INVITED PAPER

Faster-Than-Nyquist Signaling

In a bandwidth scarce world, high-rate digital radio is moving beyond the older ideas of large-alphabet modulation, independent data pulses, and straightforward error-correcting codes.

By John B. Anderson, *Fellow IEEE*, Fredrik Rusek, and Viktor Öwall, *Member IEEE*

ABSTRACT | In this paper, we survey faster-than-Nyquist (FTN) signaling, an extension of ordinary linear modulation in which the usual data bearing pulses are simply sent faster, and consequently are no longer orthogonal. Far from a disadvantage, this innovation can transmit up to twice the bits as ordinary modulation at the same bit energy, spectrum, and error rate. The method is directly applicable to orthogonal frequency division multiplex (OFDM) and quadrature amplitude modulation (QAM) signaling. Performance results for a number of practical systems are presented. FTN signaling raises a number of basic issues in communication theory and practice. The Shannon capacity of the signals is considerably higher.

KEYWORDS | Bandwidth-efficient coding; coded modulation; constrained capacities; faster than Nyquist (FTN); intersymbol interference (ISI)

I. SOME INTRODUCTORY IDEAS

This is a tutorial and survey about a new method of data transmission called faster-than-Nyquist (FTN) signaling. Its roots trace back to the 1970s, but it has attracted interest in our bandwidth-starved world because it can pack 30%–100% more data in the same bandwidth at the same energy per bit and error rate compared to traditional methods. Its signals make available a higher Shannon capacity as well, and this leads to new attitudes about capacity. The survey will introduce FTN signals, their receivers, and their Shannon capacity. We will show how the signals relate to standard methods like quadrature amplitude modulation (QAM) and orthogonal frequency division multiplex (OFDM). We will look at tests of software implementations and some early hardware implementations. FTN comes at a price of higher computational complexity but new silicon technology is making FTN systems a feasible alternative.

The FTN idea stems from basic ideas of bandwidth, detection and capacity, but it looks at them in interesting new ways. We will begin with some of these ideas.

Most data transmission uses *linear modulation*, which is a simple adding up of a sequence of data pulses, with the form $s(t) = \sqrt{E_s}\sum_n a_n h(t - nT)$. Here a_n is a sequence of independent M-ary data symbols, each with energy E_s, and a new unit-energy pulse $h(t)$ appears each T seconds, the symbol time. The transmission rate is $\log_2 M/T$ bits per second (b/s). In what follows we will refer often to another energy, the energy E_b allocated to a *data bit*, which for M-ary linear modulation is given by $E_s/\log_2 M$. The sequence $\{a_n\}$ and h can be real, which produces baseband FTN, or complex, which produces carrier FTN. For simplicity, we mostly treat baseband FTN, but a carrier system appears in Section VI. In this survey the channel will have additive white Gaussian noise (AWGN), with density $N_0/2$ watts per hertz (W/Hz), and the ratio E_b/N_0 is the signal-to-noise ratio (SNR) of the transmission.

The pulse $h(t)$ in applications is almost always *orthogonal*—or "Nyquist"—with respect to shifts by nT. The mathematical meaning of this is that the product of two pulses offset by an integer number of symbol times has integral zero; the practical meaning is that pulses are invisible to each other. This makes optimal detection possible in a symbol-by-symbol way, with just a filter matched to h and a sampler. In binary antipodal signaling at baseband, h is real and $M = 2$; in QAM, both h and a_n are complex and M is often large. In OFDM, there are a number of $s(t)$ packed together on noninterfering parallel frequencies. Almost all popular coding methods, including parity-check codes, trellis-coded modulation (TCM), low-density parity check (LDPC), and bit-interleaved coded modulation (BICM), are based on orthogonal pulses.

Into this mix has come more and more need for bandwidth: Modern communication demands more bits be

Manuscript received August 22, 2012; accepted November 12, 2012. Date of publication March 14, 2013; date of current version July 15, 2013. This work was supported by the Swedish Research Council (VR) under Grant 621-2003-3210, and by the Swedish Foundation for Strategic Research (SSF) through its Strategic Center for High Speed Wireless Communication in Lund, Sweden.
The authors are with the Swedish Strategic Center for High Speed Wireless Communication and the Department of Electrical and Information Technology, Lund University, Lund SE-221 00, Sweden (e-mail: anderson@eit.lth.se; fredrikr@eit.lth.se; vikt@eit.lth.se).
Digital Object Identifier: 10.1109/JPROC.2012.2233451

carried. The easy ways to do this are to send longer or faster data streams, that is, to consume more time or more bandwidth. We can also trade one of these for the other. But time and bandwidth are limited, often costly, resources. How can more information be carried *per* hertz and second? The proper measure of efficiency here is the *data bit density*, in b/Hz-s. A transmission can trade hertz and seconds to varying degrees, employ different modulations, and use coding or not, but what counts most is the number of income-producing data bits it carries, per Hz-s. Bit density gives us a measure of efficiency, but it plays a second role in this survey because the FTN method grows more effective at higher densities, and they point to where FTN is most useful.

Transmission methods whose bit density is below 2 b/Hz-s are called wideband because their bit energy is small and they achieve a high rate in bits per second chiefly by raising the bandwidth; space communication is an example. Methods that work above 2 are called high energy because they achieve the same end chiefly by increasing E_s/N_0, that is, by means of a higher symbol energy in the channel; short range wireless communication, which can afford a high E_s/N_0, provides many examples. These generally require about 3 dB more E_s/N_0 in the channel per unit increase in their bit density.

How can bit density be increased? If the transmission method is a simple M-ary modulator, a direct way is to increase M. With Nyquist pulses, there are about $2\mathcal{W}\mathcal{T}$ independent modulator uses available in $\mathcal{T}$ seconds and $\mathcal{W}$ hertz, that is, 2 per Hz-s. The bit density for a simple modulator is thus $2\log_2 M$ b/Hz-s. Detection theory shows that to increase this by 1 requires that E_b approximately double at large M, if the error probability is not to change. In Section III, we will see that the relevant Shannon capacity follows the same trend at high densities.

For a given transmission method, more bits can be carried by spending more time or bandwidth. If these are scarce, the method until now in both coded and uncoded communication has been to employ the M-ary modulator with a large M, which in turn requires large E_s/N_0. Telephone modems and the coded 64-ary QAM used in digital television are two examples. FTN signaling offers an alternative that follows a milder energy growth at useful densities.

For reference in what follows, here are the bit densities of common modulations. In the case of carrier modulations, the total bandwidth is taken as the sum of the positive-hertz widths of the I and Q baseband signals. Simple antipodal signaling and quaternary phase-shift keying (PSK), with the most bandwidth-efficient pulse, has bit density 2 b/Hz-s; 16QAM modulation has 4 b/Hz-s. Adding a rate 1/2 error-correcting code reduces these by half.

An important issue in coded communication is computational complexity. We lack the space for a detailed exploration, but it is useful to set up three orders of complexity in Sections V–VII: *simple*, meaning a demodulator; *trellis*, meaning that the scheme needs a trellis (Viterbi) decoder; and *iterative*, meaning repeated iterations of two soft-output trellis decoders, probably with a very long block length. The meanings of these terms will become clear as we describe FTN systems. As receiver performance moves toward Shannon capacity, both storage and computation grow rapidly, and it is important to think about whether a given closeness is worth the penalty.

While traditional signaling systems enjoy great success, some interesting questions about them lie just below the surface in this discussion. Does the assumption of orthogonal pulses have an energy or bandwidth cost? Does linear modulation achieve the Shannon performance? Is a large modulation alphabet the only way to increase bit density? Is the performance of an alternative worth its complexity? Research with FTN signals has shown that orthogonal pulses and large M do indeed have weaknesses. Linear modulation, however, does not. The weaknesses grow more dramatic as the bits per Hz-s grow. This last point means these fundamental questions are closely linked to the high-energy narrowband transmission methods now being applied in wireless, satellite, and other applications.

II. FTN SIGNALS

With this background, we turn to FTN signaling. The key aspect of the FTN method is that $h(t)$ is no longer orthogonal with respect to the symbol time. The same h is employed but the symbol time is τT, $\tau < 1$. The signal becomes

$$s(t) = \sqrt{E_s}\sum_n a_n h(t - n\tau T). \tag{1}$$

The factor τ can be thought of as a time acceleration factor since now the pulses come too fast by a factor $1/\tau$. If a filter matched to $h(t)$ is used in the detection, its samples are no longer the M-ary values a_n plus noise, but contain intersymbol interference (ISI) as well. It can be shown that the average power spectral density (PSD) shape for $s(t)$ is the same—with uncorrelated and identically distributed a_n it is in fact $|H(f)|^2$ no matter what τ is, where $H(f)$ is the Fourier transform of $h(t)$. Fig. 1 shows an example of sinc pulse FTN with orthogonal symbol time $T = 1$ and $\tau = 0.8$. Above is ordinary orthogonal linear modulation with $h(t) = (\sqrt{T}/\pi t)\sin(\pi t/T)$, with the lighter sinc pulses representing symbols $+1, -1, +1, -1, -1$ that add up to the heavier curve $s(t)$. Below is FTN with the same $T = 1$ but $\tau = 0.8$; it can be seen that the five sinc pulses are now advanced in time by 0, 0.2, 0.4, etc.

A basic parameter of signals like (1), that determines their error probability, is their *minimum distance* $d_{\min}$. It plays a fundamental role in understanding FTN. Consider

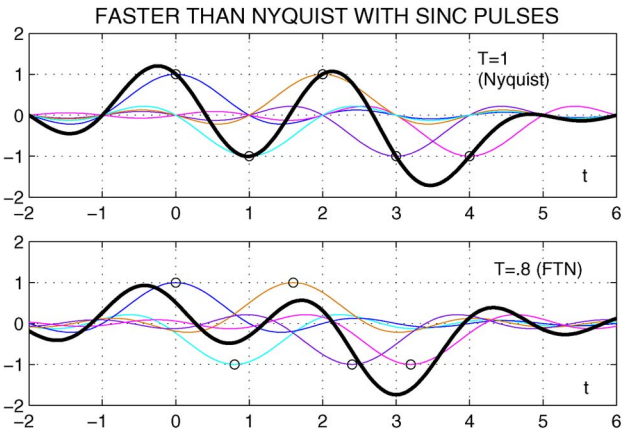

Fig. 1. ***Illustration of FTN signaling with unit-T sinc pulses and $\tau = 1$ and 0.8.***

signals $s(t)$ generated by all the different sequences $\{a_n\}$. Let $s_i(t)$ and $s_j(t)$ be two of these, such that their symbols a_n are the same up to some n_o and different thereafter at least at position $n_o + 1$. Then, $d^2_{\min}$ is the least square Euclidean distance

$$\left(\frac{1}{2E_b}\right) \int_{-\infty}^{\infty} \left|s_i(t) - s_j(t)\right|^2 dt, \qquad i \neq j \tag{2}$$

between any such pair. The error rate for the symbols a_n, with the best detection, tends to

$$Q\left(\sqrt{\frac{d^2_{\min} E_b}{N_0}}\right) \tag{3}$$

as the SNR grows. Here $1/2E_b$ is a normalizing constant, and $Q(u)$, $u > 0$, is the integral of the unit Gaussian density over $[u, \infty)$.

The square minimum distance $d^2_{\min}$ in (2) with binary orthogonal pulses is always 2, no matter what the pulse shape. This quantity is called the matched filter bound, and it and the corresponding error rate $Q(\sqrt{2E_b/N_0})$ are fundamental quantities in communication theory. A similar fixed value exists for other modulation alphabets; for example, $M = 4$ equi-spaced values lead to $d^2_{\min} = 0.8$ for any real orthogonal $h(t)$, which implies a much higher error rate at a given E_b/N_0.

The history of FTN signaling began with the 1975 paper of James Mazo [1], who investigated the binary sinc-pulse case in Fig. 1. He accelerated the pulses and discovered an interesting fact: $d^2_{\min}$ does not change when τ drops below 1 and the pulses become nonorthogonal. He found that it remains 2 for τ in the range [0.802, 1], despite the ISI. That is, $1/0.802 \approx 25\%$ more bits could be carried in the same bandwidth, without damage to the error rate. He was barely believed by some, since the result seemed to contradict Nyquist's limit of $1/2T$ Hz for orthogonal signals that carry $1/T$ b/s. Indeed, a certain quack atmosphere surrounded FTN until recently. But any apparent paradox here comes from confusing orthogonality with the distance-2 threshold: The bandwidths at which these two cease to be possible are not the same. Today we call the second—the bandwidth at which $d^2_{\min}$ first falls below its orthogonal value—the *Mazo limit*. The Nyquist limit is the bandwidth below which orthogonality can no longer exist. For the binary sinc pulse, the first is 0.401 Hz-s, in symbol-normalized terms, whereas the Nyquist limit is 0.5 Hz-s.

Mazo's result lay quiet for many years but interest grew after 1990. It was shown that the same phenomenon occurred with root raised cosine (rRC) pulses [4], the ones most commonly used in applications. An example is the binary 30% excess bandwidth rRC pulse, which ceases to have distance 2 at $\tau = 0.703$, a 42% increase in the bit density. A Mazo limit applies to nonbinary transmission [2], to pulses that are not orthogonal for any symbol time, and applies in a general way even to nonlinear modulation [3]. In all cases, the bandwidth at which the distance first drops is noticeably smaller than the Nyquist bandwidth and leads to significant bandwidth savings compared to typical practical systems.

What price is paid for this new bandwidth efficiency? Since there is now trellis-structured ISI in the matched filter samples, a trellis decoder is needed to decipher the data. Signal decoding near the Mazo limit turns out to be quite simple, on the "trellis" order of computation, but smaller τ than the limit lead to attractive combinations of bandwidth and energy efficiency, and in particular, they lead to good high bit density schemes. These require a larger trellis decoder. FTN signals can also be coded, and this increases detection complexity still more, to the "iterative" order. All this will be illustrated by the implementations in Sections V–VII.

A. FTN and Partial Response Signaling

The FTN method can be viewed as part of the larger development of partial response (PR) signals. PR introduces intentional ISI to achieve some end. It originated in the 1960s as a method of shaping and lengthening the pulse h in signals of form (1) so as to achieve properties such as spectral zeros (see [5]). The technique has evolved continuously since then, to embrace a more general view of the energy and bandwidth of signals and to study equalizers and trellis detectors like those in this survey. One object has been the discovery of signal sets of the form (1) with maximum $d_{\min}$ for a given bandwidth criterion or minimum bandwidth for a given $d_{\min}$ [6]. The more recent subject is summarized in [7, Ch. 6]. FTN signals are a subset of PR signals: They have a special type of ISI, which

stems from an agreed-upon spectrum shape, and they are not necessarily optimum in any sense.

B. Frequency FTN

An important extension of the FTN idea that appeared in 2005 [8] is squeezing signals together in frequency, just as they were accelerated in time. Consider N signals, each of bandwidth W hertz, sent in N adjacent channels with total bandwidth NW. In OFDM, the signals are orthogonal by virtue of frequency separation. Now bring the signals closer in frequency, so that the "orthogonal" in OFDM no longer holds. There is now cochannel interference, and just as with the time-FTN signal, the (binary) square minimum distance continues to be 2 up to a limit, this one a frequency Mazo limit. More bits can be carried per Hz-s, without damaging the error rate. The improvement in bit density is about the same as with time FTN, but time FTN and frequency FTN give *independent* improvements; this has been verified by finding $d_{\min}$ for various combinations of time and frequency squeezing [8], [9]. If both techniques are applied, bandwidth efficiency can more than double while $d^2_{\min}$ remains 2. We will take up a multicarrier FTN system in Section VII.

For both time FTN and frequency FTN, one can investigate the pulse shape h that leads to the least Mazo limit. Under reasonable assumptions, this pulse turns out to be nearly Gaussian in shape [10]. Note that the Gauss pulse is not orthogonal for any symbol time; the idea here is only to accelerate pulses until the $d^2_{\min}$ of the signal set falls below 2.

III. CAPACITY OF TIME FTN SIGNALS

Since ordinary modulation with FTN signals is more efficient, it is of interest whether the Shannon capacity of the single-carrier FTN channel is in some sense higher. It is indeed considerably higher, and FTN has brought with it a somewhat different way to view capacity. This section begins the discussion by reviewing how capacity works in narrowband transmission.

The concept of capacity applies to coded signals, and before beginning we need to say what coding means when applied to FTN signals. One view is as follows. The uncoded FTN signal set in (1) is that produced by a modulator; the $\{a_n\}$ are data symbols and the modulator signal sets have a certain performance, meaning here that for a certain spectrum and error rate they need a certain SNR E_b/N_0. It is much higher than Shannon capacity predicts, generally 8–10 dB more. By coding is meant the selection of a *subset* of the FTN signals, which, having fewer signals, will carry fewer data bits. The $\{a_n\}$ are no longer data symbols. Each signal in the subset corresponds to a set of data bits. For a clever enough selection of the subset, less SNR will be needed for the same per-data-bit spectrum and error rate. Shannon's capacity is the ultimate limit to this game.

An important issue is how to incorporate spectrum into the capacity calculation. In any practical transmission, the pulse h as well as the PSD will have a certain smooth spectral shape. However, capacity references for signaling are almost always computed for sinc pulses, meaning that a "square" PSD is assumed. That is, the reference assumes one spectrum, while the signals have another. Sinc pulses are difficult to use for many reasons,[1] but the issue goes deeper. The information carrying ability of the actual nonsinc pulses is generally *higher*; they leave room for higher rate transmission.

This mismatch in the benchmark is a fundamental issue, but hidden within it is a more surprising result, to which we will return at the end of the section: coding based on orthogonal pulses can only achieve the sinc capacity, not the higher nonsinc one. The disparity grows worse as the data bit density grows. The coded schemes we now employ, including "near-capacity" turbo coding, can fall significantly short of the capacity that applies with the pulses that we really use.

Now we discuss the capacity calculation. We begin with Shannon's classic Gaussian capacity.[2] Let the signals have power P watts and a *square* PSD shape on $[-W, W]$; their capacity is then

$$C_{\text{sq}} = W \log_2(1 + P/WN_0) \text{ b/s}. \tag{4}$$

The proof assumes sinc pulses. We can note about (4) that if the signal spectrum is scaled wider by κ, then P is also scaled by κ, and since P/W is then constant, C_{sq} must be scaled by κ as well. The energy per information bit carried E_b is P/C_{sq} and, consequently, E_b is unchanged. This expresses the fact that (4) is transparent to a "fair" bandwidth scaling, and except for that depends only on P/N_0. If E_b is what is of interest, a fair scaling has no effect. By taking $W = 1$, we convert the dimensions of C_{sq} to bits per Hz-s. Each E_b allows a certain bits per Hz-s to be carried.

The FTN signals in (1) have in general a nonsquare PSD. How do we add the PSD constraint to the calculation? A straightforward argument works as follows. The smooth spectrum can be approximated by many rectangular pieces, and information is then sent through these small channels. By the fundamental theorem of integral calculus, the total capacity is

$$C_{\text{PSD}} = \int_0^\infty \log_2\left[1 + \frac{2P}{N_0}|H(f)|^2\right] df \text{ b/s}. \tag{5}$$

[1]Among the reasons are: synchronization is difficult, synthesis of signal alternatives requires too much computation and storage, and sinc pulses make inefficient use of the combined time and bandwidth resource.

[2]This formula first appeared in this very Proceedings, and that paper remains an excellent source: C. E. Shannon, "Communication in the presence of noise," Proceedings of the IRE, Vol. 37, No. 1, pp. 10–21, Jan. 1949.

Here P is the signal power in watts and $|H(f)|^2$ is its spectral distribution, normalized to unit integral over the real line. For a fixed bandwidth $[-W, W]$, this rate is largest when the PSD is square, and then it is (4). Technically, C_{PSD} is called a constrained information rate, in this case the highest rate that *signals constrained to the spectral density* $H(f)$ can carry information, whether made from orthogonal pulses or not, and for any probability distribution on the input symbols a_n.

The same Hz-s argument as above applies to (5): a fair bandwidth scaling by κ increases C_{PSD} by κ and the capacity-achieving E_b is fixed at P/C_{PSD}. Now it is total P and the *shape* of the PSD that determine E_b. We can convert (5) to bits per Hz-s by adopting a calibration framework for the bandwidth. The PSD can in principle have infinite support, so it is best to measure its width by taking some characteristic point on the PSD shape. A useful one that families of shapes often have in common is their 3-dB power bandwidth; for all T-orthogonal pulses this is in fact $1/2T$ Hz. The aim of this survey is to see how much capacity opens up when sinc is replaced by another orthogonal pulse, and so this is a convenient measure.

Therefore, we scale the $H(f)$ width so that its 3-dB point lies at 1 Hz and compute (5). The result is the capacity in bits per Hz-s for this pulse shape P and spectral calibration. When sent for $\mathcal{T}$ seconds, signals with bandwidth $\mathcal{W}$ in the measurement system will have capacity $\mathcal{W}\mathcal{T}C_{\text{PSD}}$ bits. It can be seen from (4) that each increase in C_{sq} by 1 b/Hz-s requires about a 3-dB increase in P at large P. But this is not true at moderate P. Calculations with (5) and a 30% RC spectrum show that C_{PSD} requires only about 1.9 dB more P per unit increase, when C_{PSD} is in the range 4–12 b/Hz-s.

Here is an example how different the rates can be. Let $h(t)$ be a 30% rRC T-orthogonal pulse with $T = 1/2$ (hence $1/2T = 1$ Hz). Let $P/N_0 = 1$, which means a wideband, low energy channel. Both (4) and (5) yield capacity close to 1 b/s. In bit density terms this is 1 b/Hz-s, since 1 Hz was consumed. The ratio E_b/N_0 is $P/CN_0 \approx 1$ as well. Now let $P/N_0 = 1000$; this ratio is typical of digital TV and lower than a local telephone loop. Now the two capacities C_{PSD} and C_{sq} are 11.81 and 9.97 b/s, and the bit densities are the same values in bits per Hz-s. An alternate view is that E_b/N_0 at capacity is 85 in the first case but 100 in the second. Yet another calculation shows that E_b needs to rise by 5.6 dB if the capacity of orthogonal-pulse coding is to achieve the C_{PSD} value 11.81 b/s. The first numbers mean that coding based on 30% orthogonal pulses, including present-day turbo coding, is limited to a capacity 15% lower than the spectrum actually allows. In the limit of large P/N_0, the rate C_{PSD} is 30% higher than C_{sq}.

The foregoing discussion relates primarily to time FTN. In frequency FTN, many carriers would presumably be stacked, as in OFDM, and occupy a wide spectrum. A gain in capacity due to the two outer spectral skirts would have little effect on the total capacity. The capacity would essentially be (4) with the OFDM bandwidth.

To summarize the time FTN case, the sinc pulse capacity (4) is traditionally used to benchmark systems, but it has the logical flaw that it does not include the PSD constraint of the signals employed. Equation (5), on the other hand, depends on a spectrum shape. Research in recent years [14], [15] has shown some interesting results about (5).

- When a nonsinc T-orthogonal pulse is substituted for $h(t)$, (5) always grows, and can give a much larger rate.
- Codes assembled from FTN signals ($\tau < 1$) can in principle actually reach the larger rate (5). They can come near it even with binary $M = 2$ pulses. Large modulation alphabets are unnecessary.
- The higher capacities seem to be reflected in practical systems.
- Codes assembled from $s(t)$ in (1) built from orthogonal pulses are always limited to the lower rate (4), whether or not the pulse is a sinc.

We finish by expanding on the last item, the question of why the orthogonal pulse/linear modulation combination limits bit rates to the lower capacity (4), even when the pulses are not sincs. The poorer capacity, denoted "Nyquist" in figures that follow, applies to, for example, TCM and LDPC coding. Sequences of coded orthogonal pulses are so popular because the pulses act during matched filter detection like a sequence of independent symbol transmissions. Each symbol can be identified with a letter in a traditional codeword. TCM coding is built on large-alphabet QAM symbol values, but here again pulses correspond to values, this time QAM values. But this correspondence can also be used to show that capacity is limited. Imagine now that we employ an error correction scheme and orthogonal pulses are used to carry the code symbols; these can be sinc or any other kind of orthogonal pulses. Our code's performance will be limited by (4) no matter what orthogonal pulse it is coupled with, since the code mechanics are dissociated from the pulse modulation. The information carrying ability of this code thus cannot exceed capacity (4).

If all our transmissions used sinc pulses, there would be no better capacity to aspire to. But real systems use other pulse shapes, and the test results in the next sections suggest that a higher capacity does indeed govern these real-life systems.

IV. UNCODED FTN AND ITS DETECTION

How can FTN signals be detected? The pulses in (1) have longer duration and interfere with each other. Somehow each needs to be separated, by viewing its neighbors before and after. A useful general transmit/receive structure is as shown in (6) at the bottom of the next page. It starts with the data-bearing symbols $\{a_n\}$ and ends with a sequence $\{y_n\}$ that contains $\{a_n\}$ plus ISI and Gaussian noise. The

essential elements in this flow are a continuous-time matched filter plus a sampler, which convert the noisy continuous FTN signal to a discrete-time sequence. There are several variations of this basic structure, but the main point is that the end result y_n is a convolution of $\{a_n\}$ and a sequence $\{v_n\}$ plus Gaussian noise, that is $y_n = \{a_n\} * \{v_n\} + \eta_n$.

A few facts are the following. The continuous-time filter can be matched to the whole pulse $h(t)$, as in the whitened matched filter (WMF) receiver, or to a shorter basis pulse, for example, a pulse orthogonal to τT shifts. In the second case, the noise in the samples will then automatically be white. Many of the interesting $h(t)$ are zero in certain spectral regions, and the WMF is then not well defined, so the second approach is the more attractive one. The postprocessing filter $B(z)$ has a variety of uses: whitening the noise η_n, shortening the impulse response length, and making the $\{v_n\}$ minimum phase are three of them. Minimum phase is important, and to achieve it in a causal way the structure needs to reverse the sequence $\{y_n\}$ and detect it backwards. There are also detector structures that let the noise in $\{y_n\}$ remain colored [20]. Finally, prefiltering can be performed before the matched filter; we will return to this at the end of the section.

It remains to perform the Detect $\{a_n\}$ block. When $\{v_n\}$ is a finite sequence, we can say that the $\{a_n\}$ are "trellis coded" by $\{v_n\}$. One basic way to implement the Detect $\{a_n\}$ block is to "decode" them with a standard Viterbi algorithm (VA), a scheme that compares the received signal to the trellis signal structure to find the most likely $\{a_n\}$ sequence. The $\{a_n\}$ are viewed as the state of the ISI. Another way is the BCJR algorithm,[3] a method that computes the likelihoods of the $\{a_n\}$ rather than deciding them. The BCJR method turns out to be essential in iterative decoding, which is at present the only practical way to detect coded FTN schemes. The VA and BCJR are indeed attractive when the FTN is mild (τ is not too small). But under severe FTN the trellis is gigantic and other methods are called for. These may be classified as follows.

- More efficient variations of the basic VA, so-called *reduced trellis* methods. These generally assume that $\{y_n\}$ contains white noise. One kind, the Duel–Hallen, or "offset," schemes break the trellis state variable (the recent a_n) into a part that is trellis-searched and a prehistory that is not.
- Modifications to the VA idea that work with colored-noise $\{y_n\}$. We look at one of these at the end of the section.
- Searching only tiny parts of the whole trellis, *reduced search* methods. A minimum phase input $\{y_n\}$ is essential for these. We will report tests with one, the *M*-algorithm, a scheme that views a fixed, small number *M* of the trellis paths.
- The BCJR algorithm, in reduced-trellis or reduced-search variants. We will report on an *M*-algorithm variant.
- Successive interference cancellation (SIC), in which a soft estimate of the ISI is subtracted after one or more iterations. One of these is explored at the end of Section VII.
- Equalization. FTN detection is ISI removal, and one of the many simplified equalizers can be used. Any good equalizer will partially detect FTN, but if the ISI is severe, a simple scheme such as zero forcing is usually not good enough.

Detection of uncoded FTN uses algorithms of coding and is often referred to as decoding, but technically it is demodulation. We will take up true coding in the next section and devote the rest of this section to uncoded signals.

Fig. 2 shows the frequency of error events[4] for FTN with $\tau = 0.5, 0.35, 0.25$. These values lie well below the Mazo limit and represent strong ISI and narrowband transmission. The base pulse $h(t)$ is the 30% rRC pulse. This h has spectrum zero outside $1.3/2T$ Hz and so the sequence $\{v_n\}$ is in theory infinite, but a good approximation results when $\boldsymbol{v}$ has length 13, 17, and 27, respectively. The last FTN runs at density 8 b/Hz-s and has a huge memory; no variant of the full VA or BCJR is practical with this ISI. The heavy solid curves are error estimates based on a minimum distance analysis. Immediately above are performances of an *M*-algorithm BCJR (henceforth called the M-BCJR). These show that this reduced-search detector can limit the BCJR recursions respectively to 4, 7, 20

[3]Named after its authors Bahl–Cocke–Jelinek–Raviv and published in 1974 [17]. BCJR is based on the earlier Baum–Petrie algorithm for identifying models. The algorithm lay dormant many years, was recognized as the critical element in turbo decoding in the early 1990s, and won a paper prize, finally, in 1998.

[4]Errors in FTN and error-correction decoding do not occur in isolated bit errors, but rather in multibit events. Error statistics and distance analysis relate to these events. Details about the tests and the receiver algorithm can be found in [11] and [13].

$$\begin{aligned} &\{a_n\} \longrightarrow \\ &\text{Linear modulation by } h(t) \text{ at rate } \frac{1}{\tau T} \longrightarrow \text{AWGN channel} \longrightarrow \\ &\text{Matched filter} \longrightarrow \text{Sample at } n\tau T \longrightarrow \text{Discrete-time filter } B(z) \longrightarrow \{y_n\} \\ &\longrightarrow \text{Detect } \{a_n\}. \end{aligned} \tag{6}$$

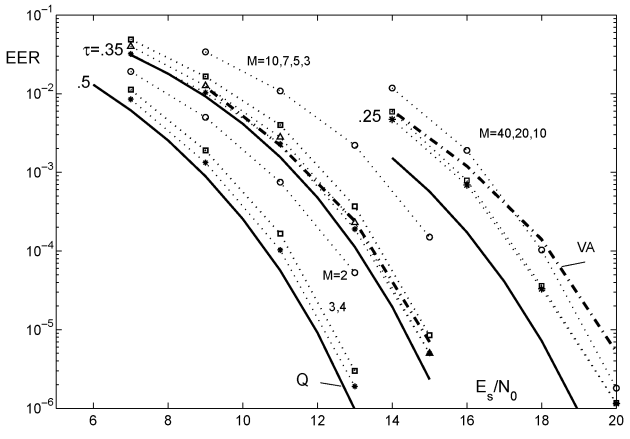

Fig. 2. *Error event rates for simple rRC FTN detection versus E_s/N_0 in decibels; an M-algorithm BCJR (dotted) is compared to a reduced-trellis VA (dash–dot) and a distance-based estimate. (From [11].)*

signal paths through the ISI trellis and still achieve near-optimal performance.[5] A reduced-trellis Viterbi decoder of size 256 and 4096 states is shown for comparison at $\tau = 0.35, 0.25$; the 4096 state reduction is much too severe, but the normal reduced VA is impractical.

It is interesting to observe that the $\tau = 0.25$ FTN case has the same PSD shape as 256-ary QAM modulation with the same $h(t)$. But 256QAM needs $E_b/N_0 \approx 24$ dB to achieve error rate $\approx 10^{-6}$, 4–5 dB more than the $\tau = 0.25$ plot. The Shannon capacity C_{PSD} crosses the x-axis at about 12 dB. In a similar way, ordinary 16QAM compares to the $\tau = 0.5$ case, both having bit density 4 b/Hz-s. Error rate 10^{-6} is reached at about 15 and 13 dB, respectively, with C_{PSD} at 4.7 dB. These comparisons show that uncoded FTN is significantly more energy efficient than QAM at the same error rate and bandwidth, but that it still lies well above capacity.

All the τ here lie well below the Mazo limit $\tau = 0.703$. With 0.703, 16-state reduced-trellis VA designs exist that perform near the theoretical error probability $Q(\sqrt{2E_b/N_0})$ for ISI-free transmission. The M-algorithm BCJR requires only three trellis paths. These are simple detectors, and they make possible a system with 30% less bandwidth.

Receivers That Work With Colored Noise: For many years most of detection theory has been concerned with receivers where the processing in the Detect $\{a_n\}$ block in (6) assumes the noise is white. The processing is then more straightforward, but it may be more for a given BER than the nonwhite alternatives, and resolving this question is a present research question. Placing a prefilter either before or after the Matched filter/Sampler blocks in (6) will in general color the noise. It can also *reduce the signal set* $d_{\min}$, which will damage the bit error rate (BER) performance, but there can be a complexity reduction for a given BER and E_b/N_0.

Examples of recent research in this sort of colored-noise receiver are Colavolpe *et al.* [20], who adapt the BCJR algorithm to colored $\{y_n\}$, and Dang *et al.* [21], who analyze the effects of prefiltering. A much older approach is combined linear Viterbi detection (CLVD), which originated in [22]–[24]. The rationale behind CLVD is to apply the filter $B(z)$ with taps $[b_0, b_1, \ldots, b_N]$ to the receiver filter samples, call them $\{x_n\}$, in order to shorten the memory of the filtered impulse response $\{v_n\} * \{b_n\}$. The shortened response is referred to as the target impulse response $\boldsymbol{v}^{\mathrm{tar}}$ and is constrained to have finite duration L. The two filters $\{b_n\}$ and $\{v_n^{\mathrm{tar}}\}$ are jointly optimized so that

$$\{b_n\} * \{x_n\} \approx \{v_n^{\mathrm{tar}}\} * \{a_n\} + \{\eta_n\}.$$

The noise $\{\eta_n\}$ here is not necessarily white. Details of this optimization can be found in [22]–[25]. A trellis detector follows that operates as if the true impulse response were $\boldsymbol{v}^{\mathrm{tar}}$ and the noise were white. With binary pulses, the number of states has been found to be reasonable with some FTN pulses.

Since the noise may be colored and $\boldsymbol{v}^{\mathrm{tar}}$ does not exactly model the signal convolution, the decoder is mismatched with its input signal. This is the classical setting of *mismatched decoding*. The concept has emerged as a powerful tool for analyzing practical FTN systems.

V. CODED FTN AND TURBO DECODING

In the last section, we saw that FTN with low-complexity receivers improves upon simple modulation with the same bit density, but performance still lies well short of capacity. What coded systems are available that reach closer to C_{PSD}? One that has been fully researched at this writing is a rate R convolutional code driving a binary FTN modulator. Now the $\{a_n\}$ sequences in (1) are convolutional codewords, and the data bit density is reduced by a factor R.[6] Standard practice for many years is to place an interleaver before the FTN modulation, so that the receiver can break up the error bursts that are characteristic of ISI. The convolutional encoder defines an allowed subset of sequences from (1), and we have a true coded system.

In principle a maximum-likelihood decoder, perhaps a VA, could estimate which word from the code was sent. As a trellis structure, the code would have a state space that was the product of the convolutional and FTN state spaces.

[5]The M-algorithm VA will perform a little better in this application. The required trellis decision depth is 15–40, depending on the value of τ.

[6]An equivalent view is that the encoder drives an orthogonal modulation, which then passes through an ISI channel.

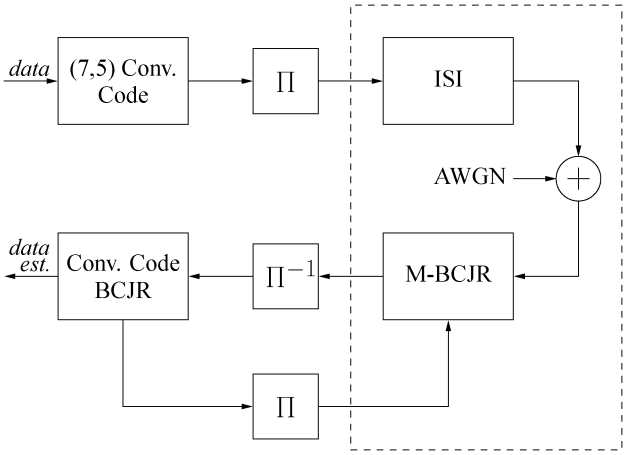

Fig. 3. *Iterative detection of FTN signals. The simple detection in Section IV includes just the dashed box.* Π *denotes an interleaver.*

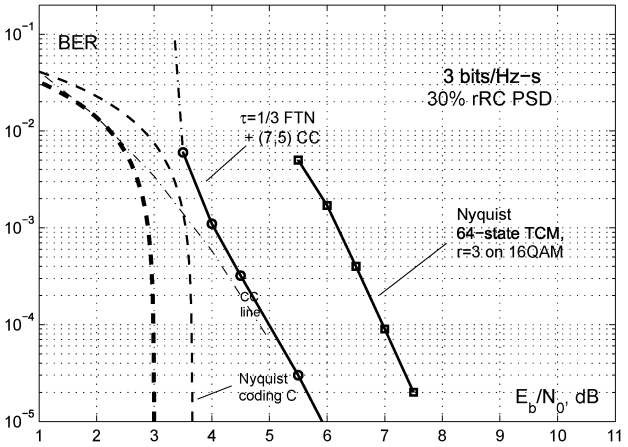

Fig. 4. *BERs for Nyquist-pulse TCM and convolutionally coded* $\tau = 1/3$ *FTN versus* E_b/N_0 *in decibels, at 3 b/Hz-s. All signals have 30% rRC PSD. rRC and sinc pulse capacities for 3 b/Hz-s shown for comparison. "CC line" denotes (7, 5) code BER over ISI-free AWGN channel (capacities do not apply to this line).*

But with convolutionally coded FTN and small τ, the VA would be far too large. A way out of this called turbo equalization was proposed by Douillard *et al.* [16] and is sketched in Fig. 3. The transmitter and the channel are the traditional convolutional code/interleaver/ISI, but the receiver is an iterative detector consisting of two soft decoders in a feedback loop, one for the ISI and one for the convolutional code. An interleaver and deinterleaver reassemble the transmission in the right order, and they also make the inputs to the soft decoders quasi-independent, which is essential to proper convergence of the iterations. The soft information needs to flow around the loop 5–50 times, depending on the SNR and the ISI.

The critical elements in the feedback loop are the soft decoders, and the design of these is a major subject. It is crucial that the information fed around is *soft*. Instead of decisions about bits, it must be a statement of their probabilities: in turbo equalization, the log-likelihood ratio (LLR) is used.[7] The standard algorithm to compute soft information values for signals that can be organized in a trellis structure is the aforementioned BCJR algorithm. The algorithm consists of forward and backward linear recursions that work along the received sequence; there is no add/compare/select as in the VA. It is still true that the ISI has a very large trellis structure, and like the VA, the BCJR needs to be simplified and focused, either by reducing its trellis or by reducing the regions of the trellis where it performs calculations. These algorithms are active research areas.

The convolutional encoder can be simple, and most research has in fact studied the memory-2 feedforward (7, 5) encoder, the standard introductory example in many textbooks. The best convolutional code depends on the operating E_b/N_0 [19], but its BCJR is in any case small. The interleavers, however, need to be long and this sets the coding scheme's block length. The overall coded FTN structure includes the interleaver, as does a turbo code, and is thus neither simple nor short.

Fig. 4 shows the BER of the (7, 5) 4-state convolutional code and the $\tau = 1/3$ FTN. The BER plot has two fundamental properties. At some low threshold E_b/N_0 ($\approx$3.5 dB in the figure), the BER suddenly drops from a high value. Thereafter, the data BER tracks that of the convolutional code over an ISI-free AWGN channel with the same E_b/N_0; this "CC line" appears in the figure. The ISI–BCJR application in the first iteration is the simple detection in Section IV, and it produces error rate $\approx Q(\sqrt{d^2_{\min} E_s/N_0})$. Shown for comparison is a Nyquist-pulse 64-state TCM system with the same PSD shape and same 3 b/Hz-s. It has a fixed complexity for all E_b/N_0, about the same as the turbo decoder operating at E_b/N_0 above 5–6 dB. It depends for its operation on orthogonal pulses.

When the FTN-induced ISI is this long or longer, the ISI–BCJR can be replaced by an M-algorithm BCJR [12], [13]. This greatly reduces computation, and under these conditions the ISI–BCJR in Fig. 4 can limit its computation to 5–100 trellis paths and the turbo decoder needs 3–40 turbo iterations and block length 1500–40 000, all depending on how far E_b/N_0 lies from capacity. The high figures will bring BER performance to within 1 dB of capacity, as shown in the figure.

A special version of the capacity C_{PSD} called the Shannon BER capacity is shown as a heavy dashed line, for systems with the 30% rRC PSD that carry 3 b/Hz-s. This useful curve shows the limit to the BER of any system working at the E_b/N_0 on the x-axis, for the given PSD and

[7] Receivers like that in Fig. 3 were studied as early as 1970, but failed because they fed around hard bit decisions.

bit density. We will skip the details of the calculation.[8] The coded FTN system lies as close as 1 dB to the BER version of C_{PSD} and the TCM system lies about 2 dB further away. The lighter dashed curve gives the BER capacity when $h(t)$ is replaced by the sinc pulse, or alternately, when coding is limited to Nyquist pulses. This curve is a BER capacity, but starting from C_{sq} in (4). The curve lies 0.7 dB closer to both schemes, but it is not the correct reference for the FTN scheme.

The complexity of coded FTN is thus at the "iterative" level. ISI–BCJRs with much reduced complexity have been developed, and there is much research on turbo decoding hardware, but even so the total complexity and block length are not small. Reducing computation and interleaver length, or using a shortened ISI model $\boldsymbol{v}$ easily drive performance away from Shannon capacity, and diminish the reward for using iterative detection in the first place.

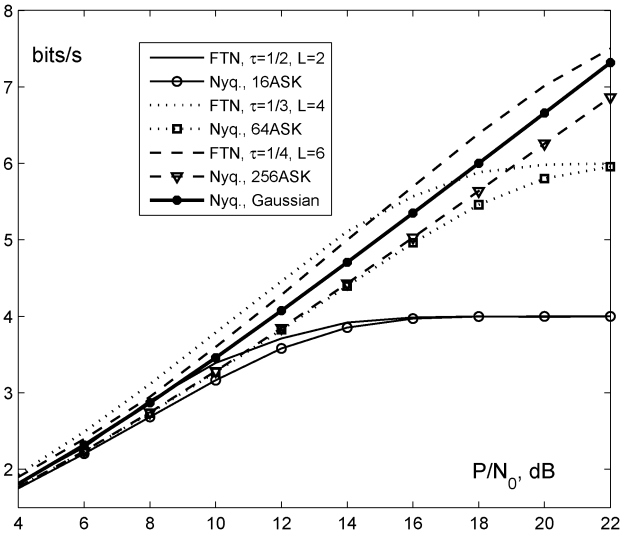

Fig. 5. *Plot of underbound to achievable rates I_{CLVD} versus P/N_0 for FTN and Nyquist schemes. Circles, squares, and triangles denote M-ary ASK Nyquist systems with $M = 16, 64, 256$. The heavy solid line uses Gaussian inputs and marks the ultimate limit of Nyquist transmission. Curves with no markers denote binary-input FTN systems with $\tau = 0.5, 0.35, 0.25$, which have the same bits per Hz-s and PSD shape as the ASK systems. There exist binary FTN systems that outperform the ultimate Nyquist transmission.*

VI. MORE ON CAPACITY

In Section III, we gave the fundamental capacity integral for signals that, like FTN, have a certain PSD. Although (5) states the ultimate information rate for a PSD, the promised rate can only be achieved under the ideal assumptions of Gaussian-distributed data and optimal detection. This section explores how capacity is affected by practical constellations of the input values, such as occur in QAM or PSK, or by processing, such as occurs in the suboptimal CLVD receiver. Aside from being interesting in itself, the CLVD idea also simplifies capacity calculation, and helps answer the question of how close coded FTN can come to C_{PSD}.

For true maximum-likelihood (ML) detection, the capacity constrained to a certain input constellation is the highest mutual information, denoted I_{ML}, between the input and output of the channel

$$I_{\mathrm{ML}} = \max I(\{y_n\}, \{a_n\}). \tag{7}$$

Here the max is over the probability distributions of $\boldsymbol{y}$ and $\boldsymbol{a}$, given AWGN, the FTN ISI and the constellation constraint. The long memory of the ISI makes I_{ML} difficult to evaluate, and one way to simplify the problem is to add the constraint that CLVD processing is present. This shortens the channel memory, and as a side benefit suggests a potentially practical low-complexity detector. We denote the new highest mutual information by I_{CLVD}. It must be true that $C_{\mathrm{PSD}} \geq I_{\mathrm{ML}} \geq I_{\mathrm{CLVD}}$, where the first inequality holds since the Gaussian input assumption has been relaxed and the second holds because ML-detection has been relaxed. The rate I_{CLVD} is derived in [25], based on the results of [26] and [27].

Fig. 5 shows the results of some of these capacity calculations. We show the max I_{CLVD}, expressed as bits per second, for optimized CLVD reception for $\tau = 0.5$, 0.35, 0.25. The memory of the target impulse response L is 2, 4, 6, which correspond to a complexity of 4, 16, 64 states in the model. The three curves marked with circles, squares, and triangles show the constrained capacities of the competing Nyquist systems ($\tau = 1$), based on 4ASK (amplitude shift keying), 8ASK and 16ASK (with complex-valued constellations, these correspond to 16QAM, 64QAM, 256QAM). The heavy line shows the capacity of a Nyquist system with random Gaussian-distributed input values, which is the optimal strategy from a capacity point of view. As can be seen, the FTN systems are always superior to the competing Nyquist systems. In fact, $\tau = 0.35$ and 0.25 FTN outperforms the optimal Gaussian-based Nyquist system even though it works with a small fixed set of channel input values and its curve is an underbound. This has even been shown for binary ASK with small enough τ [14].

It may appear that coded FTN violates the Shannon limit, but this is, of course, not the case. An FTN system has the ability to exploit the excess bandwidth of the base pulse $h(t)$. This is not the case for systems based on othogonal pulses, whose performance is independent of the excess bandwidth in the orthogonal pulse. Some study of (4) and (5) shows that the dependence of capacity on

[8]Shannon demonstrated how to compute such curves in 1959; the method is adapted to coded FTN in [18] and works from (5). In (5), the 3-dB bandwidth of the 30% RC $|H(f)|^2$ is set to 1 Hz and $C_{\mathrm{PSD}} = 3$ b/Hz-s.

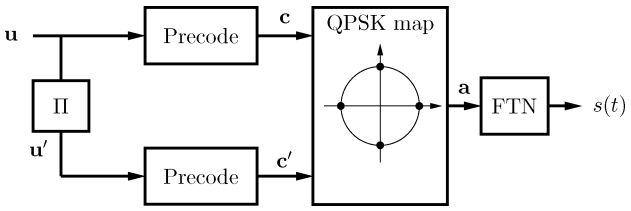

Fig. 6. *BICM-coded FTN signal generation.*

bandwidth is roughly linear, whereas it depends logarithmically on power. As P/N_0 grows, even a small fraction of power lying in a PSD stopband has a major effect. There is no FTN capacity "bonus" with $\mathrm{sinc}(t/T)$ pulses, but as soon as a practical pulse is substituted for sinc, new capacity appears.

A Carrier Modulation Example: The capacity advantage extends also to carrier modulation systems. Consider the bit-interleaved coded modulation (BICM) system in Fig. 6, where QPSK feeds an FTN system with $\tau = 1/2$. A sequence of bits $\boldsymbol{u}$ feeds two memory-2 shift-register precoders, one of which is preceded by a length-100 000 interleaver (details can be found in [28]). Mapping to QPSK and FTN modulation follow, and the receiver is an iterative decoder in the style of Section V. This FTN system can be viewed as working with complex-valued data and pulses. Test results for this BICM are shown in Fig. 7. Since $\tau = 1/2$, two QPSK symbols are transmitted every T seconds, but since the BICM rate is 1/2 this is 2 data bits every T seconds, sent over the I and Q channels. The corresponding bit density is 2 b/Hz-s, calculated as in Section V, but with a division by two to account for the I/Q modulation.

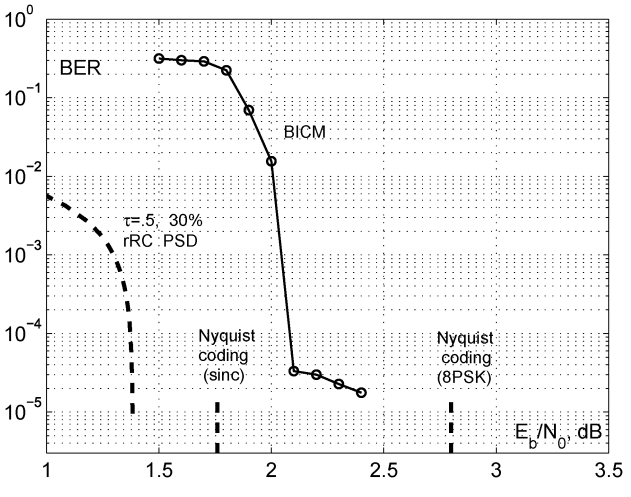

Fig. 7. *Measured BER of $\tau = 0.5$ BICM–FTN system versus E_b/N_0, compared with three capacities, with all signaling at rate 2 b/Hz-s. Capacities left to right: PSD capacity, $\tau = 0.5$, 30% rRC; Nyquist coding, (4); 8PSK Nyquist coding. All methods can be thought of as having the same PSD. The BICM–FTN outperforms any coded 8PSK system at the same bits per Hz-s.*

The BER capacity curve for 2 b/Hz-s systems with 30% rRC PSD is shown in the figure, computed in the same way as in Fig. 4. It lies as close as 0.8 dB. The capacity for sinc and other Nyquist pulse coding at 2 b/Hz-s, computed from (4), crosses the x-axis at 1.8 dB and is shown by a short mark. It is 0.4 dB closer to the BICM.

By comparing the 2-b/Hz-s BER capacity curve in Fig. 7 to the 3-b/Hz-s curve in Fig. 4, we can see that the extra bit per Hz-s costs 1.6 dB extra in E_b.

An interesting comparison to this BICM is a rate 2-b/T code that takes on Nyquist 8PSK modulator values, for example, a rate 2/3 convolutional code followed by 8PSK modulation. The BER $\approx$0 Shannon limit for such systems is at $E_b/N_0 = 2.8$ dB, which is shown by another x-axis mark. No matter how good the 8PSK coded modulation is, its low-BER performance must lie to the right of 2.8 dB. As can be seen, the BER curve of the concrete BICM–FTN scheme lies well to the *left* of 2.8 dB. What underlies this better performance is the fact that FTN exploits the excess bandwidth of practical base pulses.

VII. MULTICARRIER FTN

Many contemporary systems are based upon multicarrier modulation. Indeed, OFDM is chosen in the downlink of LTE, IEEE802.11, and WiMAX, and practically all digital subscriber lines rely on multiple carriers. It is of great practical importance that the FTN concept extends also to multicarrier modulations.

It appears that the earliest paper dealing with multicarrier FTN (MFTN) is [9], but the topic has received growing attention [29]–[35]. A number of system models exist for multicarrier FTN, but they are all variations of the form

$$s(t) = \sqrt{E_s \tau\phi} \sum_k \left[\sum_n a_{k,n} h(t - n\tau T) \right] \exp(-2\pi i k \phi t/T). \tag{8}$$

In (8), $a_{k,n}$ is the nth symbol in time on subcarrier k. The subcarrier spacing equals $F_\Delta \triangleq \phi/T$ Hz. Similarly, we define the time spacing to be $T_\Delta \triangleq \tau T$ seconds. [The normalization by $\sqrt{\tau\phi}$ in (8) ensures that the transmit power is unchanged.] Traditional multicarrier systems, which are built upon orthogonal carriers and pulses, have a unit time-frequency occupancy $T_\Delta F_\Delta = \tau\phi = 1$ Hz-s. This is a direct consequence of the sampling theorem. MFTN intentionally violates the unit time-frequency occupancy by taking $\tau\phi < 1$. Hence, the symbols $\{a_{k,n}\}$ cannot be detected in a simple way without interference, but the hope is that other

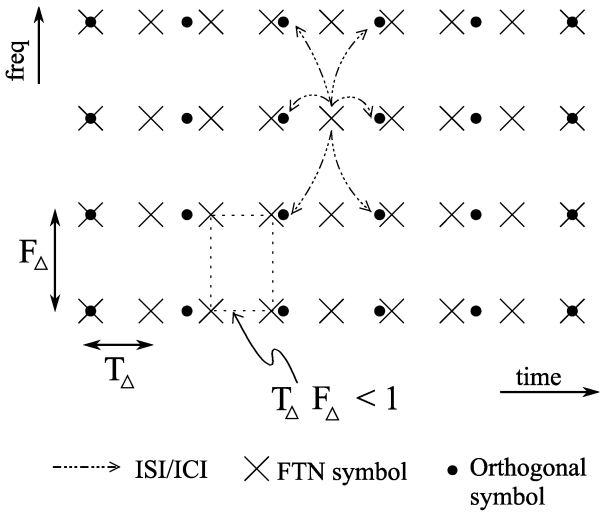

Fig. 8. *Time-frequency view of multicarrier FTN, showing ISI and intercarrier interference (ICI) paths. For simplicity, only time squeezing is employed in this example. (Courtesy D. Dasalukunte, Lund University, Lund, Sweden.)*

gains can be harvested at the expense of receiver complexity. Fig. 8 imagines the $\{a_{k,n}\}$ (crosses) superposed on the time-frequency locations of the original orthogonal transmission (dots), and shows how one pulse can interfere with its neighbors in frequency and time.[9] When $\tau\phi$ is near 1, the interference will be mostly with nearest neighbors, but lower $\tau\phi$ will lead to much more interference and to a challenging detection problem.

An alternate description of these signals appears in [42].

The Mazo limit extends naturally to MFTN. There is a smallest product $T_\Delta F_\Delta$ such that the minimum distance remains at the matched filter bound. With QPSK inputs and a 30% rRC pulse, the smallest product reported where there is no loss in minimum distance is near 0.5 [9], which corresponds to a doubling of the spectral efficiency.

Unfortunately, the state space associated with trellis detection of MFTN is far larger than for single-carrier FTN. This fact has motivated research in applicable low-complexity detection [30], [31]. In [29], a *memoryless* detector for FTN is investigated. Such a detector has the same order of complexity as a detector for orthogonal systems, regardless of the product $T_\Delta F_\Delta$. More precisely, the detector works as follows. Suppose that we want to decode the nth symbol in time on the kth subcarrier. Then, a decision variable can be formed by the inner product $r_{k,n} = \int_{-\infty}^{\infty} r(t)\psi_{k,n}^* \mathrm{d}t$, where $r(t)$ is the received signal and $\psi_{k,n}$ is an arbitrary function ($*$ denotes complex conjugate). In [29], $\psi_{k,n}$ is chosen as the matched filter $\psi_{k,n} = h^*(t - n\tau T_\Delta)\exp(2\pi i k t F_\Delta)$, but other functions such as an minimum mean square error (MMSE) filter can be used.

[9]For simplicity, Fig. 8 takes the original orthogonal frequency spacing. In reality, the new subcarriers lie closer and are phase shifts of each other; the shift is a design parameter in the FTN system.

The decision variable can be broken into a signal-dependent part, an interference-dependent part, and a noise part, in the form

$$r_{k,n} = \sqrt{E_s}a_{k,n} + \underbrace{\sum_{\ell\neq k, m\neq n} \iota_{\ell,m}a_{\ell,m}}_{\text{Interference}} + \eta_{k,n}. \tag{9}$$

Finally, the interference term is modeled as Gaussian noise so that the received signal model becomes

$$r_{k,n} = \sqrt{E_s}a_{k,n} + \tilde{\eta}_{k,n}. \tag{10}$$

The variance of $\tilde{\eta}_{k,n}$ is higher than $N_0/2$ due to the interference and $\tilde{\boldsymbol{\eta}}$ is not white, but the model is the same as that of an interference-free system and its complexity does not depend on $T_\Delta F_\Delta$. Spectral efficiency grows with decreasing $T_\Delta F_\Delta$, but at the same time the receiver performance deteriorates. According to [29], a good compromise lies near $\tau\phi = 0.8$.

A Coded MFTN Receiver: We conclude this section with a test of a coded MFTN system that has 50 carriers $\times$ 100 QPSK symbols, together with an outer (7, 5) convolutional code. We use the same system model as Fig. 3, but the "ISI" box is replaced by an MFTN modulator. The memoryless detector described above has been used, in combination with soft interference cancellation. In each iteration, soft estimates of the interference terms $\{\iota_{\ell,m}\}$ in (9) are calculated based on the LLRs from the channel decoder. The estimates are then subtracted from $r_{k,n}$ before the *a posteriori* LLRs are calculated. Interleaver size 10 000 b and seven iterations were used, although many fewer are required for products 0.5 and 0.55. The results for the products $\tau\phi = 0.45, 0.5, 0.55$ appear in Fig. 9; $\tau = 0.8$ always, so that $\phi = 0.5626, 0.625, 0.6875$. All three cases show BER similar to the no-ISI convolutional code at high SNR, but the spectral efficiency is roughly doubled.

Without FTN, the bit density here is only 1 b/Hz-s; the time-frequency MFTN roughly doubles this spectral efficiency to $1/\tau\phi$, or about 2. The BICM system in Fig. 7 has the same spectral efficiency in terms of 3-dB power bandwidth and is much more energy efficient, but it requires a far more complex receiver.

VIII. HARDWARE IMPLEMENTATIONS

A main reason that FTN-like communication is not yet in wide use is its computational complexity. The situation can be compared to LDPC codes that needed some 40 years from their initial proposal by Gallager in 1963 to their

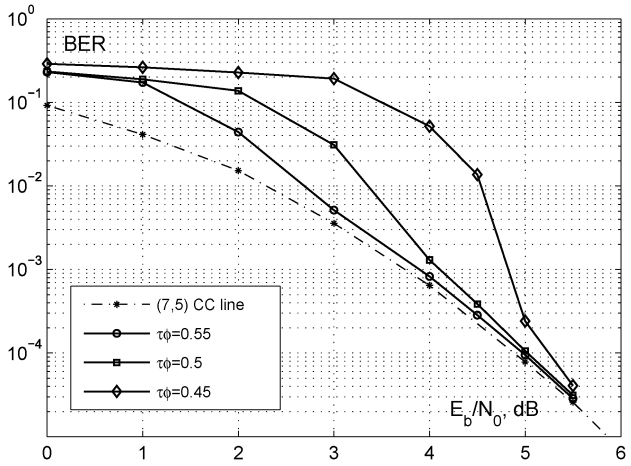

Fig. 9. *Simulation of (7, 5) encoded MFTN systems with different products $T_\Delta F_\Delta$ and memoryless SIC detection of the MFTN system. "CC line" denotes (7, 5) reference performance with orthogonal signaling. All three MFTN systems converge toward the reference. Interleaver size 10 000 b and seven iterations.*

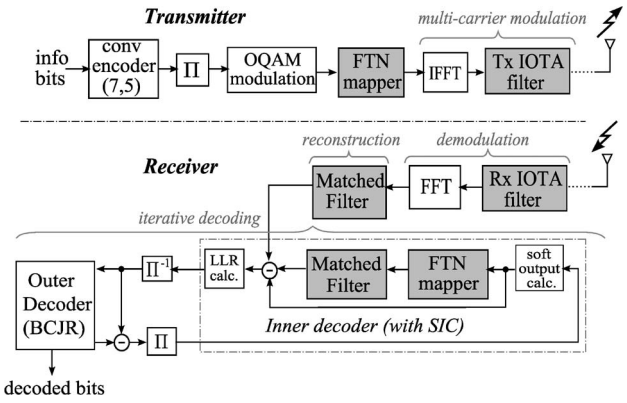

Fig. 10. *Block diagram of FTN coded transmitter/receiver implemented in hardware in [43]. (Courtesy D. Dasalukunte, Lund University.)*

hardware implementation in the 21st century. Bandwidth efficiency cannot after all be achieved by introducing a computational overhead that consumes most of the benefits. Hardware for FTN signaling began to appear around 2009, when integrated circuits matured enough to cope with the complexity demands. Tradeoffs have to be made between energy and bandwidth performance on the one hand and hardware complexity on the other. From a practical perspective, it is also important that FTN systems utilize hardware resources that are similar to those in conventional systems.

The first FTN hardware implementation papers [36]–[38] mainly investigated complexity issues related to the transmitter. Due to their popularity, multicarrier systems have been the primary focus, and one issue has been how FTN-based signaling can be combined with a traditional OFDM system without wasting resources. References [37], [38] investigate a hardware transmitter architecture based on lookup tables that projects the FTN symbols onto an orthogonal basis function set (represented, for example, by the dots in Fig. 8). IOTA pulses[10] are used to reduce the projections onto neighboring pulses. Since IOTA filters are already used to help remove the cyclic prefix of OFDM systems, this is an attractive solution. Another group at the University College London (London, U.K.) has proposed an implementation based on several phase-shifted inverse fast Fourier transforms (IFFTs), whose signals are combined to form frequency compressed FTN signals [40]–[42]. These works studied transmitters only and were mainly verified by field-programmable gate array (FPGA)-based implementations.

The major challenges in FTN hardware implementation are associated with the receiver. As the product $\tau\phi$ drops well below one and compression increases, hardware complexity, memory requirements, and energy consumption grow rapidly in a receiver that has near-ML error performance. With time-frequency FTN, the processing is a 2-D exchange between time and frequency detectors. Reception becomes very complex in the region $0.5 < \tau\phi < 0.7$, even though it lies above the Mazo limit, and the search for efficient detection algorithms has only just begun in this region. Still, good receivers are a worthwhile goal because they offer large bandwidth compaction with little loss in BER.

To our knowledge, the first hardware architecture implemented in silicon for an FTN receiver based on multicarrier modulation was presented in [43]–[45]. The $T_\Delta F_\Delta$ was kept as an adjustable parameter, i.e., in a bad channel with strong ISI, the system can back off to traditional OFDM, while in a good channel, compression can be high. Fig. 10 gives a block diagram of the transceiver. A complete iterative decoder based on the successive interference cancellation in the last section is implemented in hardware. The system includes an outer convolutional code, and a max-log-MAP BCJR convolutional decoder in the receiver. Several hardware optimization techniques were employed to lower the complexity without severely damaging the receiver performance. The architecture was implemented in 65-nm complementary metal–oxide–semiconductor (CMOS) and occupied 0.8 mm^2; clock speed was 100 MHz and power consumption was 9.6 mW at supply voltage 1.2 V. Up to 16 turbo iterations can be executed and $T_\Delta F_\Delta$ is selectable over $\{0.4, 0.5, 0.6, 0.7, 0.9\}$. At eight iterations, the data throughput was 1 Mb/s. This early work shows that the complexity introduced by FTN signaling can be handled in modern silicon

[10]The isotropic orthogonal transform algorithm (IOTA) pulse is an orthogonalization of the Gaussian pulse that is isotropic in time and frequency. They thus reduce the total spread of ISI and ICI. See [39] and [46].

processes. Further hardware implementations of FTN systems seem just around the corner.

IX. CONCLUSION

FTN signaling, either coded or uncoded, can provide up to twice the bandwidth efficiency of ordinary modulation without consuming more transmitter energy per bit. Below a threshold called the Mazo limit, more energy is needed, but the signaling still offers attractive combinations of bandwidth and energy efficiency, which allow operation in new regions of the energy–bandwidth plane. We have demonstrated these facts with tests of a number of software implementations of both baseband and carrier systems. One explanation for this better performance is that unlike Nyquist-pulse signaling, FTN takes advantage of the higher Shannon information rate that practical pulse transmission makes available. At the same time, FTN is a direct extension of techniques such as QAM and OFDM that are already in place. New FTN hardware chips that take advantage of this have already begun to appear.

As higher bit rate wireless systems grow in importance and become shorter range, they gain SNR and need to carry more data bits in the same spectrum. The same can be said for single-carrier systems such as satellite digital television, which find themselves with more SNR but still must work in the same radio-frequency (RF) channel. We have quantified the idea of more bits per unit of bandwidth by introducing the concept of bit density. New systems need to operate at densities of 3–8 b/Hz-s, rather than the lower values that have applied until now. These densities are available with uncoded large-alphabet QAM modulation, but fewer options are known for coded systems. As a general technique coding offers energy and bandwidth gains at high densities as well as low, and it is precisely in the relatively unexplored high densities that coded FTN offers new possibilities. ■

REFERENCES

[1] J. E. Mazo, "Faster-than-Nyquist signaling," *Bell Syst. Tech. J.*, vol. 54, pp. 1451–1462, Oct. 1975.

[2] C.-K. Wang and L.-S. Lee, "Practically realizable digital transmission significantly below the Nyquist bandwidth," in *Proc. IEEE Global Commun. Conf.*, Phoenix, AZ, USA, Dec. 1991, pp. 1187–1191.

[3] N. Seshadri, "Error performance of trellis modulation codes on channels with severe intersymbol interference," Ph.D. dissertation, Dept. Electr., Comput. Syst. Eng., Rensselaer Polytechnical Inst., Troy, NY, USA, Sep. 1986.

[4] A. D. Liveris and C. N. Georghiades, "Exploiting faster-than-Nyquist signaling," *IEEE Trans. Commun.*, vol. 51, no. 9, pp. 1502–1511, Sep. 2003.

[5] P. Kabal and S. Pasupathy, "Partial response signaling," *IEEE Trans. Commun.*, vol. COMM-23, no. 9, pp. 921–934, Sep. 1975.

[6] A. Said and J. B. Anderson, "Bandwidth-efficient coded modulation with optimized linear partial-response signals," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 701–713, Mar. 1998.

[7] J. B. Anderson and A. Svensson, *Coded Modulation Systems*. New York, NY, USA: Kluwer/Plenum, 2003.

[8] F. Rusek and J. B. Anderson, "The two dimensional Mazo limit," in *Proc. IEEE Int. Symp. Inf. Theory*, Adelaide, Australia, Sep. 2005, pp. 970–974.

[9] F. Rusek and J. B. Anderson, "Multi-stream faster than Nyquist signaling," *IEEE Trans. Commun.*, vol. 57, no. 5, pp. 1329–1340, May 2009.

[10] F. Rusek and J. B. Anderson, "Optimal sidelobes under linear and faster than Nyquist modulation," in *Proc. IEEE Int. Symp. Inf. Theory*, Nice, France, Jun. 2007, pp. 2301–2304.

[11] J. B. Anderson, A. Prlja, and F. Rusek, "New reduced state space BCJR algorithms for the ISI channel," in *Proc. IEEE Int. Symp. Inf. Theory*, Seoul, Korea, Jun. 2009, pp. 889–893.

[12] J. B. Anderson and A. Prlja, "Turbo equalization and an M-BCJR algorithm for strongly narrowband intersymbol interference," in *Proc. Int. Symp. Inf. Theory Appl.*, Taichung, Taiwan, Oct. 2010, pp. 261–266.

[13] A. Prlja and J. B. Anderson, "Reduced-complexity receivers for strongly narrowband intersymbol interference introduced by faster-than-Nyquist signaling," *IEEE Trans. Commun.*, vol. 60, no. 9, pp. 2591–2601, Sep. 2012.

[14] Y. G. Yoo and J. H. Cho, "Asymptotic optimality of binary faster-than-Nyquist signaling," *IEEE Commun. Lett.*, vol. 14, no. 9, pp. 788–790, Sep. 2010.

[15] F. Rusek and J. B. Anderson, "Constrained capacities for faster than Nyquist signaling," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 764–775, Feb. 2009.

[16] C. Douillard, A. Picart, P. Didier, M. Jezequel, C. Berrou, and A. Glavieux, "Iterative correction of intersymbol interference: Turbo equalization," *Eur. Trans. Telecommun.*, vol. 6, pp. 507–511, Sep./Oct. 1995.

[17] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 2, pp. 284–287, Mar. 1974.

[18] J. B. Anderson and F. Rusek, "The Shannon bit error limit for linear coded modulation," in *Proc. Int. Symp. Inf. Theory Appl.*, Parma, Italy, Oct. 2004, pp. 9–11.

[19] J. B. Anderson and M. Zeinali, "Best rate 1/2 convolutional codes for turbo equalization with severe ISI," in *Proc. IEEE Int. Symp. Inf. Theory*, Cambridge, MA, USA, Jul. 2012, pp. 2366–2370.

[20] G. Colavolpe and A. Barbieri, "On MAP symbol detection for ISI channels using the Ungerboeck observation model," *IEEE Commun. Lett.*, vol. 9, no. 8, pp. 720–722, Aug. 2005.

[21] U. L. Dang, W. H. Gerstacker, and S. T. M. Slock, "Maximum SINR prefiltering for reduced-state trellis-based equalization," in *Proc. IEEE Int. Conf. Commun.*, Kyoto, Japan, Jun. 2011, DOI: 10.1109/icc.2011.5963034.

[22] D. D. Falconer and F. R. Magee, "Adaptive channel memory truncation for maximum likelihood sequence estimation," *Bell Sys. Tech. J.*, vol. 52, pp. 1541–1562, Nov. 1973.

[23] S. A. Fredricsson, "Joint optimization of transmitter and receiver filter in digital PAM systems with a Viterbi detector," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 2, pp. 200–210, Mar. 1976.

[24] C. T. Beare, "The choice of the desired impulse response in combined linear-Viterbi algorithm equalizers," *IEEE Trans. Commun.*, vol. COMM-26, no. 8, pp. 1301–1307, Aug. 1978.

[25] F. Rusek and A. Prlja, "Optimal channel shortening of MIMO and ISI channels," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 810–818, Feb. 2012.

[26] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai Shitz, "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 46, no. 6, pp. 1953–1967, Nov. 1994.

[27] A. Ganti, A. Lapidoth, and I. E. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 46, no. 6, pp. 2315–2328, Nov. 2000.

[28] F. Rusek and J. B. Anderson, "Serial and parallel concatenations based on faster than Nyquist signaling," in *Proc. IEEE Int. Symp. Inf. Theory*, Seattle, WA, USA, Jul. 2006, pp. 1993–1997.

[29] A. Barbieri, D. Fertonani, and G. Colavolpe, "Time-frequency packing for linear modulations: Spectral efficiency and practical detection schemes," *IEEE Trans. Commun.*, vol. 57, no. 10, pp. 2951–2959, Oct. 2009.

[30] I. Kanaras, "Spectrally efficient multicarrier communication systems: Signal detection, mathematical modelling and optimisation," Ph.D. dissertation, Dept. Electron. Electr. Eng., Univ. College London, London, U.K., 2010.

[31] A. Chorti, I. Kanaras, M. R. D. Rodrigues, and I. Darwazeh, "Joint channel equalization and detection of spectrally efficient FDM signals," in *Proc. IEEE Conf. Pers. Indoor*

Mobile Radio Commun., Istanbul, Turkey, Sep. 2010, pp. 177–182.

[32] S. Isam and I. Darwazeh, "Precoded spectrally efficient FDM system," in *Proc. IEEE Conf. Pers. Indoor Mobile Radio Commun.*, Istanbul, Turkey, Sep. 2010, pp. 99–104.

[33] I. Kanaras, A. Chorti, M. Rodrigues, and I. Darwazeh, "Investigation of a semidefinite programming detection for a spectrally efficient FDM system," in *Proc. IEEE Conf. Pers. Indoor Mobile Radio Commun.*, Tokyo, Japan, Sep. 2009, pp. 2827–2832.

[34] M. He, D. Liang, and Q. Cao, "A modulation with higher bandwidth efficiency than OFDM," in *Proc. Int. Conf. Signal Process. Syst.*, Dalian, China, Jul. 2010, pp. 393–397.

[35] F.-M. Han and X.-D. Zhang, "Wireless multicarrier digital transmission via Weyl-Heisenberg frames over time-frequency dispersive channels," *IEEE Trans. Commun.*, vol. 57, no. 6, pp. 1721–1733, Jun. 2009.

[36] I. Kanaras, A. Chorti, M. R. D. Rodrigues, and I. Darwazeh, "Spectrally efficient FDM signals: Bandwidth gain at the expense of receiver complexity," in *Proc. IEEE Int. Conf. Commun.*, Dresden, Germany, Jun. 2009, DOI: 10.1109/ICC.2009.5199477.

[37] D. Dasalukunte, F. Rusek, J. B. Anderson, and V. Owall, "A transmitter architecture for faster-than-Nyquist signaling systems," in *Proc. IEEE Int. Symp. Circuits Syst.*, Taipei, Taiwan, May 2009, pp. 1028–1031.

[38] D. Dasalukunte, F. Rusek, V. Owall, K. Ananthanarayanan, and M. Kandasamy, "Hardware implementation of mapper for faster-than-Nyquist signaling transmitter," in *Proc. IEEE NORCHIP*, Trondheim, Norway, Nov. 2009, DOI: 10.1109/NORCHP.2009.5397801.

[39] B. Le Floch, M. Alard, and C. Berrou, "Coded orthogonal frequency division multiplex," *Proc. IEEE*, vol. 83, no. 6, pp. 982–996, Jun. 1995.

[40] M. R. Perrett and I. Darwazeh, "Flexible hardware architecture of SEFDM transmitters with real-time non-orthogonal adjustment," in *Proc. 18th Int. Conf. Telecommun.*, Cyprus, May 2011, pp. 369–374.

[41] P. N. Whatmough, M. R. Perrett, S. Isam, and I. Darwazeh, "VLSI architecture for a reconfigurable spectrally efficient FDM baseband transmitter," in *Proc. IEEE Int. Symp. Circuits Syst.*, Rio de Janeiro, Brasil, May 2011, pp. 1688–1691.

[42] P. N. Whatmough, M. R. Perrett, S. Isam, and I. Darwazeh, "VLSI architecture for a reconfigurable spectrally efficient FDM baseband transmitter," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 59, no. 5, pp. 1107–1118, May 2012.

[43] D. Dasalukunte, F. Rusek, and V. Öwall, "Multicarrier faster-than-Nyquist signaling transceivers: Hardware architecture and performance analysis," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 58, no. 4, pp. 827–838, Apr. 2011.

[44] D. Dasalukunte, F. Rusek, and V. Öwall, "Improved memory architecture for multicarrier faster-than-Nyquist iterative decoder," in *Proc. IEEE Comput. Soc. Annu. Symp. Very Large Scale Integr.*, Chennai, India, Jul. 2011, pp. 296–300.

[45] D. Dasalukunte, F. Rusek, and V. Öwall, "A 0.8 mm^2 9.6 mW implementation of a multicarrier faster-than-Nyquist signaling iterative decoder in 65 nm CMOS," in *Proc. 38th Eur. Solid State Circuits Conf.*, Bordeaux, France, Sep. 2012, pp. 173–176.

[46] M. Alard, "Construction of a multicarrier signal," U.S. Patent 6 278 686, Aug. 2001.

ABOUT THE AUTHORS

John B. Anderson (Fellow, IEEE) was born in New York State in 1945. He received the Ph.D. degree in electrical engineering from Cornell University, Ithaca, NY, USA, in 1972.

During 1972–1980, he was on the electrical engineering faculty at McMaster University, Hamilton, ON, Canada, and during 1981–1998, he was a Professor at Rensselaer Polytechnic Institute, Troy, NY, USA. Since 1998, he has held the Ericsson Chair in Digital Communication at Lund University, Lund, Sweden. He has held visiting professorships at the University of California at Berkeley (Berkeley, CA, USA), Chalmers University (Göteborg, Sweden), Queen's University (Kingston, ON, Canada), Deutsche Luft und Raumfahrt (Cologne, Germany), and Technical University of Munich (Munich, Germany). He was the Director of the Swedish Strategic Research Foundation Center for High Speed Wireless Communication at Lund University during 2005–2011. He is an author of six textbooks, including most recently *Digital Transmission Engineering* (Piscataway, NJ, USA: IEEE Press, 2005, 2nd ed.), *Coded Modulation Systems* (New York, NY, USA: Plenum/Springer, 2003), and *Understanding Information Transmission* (Piscataway, NJ, USA: IEEE Press, 2005). Since 1998, he has edited the IEEE Press book *Series on Digital and Mobile Communication*. His research work is in coding and communication algorithms, bandwidth-efficient coding, and the application of these to data transmission and compression. He has served widely as a consultant in these fields.

Dr. Anderson was a member of the IEEE Information Theory Society Board of Governors during 1980–1987 and 2001–2006, serving as the Society's Vice President and President (1985). In 1983 and 2006, he was Co-Chair of the IEEE International Symposium on Information Theory. In the IEEE publications sphere, he served on the Publications Board of the IEEE on three occasions, and was Editor-in-Chief of IEEE Press during 1994–1996 and 2012–2014. He has also served as an Associate Editor for several IEEE Transactions. He received the Humboldt Research Prize (Germany) in 1991. In 1996, he was elected Swedish National Visiting Chair in Information Technology. He received the IEEE Third Millennium Medal in 2000.

Fredrik Rusek was born in Lund, Sweden, in 1978. He received the M.S. and Ph.D. degrees in electrical engineering from Lund University, Lund, Sweden, in 2003 and 2007, respectively.

Since 2008, he has held an assistant professorship at the Department of Electrical and Information Technology, Lund University. His research interests include modulation theory, equalization, wireless communications, and applied information theory.

Viktor Öwall (Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering from Lund University, Lund, Sweden, in 1988 and 1994, respectively.

During 1995–1996, he joined the Electrical Engineering Department, University of California at Los Angeles, Los Angeles, CA, USA, as a Postdoctoral Researcher, where he mainly worked in the field of multimedia simulations. Since 1996, he has been with the Department of Electrical and Information Technology, Lund University. He is currently full Professor at the same department and since 2009 the Head of the Department. He is the Director of the VINNOVA Industrial Excellence Center in System Design on Silicon (SoS). His main research interest is in the field of digital hardware implementation, especially algorithms and architectures for wireless communication, image processing, and biomedical applications. His research projects include combining theoretical research with hardware implementation aspects in the areas of wireless communication, video processing, and digital holography.

Dr. Öwall was an Associate Editor of the IEEE Transactions on Circuits and Systems—Part II: Analog and Digital Signal Processing from 2000 to 2002 and of the IEEE Transactions on Circuits and Systems—Part I: Regular Papers from 2007 to 2009. He was Guest Editor for the Special issue on ISCAS 2010 of the IEEE Transactions on Biomedical Circuits and Systems.