Optimal sidelobes under linear and faster-than-Nyquist modulation

Rusek, Fredrik; Anderson, John B

Published in:
Proc., 2007 IEEE Inter. Symp. Information Theory

2007

Optimal Side Lobes under Linear and Faster-than-Nyquist Modulation

John B. Anderson and Fredrik Rusek
Electrical and Information Tech. Dept. and Strategic Center for High Speed Wireless Communication
Lund University
Box 118, SE-221 00 Lund SWEDEN
Email: anderson@it.lth.se, fredrikr@it.lth.se

***Abstract*— We minimize the frequency and time occupancy of multicarrier binary linear modulation based on two-dimensional faster than Nyquist (FTN) signaling. FTN analysis provides the asymptotic time–frequency consumption per bit and prolate spheroidal wave analysis minimizes the side lobe occupancy. For both problems, an excellent choice is a Gaussian pulse, with some adjustment of the side lobes.**

I. INTRODUCTION

This paper investigates how to minimize the main and side lobe power of linear modulation, with and without faster-than-Nyquist signaling. We will conclude that from either the main or side lobe point of view, the Gaussian pulse shape is an excellent choice.

Ordinary linear modulation signals have the baseband form

$$s(t) = \sqrt{E_s/T} \sum_n a_n h(t - nT), \qquad (1)$$

in which a_n are M-ary independent and identically distributed data values with zero mean and unit variance, and $h(t)$ is a unit-energy baseband pulse. This form underlies QAM, TCM, and the subcarriers in orthogonal frequency division multiplex (OFDM), as well as many other transmission systems. In these, $h(t)$ is a T-orthogonal pulse, that is, $\int h(t-nT)h^*(t-mT)dt$ is zero, $m \neq n$. In 1975 Mazo[1] pointed out that binary $\mathrm{sinc}(t/T)$ pulses in (1) could be sent "faster", with symbol time $T_\Delta < T$, without loss of signal minimum distance ($d^2_{\min} = 2$ for binary pulses). The asymptotic error probability is thus unaffected, although the receiver is more complex. This he called faster than Nyquist (FTN) signaling, because the pulses appear faster than allowed by Nyquist's orthogonality limit.

FTN signaling has since been extended in many ways, and this paper will use several. The modulation can be nonbinary and the pulses need not be $\mathrm{sinc}(\cdot)$. In fact, the pulse need not be orthogonal at any T; for reasonable pulses such as the Gaussian there will be a least T_Δ at which the minimum distance first falls below the isolated pulse value. Furthermore, the FTN concept can be applied simultaneously in time as well as frequency: Many signals of form (1) can be stacked in frequency to form the inphase and quadrature (I/Q) signal given by the real part of

$$s(t) = \sqrt{2E_s/T} \sum_{k=0}^{K-1} \sum_{n=0}^{N-1} a_{k,n} h(t - nT)\, e^{\jmath 2\pi (f_0 + f_k)t} \qquad (2)$$

This is a superposition of $2K$ linear carrier modulations that carries NK complex data values $a_{k,n} = a^I_{k,n} + \jmath a^Q_{k,n}$. If $f_k = kf_\Delta$, $k = 0, 1, \ldots K-1$, and f_Δ is equal twice the single-sided bandwidth of $h(t)$, the $\cos((f_0+f_k)t)/\sin((f_0+f_k)t)$ carrier signals are orthogonal; if $h(t)$ is T-orthogonal, all $2NK$ pulses are mutually orthogonal. In OFDM, both conditions hold at least approximately. In FTN signaling of this type, there are least combinations of f_Δ and T_Δ for which there is no loss of signal minimum distance. Collectively these make up the *Mazo limit* to signaling with this $h(t)$ and alphabet. Of particular interest in this paper is the least achievable product $f_\Delta T_\Delta$. We have introduced these generalizations in [2], [3], [4], where more details may be found.

With or without FTN, signals occupy a certain bandwidth and time. A communication pulse has a main lobe and side lobes in both frequency and time. A view that closely relates to practice is that the symbols are carried by the main lobe and will be detected correctly most of the time if the main lobe can be observed; the side lobes, on the other hand, interfere with other users, especially when the desired signal lies far and an undesired signal lies near. The *occupancy* of the signal, the time–frequency that it denies to others, is the product of the time and frequency total widths. Figure 1 shows this occupancy, together with the time and frequency of the main lobe; the time–frequency centers of some pulses are shown as dots. In this paper we seek pulses $h(t)$ that minimize the occupancy, when the side lobes count or when only the main lobe counts.

One or both of this bandwidth and time is in theory infinite, but in the practical world we only perceive signals above a certain threshold. In this paper we define the bandwidth and time frame in an energy out of band sense, namely, they are widths inside which all but γ of the signal exists. γ will be taken as a fraction of the pulse energy E_s. For example, $\gamma = .01$ leads to the 99% energy frequency and time frame, outside which lies 1% of the signal energy in frequency or time. These definitions are formally set in Section II.

The paper will first set up a time–frequency analysis frame-

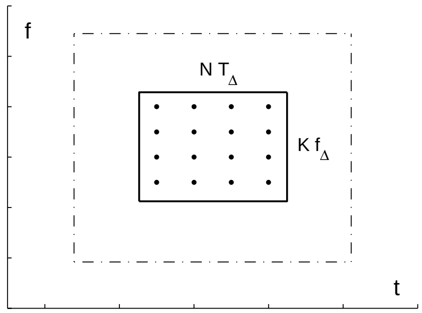

Fig. 1. Time–frequency occupancy of the main (solid) and side (dashed) lobes for a signal with N pulses and K subcarriers.

work. Section III then optimizes the length of the pulse sequence N against the number of subcarriers K. Section IV sketches FTN results for these multicarrier modulations. Numerical results for the time–frequency consumption of root RC and more optimal pulse shapes appear in Section V. Throughout, the channel will be subject to additive white Gaussian noise (AWGN) with power spectral density $N_0/2$ W/Hz, with no other disturbances.

II. Measurement of Frequency and Time

In what follows we take the standard signal space theoretical view of signals in AWGN and measure signal bandwidth and time. In this view, what matters about a signal is its *time–bandwidth product*, in Hz-s.[1] Suppose a $s(t)$ is time-scaled by a factor, keeping symbol energy constant; its Fourier transform will be scaled in frequency by the inverse factor, and its time–bandwidth product, by any reasonable measure, will be unchanged. The detection properties of a set of signals so scaled are also invariant (they depend only on E_s/N_0), as is the Shannon capacity of such signals. The scaled sets are the same set in the signal theory view. In practice, a time limit is set by the latency allowed in the communication, and the widest allowed bandwidth is set by limits of hardware and electromagnetics. But latency and hardware are not part of signal theory.

While pulses have time width and bandwidth, the signals in (1)–(2) are stacked sequences of pulses driven by random symbols, and we will describe their time and bandwidth in terms of energy densities. For simplicity take both $h(t)$ and $H(f)$ to be symmetric and real. Then the average energy spectral density (ESD) of signal $s(t)$ per symbol time is (take positive f only)

$$\begin{aligned}\mathfrak{S}_F(f) &\triangleq \frac{1}{N}\mathcal{E}\left[\left|\mathcal{F}\left\{s(t)\right\}\right|^2\right] \\ &= \frac{1}{N}\sum_{k=0}^{K-1}|H(f-kf_\Delta-f_0)|^2, \quad f_0 \gg Kf_\Delta \quad (3)\end{aligned}$$

[1]The dimension Hz-s is in a sense dimensionless but we retain it is a measure of the “changeability” of a signal; one Hz-s is the changeableness of a signal per unit of the time–frequency product, asymptotically as time and frequency occupancy grow.

The expectation is over the data symbols. Note that the average energy in $s(t)$ is $2NK$ and the integral of $\mathfrak{S}_F(f)$ is KE_s. Next define the average energy temporal density (ETD) per subcarrier as

$$\begin{aligned}\mathfrak{S}_T(t) &\triangleq \frac{1}{K}\mathcal{E}\left[|s(t)|^2\right] \\ &= \frac{1}{K}\sum_{n=0}^{N-1}|h(t-nT_\Delta)|^2, \quad \text{all } t \qquad (4)\end{aligned}$$

Note that several pulses may contribute to the ETD in a given symbol interval, and several subcarriers may contribute to the ESD around a given subcarrier $f_0 + kf_\Delta$. For large K and N, $\mathfrak{S}_F(f)$ and $\mathfrak{S}_T(t)$ settle into a steady-state region about Kf_Δ Hz wide located above f_0 and about NT_Δ s wide around time 0. These we take to be the main lobe spectral and temporal widths. On a per bit basis the asymptotic occupancy is the product per bit, $\frac{1}{2}f_\Delta T_\Delta$ Hz-s/bit. The factor 1/2 accounts for the fact that there are $\sin$ and $\cos$ subcarriers.

A useful benchmark is orthogonal signaling with $\mathrm{sinc}(\cdot)$ pulses and no FTN, for which $f_\Delta = 1/T$, $T_\Delta = T$, and the asymptotic occupancy is 1/2 Hz-s/bit.

The side lobes widths will be defined by the average energy out of band, as a fraction of one pulse's energy. This is justified by the fact that interference is mostly “local”, in the neighborhood of one pulse's frequency and time. For $0 < \gamma < 1$ let the extra upper side lobe frequency width be the ε'_F Hz that satisfies

$$\frac{1}{E_s}\int_{\varepsilon'_F+f_0+(K-\frac{1}{2})f_\Delta}^{\infty}\mathfrak{S}_F(f)\,df \;=\; \gamma \qquad (5)$$

The spectrum $[f_0 - \frac{1}{2}f_\Delta, f_0 + (K-\frac{1}{2})f_\Delta]$ is reserved for the main lobe, and because of symmetry the total extra on both sides is $2\varepsilon'_F$. Similarly, the right hand time side lobe extra width is the ε'_T s that satisfies

$$\frac{1}{E_s}\int_{\varepsilon'_T+(N-\frac{1}{2})T_\Delta}^{\infty}\mathfrak{S}_T(f)\,dt \;=\; \gamma \qquad (6)$$

The time $[-\frac{1}{2}T_\Delta, (N-\frac{1}{2})T_\Delta]$ is reserved for the main lobe, and the total extra on both sides is $2\varepsilon'_T$. The total time–frequency occupancy is taken as

$$\frac{1}{2}(N+2\varepsilon_T)(K+2\varepsilon_F)f_\Delta T_\Delta \quad \text{Hz-s} \qquad (7)$$

where $\varepsilon_F = \varepsilon'_F/f_\Delta$ and $\varepsilon_T = \varepsilon'_T/T_\Delta$.

III. Optimal Subcarrier and Symbol Numbers

For a given packet of bits, it is necessary to optimize N and K, the number of symbols in time and frequency, so as to minimize occupancy per bit for the same product NK. This is because each pulse shape $h(t)$ leads to its own side lobe expansions ε_T and ε_F; if ε_T is longer than ε_F then a larger N than K is needed, and vice versa. The best ratio N/K follows from a simple derivation. Fix $\rho = NK$ and $f_\Delta T_\Delta$ and then minimize (7) over K to obtain that Eq. (7) takes minimum value

$$\frac{1}{2}\left(\sqrt{NK}+2\sqrt{\varepsilon_T\varepsilon_F}\right)^2 f_\triangle T_\triangle \tag{8}$$

when

$$\frac{N}{K}=\frac{\varepsilon_T}{\varepsilon_F} \quad \text{or} \quad K=\sqrt{\rho\varepsilon_F/\varepsilon_T} \tag{9}$$

For extreme values of $\varepsilon_F/\varepsilon_T$, either N or K is simply 1. For large packets any ratio N/K leads to the same occupancy per bit, since only the main lobe matters; otherwise, the best ratio is set by the side lobe expansion factors. Some typical values will be given in Section V.

IV. PSWF Estimation of the Minimal Side Lobes

For a single baseband pulse $h(t)$ a well known method based on the prolate spheroidal wave function (PSWF) exists to minimize the energy outside given frequency and time widths $[-W, W]$ Hz and $[0, \tau]$ s. The method is described in a series of five papers beginning in 1965 by Slepian and coauthors, the most useful of which for our purposes are the first two [5]. Many papers have applied the PSWF method to pulse design since 1965; an early example that treats finite pulses is [6]. The optimal solutions for our problem are allowed to have infinite support. They are eigenfunctions of a certain integral operator and depend only on the time–frequency product $W\tau$ [5, p. 45ff]. Suppose the same fraction γ of the pulse energy is to lie outside both $[-W, W]$ and $[0, \tau]$; then there corresponds a principal eigenvalue and PSWF eigenfunction that can be used to construct the pulse with least $W\tau$ product subject to the double out of band constraint [5, pp. 65–80]. A further useful property for us is that the solution tends to Gaussian pulse as $\gamma \to 0$, i.e., as $W\tau$ grows.[2]

Figure 2 plots the least-$W\tau$ solutions for $\gamma = .01, .001, .00001$, which means that half these amounts lie on each side of the time and frequency bands. The $W\tau$ products are, respectively, 0.859, 1.28, 2.84. A Gaussian pulse closely fits the main part of the second two pulses.

The PSWF solution is not precisely the solution to the side lobe problem in Section II because more than one pulse contributes to the side lobe energy densities. But with rapidly falling temporal and spectral densities, the ETD and ESD are to a first approximation equal to the tails of $|h(t)|^2$ and $|H(f)|^2$. Our main interest is pulses close to Gaussian, which are particularly rapidly falling. In any case, we can calculate the actual $\mathfrak{S}_F$ and $\mathfrak{S}_T$ for the PSWF solutions and compare their out of band energies to those of other popular pulses.

V. Good FTN Parameters and Numerical Calculations

In this section we compute numerical values for the main and side lobe occupancies of classic 10 and 30% root RC pulse signaling systems, as well as a PSWF example and a simple Gaussian pulse system than lies close to the PSWF example.

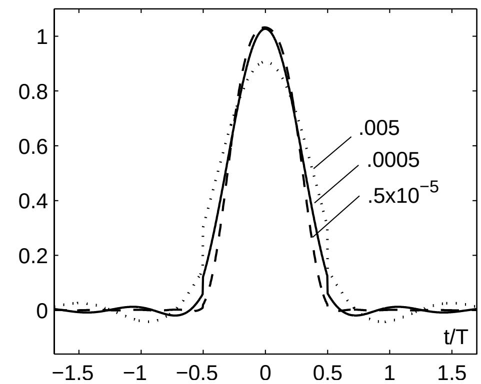

Fig. 2. PSWF solutions for the pulse with least $W\tau$ product, among those with energy fraction $\frac{1}{2}\{.01, .001, .00001\}$ outside each side of $[-W, W]$ and $[-\tau/2, \tau/2]$. Symbol time is set to 1.

The main lobe calculation depends on an FTN analysis. It has been taken up in our earlier papers [3], [4] and is not repeated here for space reasons. In brief, for a given pulse one computes the Euclidean minimum distance over a large number of error events in order to obtain a tight distance estimate. This is repeated for a number of combinations of $f_\triangle$ and $T_\triangle$, searching for the one with $d^2_{\min}$ estimate 2 that has the least product. Table I shows the outcome of this for the two root RC pulses, the PSWF solution for $\gamma = .0005$ out of band average energy in both time and frequency, and the Gaussian pulse $h(t) = 1/\sqrt{2\pi\sigma^2}\exp(-t2\sigma^2)$ with $\sigma^2 = .399$.[3] The $f_\triangle$ and $T_\triangle$ in the table are the best that we presently know for the pulse. In terms of asymptotic main lobe occupancy, it can be seen that 10% root RC is better than 30%, and both consume about half the sinc pulse benchmark. The Gauss pulse is chosen because it closely resembles the PSWF solution, and the product $f_\triangle T_\triangle$ is close to 0.30 for both. This product is somewhat worse than the 10% root RC case, but we will see next that the root RC has very poor side lobe behavior.

Pulse	$h(t)$	$f_\triangle$	$T_\triangle$	$.5 f_\triangle T_\triangle$
1	30% rtRC	.674	.89	.300
2	10% rtRC	.660	.843	.278
3	Gauss, .399	.706	.86	.303
4	PSWF, .0005	1.154	.52	.300

TABLE I

Best known FTN parameters for four pulses.

In Table II is shown the outcome of the side lobe calculations for the same pulses. First are shown the time and frequency expansions ε for the four cases in units of $f_\triangle$ and $T_\triangle$. These are very small for the two root RC pulses, which are designed to have a narrow, finite bandwidth; their frequency

[2]This is because the eigenfunction of the basic PSWF operator with some $W > 0$ tends to a Gaussian pulse as τ grows.

[3]This is the pulse whose Fourier transform is identical to itself; it thus has balanced time and frequency spread. When scaled wider by 1.54, the PSWF solution here is approximately this Gaussian pulse.

side lobes are only 0.81 and 0.61 f_Δ-units larger than their main lobes. But this is bought at a huge price in the time side lobes, which are $27T_\Delta$ in the 10% case. In order to achieve a small occupancy, the time and frequency expansions must be kept in balance; it is $\sqrt{\varepsilon_T \varepsilon_F}$ in Eq. (8) that sets the occupancy for a given packet size. The balance is performed much better by the Gauss and PSWF pulses.

Pulse	$2\varepsilon_T$	$2\varepsilon_F$	Opt. N/K	$W\tau/100$ Opt.	$W\tau/100$ $K=1$	$W\tau/10000$ Opt.	$W\tau/10000$ $K=1$
1	8.2	.82	10.0	.48	.59	.32	.55
2	27.0	.61	44.0	.55	.57	.30	.45
3	1.16	1.63	.71	.39	.81	.31	.80
4	.99	1.26	.79	.32	.68	.31	.68

TABLE II

SIDE LOBE WIDTHS, OPTIMAL N/K RATIOS, AND SIDE LOBE OCCUPANCY (HZ-S/BIT) FOR 0.0005 FRACTIONAL ENERGY OUT OF BAND AND SELECTED PULSES. NK IS 100 AND 10000.

Next shown in the table is the optimal pulse-to-subchannel ratio N/K. It needs to be 44 in the 10% root RC case to overcome the long time side lobes there. N/K is close to unity in the Gauss and PSWF cases because of the excellent balance of their side lobes. When the optimal ratio is employed the side lobe occupancy in Hz-s/bit is the columns marked "Opt." in the table. One cannot of course have a fractional K or N, but the optimum is quite flat, so that nearby integers that multiply to ρ provide near-"Opt." occupancy. There are two such columns, one for packet size equal twice $\rho = 100$ bits and one for $\rho = 10000$. Next to them is the occupancy for $K = 1$, when there is only one carrier. Comparing the outcomes shows the effect of optimal frequency/time allocation; without this, the side lobe occupancy is in every case much larger.

With small packets, the Gauss and PSWF pulses are clearly superior when an optimal N/K is used, and the PSWF is nearly twice as efficient as the 10% root RC pulse. With very large packets, the main lobe dominates, and with optimal N/K all pulse have similar efficiency, reflecting the fact that their FTN performance is similar.

When the energy out of band is restricted to a low value like $\gamma = .00001$, the effects just reported are more extreme. The root RC pulses especially have long time side lobes, so long that the best allocation of N and K needs $K = 1$.

The FTN and minimal side lobe problems are distinct, and the PSWF solution to the second is not necessarily the least time–frequency solution to the FTN problem. It is perhaps fortuitous that a Gaussian-like pulse nearly solves both problems. However, a closer study of pulses with tight two-dimensional FTN packing shows that FTN behavior is strongly affected by the main temporal lobe of the pulse, and the major pulse candidates all have roughly similar main lobes. If the near side lobes are reasonably small in frequency and time, they have little effect on minimum distance. The side lobe optimization, on the other hand, has to do with small modifications to the side lobe structure. The fortuitous aspect is that the PSWF solution does not affect much the compact main pulse shape.

Although we have not taken it up in this paper, receiver design is a challenging and interesting problem. With one subcarrier, FTN can be closely modeled as intersymbol interference, and a simple Viterbi algorithm is an effective solution. With more than one, error patterns are two dimensional, in a manner similar to what occurs in magnetic recording. We have explored several iterative schemes. In general, as the $f_\Delta T_\Delta$ product moves toward the Mazo limit, decoding becomes much harder.

VI. CONCLUSION

We have solved two problems in the time–frequency occupancy of binary multicarrier linear modulation. One, the main lobe occupancy, was minimized by faster than Nyquist analysis, that is, by seeking the closest packing in time and frequency that retains asymptotically the error performance of binary antipodal signaling. The second, the side lobe occupancy, was solved by a classical prolate spheroidal wave approach. Fortunately, the solution to each is close to a Gaussian pulse, so that we can say that a prolate function is close to optimal for FTN signaling from either point of view. FTN signaling does require a more complex receiver, but if this is accepted, the choice of pulse has a small effect on hardware, as long as its side lobes are reasonable. The prolate pulse is thus clearly the attractive one.

The result of this work is a binary modulation that runs for reasonable packets at around 0.30 Hz-s/bit. With a prolate pulse, packets as small as 200 bits do not much increase this consumption. In future work we will extend the calculations given here to nonbinary signaling and explore more fully the receiver problem.

ACKNOWLEDGMENTS

This work was supported by the Swedish Research Council (VR) through Grant 621-2003-3210, and by the Swedish Foundation for Strategic Research (SSF) through its Strategic Center for High Speed Wireless Communication at Lund.

REFERENCES

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

[2] F. Rusek and J.B. Anderson, "M-ary Coded Modulation by Butterworth Filtering," *Proc.*, 2003 Int. Symp. Information Theory, Yokohama, p. 184, June 2003.

[3] F. Rusek and J.B. Anderson, "The Two Dimensional Mazo Limit," *Proc.*, 2005 Int. Symp. Information Theory, Adelaide, pp. 970–974, Sept. 2005.

[4] F. Rusek and J.B. Anderson, "Non binary and precoded faster than Nyquist signaling," accepted, *IEEE Trans. Commun.*, Oct. 2006.

[5] D. Slepian, H.O. Pollak and H.J. Landau, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty," Parts I and II, *Bell Sys. Tech. J.*, pp. 43–63 and 65–84, Jan. 1961.

[6] P.H. Halpern, "Optimum Finite Duration Nyquist Signals," *IEEE Trans. Commun.*, vol. 37, pp. 884–888, June 1979.