

Fig. 6. The BER of the RAKE receiver with $L = 1, 2,$ and 4 fingers used in a multipath Rayleigh channel versus the signal-to-interference ratio E_b/J_0 . In all cases $\rho = 0.25$. Dashed line = complete coherence; solid line = using phase control with $\rho = 0.25$.

first-order closed loop. In a noiseless case, a first-order loop never achieves the zero error condition since the phase adjustments are proportional to the current phase error.

The bit-error rate for the system for different paths L as a function of the signal energy to interference ratio E_b/J_0 is depicted in Fig. 6. The dashed lines show the BER for a receiver with exact coherence, while the solid lines show the BER for a receiver using the phase algorithm in the previous section. We see that the performance converges to that of the receiver with complete coherence rapidly as E_b/J_0 increases.

In order to see the implications of a different number of users in the system, we may use the relation (46) to compute the ratio E_b/N_0 , but recall that we have used a Gaussian approximation which is accurate only if the product $LKN \gg 1$.

VI. CONCLUSIONS

We have analyzed the performance of the coherent reception in different channel scenarios, the AWGN channel, and the multipath Rayleigh channel. In the analysis we have taken into consideration imperfections in coherence, since due to the system noise and multiple-access interference perfect phase coherence is not often available. However, at levels of the signal-energy-to-interference ratios of interest we have seen that the receivers behave as if there were complete coherence.

One topic that we have not considered is the phase acquisition process, which is of interest in a fast fading channel where the carrier phases change faster. In such a channel, the algorithm described must be changed to a loop of higher order. Otherwise, the algorithm may not be able to follow the faster changes of the phase.

ACKNOWLEDGMENT

The authors wish to thank Dr. A. J. Viterbi for his insightful comments on this work and the anonymous reviewers for their very useful suggestions.

REFERENCES

- [1] A. J. Viterbi, *CDMA Principles of Spread Spectrum Communication*. Reading, MA: Addison-Wesley, 1995.
- [2] TIA/EIA/IS-95, "Mobile station-base station standard for dual-mode wideband spread spectrum cellular system," Telecommun. Industry Assoc., July 1993.
- [3] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, Jr., and C. E. Wheatley, III, "On the capacity of a cellular CDMA system," *IEEE Trans. Veh. Technol.*, vol. 40, pp. 303–312, May 1991.
- [4] S. Verdú, *Multiuser Detection*. Cambridge, UK: Cambridge Univ. Press, 1998.
- [5] R. L. Peterson, R. E. Ziemer, and D. E. Borth, *Introduction to Spread Spectrum Communications*. Englewood Cliffs, NJ: Prentice-Hall, 1995.
- [6] A. J. Viterbi, *Principles of Coherent Communication*. New York: McGraw-Hill, 1966.
- [7] W. C. Lindsey and C. M. Chie, *Phase-Locked Loops*. New York: IEEE Press, 1986.
- [8] M. T. Wasan, *Stochastic Approximation*. Cambridge, U.K.: Cambridge Univ. Press, 1969.
- [9] J. G. Proakis, *Digital Communications, 2nd Ed.* New York: McGraw-Hill, 1989.
- [10] W. Feller, *An Introduction to Probability Theory and Its Applications*, vols. I and II. New York: Wiley, 1957.

Shaping and Nonequiprobable Signaling for Intensity-Modulated Signals

Da-shan Shiu, *Student Member, IEEE*, and
Joseph M. Kahn, *Senior Member, IEEE*

Abstract—The theory of shaping and nonequiprobable signaling, which has been developed for conventional electrical signals, must be modified to treat intensity-modulated (IM) signals. We show that for IM signals, the optimum shape of the constellation bounding region in N -dimensional (N -D) space is an N -D simplex. As $N \rightarrow \infty$, the maximum achievable shape gain is 1.33 dB (in terms of transmitted power), and the resulting marginal signaling distribution on the one-dimensional (1-D) constituent constellation is exponential. We also investigate the tradeoffs between shaping and its negative consequences, and find that a 1-dB shape gain can be achieved while incurring reasonable increases in peak-to-average power ratio and constellation expansion ratio.

Index Terms—Intensity modulation, nonequiprobable signaling, optical communications, shaping.

I. INTRODUCTION

Many signaling schemes are based on finite-dimensional lattices [1]. Usually the signal constellation consists of all lattice points within a bounding region. If the constellation points are used equiprobably, the shape of this bounding region determines the average signal power. By selecting a certain shape other than a cube for the con-

Manuscript received July 29, 1998; revised April 30, 1999. This work was supported in part by the National Science Foundation under Grant ECS-9710065, LG Electronics, Hewlett-Packard, and the University of California MICRO Program. The material in this correspondence was presented in part at GLOBECOM'98, Sydney, Australia, October 1998.

The authors are with the Department of Electrical Engineering and Computer Sciences, University of California Berkeley, CA 94720 USA (e-mail: {dsshui;jmk}@eecs.Berkeley.edu).

Communicated by F. R. Kschischeng, Associate Editor for Coding Theory. Publisher Item Identifier S 0018-9448(99)08283-8.

stellation bounding region, it is possible that the average transmitted signal power can be reduced while maintaining the communication bit rate. It has been shown that for electrical signals spherical bounding regions minimize average signal power. In an N -dimensional (N -D) space, as $N \rightarrow \infty$, the shape gain of the N -sphere over the N -cube approaches the ultimate shape gain of 1.53 dB [2]. Employing a spherical bounding region the marginal signaling distribution induced on the two-dimensional (2-D) constituent constellation approaches Gaussian as $N \rightarrow \infty$.

Nonequiprobable signaling refers to selecting more frequently those signal points whose energies are lower [3]. One can regard constellation bounding region shaping and nonequiprobable signaling as two different but related ways to achieve the same goal. Note that the average signal energy is completely determined by the marginal signal distribution on the 2-D constituent constellation; both shaping and nonequiprobable signaling result in nonuniform distributions on the 2-D constituent constellation that are more average-power efficient. Practical considerations influence the use of constellation bounding region shaping or nonequiprobable signaling, or their combined use. For example, constellation bounding region shaping is only effective when N is reasonably large.

Adverse effects of shaping and nonequiprobable signaling include an increase in peak-to-average-power ratio (PAR) and an expansion in size of the constituent constellation, which is described by the constellation expansion ratio (CER). Addressing of the constellation points is another important issue. Techniques have been developed to achieve a good tradeoff between shape gain, the adverse effects of shaping, and addressing complexity. Examples of shaping techniques include the generalized cross constellation [2], trellis shaping [4], and shell mapping [5]; examples of nonequiprobable signaling techniques include equal-size constellation partitioning [3] and unequal-size constellation partitioning [6]. A particularly simple addressing scheme for nonequiprobable signaling is to use codewords of unequal lengths for constellation points of unequal probabilities [7]. However, this results in a communication bit rate that is not constant, which can be problematic in certain applications. If the bit-rate variation needs to be controlled, techniques such as balanced codes [8] exist that can be used to ensure a constant bit rate.

In this correspondence, we examine shaping and nonequiprobable signaling for another class of signals: the intensity-modulated (IM) signals. IM signals are completely different from the more typical *electrical* signals in several key respects. An example of an electrical signal is a voltage signal transmitted along a transmission line. The baseband equivalent of the transmitted signal, $x(t)$, is a complex function of time, and the instantaneous transmitted power is proportional to $|x(t)|^2$.

In intensity modulation (IM), the instantaneous power output of the transmitter is modulated in proportion to some function of the modulating signal. IM is widely used in optical communications. An application that illustrates the use of IM signals is wireless infrared communication [9]. A wireless infrared communication system consists of an optical transmitter, a linear time-invariant channel having impulse response $h(t)$, and a receiving photodetector with responsivity R . Such a system has an equivalent baseband model that hides its carrier frequency. Let $x(t)$ represent the instantaneous optical power of the transmitter. The transmitted signal $x(t)$ and the impulse response $h(t)$ must always be real and nonnegative. In this correspondence, we consider specifically pulse amplitude modulation (PAM)-like input signals, i.e., $x(t) = \sum_k x_k \zeta(t - kT)$, where $\zeta(t)$ is a nonnegative pulse, T is the signaling period, and the nonnegative sequence $\{x_k\}$ encodes the information. The received photocurrent is

$$y(t) = Rh(t) \otimes x(t) + n(t). \quad (1)$$

The term $n(t)$ represents noise, and may include contributions from thermal and/or shot noise. In many applications, the noise $n(t)$ can be described as white and Gaussian. For example, in fiber-optic communications, the thermal noise of the receiver, which is Gaussian and approximately white, usually dominates [10]. In a wireless optical communication link, the receiver collects background radiation, which induces shot noise that can be modeled as white, Gaussian, and signal-independent if its intensity is high [9]. Unlike the transmitted signal $x(t)$, both the noise $n(t)$ and the received signal $y(t)$ can be negative. This is consistent with the fact that they represent current, not optical power.

When the term "power" is used, one must be very specific about whether the transmitted optical or the received electrical power is referred to. The instantaneous transmitted optical power is $x(t)$. Assuming that the sequence $\{x_k\}$ is ergodic, the average transmitted optical power, i.e., the time average of $x(t)$, is

$$\langle x_k \rangle T^{-1} \int_{-\infty}^{\infty} \zeta(t) dt.$$

The instantaneous received electrical power, assuming noise is absent, is proportional to $[h(t) \otimes x(t)]^2$, and the average received electrical power is proportional to the time average of this quantity. Therefore, the average received signal power is not proportional to the average transmitted signal power. For example, comparing transmission of the signal $x(t)$ to that of $g \cdot x(t)$, $g > 0$, the latter requires a change of transmitted optical power of $10 \log_{10} g$ decibels over the former. In response to this change in transmitted power, the received electrical signal-to-noise ratio (SNR) changes by $20 \log_{10} g$ decibels. To avoid potential confusion, hereafter in this correspondence, decibel units are used exclusively to represent differences in the transmitted optical power.

In summary, while the concepts of shaping and nonequiprobable signaling also apply to IM signals, the results and techniques developed for electrical signals are not directly applicable to IM signals for the following two reasons:

- Transmitted IM signals are always nonnegative. Therefore, the coordinates of every constellation point must be nonnegative and, as a result, the constellation bounding region must not enclose any lattice points with negative coordinates. For example, N -spheres obviously violate this requirement.
- The transmitted energy of IM signals is proportional to the amplitude, not the square of the amplitude, of the transmitted signal. Therefore, to minimize the average transmitted power, the average L^1 norm of the constellation points should be minimized instead of the average L^2 norm.

The remainder of this correspondence is organized as follows. In Section II, we define the concepts and parameters employed in the correspondence. In Section III, we present the shape of the constellation bounding region that achieves the highest shape gain, and derive the ultimate shape gain of IM signal constellations. In Section IV, the tradeoffs between shaping gain and CER, PAR, and complexity are analyzed, and examples of shaping codes are provided. Concluding remarks can be found in Section V.

II. DEFINITIONS

In this section, we define the concepts and parameters used throughout this correspondence. Most of these parameters have been defined in the literature for electrical signals [2], [3]. They are appropriately modified here based on the characteristics of IM signals.

A. Constellation, Signaling Probability, and Normalized Bit Rate

Throughout this correspondence, we deal with constellations Ω embedded in an N -dimensional (N -D) vector space having well-defined Euclidean distance and norm. A constellation Ω often comprises a collection of points that belong to the intersection of a lattice Λ , or a translate $a + \Lambda$, and a finite region R . The transmitter is assumed to emit a sequence of symbols drawn independently from Ω according to a signaling probability mass function $p(\mathbf{r})$, $\mathbf{r} \in \Omega$. The symbols are emitted at some fixed symbol rate. The entropy rate of the sequence of output symbols from Ω with signaling probability mass function $p(\mathbf{r})$ is

$$H(p) = - \sum_{\mathbf{r} \in \Omega} p(\mathbf{r}) \log_2 p(\mathbf{r})$$

bits per N -D symbol.

Because there is no quadrature-phase component in an IM signal, an N -D IM symbol is obtained by concatenating N PAM-like (1-D) symbols. Accordingly, we define the *basic dimension* of IM signals to be 1-D. We define a normalization coefficient η as $\eta = 1/N$ for IM signals. The normalized bit rate per basic dimension β is the entropy rate per basic dimension and is defined as $\beta = \eta H(p)$. The highest normalized bit rate for a given constellation size is $\eta \log_2 |\Omega|$, which is achieved only with uniform signaling probability. By contrast, for electrical signals, an N -D symbol is obtained by concatenating $N/2$ quadrature amplitude modulated (QAM)-like (2-D) symbols; the basic dimension and η for electrical signals are 2-D and $2/N$, respectively.

Henceforth, in this correspondence, the notation (Ω, p, β) is used to denote a constellation Ω with signaling probability mass function $p(\mathbf{r})$ and normalized bit rate β . A broader meaning is given to the word “constellation;” not only does it represent a collection of signal points, but it also specifies the signaling probability mass function. Because in this correspondence it is only meaningful to compare constellations having the same normalized bit rate, β is added to the notation even though it can be determined from Ω and p . If the constellation points are selected equiprobably, the constellation is denoted by $(\Omega, \text{uniform}, \beta)$.

B. Average Constellation Energy and the Continuous Approximation

For IM signals, the energy of an N -D constellation point is $\mathbf{r} = (x_1, \dots, x_N) \in \Omega$ is

$$E(\mathbf{r}) = \sum_{i=1}^N x_i^2.$$

The average constellation energy per basic dimension is

$$E(\Omega) = \eta \sum_{\mathbf{r} \in \Omega} p(\mathbf{r}) E(\mathbf{r}).$$

By contrast,

$$E(\mathbf{r}) = \sum_{i=1}^N x_i^2$$

for electrical signals.

If an N -D constellation (Ω, p, β) is based on a lattice and $\eta|\Omega|$ is large, the average constellation energy can be estimated conveniently by approximating the discrete signaling probability mass function $p(\mathbf{r})$, $\mathbf{r} \in \Omega$ by a continuous probability density function $p_c(\mathbf{r})$, $\mathbf{r} \in R^N$. Roughly speaking, integrating $p_c(\mathbf{r})$ over a *fundamental volume*¹ about a lattice point \mathbf{r} yields $p(\mathbf{r})$. The function $p_c(\mathbf{r})$ is zero if \mathbf{r} is

outside the bounding region R . The average constellation energy of (Ω, p, β) per basic dimension is approximated by

$$E(\Omega) \approx \eta \int E(\mathbf{r}) p_c(\mathbf{r}) d\mathbf{r}. \quad (2)$$

The use of this technique to estimate the average energy of a constellation is called the *continuous approximation*. It is a useful simplification because without it one has to do case-by-case study for each normalized bit rate β and each underlying lattice of interest.

C. Constellation Figure of Merit

A common measure for the reliability of digital communication is the minimum Euclidean distance of the constellation. For a given normalized bit rate β , it is typically desired to maximize $d_{\min}(\Omega)$ for a given average constellation energy per basic dimension $E(\Omega)$. The constellation figure of merit (CFM) is a dimensionless, scale-invariant quantity relating the average constellation energy to the minimum Euclidean distance. For IM signals, CFM is defined as

$$\text{CFM}(\Omega) \equiv \frac{d_{\min}(\Omega)}{E(\Omega)}. \quad (3)$$

Suppose that the only system impairment is additive white Gaussian noise (AWGN) with one-sided power spectral density N_0 . The following approximation relates CFM to the symbol error probability:

$$P_e \approx \bar{N} Q \left(\sqrt{\frac{d_{\min}^2(\Omega)}{2N_0}} \right) = \bar{N} Q \left(\frac{E(\Omega) \text{CFM}(\Omega)}{\sqrt{2N_0}} \right) \quad (4)$$

where \bar{N} is the error coefficient [11]. (In Section I, we used wireless infrared communication as an example of an application employing IM signals. The noise component $n(t)$ in (1) is indeed modeled as white, Gaussian, and independent of signal $x(t)$ [9].) Therefore, the gain in average constellation energy, or average transmitted power, of constellation (Ω, p, β) over constellation (Ω', p', β) is

$$\gamma = 10 \log_{10} \frac{\text{CFM}(\Omega)}{\text{CFM}(\Omega')} \text{ dB}. \quad (5)$$

By contrast, for electrical signals CFM is defined by $\text{CFM}(\Omega) \equiv d_{\min}^2(\Omega)/E(\Omega)$. With this definition, (5) still applies [2].

D. Baseline Constellation

For comparison purposes, the baseline constellation in N -D space is defined as the N -D constellation constructed over the simple cubic lattice with cubic-shaped bounding region and uniform signaling probability mass function. It is denoted by $(\Omega_b, \text{uniform}, \beta)$. For example, in 1-D space, the baseline constellation with $d_{\min} = 1$ consists of points $\{0, 1, 2, \dots, 2^\beta - 1\}$. With the continuous approximation, the 1-D baseline constellation with a normalized entropy rate β is a uniform distribution in $[0, 2^\beta]$. The N -fold Cartesian product of it with itself is the baseline constellation in N -D space. We define the baseline CFM, $\text{CFM}(\Omega_b)$, as the CFM of the baseline constellation.

E. Coding Gain, Shape Gain, and Bias Gain

It has been shown that the CFM of constellation (Ω, p, β) can be related to the baseline CFM by

$$\text{CFM}(\Omega) \approx \text{CFM}(\Omega_b) \gamma_c(\Lambda) \gamma_s(R, p) \quad (6)$$

where $\gamma_c(\Lambda)$ is the coding gain of the lattice Λ and $\gamma_s(R, p)$ is the *shaping gain* [2]. Roughly speaking, coding gain describes the density of points packed in a unit volume for a given minimum Euclidean distance. It is a property of the underlying lattice structure and is

¹The fundamental volume is the N -D volume per lattice point.

not the subject of this correspondence. Shaping gain $\gamma_s(R, p)$ is the product of two components, the *shape gain* of the bounding region R and the *bias gain* of the signaling probability $p(\mathbf{r})$.

Shape gain is defined as the reduction of average constellation energy due to the shape of the constellation bounding region R , as compared to the baseline (cubic) shape. The highest shape gain is achieved if R has the property that any lattice point inside R has energy no greater than that of any lattice point outside R . For electrical signals, the optimal shaping region in N -D is an N -sphere; for IM signals, it is the region in the nonnegative orthant² bounded by a plane, as will be shown in Section III. The highest achievable shape gain in an N -D space is monotonically increasing in its dimensionality; using the optimal shaping region, the ultimate shape gain is achieved when $N \rightarrow \infty$. However, to achieve a high shape gain, it is necessary to design the constellation in a high-dimensional space, which may lead to high complexity. Furthermore, one might not have much freedom in choosing N because the dimensionality of the underlying lattice Λ usually is a more important concern.

Bias gain is the reduction of average constellation energy due to more frequent use of lower energy constellation points, as compared to equiprobable signaling. Specifically, it is the reduction of average constellation energy of (Ω, p, β) relative to $(\Omega', \text{uniform}, \beta)$ when the bounding regions of Ω and Ω' , $R(\Omega)$, and $R(\Omega')$, have the same shape, i.e., $R(\Omega)$ is obtained by scaling $R(\Omega')$. What makes bias gain important is that ultimate shape gain can be obtained in a space of any finite dimensionality in the form of combined bias gain and shape gain [3].

Hereafter in this correspondence, when there is no need to distinguish between constellation bounding region shaping and nonequiprobable signaling, we will use the term *shaping* to refer to both of them. The codes employed to achieve shaping gain are called *shaping codes*.

F. Constellation Expansion Ratio and Peak-to-Average Power Ratio

A constituent constellation is the projection of an N -D constellation Ω onto a given M -D constellation, where M divides N . We often use $M = 1$ for IM signals and $M = 2$ for electrical signals.

One drawback of shaping is that it requires the size of the constituent constellation to be expanded. Intuitively, expansion of the constituent constellation occurs because the shaped constellation bounding region encloses some lattice points that are far away from the origin in only a small number of coordinates. On the other hand, nonequiprobable signaling requires more constellation points than equiprobable signaling to convey a given information rate. As a result, it requires the transmitter and receiver to support a wider dynamic range. The constellation expansion ratio (CER) of a constellation (Ω, p, β) is defined to be the ratio of the size of the constituent constellation of (Ω, p, β) to the size of the constituent constellation of the baseline constellation $(\Omega_b, \text{uniform}, \beta)$.

A measure of the sensitivity of a signal constellation to nonlinearities and other signal-dependent perturbations is the peak-to-average-power ratio (PAR). PAR is the ratio of the value of the largest coordinate among all constellation points to the average constellation energy per dimension. Shaping also results in an increase in PAR, partly because the average constellation energy is reduced and partly because the peak power may be increased. It is important to keep CER and PAR low when designing shaping codes.

III. OPTIMAL SHAPING FOR INTENSITY-MODULATED SIGNALS

To achieve the highest shape gain, the shape of the bounding region R should be chosen such that every lattice point inside R has energy

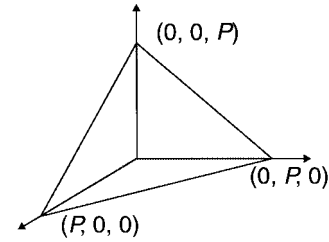


Fig. 1. The optimal constellation bounding region for 3-D IM signal constellations.

no greater than that of any lattice point outside R . For IM signals, the coordinate of a constellation point represents the energy of that point along the corresponding dimension, and the overall energy of a constellation point is the sum of its coordinates. Therefore, the optimal constellation bounding region must minimize the quantity $\max_{\mathbf{r} \in \Omega} E(\mathbf{r})$, and it cannot enclose any constellation point with negative coordinates. The optimal N -D constellation region is thus bounded inside the $N + 1$ planes defined by $x_i = 0$, $i = 1, 2, \dots, N$ and $\sum_{i=1}^N x_i = L$, where L is the highest energy of any constellation point. This region is denoted by $R_N(L)$, i.e.,

$$R_N(L) = \left\{ \mathbf{r} = (x_1, \dots, x_N) \mid x_i \geq 0, \right. \\ \left. \text{for } i = 1, 2, \dots, N \text{ and } \sum_{i=1}^N x_i \leq L \right\}. \quad (7)$$

The $N + 1$ vertices that define $R_N(L)$ are $(0, 0, \dots, 0)$, $(L, 0, \dots, 0)$, $(0, L, 0, \dots, 0)$, \dots , $(0, \dots, 0, L)$. Excluding the origin, the N vertices form a simplex in the N -D space. For example, the shape of the optimal bounding region is an isosceles triangle in 2-D and a tetrahedron in 3-D, as shown in Fig. 1.

To calculate the shape gain of $R_N(L)$ over the baseline cubic shape, we compare the average energy of $R_N(L)$ to the average energy of an N -D cube with the same volume. Note that under the continuous approximation, if the signaling distribution is uniform, two constellations have the same normalized bit rate if they have the same volume per dimension. The volume bounded by $R_N(L)$ is $V_N(L) = L^N/N!$, and the average constellation energy per basic dimension is $P = L/(N + 1)$. An N -cube whose volume is $V_N(L)$ has average energy per 1-D $2^{-1}(L^N/N!)^{1/N}$. The highest achievable shape gain in N -space is thus

$$\frac{1}{2} \left(\frac{L^N}{N!} \right)^{1/N} \div \frac{L}{N+1} = \frac{1}{2} \left(\frac{(n+1)^N}{N!} \right)^{1/N}. \quad (8)$$

As $N \rightarrow \infty$, the shape gain increases monotonically to the limit

$$\lim_{N \rightarrow \infty} \frac{1}{2} \left(\frac{(N+1)^{1/N}}{N!} \right)^{1/N} = \frac{e}{2} = 1.33 \text{ dB}.$$

This is called the ultimate shape gain for IM signals.

A uniform signaling distribution inside the optimal N -D shaping region induces a nonuniform signaling distribution in the 1-D constituent constellation space. This distribution is of interest because it indicates the marginal signaling distribution that nonequiprobable signaling should seek to achieve. We denote the marginal signaling probability density function in the 1-D constituent constellation induced by a uniform signaling distribution inside $R_N(L)$ by $f(x)/f(0)$, $x \in [0, L]$. The probability that the first coordinate of a randomly selected constellation point lies in $[x, x + dx]$ is the ratio of the volume of the intersection of $R_N(L)$ and the planar slice

$$\{\mathbf{r} = (x_1, \dots, x_N) \mid x \leq x_1 \leq x + dx\}$$

²The set of all nonnegative real n -tuples is called the nonnegative orthant.

which is $V_{N-1}(L-x) dx$, to the total volume of $R_N(L)$. Thus

$$f(x)/f(0) = V_{N-1}(L-x)/V_{N-1}(L).$$

As $N \rightarrow \infty$,

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{f(x)}{f(0)} &= \lim_{N \rightarrow \infty} \left(\frac{(L-x)}{L} \right)^{N-1} \\ &= \lim_{N \rightarrow \infty} \left(1 - \frac{x}{(N+1)P} \right)^{N-1} = e^{-x/P}. \end{aligned} \quad (9)$$

Thus the induced signaling distribution in 1-D constituent constellation space is exponential.

There are lattice points that have low average energies but have high peak energies, e.g., $(L, 0, \dots, 0)$. These points are included in optimally shaped constellations but not the baseline constellations, resulting in the increases in CER and PAR. One may consider intentionally choosing the shape of the constellation bounding region to be nonoptimal—to smooth out the corners of R_N —in order to achieve desired tradeoffs between shape gain and CER/PAR penalties. This is done in some constellation designs for electrical signals, such as the generalized cross constellation [2].

The ultimate shape gain and the marginal distribution in the constituent constellation space induced by optimal shaping can also be derived from the perspective of nonequiprobable signaling. Note that the shape gain is completely determined by the induced marginal distribution. Given a fixed volume that a constellation bounding region must enclose, the optimum constellation bounding region must induce a marginal distribution that has the lowest average energy. The dual problem to this is that, given that the first moment of the marginal distribution $f(x)$ is constrained by $\int_0^\infty x dx \leq P$, the optimal marginal distribution should have the largest possible differential entropy. Using standard optimization techniques, such a distribution is easily shown to be exponential. An exponential distribution with mean (average constellation energy) P has differential entropy (normalized bit rate) $\beta = \log_2 e + \log_2 P$. On the other hand, a uniform (baseline) probability density function having the same differential entropy is over $[0, e^{1+\ln P} = eP]$. The highest achievable gain using only nonequiprobable signaling is thus the ratio of $eP/2$ to P , which is $e/2$. This is an upper bound on the ultimate shape gain, and is indeed achieved by infinite-dimensional shaping. An analogous argument to this has been used by Forney and Wei [2] to prove that optimal shaping induces a Gaussian distribution on the 2-D constituent constellation in the infinite-dimension limit.

IV. DESIGN OF SHAPING CODES

A practical shaping code should yield a reasonable shaping gain while incurring acceptable penalties. In this section, we develop upper bounds on shaping gain versus CER and PAR. We use a nonequiprobable signaling technique adapted from [3] to demonstrate the tradeoff between shaping gain and complexity. An example of a shaping code is then provided.

Before proceeding, we note that the “shell mapping” proposed in [5] is an alternative to the approach we employ here. The shell mapping technique can often be used to map a given problem with IM signals to an equivalent problem with electrical signals, thus making it possible to transfer known results for electrical signals to the IM problem. Here, we prefer to solve the IM problem directly, since we feel it is more clear and intuitively appealing.

A. Bounds on CER and PAR Versus Shaping Gain

We focus on the marginal signaling probability density function in 1-D to study bounds on CER and PAR versus shaping gain. The nonuniform marginal signaling probability can result from both

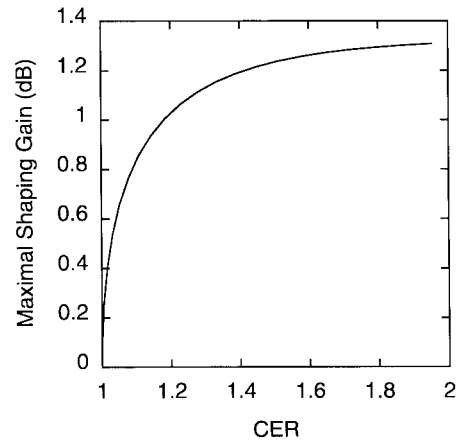


Fig. 2. Best possible tradeoff between shaping gain and CER.

high-dimensional shaping of the constellation bounding region and intentional nonequiprobable signaling.

To derive the tradeoff between CER and shaping gain, consider a peak-power-limited 1-D constellation. Let $p_c(x)$, $x \in [0, L]$, denote the signaling probability density function. We assume $p_c(0)$ and $p_c(L)$ are nonzero. Let H and E denote the differential entropy and the expectation of $p_c(x)$, respectively. The quantities H and E represent the normalized bit rate of the constellation and the average constellation energy, respectively. The baseline signaling probability density function whose differential entropy equals H is uniform in $[0, H]$ and has a mean $H/2$. Thus the shaping gain of $p_c(x)$ is $\gamma_s = (H/2)/E$ and the CER is L/H . By standard variational arguments using a Lagrange multiplier λ , the probability distribution $p_c^*(x)$ that maximizes the shaping gain $(H/2)/E$ for a fixed CER L/H is a truncated exponential distribution

$$p_c^*(x) = \begin{cases} \left(\frac{\lambda}{1 - e^{-\lambda L}} \right) e^{-\lambda x}, & x \in [0, L] \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

where λ controls the tradeoff between shaping gain and CER. The PAR is determined by CER and shaping gain

$$\text{PAR} = 2 \cdot \text{CER} \cdot \text{shaping gain}. \quad (11)$$

For a fixed shaping gain, minimizing the CER is equivalent to minimizing the PAR. Fig. 2 illustrates the upper bound on shaping gain as a function of the CER, which is achievable only by the truncated exponential signaling distribution. Insofar as the continuous approximation is accurate, we can conclude that, in principle, it is possible to achieve shaping gains of more than 1 dB while keeping CER less than 1.2 (PAR below 3), or to achieve 90% of the ultimate shape gain while keeping CER at about 1.4 (PAR about 3.7).

B. Bounds on Bias Gain Versus the Number of Values of Signaling Probability

One nonequiprobable signaling scheme is performed by partitioning a constellation into T equal-size, nonoverlapping, contiguous subconstellations $\Omega_0, \Omega_1, \dots, \Omega_{T-1}$. A shaping code selects the subconstellation Ω_i with frequency f_i , and signal points belonging to the same subconstellation are to be used equiprobably [3]. In 1-D space, without loss of generality, Ω_i can be assumed to be $[i, i+1]$. The highest achievable bias gain, given that only T different values of signaling probability are available, can be derived as follows. The

TABLE I
THE HIGHEST ACHIEVABLE BIAS GAIN AND THE CORRESPONDING BIAS FACTOR x , CER, AND PAR
WHEN T EQUAL-SIZED SUBCONSTELLATIONS ARE USED

T	Highest Achievable Bias Gain (dB)	Bias Factor x	CER	PAR
2	0.71	0.23	1.23	2.90
4	1.07	0.37	1.53	3.93
8	1.23	0.53	1.90	5.04
16	1.30	0.68	2.30	6.21

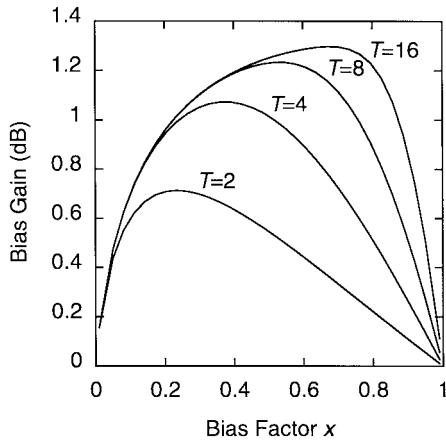


Fig. 3. Bias gain as a function of bias factor x and the number of subconstellations T .

bias gain as a function of f_i is

$$\frac{2^{H(f_0, f_1, \dots, f_{T-1})}^{-1}}{\sum_{i=0}^{T-1} \left(i + \frac{1}{2}\right) f_i}$$

where

$$H(f_0, f_1, \dots, f_{T-1}) = \sum_{i=0}^{T-1} f_i = 1.$$

To maximize the bias gain, we need to maximize

$$J = H(f_0, f_1, \dots, f_{T-1}) - \log_2 \left[\sum_{i=0}^{T-1} \left(i + \frac{1}{2}\right) f_i \right]$$

subject to the constraint $\sum_{i=0}^{T-1} f_i = 1$. The solution for the signaling probability f_i can be shown to be the truncated geometric distribution having a critical parameter x . We refer to x as the bias factor

$$f_i^* = \left(\frac{1-x}{1-x^T} \right) x^i, \quad i = 0, 1, \dots, T-1. \quad (12)$$

This can be interpreted as a staircase approximation to the truncated exponential distribution in (10), which is the optimal continuous marginal distribution when the peak energy is limited. If the optimal signaling probability in (12) is used, then

$$\text{CER} = T 2^{-H(f_0^*, f_1^*, \dots, f_{T-1}^*)}$$

and

$$\text{PAR} = T \left[\sum_{i=0}^{T-1} \left(i + \frac{1}{2}\right) f_i^* \right]^{-1}.$$

The bias gain, CER, and PAR as functions of the bias factor x are shown in Figs. 3–5, respectively. The highest achievable bias

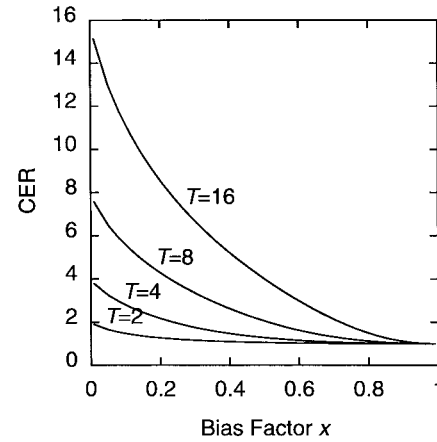


Fig. 4. Constellation as a function of bias factor x and the number of subconstellations T .

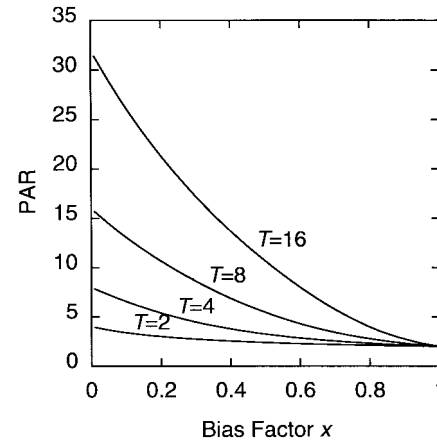


Fig. 5. Peak-to-average power ratio (PAR) as a function of bias factor x and the number of subconstellations T .

gains for $T = 2, 4, 8, 16$, and the corresponding values of x , CER and PAR are shown in Table I. It is desirable to choose x larger than the value that yields the highest bias gain to realize a favorable tradeoff between bias gain and CER/PAR penalties. We find that a four-valued probability distribution suffices to achieve a bias gain of over 1 dB for CER = 1.53 and PAR = 3.93, and a 16-valued probability distribution suffices to achieve a shape gain that is only 0.03 dB lower than the ultimate shape gain of $e/2$.

C. Examples of Shaping Codes

Suppose that the design objective is to achieve at least 1-dB shaping gain at a normalized bit rate of approximately 2 bits per basic

TABLE II
THE HUFFMAN PROCEDURE USED TO MAP
INPUT-BIT SEQUENCES TO SUBCONSTELLATIONS

Source Bit Sequence	Subconstellation Chosen	A Priori Probability
0	Ω_0	1/2
10	Ω_1	1/4
110	Ω_2	1/8
1110	Ω_3	1/16
1111	Ω_4	1/16

dimension via nonequiprobable signaling in 1-D signal space. The number of subconstellations T can be determined from the product of CER and the number of subconstellations needed in the baseline constellation. From Fig. 4, CER is at least 1.2 at 1-dB shaping gain. Using the baseline constellation, $2^\beta = 4$ subconstellations are needed to support a normalized bit rate $\beta = 2$. Therefore, we choose $T = 5 \geq 4 \times 1.2$. Let subconstellation $\Omega_i, i = 0, 1, \dots, 4$ be selected with probability f_i . The truncated geometric distribution that achieves $H(f) = 2$ is $f_i^* = 0.6^{i-1} f_0^*, f_0 = 0.4337$. It may be desirable to round up these probabilities to dyadic numbers, i.e.,

$$(f_0, f_1, f_2, f_3, f_4) = (1/2, 1/4, 1/8, 1/16, 1/16)$$

so that a straightforward Huffman code such as the one shown in Table II can be used for the addressing task. The normalized bit rate and shaping gain of this shaping code are 1.875 bits per 1-D and 1.154 dB, respectively. The values of CER and PAR (1.25 and 5.41 dB, respectively) are very close to the values predicted in Section IV-A. With this addressing scheme, the data rate is probabilistic; an additional rate-control technique such as a balanced code can be used to ensure a constant bit rate.

To seek higher shaping gain than that achieved by this 1-D non-equiprobable signaling shaping code, one can employ a higher dimensional signal space. In Fig. 6, a 2-D shaping code is designed based on the previous 1-D shaping code. First one forms a 2-D constellation as the two-fold Cartesian product of the 1-D constellation with itself. The Cartesian product does not change the marginal signaling distribution on the 1-D constituent constellation; therefore the shaping gain, CER, and PAR are not changed. The bounding region of this 2-D constellation is a square. To utilize the shape gain on the 2-D space, the subconstellations located near the upper right corner of the square are replaced by the subconstellations that are closer to either one of the two axes, thus reducing the average constellation energy (while increasing the peak energy). Note that the shape of the constellation bounding region of the shaping code is similar to the shape of the 2-D optimal shaping region, which is an isosceles right triangle. The shaping gain, CER, and PAR are 0.05 dB, 1.2 times, and 0.84 dB higher than their counterparts for the previous 1-D shaping code. This illustrates that it is possible to achieve higher shaping gain by designing the code in a higher dimensional space, but the increased shaping gain will generally be accompanied by increased PAR and CER.

V. CONCLUSIONS

For communication systems employing intensity-modulated signals, the average transmitted power can be reduced by shaping the bounding region of the constellation and/or by employing nonequiprobable signaling. Methods developed for conventional electrical signals cannot be applied directly to IM signals because

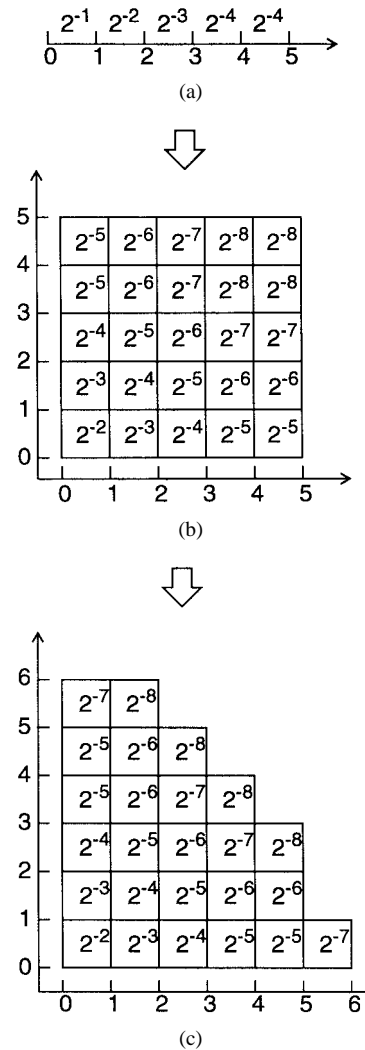


Fig. 6. Construction of the shaped 2-D constellation.

transmitted IM signals are nonnegative and the transmitted power is proportional to the instantaneous amplitude of the transmitted IM signal. In light of the differences between IM signals and electrical signals, we found it necessary to redefine various parameters for IM signals, such as the constellation figure of merit.

We found that to achieve the highest possible shape gain in N -D space, the shaped constellation should lie within the region in the nonnegative orthant bounded by the N -simplex, where L is the largest coordinate of any constellation point. As $N \rightarrow \infty$, the shape gain over an N -cube approaches the ultimate shape gain of $e/2 = 1.33$ dB. Equiprobable signaling in the optimally shaped N -D constellation induces an exponential signaling probability density function on the 1-D constituent constellation as $N \rightarrow \infty$. The ultimate shape gain can be achieved in 1-D in the form of bias gain using nonequiprobable signaling if the resulting marginal signaling distribution can be made exponential.

The major drawback of shaping is that increased shaping gain is accompanied by increases in CER and PAR. Furthermore, the dimensionality of the signal space and/or the number of different values of signaling probability must be large to support a high shaping gain. We investigated the tradeoffs between shape gain and these adverse effects. We found that a 1-dB shaping gain can be achieved with reasonable penalties. We have found that shaping codes that achieve gains closer to the ultimate limit incur sharply higher CER, PAR, and complexity penalties.

REFERENCES

- [1] G. D. Forney, "Coset codes—Part I: Introduction and geometrical classification," *IEEE Trans. Inform. Theory*, vol. 34, pp. 1123–1151, Sept. 1998.
- [2] G. D. Forney and L. Wei, "Multidimensional constellations—Part I: Introduction, figures of merit, and generalized cross constellations," *IEEE J. Select. Areas Commun.*, vol. 7, pp. 941–958, Aug. 1989.
- [3] A. R. Calderbank and L. H. Ozarow, "Nonequiprobable signaling on the Gaussian channel," *IEEE Trans. Inform. Theory*, vol. 36, pp. 726–740, July 1990.
- [4] G. D. Forney, "Trellis shaping," *IEEE Trans. Inform. Theory*, vol. 38, pp. 281–300, Mar. 1992.
- [5] A. K. Khandani and P. Kabal, "Shaping multidimensional signal spaces—Part I: Optimum shaping, shell mapping," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1799–1808, Nov. 1993.
- [6] J. N. Livingston, "Shaping using variable-size regions," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1347–1353, July 1992.
- [7] F. R. Kschischang and S. Pasupathy, "Optimal nonuniform signaling for Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 39, pp. 913–929, May 1993.
- [8] A. R. Calderbank and M. Klimesh, "Balanced codes and nonequiprobable signaling," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1119–1122, May 1992.
- [9] J. M. Kahn and J. R. Barry, "Wireless infrared communications," *Proc. IEEE*, vol. 85, pp. 265–298, Feb. 1997.
- [10] G. P. Agrawal, *Fiber-Optic Communication Systems*. New York: Wiley, 1997.
- [11] J. G. Proakis, *Digital Communications*, 3rd ed. New York: McGraw-Hill, 1995.

On the Inverse Windowed Fourier Transform

Laura Rebollo-Neira and Juan Fernandez-Rubio

Abstract—The inversion problem concerning the windowed Fourier transform is considered. It is shown that, out of the infinite solutions that the problem admits, the windowed Fourier transform is the "optimal" solution according to a maximum-entropy selection criterion.

Index Terms—Gabor transform, inversion problems, maximum entropy, windowed Fourier transform.

I. INTRODUCTION

The use of a generalized Fourier integral to convey simultaneous time and frequency information was first introduced by Gabor (1946). In [2], he defines a windowed Fourier integral, using a Gaussian window. Later, the window was generalized to any function in $L^2(R)$, the space of square integrable functions. The so generalized Gabor

Manuscript received March 1, 1997; revised February 17, 1999. This work was supported by CIRIT of Catalunya, CICYT of Spain (TIC96-0500-C10-01, TIC98-0412), and CICIPBA of Argentina.

L. Rebollo-Neira is with CICIPBA (Comisión de Investigaciones Científicas de la Provincia de Buenos Aires), Departamento de Física, Universidad Nacional de La Plata C.C. 727, 1900 La Plata, Argentina.

J. Fernandez-Rubio is with the Department de Teoria del Senyal i Comunicacions, Escola Tecnica Superior, d'Enginyers de Telecomunicació, Campus Nord, UPC, Edifici D-4, c/. Gran Capita s/n. 08034 Barcelona, Spain (e-mail: juan@gps.tsc.upc.es).

Communicated by C. Herley, Associate Editor for Estimation.
Publisher Item Identifier S 0018-9448(99)08127-4.

transform is mostly referred to as the windowed Fourier transform (WFT).

Restricting the space of signals to $L^2(R)$, the WFT is a mapping from $L^2(R)$ to $L^2(R^2)$ which is not bijective. As a consequence, lack of uniqueness of the inverse problem must be expected. In this contribution, we focus on a statistical analysis of the inversion problem. First the problem is shown to admit an infinite number of solutions. We then work on the space of possible solutions adopting a statistical description as the essential tool. The possible solutions are considered as a stochastic process distributed according to a (to be determined) probability density. The desired solution is estimated as the mean value of the random process. Among all the probability densities capable of yielding admissible mean-value solutions we single out one, adopting the maximum-entropy principle (MEP). Finally, we show that, from the maximum-entropy (ME) probability density a mean-value solution is inferred which is identical to the WFT. Thereby the WFT is shown to be an "optimal" solution according to an ME selection criterion. This result also holds as a property within the *Frame Theory* [9].

II. THE WFT INVERSE PROBLEM

Definition: Let $f(x) \in L^2(R)$ be a given signal and $g(x) \in L^2(R)$ be any fixed function in $L^2(R)$. The WFT of $f(x)$ is a function $F(\omega, t) \in L^2(R^2)$ defined by

$$F(\omega, t) = \langle e^{i\omega x} g(x-t) | f(x) \rangle = \int_R e^{-i\omega x} g^*(x-t) f(x) dx \quad (1)$$

where $g^*(x)$ denotes the complex conjugate of $g(x)$.

The signal can be reconstructed from its WFT through the inversion formula [1], [3], [6]

$$f(x) = \frac{1}{C_g} \int_{R^2} e^{i\omega x} g(x-t) F(\omega, t) d\omega dt \quad (2)$$

where

$$C_g = \|g\|^2 = \int_R |g(x)|^2 dx.$$

Although the inversion formula (2) allows the recovery of a signal from its WFT, the inversion is not unique. Let us denote \mathcal{W} to the image of the WFT, i.e.,

$$\mathcal{W} = \left\{ F(\omega, t); F(\omega, t) = \int_R e^{-i\omega x} g^*(x-t) f(x) dx; \right. \\ \left. \text{for some } f(x) \in L^2(R) \right\}. \quad (3)$$

\mathcal{W} is only a closed subspace, not all of $L^2(R^2)$ (not every function $h(\omega, t) \in L^2(R^2)$ belongs to \mathcal{W}). The next theorem, whose proof is given in [6, p. 56], provides the necessary and sufficient condition for $h(\omega, t) \in \mathcal{W}$.

Theorem 1: A function $h(\omega, t)$ belongs to \mathcal{W} if and only if it is square integrable and, in addition, satisfies

$$h(\omega', t') = \frac{1}{C_g} \int_{R^2} K(\omega', t', \omega, t) h(\omega, t) d\omega dt \quad (4)$$

where

$$K(\omega', t', \omega, t) = \langle e^{i\omega' x} g(x-t') | e^{i\omega x} g(x-t) \rangle \\ = \int_R e^{-i\omega' x} g^*(x-t') e^{i\omega x} g(x-t) dx. \quad (5)$$