On the Combination of Multilevel Coding and Signal Shaping

Robert Fischer, Johannes Huber, Udo Wachsmann

Lehrstuhl für Nachrichtentechnik II, Universität Erlangen–Nürnberg
Cauerstraße 7/NT, 91058 Erlangen, Germany
Phone: +49-9131-857114, Fax: +49-9131-858849, Email: fischer@nt.e-technik.uni-erlangen.de

Abstract — **In this paper, the combination of signal shaping and coding in multilevel coding schemes is discussed. It is shown that for optimized discrete Gaussian distributed signal points the gap to the Shannon limit is negligible. Since for finite constellations coding and shaping are not separable, the optimum sharing of the total redundancy between coding and shaping is derived. It turns out that for combined coding and shaping 1 bit redundancy per dimension is necessary to achieve optimum performance. Furthermore, the optimal rate assignment in a multilevel coding scheme when shaping is active is presented. In practice, noticeable shaping gains are only achievable if the presented rules are taken into account. Simulation results show that, for an optimized shaping redundancy and optimum rate design, multilevel coding with large block length achieves BER = 10^{-5} exactly at the capacity limit for equiprobable signal points.**

1 Introduction

In [WFH97a, WH95, HWF98] it is shown that power and bandwidth efficient digital communication is possible via multilevel coding (MLC) together with multistage decoding (MSD) if and only if the individual rates of the component codes are chosen properly. Here, for further improvement, the combination of multilevel coding and signal shaping is discussed. It is well known that besides channel coding signal shaping provides additional gain by replacing a uniformly distributed signal by a (discrete) Gaussian distributed one and hence reducing average transmit power. In many situations it is easier to realize shaping gain than to apply comparably more powerful coding. In order to approach the Shannon limit shaping is indispensable.

The work was supported in part by Deutsche Forschungsgemeinschaft under contract Hu 634/1-1

First, in Section 2, the achievable shaping gain is derived, revealing that a power gain does not translate directly into a gain in capacity. With the restriction to discrete Gaussian distributions, the optimization of the a–priori probabilities of the signal points is addressed in Section 3. This directly leads to fundamental statements concerning the optimum sharing of redundancy between coding and shaping as well as the necessary total amount of redundancy to achieve optimality when shaping is active. A brief introduction to the rate design for MLC schemes is given in Section 4. This key point is shown exemplarily for signal sets with optimized Gaussian distributions operating on the additive white Gaussian noise (AWGN) channel. The paper closes in Section 5 with implementation issues and simulation results. Because of the results of [FHW96, WFH97b], where we have shown that MLC based on one–dimensional constellations are most power efficient and because shaping algorithms automatically impose proper spherical boundaries in many dimensions we restrict ourselves to uniformly spaced M–ary one–dimensional constellations.

2 Fundamental Limits

In literature, the gain due to *shaping*, i.e., the gain of replacing a uniformly distributed constellation by a signal with optimized distribution, is usually derived only for very large constellations, which leads to the following two statements [FW89]:

A: The maximum shaping gain, i.e. the maximum reduction in average transmit power is given by:

$$G_{s,\max} = \frac{\pi \mathrm{e}}{6} \quad \hat{=} \quad 1.53 \ \text{dB}.$$

But for situations most relevant in practice, using "small" signal sets, the limit of 1.53 dB can never be achieved. In [KP93] the shaping gain of finite constellations is calculated to be approximately

$$G_s \approx \frac{\pi \mathrm{e}}{6}(1-2^{-2R}), \qquad (1)$$

where R is the transmission rate per dimension.

B: Coding gain and shaping gain are separable.

This statement is only true asymptotically for very large constellations. Contrary to many authors (e.g. [FW89, For96]) we are interested in the analysis of finite constellations where coding and shaping interact and the respective gains cannot simply be added (in dB). The reason is that, on the one hand, signal power is decreased, leading to a gain. But, on the other hand, a loss in channel output entropy $H(Y)$ and, hence, in mutual information[1] $I(A;Y) = H(Y) - H(N)$ is recognizable, because shaping fixes the entropy $H(A)$ of the transmit symbols A instead of the entropy of the channel output symbols Y. Thus we have to distinguish between the pure *shaping* (power) *gain* G_s (fixing $H(A)$) and the gain G_c for fixed mutual information $I(A;Y)$, which we will denote *capacity gain*.

In Fig. 1 G_c is plotted versus the desired capacity C. G_c was calculated as the difference between the required signal–to–noise ratio (SNR) for transmission of the rate C over the AWGN channel using a continuous uniformly distributed channel input signal and the SNR at the Shannon limit. Additionally, the maximum shaping gain of discrete constellations (Eq. (1)) is shown[2]. As one can see, in a wide range the shaping gain is much greater than the gain in capacity. Only asymptotically, for high rates the whole shaping gain directly translates to a gain in capacity, approaching the ultimate shaping gain of $\frac{\pi \mathrm{e}}{6}$. Contrary, for $C \to 0$ the capacity gain completely vanishes.

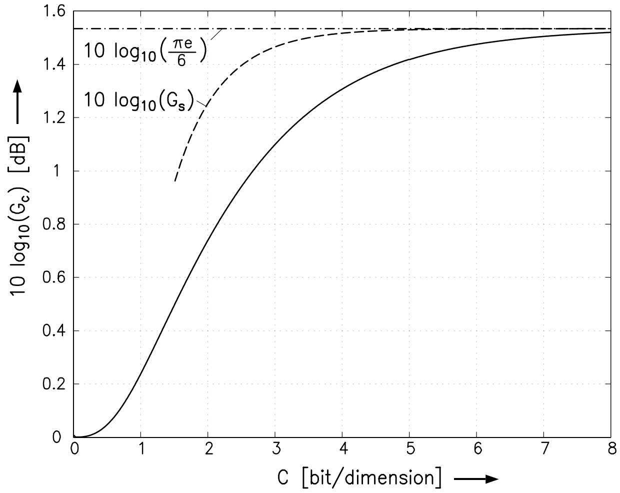

Figure 1: Capacity gain G_c (solid line) and shaping gain G_s for discrete constellation (Eq. (1)) (dashed line), resp., vs. capacity.

3 Optimization of Signal Constellation

In order to maximize the mutual information $I(A;Y)$, i.e., to approach channel capacity, an optimization over the a–priori probabilities of the signal points of finite constellations has to be performed. This procedure is quite difficult and no algorithm is available to achieve the optimum distribution in practice. Therefore, we force the channel input symbols a_m to be (discrete) Gaussian distributed, i.e.

$$\begin{aligned} &\Pr\{a_m\} = K(\lambda)\cdot \mathrm{e}^{-\lambda |a_m|^2}, \qquad \lambda \geq 0, \\ &a_m \in \{\pm 1, \pm 3, \ldots, \pm(M-1)\}, \quad M \text{ even,} \end{aligned} \qquad (2)$$

where $K(\lambda) = \left(\sum_{a_m} \mathrm{e}^{-\lambda|a_m|^2}\right)^{-1}$ normalizes the distribution. The parameter λ governs the trade–off between average power σ_a^2 of signal points and entropy $H(A)$. For $\lambda = 0$ a uniform distribution results, whereas for $\lambda \to \infty$ only the two minimum energy signal points are used. As we will see later, by selecting λ (and hence $H(A)$) properly, the performance of the optimum (not necessarily Gaussian) distribution can be approached. For a given M–ary constellation and target transmission rate $C < \mathrm{ld}(M)$, this variation of entropy moreover directly leads to the optimum partitioning of total redundancy $\mathrm{ld}(M) - C$ into coding redundancy $r_c = H(A) - C$ and shaping redundancy $r_s = \mathrm{ld}(M) - H(A)$. As an example the transmission of $C = 2$ bit/dimension using a $M = 8$–ary ASK signal constellation is considered. In Fig. 2 the SNR–gap

$$\begin{aligned} &10 \cdot \log_{10}(\Delta\mathrm{SNR}) = \\ &\quad 10 \cdot \log_{10}(E_b/N_0)\,|_{\text{discrete Gaussian}} - \\ &\quad 10 \cdot \log_{10}(E_b/N_0)\,|_{\text{Shannon limit}} \end{aligned} \qquad (3)$$

is plotted over the entropy $H(A)$. There are three important points: First, for $H(A) = 3$ a uniformly distributed 8ASK constellation results

[1] $H(N)$ denotes the differential entropy of the additive Gaussian noise: $N = Y - A$.

[2] The approximation is only tight for $C > 1.5$ bit/dim.

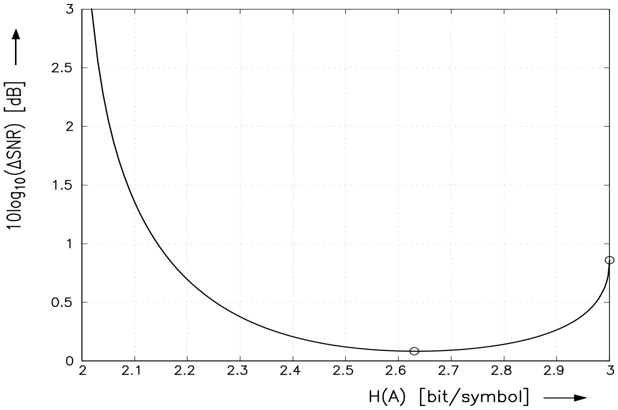

Figure 2: SNR–gap (capacity limit to Shannon limit, see Eq. (3)) for desired rate $C = 2$ bit/dimension and 8ASK constellation over entropy $H(A)$ of constellation.

where only coding is active. Second, as $H(A)$ approaches 2 only signal shaping is used. Because, in principle, for $C = H(A)$ error free transmission is only possible for a noise–less channel, the gap goes to infinity. Third, the minimum is obtained for $H(A) = 2.63$. Thus, in the optimum point redundancy of $\mathrm{ld}(M) - C = 1$ bit has to be divided into $r_c = 0.63$ bit coding redundancy and $r_s = 0.37$ bit shaping redundancy for this specific example. Here, the SNR–gap is only approx. 0.05 dB, revealing that discrete Gaussian distributed symbols are a very good choice. An optimization for an arbitrary discrete distribution was performed using a modified version of the Blahut–Arimoto algorithm [Bla72]. The resulting optimum is not exactly Gaussian, but the difference to the optimum shown in Fig. 2 is not visible at all (below 0.001 dB). It should be noted that for $2.5 \leq H(A) \leq 2.8$ bit/symbol ΔSNR differs only slightly. Thus, the selection of the optimum entropy is not very sensitive. In the optimum, without any extra constellation expansion, a capacity gain G_c of about 0.78 dB over solely channel coding results.

For each target transmission rate or equivalently for each SNR value the described optimization has to be performed individually, resulting in different distributions. For a 8ASK constellation with Ungerböck set partitioning the results are summarized in Fig. 3. On the top, the capacity C is plotted versus the signal–to–noise ratio E_s/N_0 in dB. The solid line is valid for uniform signaling, whereas the dashed line represents the optimized discrete Gaussian constellations. Additionally, the Shannon limit is shown (dotted line). As one can see, over a wide range the Shannon limit can be approached. The plot in the middle displays the optimal sharing of total redundancy $3 - C$ between coding and shaping. As a rule of a thumb, we can state that one bit total redundancy should be divided into 2/3 bit coding redundancy and 1/3 bit shaping redundancy.

Finally, in Fig. 4 the discussed optimization is performed for $M = 2$, 4, 8, 16 and 32–ary ASK constellations. It is well known that for 2ASK no shaping gain can be achieved. The gain increases as the size of the constellation increases (cf. Fig. 1). Ungerböck already stated [Ung82] that by introducing 0.5 bit redundancy per dimension, almost the entire coding gain can be realized if no shaping is active. Going to still larger constellations is not rewarding. Less redundancy causes an inevitable loss from the maximum coding gain for equiprobable signal points. Here, for combined coding and shaping 1 bit redundancy per dimension should be introduced. Then, the Shannon limit curve is approached. Reducing redundancy to 0.5 bit/dimension already causes an inevitable loss of approx. 0.5 dB to the Shannon limit curve, although even here, the SNR–gain for shaping and coding is already much greater than for coding solely. For $E_b/N_0 \to \infty$ shaping becomes inactive and the curves merge.

4 Rate Design for Multilevel Codes

For $(M = 2^\ell)$–ary digital modulation schemes, MLC together with MSD is a theoretically optimum approach to coded modulation, cf. [Hub94, HW94, WFH97a, HWF98]. Key point is the proper assignment of code rates to the component codes of a multilevel code. The code rates R^i have to be chosen equal to the capacities C^i of the equivalent channels defined for each binary coding level i, $i = 0, \ldots, \ell - 1$.

For a 8ASK constellation with Ungerböck set partitioning the capacities of the individual levels are comprised in Fig. 3 (bottom). Again, the solid lines correspond to uniform signaling. The dashed lines are valid for Gaussian distributions, optimized for each SNR value. It is important to notice that rate design completely changes when shaping is active. In particular, the rate of the highest level decreases strongly because this level has to carry the entire shaping redundancy. This fact directly leads to a simple construction of MLC schemes with signal shaping, which is shown in the next section.

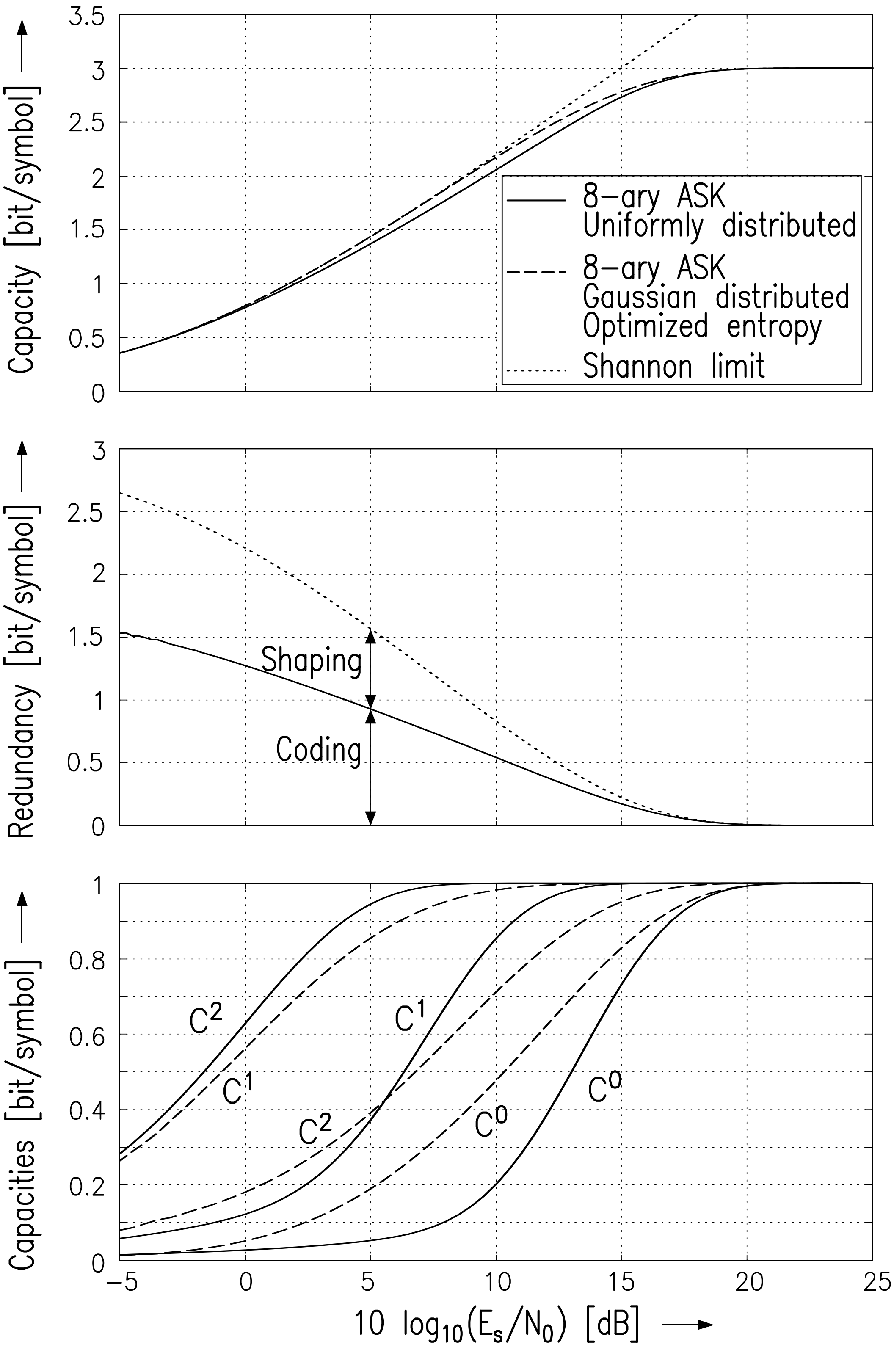

Figure 3: Top: Shannon limit and capacity C of 8ASK (AWGN channel, Ungerböck's set partitioning) with uniform signaling (solid) and with a optimized discrete Gaussian constellation (dashed). Middle: Sharing of coding and shaping redundancy. Bottom: Capacities C^i of the equivalent channels of the corresponding MLC scheme.

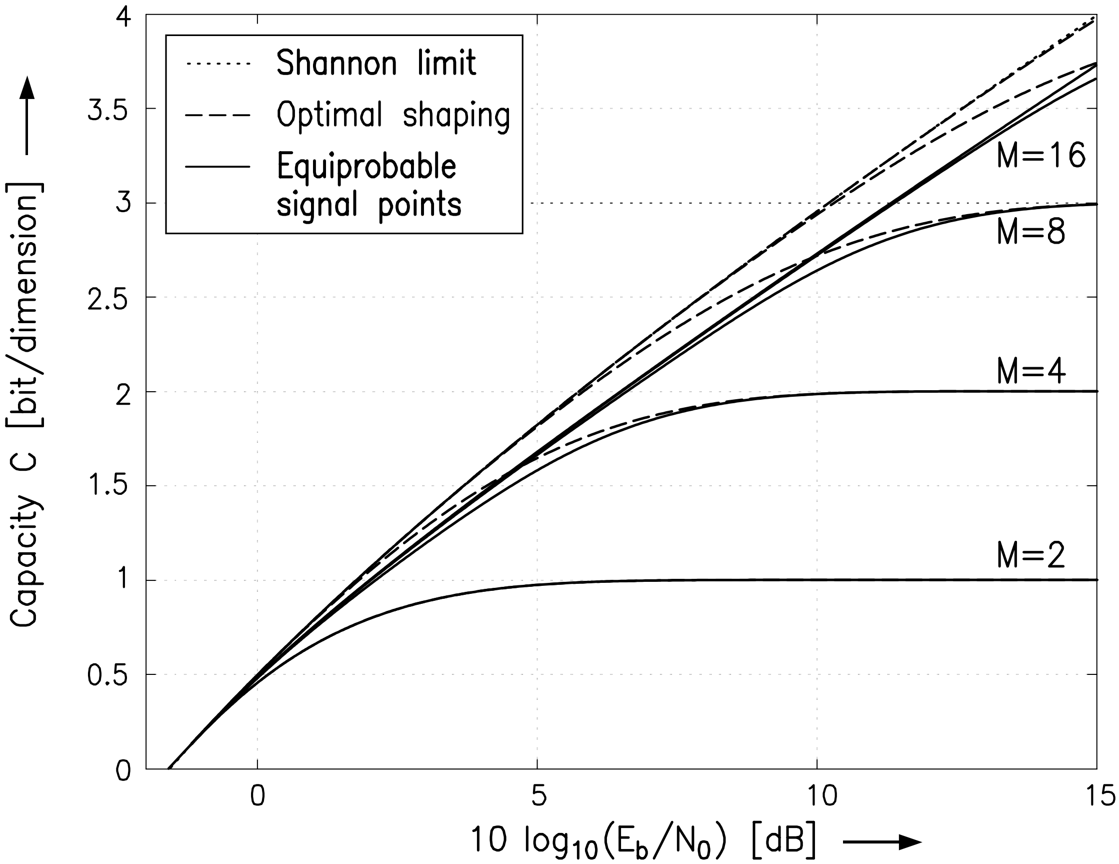

Figure 4: Capacity of ASK schemes with and without shaping.

5 Implementation and Simulation Results

In principle all shaping techniques can be combined with MLC schemes. Here, we prefer *trellis shaping* [For92], because in fact trellis shaping takes the lower levels into account but does not modify them. Thus, the MLC approach of using independent binary levels can be preserved. The idea how to combine MLC with shaping is to use all levels but the highest for coding as usual in a MLC scheme, e.g. by employing block codes of length N and appropriate rate. Only the highest level is involved in shaping, resulting in independent levels as desired. N_s consecutive symbols form one shaping step (N does not need to be a integer multiple of N_s). For trellis shaping a rate $1/N_s$ shaping convolutional code is used at the shaping level. Hence, $N_s - 1$ data bits are scrambled with one shaping bit, leading to a N_s–dimensional shaping scheme. A generalization to other shaping code rates is straightforward.

Simulations for multilevel coded 8ASK using terminated ($N = 4000$) 64–state convolutional codes (CC) as components and Ungerböck set partitioning were performed (see Fig. 5). Due to the capacity arguments in Fig. 2, a three–dimensional shaping scheme ($H(A) = 2.66$), which is very close to the optimum, is expected to perform better than a two–dimensional one ($H(A) = 2.5$). Additionally, the scheme without shaping is investigated for reference. The individual rates are chosen according to the capacities of the individual levels, considering the nonuniform distribution of signal points introduced by shaping (see Fig. 3). This results in quite different rate distributions $R^0/R^1/R^2$ for all three schemes (see caption of Fig. 5). As predicted from capacity arguments the scheme with $H(A) = 2.66$ performs best. Please notice, while using shaping the nonuniform distribution of signal points has to be taken into account in the decoding procedure, i.e. maximum–likelihood decoding has to be replaced by maximum–a–posteriori decoding. If the convolutional codes in the above example are replaced by turbo codes of large block length $N = 40000$ for $H(A) = 2.66$ BER $= 10^{-5}$ can be achieved at the capacity limit $10 \cdot \log_{10}(E_b/N_0) = 6.6$ dB valid for equiprobable signal points [WFH97b].

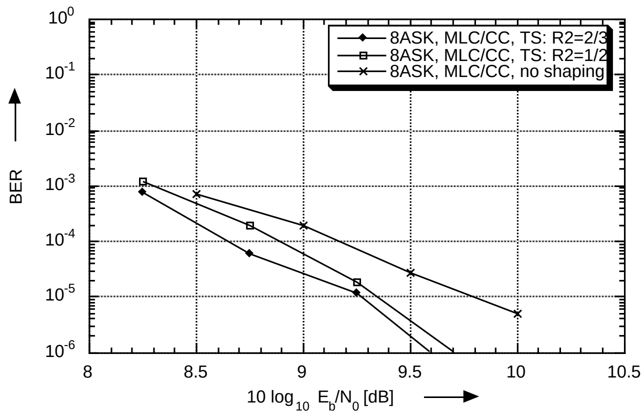

Figure 5: Simulation results: 8ASK with MLC employing (64–state) convolutional codes as components, total rate $R = 2$ bit/symbol, AWGN channel. Different shaping strategies: 1. Three-dimensional trellis shaping (TS) at highest level ($R^0/R^1/R^2 = 0.37/0.96/0.67$). 2. Two-dimensional TS at highest level ($R^0/R^1/R^2 = 0.52/0.98/0.5$). 3. No shaping is applied ($R^0/R^1/R^2 = 0.18/0.82/1.0$). The desired rates are achieved by puncturing.

To summarize, the presented rules lead to very powerful transmission systems bridging the gap between signaling with uniformly distributed signal points and Shannon limit.

6 Conclusions

Besides channel coding signal shaping provides further gain by reducing average transmit power. Since for finite constellations coding and shaping are not separable, the interaction has to be taken into account when designing coded modulation schemes. The key design point is the optimum sharing of redundancy between coding and shaping. Assuming discrete Gaussian constellations it turns out that a redundancy of 1 bit/dimension is sufficient to approach to the Shannon limit. The combination of multilevel coding *and* shaping results in completely different rates when compared to the case without shaping. In order to achieve noticeable shaping gains in practice, it is indispensable to take into account the presented rate design. Then, very powerful transmission systems result, bridging the gap between signaling with uniformly distributed signal points and Shannon limit.

References

[Bla72] R.E. Blahut. Computation of Channel Capacity and Rate–Distortion Functions. *IEEE Trans. Inf. Theory*, vol.18:pp.460–473, 1972.

[FHW96] R. Fischer, J. Huber, and U. Wachsmann. Multilevel coding: Aspects from information theory. In *Proc. IEEE Inf. Theory Workshop (ITW'96)*, page 9, Haifa, Israel, June 1996.

[For92] G.D. Forney, Jr. Trellis Shaping. *IEEE Trans. Inf. Theory*, vol.38:pp.281–300, March 1992.

[For96] G.D. Forney, Jr. Approaching the Capacity of the AWGN Channel with Coset Codes and Multilevel Coset Codes. Submitted to IEEE Trans. Inf. Theory, 1996.

[FW89] G.D. Forney, Jr. and L.-F. Wei. Multidimensional constellations—part I: Introduction, figures of merit, and generalized cross constellations. *IEEE J. Select. Areas Commun.*, vol.SAC-7:pp.877–892, Aug. 1989.

[Hub94] J. Huber. Multilevel Codes: Distance Profiles and Channel Capacity. In *ITG-Fachbericht 130*, pages 305–319, München, Oct. 1994.

[HW94] J. Huber and U. Wachsmann. Capacities of Equivalent Channels in Multilevel Coding Schemes. *Electronics Letters*, vol. 30:pp. 557–558, March 1994.

[HWF98] J. Huber, U. Wachsmann, and R. Fischer. Coded Modulation by Multilevel–Codes: Ovreview and State of the the Art. In *ITG–Fachberichte*, Mar. 1998. Same conference record.

[KP93] F.R. Kschischang and S. Pasupathy. Optimal Nonuniform Signaling for Gaussian Channels. *IEEE Trans. Inf. Theory*, vol.39:pp.913–929, May 1993.

[Ung82] G. Ungerböck. Channel coding with multilevel/phase signals. *IEEE Trans. Inf. Theory*, vol.28:pp.55–67, Jan. 1982.

[WFH97a] U. Wachsmann, R. Fischer, and J. Huber. Multilevel Coding: Basic Concepts, Capacity, and Rate Design. Submitted to IEEE Trans. Inf. Theory, 1997.

[WFH97b] U. Wachsmann, R. Fischer, and J. Huber. Multilevel Coding: Dimensionality of the Constituent Signal Set, Labeling, and Hard Decision Decoding. Submitted to IEEE Trans. Inf. Theory, 1997.

[WFH97c] U. Wachsmann, R. Fischer, and J. Huber. Multilevel Coding: Distance Profile and Signal Shaping. Submitted to IEEE Trans. Inf. Theory, 1997.

[WH95] U. Wachsmann and J. Huber. Power and bandwidth efficient digital communication using turbo codes in multilevel codes. *Europ. Trans. Telecommun. (ETT)*, vol. 6:pp. 557–567, Sept.-Oct. 1995.