustrate the method using the example of carrier phase estimation on a DSB signal where the modulating signal $a(t)$ is a zero-mean, Gaussian random process with a substanti-

ally flat spectrum bandlimited to W Hz. Finding the expectation of $L(\hat{\theta})$ with respect to the Gaussian message process can be done without great difficulty by making a Karhunen-Loeve expansion of the process to give a series representation with independent coefficients [18]. The result of this averaging gives a log-likelihood function closely approximated by

$$\Lambda(\hat{\theta}) = \int_{T_0} [\mathrm{Re}\, \alpha(t) \exp(-j\hat{\theta})]^2\, dt$$
$$= \tfrac{1}{2}\, \mathrm{Re}\left[\exp(-j2\hat{\theta}) \int_{T_0} \alpha^2(t)\, dt\right]$$
$$+ \frac{1}{2} \int_{T_0} |\alpha(t)|^2\, dt \qquad (30)$$

where $\alpha(t)$ is the complex envelope, relative to f_0, of the received signal. We ignore the second integral in (30) as it is independent of $\hat{\theta}$. (30) suggests a practical implementation of the ML phase estimator. Consider a receiver structure which produces the complex signal

$$\lambda(t) = \int_{t-T_0}^{t} \alpha^2(s)\, ds \triangleq \rho(t) \exp(j2\tilde{\theta}(t)). \qquad (31)$$

The integral in (31) is the convolution product of $\alpha^2(t)$ and a T_0-second rectangle, hence $\lambda(t)$ can be regarded as the complex envelope of the output of the squarer/bandpass filter configuration of Fig. 1. In this case, $H(f)$ corresponds to a sinc $(T_0 f)$ shape centered at $2f_0$. Writing $\lambda(t)$ in polar form as shown in (31), we see that the corresponding term in (30) is maximized, at any t, by choosing $\hat{\theta} = \tilde{\theta}(t)$. In other words, by suitably designing the shape of the bandpass transfer function, the simple structure of Fig. 1 is a ML phase estimator, in the sense that the instantaneous phase of the timing wave output is the best estimate of the DSB carrier phase based on observations over only the past T_0 seconds. The phase jitter can be evaluated approximately by the same method leading to (11), with the result that

$$\mathrm{var}\, \phi = \frac{2N_0}{ST_0} = \frac{1}{WT_0}\left(\frac{S}{N}\right)^{-1} \qquad (32)$$

where $S = k_{aa}(0)$ is the signal power and $N = 2N_0 W$ is noise power over the signal band.

The tracking loop version of this phase estimator is developed by forming a loop error signal proportional to the derivative of Λ with respect to $\hat{\theta}$. Then, as the loop action tends to drive the error signal to zero, the resulting value of $\hat{\theta}$ should correspond to a maximum of Λ. Since the control voltage $v(t)$ for a VCO normally controls frequency, rather than phase, we suppress the integration in (30) and let

$$v(t) = \frac{1}{8}\frac{\partial}{\partial\hat{\theta}} [\mathrm{Re}\, \alpha(t) \exp(-j\hat{\theta})]^2$$
$$= \tfrac{1}{8}\, \mathrm{Re}\, [\alpha^2(t) \exp(-j2\hat{\theta} - j\pi/2)] \qquad (33)$$

which is the same control voltage that appears in the Costas loop (9) and Fig. 2(b), with the normalization, $A = 1$. If the VCO had a voltage-controlled phase, rather than frequency, we would include the T_0 -second integration effect by means of the loop filter. However, in this case we simply let $F(f) = 1$ and rely on the integration inherent in the VCO. The parameter T_0 is related to loop performance by adjusting the loop gain factor M, which is proportional to loop bandwidth, so that the steady-state jitter variance for the loop is identical to (32) for the nontracking implementation.

For the DSB signal with additive noise, we have

$$\alpha(t) = [a(t) + u_I(t) + ju_Q(t)] \exp(j\theta) \qquad (34)$$

where u_I and u_Q are the I and Q components of noise relative to the carrier phase θ. Letting the VCO gain constant be M (hertz/volt) so that $\dot{\hat{\theta}}(t) = 2\pi M v(t)$ and assuming a high signal-to-noise ratio so that the second-order noise effects can be neglected, a linearized loop equation for phase error, $\phi = \hat{\theta} - \theta$, appears as

$$\dot{\phi}(t) + 2\pi B_L \phi(t) = \frac{2\pi B_L}{S} u_Q(t) a(t) - \frac{2\pi B_L}{S} b(t)\phi(t). \qquad (35)$$

The difficult part of solving this equation to get the steady-state variance of ϕ is the second driving term where $b(t) \triangleq a^2(t) - S$ and $\phi(t)$ are clearly not independent. It turns out however, that if the loop bandwidth parameter $B_L = \frac{1}{4}$ MS is sufficiently small compared to signal bandwidth W, then this term can be neglected. The other excitation term $u_Q a$ can be treated as white noise with a spectral density of $2N_0 S$, and the steady-state variance of ϕ can be determined by conventional frequency-domain techniques. The result is

$$\mathrm{var}\, \phi = 2\pi B_L N_0 / S \qquad (36)$$

and, equating (36) and (32) we find that $B_L = 1/\pi T_0$ is the relation sought between observation interval and tracking loop bandwidth.

Turning now to ML timing recovery for the baseband PAM signal, with $\hat{\tau}$ replacing $\hat{\theta}$, and using (18) for $y(t, \hat{\tau})$, the log-likelihood function for the case of a known signal (29) becomes

$$\Lambda(\hat{\tau}) = \frac{1}{N_0} \sum_{k=-\infty}^{\infty} a_k q_k \qquad (37)$$

where

$$q_k(\hat{\tau}) = \int_{T_0} z(t) g(t - kT - \hat{\tau})\, dt.$$

It is possible to use this expression directly for timing recovery in a situation where a relatively long sequence, say K, of known symbols is transmitted as a preamble to the actual message sequence. The receiver would store the K-symbol sequence and attempt to establish the correct timing before the end of the preamble. The idea can also be used during message transmission if the symbols are digitized, so that the receiver makes decisions as to which of the finite number of possible symbols

have been transmitted. The receiver decisions are then assumed to be correct, at least for the purposes of timing recovery. The bootstrap type of operation is referred to as *decision-directed* or *data-aided* timing recovery and it has received extensive study for both symbol timing and carrier phase recovery [19]-[22]. In the following, we shall use the term "data-aided" to refer to both modes of operation, i.e., the start-up mode where a known data sequence is being transmitted and the tracking mode where the symbol detector output sequence is used.

For recovery strategies which are not data aided, we need to average the likelihood function (37) over the random data variables. If we assume that the $\{a_k\}$ are independent Gaussian random variables and also that the data pulses have unit energy and are orthogonal over the T_0 interval, i.e.,

$$\int_{T_0} g(t-kT)g(t-mT)\,dt = \delta_{km} \tag{38}$$

then the log-likelihood function is given by

$$\Lambda(\hat{\tau}) = \frac{1}{2N_0} \sum_{k=-\infty}^{\infty} q_k{}^2(\hat{\tau}) \tag{39}$$

where the q_k are the same quantities defined in (37). Although the Gaussian density is obviously not an accurate model for digital data signals, we want to consider it here because it provides the link between the ML estimators and the estimators of Sections II and III based on statistical moment properties. It is the Gaussian assumption that leads to the square-law type of nonlinearity. If we consider equiprobable binary data, for example, the corresponding log-likelihood function is [23]-[25]

$$\Lambda(\hat{\tau}) = \sum_k \ln\cosh\frac{1}{N_0} q_k(\hat{\tau}) \tag{40}$$

and since $\ln\cosh x \cong \frac{1}{2}x^2$ for small x, the square-law nonlinearity is near optimum at the lower signal-to-noise ratios. The log-likelihood function for equiprobable independent multilevel data has also been derived [25], [26]. When the Gaussian assumption is used for the data, it is also possible to consider correlated data as well as nonorthogonal pulses, i.e., when (38) does not hold. Both of these effects can be dealt with by replacing the q_k-sequence in (39) by a linear discrete-time filtered version of this sequence [27]. In summary, we find that recovery circuits based on the Gaussian-distributed data assumption are somewhat simpler than the optimum circuits and in most situations the jitter performance is not appreciably worse. We note that the method for evaluating rms jitter, presented in Section VI, does not depend on the particular kind of density function used to characterize the data.

When it comes to implementation of receivers based on (37) and (39) for the data-aided (DA) or nondata-aided (NDA) strategies, we usually resort to an approximation which involves replacing the infinite sum by a K-term sum, where $KT = T_0$, and replacing the finite integration interval by an infinite interval. Then the approximate implementable, log-likelihood function in the DA case is taken as

$$\tilde{\Lambda}(\hat{\tau}) = \frac{1}{N_0} \sum_{k=0}^{K-1} a_k \tilde{q}_k \tag{41}$$

where

$$\tilde{q}_k(\hat{\tau}) = \int_{-\infty}^{\infty} z(t)g(t-kT-\hat{\tau})\,dt.$$

With this approximation, the integral is a convolution integral, and $\tilde{q}_k$ can be interpreted as the sampled (at $t = kT + \hat{\tau}$) output of a matched filter having the impulse response $g(-t)$. The same approximation is used for the NDA case (39) and the orthogonality condition (38) can be interpreted to mean that the matched filter response to a single data pulse is a pulse satisfying the Nyquist criterion. This approximation, which leads to relatively simple implementations for the recovery circuits, does introduce a degradation from the idealized ML performance. An interesting interpretation of the effect of the approximation is that it introduces a pattern-dependent component of jitter, as discussed in Section VI.

For tracking loop implementation of these timing recovery strategies, we use a voltage-controlled clock (VCC) driven by

$$v(t) = \frac{\partial}{\partial\hat{\tau}}[a_k\tilde{q}_k] = -a_k \int_{-\infty}^{\infty} z(t)\dot{g}(t-kT-\hat{\tau})\,dt \tag{42}$$

for the DA case, and

$$v(t) = \frac{\partial}{\partial\hat{\tau}}[\tfrac{1}{2}\tilde{q}_k{}^2] = -\left[\int_{-\infty}^{\infty} z(t)g(t-kT-\hat{\tau})\,dt\right] \cdot \left[\int_{-\infty}^{\infty} z(t)\dot{g}(t-kT-\hat{\tau})\,dt\right] \tag{43}$$

for the NDA case. The K-term summation is suppressed, being replaced by the integration action of the VCC as in the case of the Costas loop phase recovery circuit discussed earlier. Similar also is the relation between T_0 and the loop bandwidth, the loop gain being adjusted so that the steady-state variance of timing jitter is the same as for one-shot estimation in a single observation interval. The result is also $B_L = 1/\pi T_0$ [26]. The tracking loop configurations are evident from inspection of (42) and (43).

One structure will serve for both strategies by incorporating a DA/NDA mode switch as shown in Fig. 4. This could be quite useful in a system that uses the DA strategy on a message preamble, then switches to the NDA strategy when the message symbols begin. Notice that the NDA configuration is remarkably like a Costas loop, which suggests the existence of an equivalent realization using a square-law device. This alternative and equivalent form is shown in Fig. 5. The corresponding implementation of (40) for NDA recovery with binary data involves the same structure as shown in Fig. 4, except that a $\tanh(\cdot)$ nonlinearity is incorporated into the upper path of the NDA loop [25].

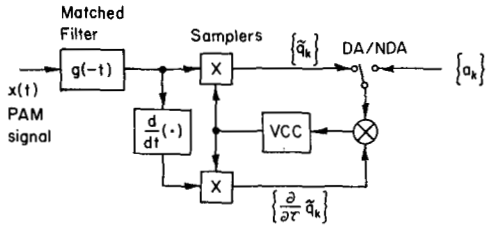

Fig. 4. ML baseband PAM timing recovery circuit.

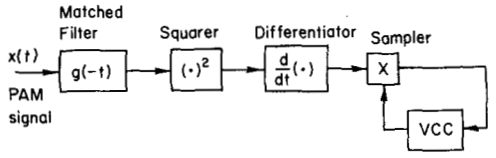

Fig. 5. Alternative implementation of NDA baseband timing recovery.

V. JOINT RECOVERY OF CARRIER PHASE AND SYMBOL TIMING

When a carrier system, such as VSB/PAM or QAM/PAM is used to transmit a digital data signal, we have the possibility of jointly estimating the carrier phase and symbol timing parameters. Such a strategy certainly cannot be worse than estimating the parameters individually and, in some cases, join estimation gives remarkable improvements. Some authors have extended the idea to joint estimation of the data sequence and the two timing parameters [28]-[30]. We shall not consider this latter possibility here, but shall consider both DA and NDA joint parameter estimation. Our conjecture is that, in the majority of applications, DA recovery performance differs little from that of joint estimation of data and timing parameters.

We consider first the QAM/PAM data signal case where we want to estimate θ and τ in

$$y(t;\theta,\tau) = \mathrm{Re}\left[\left\{\sum_k a_k g(t-kT-\tau) + jb_k h(t-kT-\tau)\right\} \cdot \exp(j\theta)\exp(j2\pi f_0 t)\right] \tag{44}$$

from receiver measurements on $z(t) = y(t) + n(t)$ over a T_0-second observation interval. The implementable version of the log-likelihood function for the DA case is

$$\tilde{\Lambda}(\hat{\theta},\hat{\tau}) = \frac{1}{N_0}\sum_{k=0}^{K-1} a_k\tilde{q}_k + b_k\tilde{p}_k \tag{45}$$

where

$$\tilde{q}_k(\hat{\theta},\hat{\tau}) = \mathrm{Re}\left[\exp(-j\hat{\theta})\int_{-\infty}^{\infty}\alpha(t)g(t-kT-\hat{\tau})\,dt\right]$$

$$\tilde{p}_k(\hat{\theta},\hat{\tau}) = \mathrm{Re}\left[-j\exp(-j\hat{\theta})\int_{-\infty}^{\infty}\alpha(t)h(t-kT-\hat{\tau})\,dt\right]$$

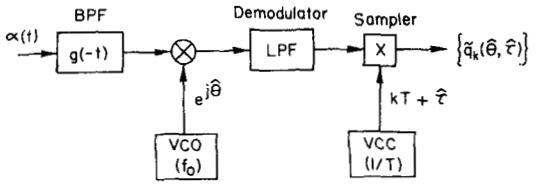

Fig. 6. Receiver implementation of the $\tilde{q}_k(\hat{\theta},\hat{\tau})$ test statistic.

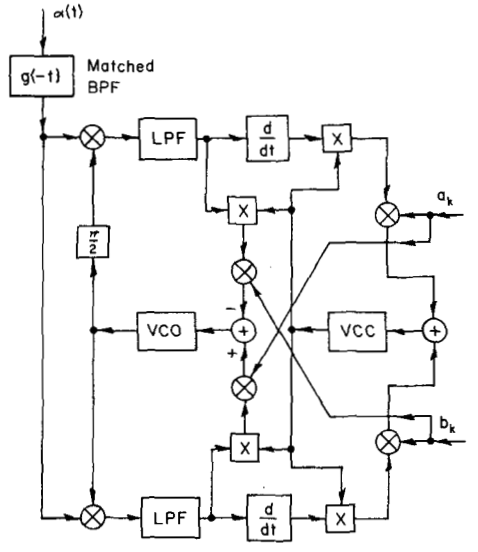

Fig. 7. Data-aided QAM joint tracking loop for carrier phase and symbol timing.

and $\alpha(t)$ is the complex envelope of the received signal. The $\tilde{q}_k$ quantities are interpreted as the sampled (at $t = kT + \hat{\tau}$) output of a coherent demodulator (operating at a phase $\hat{\theta}$) whose input is a bandpass filtered version of the received signal. The receiver implementation for these quantities is shown in Fig. 6, and a similar implementation would provide the $\tilde{p}_k$ quantities.

For the joint tracking loop the partial derivatives of $\tilde{\Lambda}$ with respect to $\hat{\theta}$ and $\hat{\tau}$, without the K-term summation, are used to update the VCO and VCC frequencies once every T seconds. For the normal QAM case we let $h(t) = g(t)$ and some simplifications result, for then $\partial\tilde{q}_k/\partial\hat{\theta} = \tilde{p}_k$ and $\partial\tilde{p}_k/\partial\hat{\theta} = -\tilde{q}_k$. The $\partial\tilde{q}_k/\partial\hat{\tau}$ and $\partial\tilde{p}_k/\partial\hat{\tau}$ quantities are obtained by differentiating the I and Q baseband signals before sampling. The complete tracking loop implementation for the DA case is shown in Fig. 7.

For balanced QAM/PAM with identical pulse shapes and statistically identical independent data in the I and Q channels, the NDA mode of recovery fails [27] because the NDA log-likelihood function is

$$\tilde{\Lambda}(\hat{\theta},\hat{\tau}) = \frac{1}{2N_0}\sum_{k=0}^{K-1}\tilde{q}_k^2 + \tilde{p}_k^2 \tag{46}$$

and this is independent of $\hat{\theta}$ under the previous assumptions. Fortunately, a simple modification makes the NDA mode effective. This modification is $h(t) = g(t \pm T/2)$ and the format is called *staggered* QAM (SQAM). The implementation is similar, but somewhat more complex, to that shown in Fig. 7

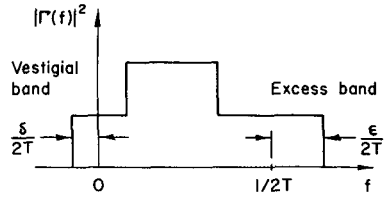

Fig. 8. Energy spectral density for VSB data pulse.

because additional samplers are needed for sampling at both $kT + \hat{\tau}$ and $kT + \hat{\tau} \pm T/2$.

Considering now the one-dimensional VSB/PAM case, the log-likelihood functions are the same as (45) and (46) with the b_k and $\tilde{p}_k$ quantities omitted, and with

$$\tilde{q}_k(\hat{\theta}, \hat{\tau}) = \mathrm{Re}\left[\exp(-j\hat{\theta}) \int_{-\infty}^{\infty} \alpha(t)\gamma * (t - kT - \hat{\tau})\, dt \right]. \tag{47}$$

In (47), $\gamma(t) = g(t) + j\,\tilde{g}(t)$ is the complex envelope of a single, unit-amplitude carrier data pulse. The orthogonality condition corresponding to (38) for the NDA case can be satisfied by a pulse whose energy spectrum $|\Gamma(f)|^2$ has a shape of the form shown in Fig. 8, exhibiting a Nyquist-type of symmetry in both of its rolloff regions.

In contrast to the QAM case, we find that the VCO and VCC frequency-control voltages should be derived as linear combinations of the partial derivatives of $\tilde{\Lambda}$ with respect to $\hat{\theta}$ and $\hat{\tau}$, i.e., there is a coupling between the parameter estimates. To show this, we consider the approximate solution for the one-shot estimator based on a Taylor series expansion of $\tilde{\Lambda}$ about trial values of θ_0 and τ_0. If these values are sufficiently close to the true values, then we can take as refined estimates, θ_1 and τ_1, the solutions of

$$\begin{bmatrix} \tilde{\Lambda}_{\theta\theta}(\theta_0, \tau_0) & \tilde{\Lambda}_{\theta\tau}(\theta_0, \tau_0) \\ \tilde{\Lambda}_{\theta\tau}(\theta_0, \tau_0) & \tilde{\Lambda}_{\tau\tau}(\theta_0, \tau_0) \end{bmatrix} \begin{bmatrix} \theta_1 - \theta_0 \\ \tau_1 - \tau_0 \end{bmatrix} = \begin{bmatrix} -\tilde{\Lambda}_{\theta}(\theta_0, \tau_0) \\ -\tilde{\Lambda}_{\tau}(\theta_0, \tau_0) \end{bmatrix} \tag{48}$$

where the subscripts in (48) denote partial derivatives. The solution of (48) is greatly simplified if the 2 × 2 matrix is replaced by its mean value, and this is valid at moderately high signal-to-noise ratio and moderately long observation intervals. Then we have a simple form of estimation given by

$$\begin{bmatrix} \theta_1 - \theta_0 \\ \tau_1 - \tau_0 \end{bmatrix} = -A^{-1} \begin{bmatrix} \tilde{\Lambda}_{\theta}(\theta_0, \tau_0) \\ \tilde{\Lambda}_{\tau}(\theta_0, \tau_0) \end{bmatrix} \tag{49}$$

where the 2 × 2 matrix A is the expected value of the matrix in (48). The A matrix can be regarded as a generalization of the A_1 quantity in (22), (24) for the single-parameter recovery problem. For the joint tracking loop, the VCO and VCC control voltages are linear combinations of the Λ_θ and Λ_τ quantities (without the K-term summation) as characterized by the inverse of the A matrix. In the QAM and SQAM cases, the A matrix is diagonal so that no coupling is required, but for VSB there are strong off-diagonal terms. It has been shown [30]-[32] that loop convergence rates can be substantially improved by incorporating this coupling on the control signals.

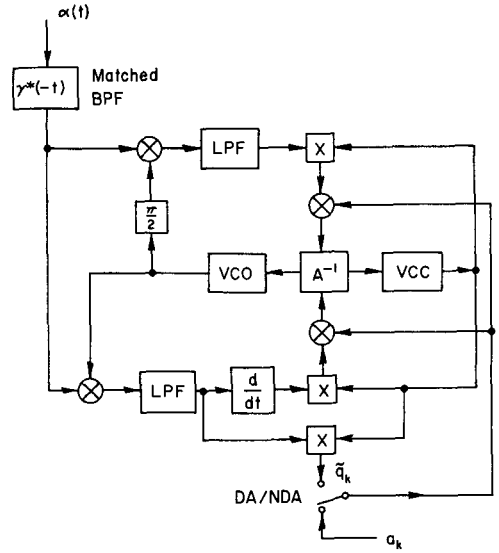

Fig. 9. One-dimensional joint tracking loop.

A block diagram for the one-dimensional joint tracking loop is shown in Fig. 9. A DA/NDA mode switch is shown in the diagram, but it must be recognized that the A^{-1} coupling matrix is a compromise value in either one or both modes because the A matrix is quite different for the DA and NDA cases, as discussed in the next section. Another practical consideration is that the configuration of Fig. 9 can also be used for QAM and SQAM with some loss in performance. Here we just eliminate the $\tilde{p}_k$ quantities in (45) or (46) and use the matched BPF for the I-channel pulse, i.e., let $\gamma^*(-t) = g(-t)$. The loss in performance will be about 3 dB or greater, depending on signal-to-noise ratio and on which parameter is considered, as there are some cancellations of pattern-dependent jitter in the configuration of Fig. 7 which are not possible in that of Fig. 9.

VI. PERFORMANCE OF TIMING RECOVERY SCHEMES

A convenient approach to evaluating timing recovery circuit performance, is to derive expressions for rms phase-and timing-jitter directly from (49), or its one-dimensional counterpart for individual estimation of θ or τ. For these calculations we assume θ_0 and τ_0 are the true values of the parameters, so that the left-hand side of (49) gives the jitter variables. The equation is linearized in the sense that the jitter variables depend linearly on the receiver measurements $\tilde{\Lambda}_\theta$ and $\tilde{\Lambda}_\tau$. As in all analyses of this type, the results are accurate if the jitter is relatively small, which in this case generally means a moderately high signal-to-noise ratio and a moderately long observation interval. This approach affords an effective means to study jitter performance with respect to the values of all system parameters, such as signal-to-noise ratio, T_0 (or K), excess bandwidth, and pulse shape. It also allows comparison of jitter performance of the various modulation formats and evaluation of the DA strategy relative to the NDA strategy.

Another important aspect that can be examined from the rms jitter calculations is the effect of the implementation approximations $\Lambda \to \tilde{\Lambda}$ and $q_k \to \tilde{q}_k$. Let us take as an illustrative example the one-dimensional case of baseband timing recovery. For the implementable DA case we get

$$\text{var}\,\hat{\tau} = \text{var}\,\tilde{\Lambda}_\tau(\tau)[E\{\tilde{\Lambda}_{\tau\tau}(\tau)\}]^{-2}$$

$$= \frac{N_0}{a^2 KD} + \frac{\sum_{k=0}^{K-1} \sum_{m \in K'} \dot{r}^2(mT - kT)}{K^2 D^2} \tag{50}$$

where

$$r(t) = \int_{-\infty}^{\infty} g(s+t)g(s)\,ds$$

and

$$D = -\ddot{r}(0) = \int_{-\infty}^{\infty} \dot{g}^2(t)\,dt$$

is the energy in the time derivative of a single data pulse. The notation $m \in K'$ in (50) means that the sum is taken for all $k < 0$ and $\geqslant K$, i.e., just the terms not used in the other sum. The first term in (50) is seen to vary inversely with K, signal-to-noise ratio, and the energy in $\dot{g}(t)$. From this it is obvious that "sharp-edged" pulses can give excellent timing recovery performance. In fact, the entire denominator in the first term can be taken approximately as the expected value of the energy of the time-derivative of the received PAM signal over the T_0 -second observation interval.

Another significant aspect of the first term in (50) is that it gives the entire jitter variance if Λ and q_k are used instead of $\tilde{\Lambda}$ and $\tilde{q}_k$. In other words, the second term in (50) gives the additional jitter variance resulting from the practical implementation considerations. This term does not depend on the noise level and it can be regarded as the effect of the pattern-dependent component of jitter. It varies inversely with K^2 since the numerator is essentially a constant even for moderately small values of K. Thus we see that if severe requirements are placed on acquisition time (small K), then the effect of this pattern-dependent term is apt to dominate. Otherwise, for larger K, the first term may be dominant and the difference between true ML estimation and its implementable approximation may be negligible.

Although the variance expressions are somewhat different, the same general conclusions about the two types of jitter terms hold for the NDA timing recovery case, and for joint timing and carrier phase recovery [26]. Another physical interpretation, in the case of carrier phase recovery, is as follows. There are two random interference components producing jitter in the carrier phase tracking loop; one is due to the additive noise on the input signal, and the other due to the message sidebands of the carrier signal. It is primarily the quadrature components of these interferences that cause the jitter. The quandrature noise has a flat spectrum about the carrier frequency, so the jitter variance due to this interference should vary in direct proportion to the loop bandwidth. The quadrature data dependent interference has a spectral null in the vicinity of the carrier frequency, so we would expect jitter variance due to this effect to increase faster than linearly with loop bandwidth. Hence, for a given signal-to-noise ratio, and for a large enough loop bandwidth (rapid acquisition) we would expect the pattern-dependent term to dominate.

The relative performance of the different modulation formats is governed primarily by the size of the elements of the A matrix. Also, the A matrix almost completely characterizes the difference in performance of the DA and NDA strategies. For example, in the NDA/VSB case, the A matrix elements contain terms proportional to the integral of the product of $|\Gamma(f)|^2$ and $|\Gamma(1/T - f)|^2$ [26], [27]. Thus the size of the terms depends on the amount of overlap of the pulse energy spectrum and its frequency-translated version, and this depends on the amount of excess bandwidth available. For the staircase shape shown in Fig. 8, the term is directly proportional to the excess bandwidth factor ϵ. As a result, the rms jitter has a $1/\epsilon$ behavior and performance is unacceptable at very small excess bandwidth. On the other hand, the A matrix for DA/VSB has a completely different dependence on $\Gamma(f)$ and it results in a finite jitter variance at $\epsilon = 0$ and a much slower rate of decrease for increasing ϵ. In fact, with excess bandwidths over about 30 percent, the difference between DA and NDA jitter is usually small enough so that it may not be worth the additional circuit complexity to implement the DA strategy [26].

Finally, the variance expressions can be used to compare performance with different pulse shapes or to solve for optimum pulse shapes. There is no universally optimal pulse shape to cover the variety of cases discussed here. For one thing, we can see from (50) that the optimal pulse shape can depend on the signal-to-noise ratio. It is found, however, that the staircase, or "double-jump," rolloff pictured in Fig. 8 is optimal in certain cases and tends to be desirable in all cases. For example, it is better than the familiar "raised-cosine" rolloff, by a factor of approximately 2 in jitter variance [26]. It is interesting to note that such pulse shaping is also optimal from the standpoint of providing maximal immunity to timing or phase offsets [33], [34]. This fact accentuates the importance of proper pulse shaping for overall system performance.

APPENDIX

COMPLEX ENVELOPE REPRESENTATION OF SIGNALS

A straightforward extension of the familiar two-dimensional phasor representation for sinusoidal signals has proven to be a great convenience for dealing with carrier-type data signals where properties of amplitude and phase shift are of special significance. As a supplement to this paper only the most basic relationships are presented. More details and the derivations of the formulas can be found in some texts on communication systems or in [6, chs. 4 and 7].

An arbitrary signal $x(t)$ can be represented exactly by a complex envelope $\gamma(t)$ relative to a "center" frequency f_0, which for modulated-carrier signals is usually, but not necessarily, taken as the frequency of the unmodulated carrier.

$$x(t) = \text{Re}\,[\gamma(t) \exp(j2\pi f_0 t)]. \tag{A-1}$$

Expressing the complex number $\gamma(t)$ in polar form reveals directly the instantaneous *amplitude* $\rho(t)$ and *phase* $\theta(t)$ of the signal.

$$\gamma(t) = \rho(t) \exp [j\theta(t)] = c_I(t) + jc_Q(t). \tag{A-2}$$

In some situations, the rectangular form of $\gamma(t)$ in (A-2) has a more direct bearing on the problem as it decomposes the signal into its *in-phase* and *quadrature* (I and Q) components.

$$x(t) = c_I(t) \cos 2\pi f_0 t - c_Q(t) \sin 2\pi f_0 t. \tag{A-3}$$

Equation (A-1) might be regarded as one part of a transform pair. The other equation, i.e., how to get $\gamma(t)$, given $x(t)$, presents a small problem. Due to the nature of the "real part of" operator Re, there is not a unique $\gamma(t)$ for a given $x(t)$. We solve this problem by making the definition

$$\gamma(t) = [x(t) + j\hat{x}(t)] \exp (-j2\pi f_0 t) \tag{A-4}$$

where $\hat{x}(t)$ is the Hilbert transform of $x(t)$. The prescription for getting $\gamma(t)$ from $x(t)$ is especially simple in the frequency domain. The Fourier transform $\Gamma(f)$ is obtained by doubling $X(f)$, suppressing all negative-frequency values, and frequency-translating the result downward by an amount f_0. Incidentally, using this approach no narrow-band approximations concerning $x(t)$ are necessary, and an arbitrary value of f_0 can be selected.

We now characterize the two most important signal processing operations, filtering and multiplication, in terms of equivalent operations on complex envelopes. Consider first the time-invariant bandpass filtering operation in Fig. 10. We express the bandpass transfer function $H(f)$ in terms of an equivalent low-pass transfer function $\Omega(f)$, according to

$$H(f) = \Omega(f - f_0) + \Omega^*(-f - f_0) \tag{A-5}$$

$\Omega(f)$ is not necessarily a physical transfer function. If $H(f)$ exhibits asymmetry about f_0, then $\Omega(f)$ is asymmetric about $f = 0$ and the corresponding impulse response $\omega(t)$ is complex. In fact, $\omega(t)$ is precisely the complex envelope of $2h(t)$, where $h(t)$ is the real impulse response of the bandpass filter.

Straightforward manipulation shows that the input-output relation for complex envelopes is also a time-domain convolution

$$\beta(t) = [\omega \otimes \gamma](t) \tag{A-6}$$

and this result is general because of our particular method for defining the complex envelope in (A-4). If we express $\omega(t)$ in terms of its real and imaginary parts, $\omega(t) = p_I(t) + j\, p_Q(t)$, then the two-port bandpass filtering operation can be represented by a real four-port filter with separate ports for the I and Q input and output. The four-port filter is a lattice configuration involving the transfer functions $P_I(f)$ and $P_Q(f)$ as shown in Fig. 10.

$$P_I(f) = \tfrac{1}{2}\, \Omega(f) + \tfrac{1}{2}\, \Omega^*(-f)$$

$$P_Q(f) = \frac{1}{2j}\, \Omega(f) - \frac{1}{2j}\, \Omega^*(-f). \tag{A-7}$$

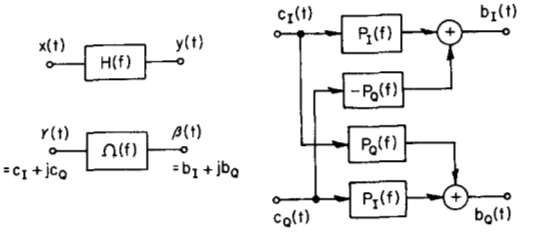

Fig. 10. Bandpass filtering and low-pass equivalent operation on complex envelope signals.

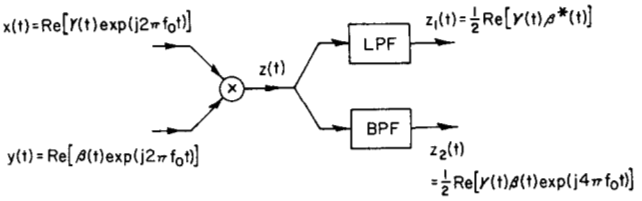

Fig. 11. Low-frequency and $2f_0$ terms of product of two bandpass signals.

Notice that if $H(f)$ is symmetric about f_0, then $P_Q(f) = 0$ (this is the definition of symmetry for a bandpass filter) and there is no cross coupling of the I and Q components in the filtering operation.

Next we consider the output of a multiplier circuit, $z(t) = x(t)y(t)$, when the two inputs are expressed in complex envelope notation. From (A-8), the multiplier output consists of two terms, one representing low-frequency components and the other representing components around $2f_0$.

$$z(t) = \text{Re } [\gamma(t) \exp (j2\pi f_0 t)] \text{ Re } [\beta(t) \exp (j2\pi f_0 t)]$$

$$= \tfrac{1}{2} \text{ Re } [\gamma(t)\beta^*(t)] + \tfrac{1}{2} \text{ Re } [\gamma(t)\beta(t) \exp (j4\pi f_0 t]. \tag{A-8}$$

In most applications a multiplier is followed by either a low-pass filter (LPF) or a bandpass filter (BPF), as shown in Fig. 11, in order to select either the first or second term in (A-8) and completely reject the other term. In our application we may regard $y(t)$ as the reference carrier; then the LPF output $z_1(t)$ is the response of a coherent demodulator to $x(t)$. If $y(t) = x(t)$, so that the multiplier is really a squarer circuit, the BPF output $z_2(t)$ can be used for carrier phase recovery. Its complex envelope, relative to $2f_0$, is proportional to $\gamma^2(t)$.

Finally, when the bandpass signal is modeled as a random process, we use the same correspondence, (A-1) and (A-4), between the real process $x(t)$ and the complex envelope process $\gamma(t)$. It is of interest to relate the statistical properties of $x(t)$ to those of its in-phase and quadrature components, relative to some f_0. First we note that $E[x(t)] = \text{Re } \{E[\gamma(t)] \exp (j2\pi f_0 t)\}$; hence for a wide-sense stationary (WSS) $x(t)$ process, $\gamma(t)$ must be a zero-mean process, in order that $E[x(t)]$ be independent of t. Proceeding to an examination of second-order moments, it is a simple matter to show that $\gamma(t)$ must be a WSS process if $x(t)$ is to be a WSS process. The converse is not true. A WSS $\gamma(t)$ may produce a nonstationary $x(t)$, indicated as follows. Rewriting (A-1) as

$$x(t) = \tfrac{1}{2}\, \gamma(t) \exp (j2\pi f_0 t) + \tfrac{1}{2}\, \gamma^*(t) \exp (-j2\pi f_0 t) \tag{A-9}$$

the autocorrelation for $x(t)$ can be expressed as

$$k_{xx}(t+\tau, t) = E[x(t+\tau)x(t)] = \tfrac{1}{2}\,\mathrm{Re}\,[k_{\gamma\gamma}(\tau) \cdot \exp(j2\pi f_0\tau)] + \tfrac{1}{2}\,\mathrm{Re}\,[k_{\gamma\gamma^*}(\tau) \cdot \exp(j4\pi f_0 t + j2\pi f_0\tau)] \quad \text{(A-10)}$$

where, for complex WSS processes, we define the autocorrelation of $\gamma(t)$ as

$$k_{\gamma\gamma}(\tau) = E[\gamma(t+\tau)\gamma^*(t)]. \quad \text{(A-11)}$$

The quantity $k_{\gamma\gamma^*}(\tau) = E[\gamma(t+\tau)\,\gamma\,(t)]$ in (A-10) can be regarded as the cross correlation between signal components centered at $+f_0$ and at $-f_0$. If $x(t)$ is WSS, then this cross correlation must vanish in order that the t-dependent term in (A-10) vanish. Otherwise $x(t)$ is a cyclostationary process.

If we let $\gamma(t) = u(t) + j\,v(t)$, where the I and Q processes, $u(t)$ and $v(t)$, are jointly WSS, then we have

$$k_{\gamma\gamma^*}(\tau) = k_{uu}(\tau) - k_{vv}(\tau) + j[k_{vu}(\tau) + k_{uv}(\tau)] \quad \text{(A-12)}$$

and the condition for stationarity of $x(t)$ requires that

$$k_{uu}(\tau) = k_{vv}(\tau) \quad \text{and } k_{vu}(\tau) = -k_{uv}(\tau). \quad \text{(A-13)}$$

Thus for a WSS bandpass process, the I and Q components are balanced in the sense that they have the same autocorrelation function. Also, the cross correlation of the I and Q components must be an odd function, since $k_{vu}(\tau) = k_{uv}(-\tau)$ for any pair of WSS processes. For example, $u(t) = v(t)$ would satisfy the autocorrelation condition in (A-13), but not the cross correlation condition. The size of $k_{\gamma\gamma^*}(\tau)$ indicates the degree of cyclostationarity of a bandpass process. In the extreme case where either the I or Q component is missing, as in DSB-AM, we would have $k_{\gamma\gamma^*}(\tau) = \pm\, k_{\gamma\gamma}\,(\tau)$, e.g., for $v(t) = 0$,

$$k_{xx}(t+\tau, t) = k_{uu}(\tau)\cos(2\pi f_0\tau)\cos(2\pi f_0 t + 2\pi f_0\tau). \quad \text{(A-14)}$$

In modeling an additive noise process $n(t)$ on received signals, we often use the white-noise assumption wherein $k_{nn}\,(\tau) = N_0\delta(\tau)$. If we let $r(t) + j\,s(t)$ be the complex envelope of the process relative to any f_0 which is significantly larger than the passband width of the signals, then the white-noise process is equivalently modeled by I and Q processes whose correlation functions are given by

$$k_{rr}(\tau) = k_{ss}(\tau) = 2N_0\delta(\tau); \qquad k_{rs}(\tau) = 0. \quad \text{(A-15)}$$

REFERENCES

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

[2] W. C. Lindsey, *Synchronization Systems in Communication and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1972.

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

[4] F. M. Gardner, *Phaselock Techniques*, 2nd ed. New York: Wiley, 1979.

[5] A. J. Viterbi, *Principles of Coherent Communication*. New York: McGraw-Hill, 1966.

[6] L. E. Franks, *Signal Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1969.

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

[8] W. A. Gardner and L. E. Franks, "Characterization of cyclostationary random signal processes," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 4–14, Jan. 1975.

[9] F. M. Gardner, "Hangup in phase-lock loops," *IEEE Trans. Commun.*, vol. COM-25, pp. 1210–1214, Oct. 1977.

[10] M. K. Simon, "Further results on optimum receiver structures for digital phase and amplitude modulated signals," presented at 1978 Int. Conf. Commun., Toronto, Canada, 1978.

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

[12] P. Kabal and S. Pasupathy, "Partial response signaling," *IEEE Trans. Commun.*, vol. COM-23, pp. 921–934, Sept. 1975.

[13] H. Sailer, "Timing recovery in data transmission systems using multilevel partial response signaling," presented at 1975 Int. Conf. Commun., San Francisco, CA, 1975.

[14] S. U. H. Qureshi, "Timing recovery for equalized partial response systems," *IEEE Trans. Commun.*, vol. COM-24, pp. 1326–1330, Dec. 1976.

[15] E. Roza, "Analysis of phase-locked timing extraction circuits for pulse code transmission," *IEEE Trans. Commun.*, vol. COM-22, pp. 1236–1249, Sept. 1974.

[16] F. M. Gardner, "Self-noise in synchronizers," this issue, pp. 1159–1163.

[17] H. L. Van Trees, *Detection, Estimation and Modulation Theory, Part I*. New York: Wiley, 1968.

[18] L. E. Franks, "Acquisition of carrier and timing data-I," in *Signal Processing in Communication and Control*. Groningen, The Netherlands: Noordhoff, 1975, pp. 429–447.

[19] W. C. Lindsey and M. K. Simon, "Data-aided carrier tracking loop," *IEEE Trans. Commun.*, vol. COM-19, pp. 157–168, Apr. 1971.

[20] R. Matyas and P. J. McLane, "Decision-aided tracking loops for channels with phase jitter and intersymbol interference," *IEEE Trans. Commun.*, vol. COM-22, pp. 1014–1023, Aug. 1974.

[21] M. K. Simon and J. G. Smith, "Offset quadrature communications with decision-feedback carrier synchronization," *IEEE Trans. Commun.*, vol. COM-22, pp. 1576–1584, Oct. 1974.

[22] U. Mengali, "Synchronization of QAM signals in the presence of ISI," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-12, pp. 556–560, Sept. 1976.

[23] 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.

[24] P. A. Wintz and E. J. Luecke, "Performance of optimum and suboptimum synchronizers," *IEEE Trans. Commun.*, vol. COM-17, pp. 380–389, June 1969.

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

[26] M. H. Meyers and L. E. Franks, "Joint carrier phase and symbol timing for PAM systems," this issue, pp. 1121–1129.

[27] L. E. Franks, "Timing recovery problems in data communication," in *Communication Systems and Random Process Theory*. Groningen, The Netherlands: Sijthoff and Noordhoff, 1978, pp. 111–127.

[28] 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.

[29] G. Ungerboeck, "Adaptive maximum likelihood receiver for carrier-modulated data transmission systems," *IEEE Trans. Commun.*, vol. COM-22, pp. 624–636, May 1974.

[30] D. D. Falconer and J. Salz, "Optimal reception of digital data over the Gaussian channel with unknown delay and phase jitter," *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 117–126, Jan. 1977.

[31] U. Mengali, "Joint phase and timing acquisition in data transmission," *IEEE Trans. Commun.*, vol. COM-25, pp. 1174–1185, Oct. 1977.

[32] M. Mancianti, U. Mengali, and R. Reggiannini, "A fast start-up algorithm for channel parameter acquisition in SSB-AM data transmission," presented at 1979 Int. Conf. Commun., Boston, MA, 1979.

[33] L. E. Franks, "Further results on Nyquist's problem in pulse transmission," *IEEE Trans. Commun.*, vol. COM-16, pp. 337–340, Apr. 1968.

[34] F. S. Hill, "Optimum pulse shapes for PAM data transmission using VSB modulation," *IEEE Trans. Commun.*, vol. COM-23, pp. 352–361, Mar. 1975.

L. E. Franks (S'48–M'61–SM'71–F'77) was born in San Mateo, CA, on November 8, 1931. He received the B.S. degree in electrial engineering from Oregon State University, Corvallis, in 1952, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1953 and 1957, respectively.

In 1958 he joined Bell Laboratories, Murray Hill, NJ, working on filter design and signal analysis problems. He moved to Bell Laboratories, North Andover, MA, in 1962 to serve as Supervisor of the Data Systems Analysis Group. In 1969 he became a Faculty Member at the University of Massachusetts, Amherst, where he is currently Professor of Electrical and Computer Engineering and is engaged in research and teaching in signal processing and communication systems. He served as Chairman of the Department between 1975 and 1978. He was Academic Visitor at Imperial College, London, England, in 1979.

Dr. Franks is the author of the textbook, *Signal Theory* (Englewood Cliffs, NJ: Prentice-Hall, 1969). He is a member of the URSI Commission C, the Communication Theory and Data Communication Systems Committees of the IEEE Communications Society, and currently is the Associate Editor for Communications for the IEEE TRANSACTIONS ON INFORMATION THEORY.

Joint Carrier Phase and Symbol Timing Recovery for PAM Systems

M. H. MEYERS, MEMBER, IEEE, AND L. E. FRANKS, FELLOW, IEEE

Abstract—**The detection of pulse amplitude modulation (PAM) carrier signals requires accurate symbol timing and carrier phase references. In most cases, it is desired to estimate these parameters directly from measurements on the received data signal. This paper adds to and unifies the theory of maximum likelihood [ML] estimation as applied to PAM timing and phase recovery.**

Several different estimation strategies are considered. Data-aided [DA] estimators are found which assume the transmitted data symbols are known at the receiver. Nondata-aided [NDA] estimators are found which require only knowledge of the statistics of the transmitted data symbols. Structures for estimation of symbol timing, carrier phase, and joint estimation of timing and phase are presented.

The estimators are evaluated on the basis of their error variances. Relatively simple approximate expressions for these error variances are presented. These expressions allow the comparison of the effects of excess bandwidth, different modulation schemes, DA versus NDA recovery, and joint estimation versus estimation of only one parameter. A practical implementation of the ML estimator, termed a pseudo-maximum likelihood (PML) estimator, is proposed and analyzed. The performance of the PML estimator is shown to include a noise-independent, data-dependent jitter which dominates in many cases of practical interest.

I. INTRODUCTION

COHERENT demodulation of a modulated-carrier signal requires the presence at the receiver of a reference carrier with a precise phase relation to the received signal. System economy usually requires that this reference be derived from the received signal itself. If there is an unmodulated component of the carrier (pilot carrier) present in the signal, then a phase-locked loop can be used to determine the phase and track slow fluctuations in its value. Even with "suppressed-carrier" signals, a refined version of this approach called the Costas loop [1], [2] can be used if there are two sidebands having some degree of correlation, as in PAM/DSB or, to some extent, in PAM/VSB systems. However, this particular method fails with a PAM/SSB signal.

Another bandwidth-efficient modulation scheme is QAM, but if the in-phase and quadrature channel signals are balanced (i.e., statistically identical), then the Costas loop fails in this case, also. Not only is phase recovery more difficult in these cases, but the accuracy requirements on phase recovery are more stringent. A phase error in SSB introduces quadrature distortion, while in QAM, crosstalk interference is introduced between the in-phase and quadrature channels. One solution to this problem is "decision-directed" or "data-aided" phase recovery [3]-[12] which can be employed when the modulating message signal is some form of digital data sequence.

With synchronous digital data signals, there is another phase recovery problem concerning the proper sampling instants for detecting the data sequence. Failure to sample at the correct instants leads to intersymbol interference, which can be especially severe if the signal bandwidth is sharply limited. Data-aided approaches to symbol timing recovery have also been proposed [4], [7], [9]-[11].

In PAM/VSB (SSB) or PAM/QAM systems, we have the possibility of joint estimation of the carrier phase and symbol timing parameters, either using data-aided (DA) or nondata-aided (NDA) strategies. Joint estimation is the topic of this paper. We examine the implementation of these recovery schemes and evaluate their performance in terms of the rms error (jitter) in the estimates of the parameters. We are particularly interested in the relative effectiveness of the DA and NDA schemes and the dependence of performance on signal-to-noise ratio, the length of time available for making the estimates, and the amount of bandwidth allocated to the signal. Previous timing recovery studies [4], [5], [13], [16]

Manuscript received April 21, 1979; revised January 14, 1980. This work was performed at the University of Massachusetts, Amherst, and was supported by the National Science Foundation under Grant ENG 76-19492.

M. H. Meyers is with Bell Laboratories, North Andover, MA 01845.

L. E. Franks is with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003.

0090-6778/80/0800-1121$00.75