RZ 3305 (# 93351) 12/18/00
Electrical Engineering 19 pages

Research Report

Rate-Splitting Multilevel Codes over an AWGN Channel

X.-Y. Hu

IBM Research
Zurich Research Laboratory
Säumerstrasse 4
8803 Rüschlikon
Switzerland

Email: xhu@ibm.zurich.com, Tel: +41 1 724-8391

IBM Research
Almaden · Austin · Beijing · Delhi · Haifa · T.J. Watson · Tokyo · Zurich

Rate-Splitting Multilevel Codes over an AWGN Channel

X.-Y. Hu

IBM Research, Zurich Research Laboratory, 8803 Rüschlikon, Switzerland

Abstract

This paper deals with a novel multilevel coding technique for an additive white Gaussian noise (AWGN) channel, referred to as rate-splitting multilevel codes (RS–MLC), which is capable of achieving the Shannon capacity limit in the high signal-to-noise regime by stacking up low-rate capacity-approaching codes. The rate-splitting concept originally occurred in a work for Gaussian multiple access channels by Rimoldi and Urbanke. RS–MLC has the advantage of allowing comparatively low-complexity successive cancelation decoding, also known as multistage decoding. Unlike classic multilevel codes where the rates at each level should be appropriately chosen via an information-theoretical approach, this scheme removes this limitation and admits a rigorous capacity achievability proof from multiuser information theory. Analysis of error propagation of multistage decoding is presented, providing guidelines for power and rate allotment among different levels. Application examples are illustrated in an analytical manner. As a byproduct, a new synchronization scheme suitable for TC and LDPC codes is described. Finally, we investigate the coding latency inherent in RS–MLC via random error exponent. It is shown that RS–MLC exhibits poor performance in terms of coding latency. The implication is that an asymptotic-optimal coding scheme may not necessarily be ideal, particularly for applications with rigid system delay.

Key Words: AWGN, Shannon Capacity, Multiple Access, Multilevel Codes, Turbo Codes, LDPC, Coding Latency

I. Introduction

Reliable communication triggered by Claude Shannon [1] in 1948 eliminated the misconception that the only way of achieving an arbitrarily small probability of error on a communication channel was to reduce the transmission rate to zero. Shannon demonstrated that it is possible to transmit information at any rate below capacity with an arbitrarily small probability of error, provided that the block length or delay can be infinite. The method to proof is random coding, where the existence of a good code is shown by averaging over all possible codes.

Ever since Shannon's famous channel capacity theorem, the challenge of coming close to capacity has generated fruitful research on modulation, coding, and equalization [2]–[6]. Notably, advances in turbo codes (TC), low-density parity check (LDPC) codes and multilevel codes (MLC) have sparked worldwide excitement and put a decisive end to the long-standing conjecture that the cutoff rate R_0 might represent the "practical capacity". Performance within tenths of a decibel of the Shannon limit is now routinely demonstrated with reasonable decoding complexity, albeit with large delay.

Turbo codes, which are originally known as parallel concatenated convolutional codes, were introduced in a paper by Berrou et al. [7]. Turbo codes combine a convolutional code along with a pseudo-random interleaver and maximum a posteriori probability (MAP) iterative decoding [8, 9] to achieve performance very close to the Shannon limit. Extensions of parallel to serially concatenated codes improve performance on the "error floor" effect [10]. Theoretical investigations to understand Turbo codes are still underway [11]-[14].

LDPC, originally introduced by Gallager [15] thirty years ago and rediscovered recently by MacKay and Neal [16], also exhibits near-Shannon-limit performance with iterative decoding on graphs [17]. Actually, LDPC are the kind of codes that can achieve the Shannon limit on the binary-symmetric channel, first proved by Gallager, then demonstrated by Richardson et al. [18] with long irregular LDPC.

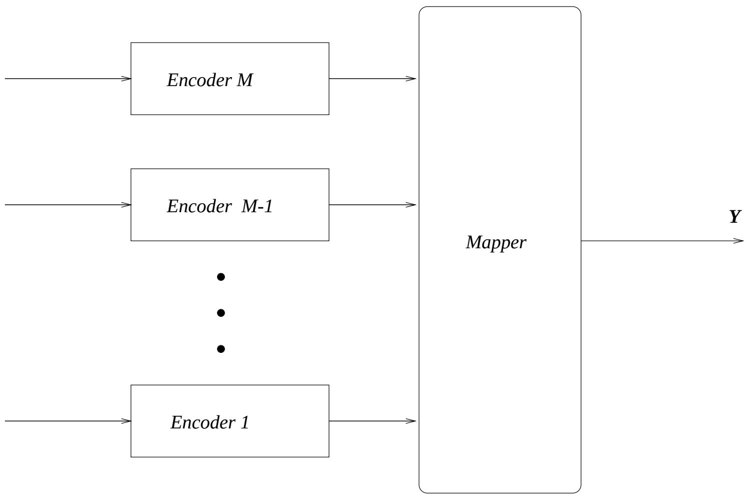

Figure 1: Multilevel encoder.

Both TC and LDPC have been confirmed to be able to approach quite closely the Shannon limit with low signal-to-noise ratios (SNR). For the high-SNR regime, MLC [20]–[28] appears to be a good way to approach the capacity limit. It has been suggested [24, 25] that a multi-level coding scheme using multistage decoding (MSD) (Block diagrams of MLC and MSD are shown in Fig. 1 and Fig. 2, respectively.) is optimal in the sense that it can approach the capacity of the underlying signal set, provided that (a) the error propagation between each level can be neglected, i.e., assuming the perfect knowledge of the previously decoded information, and (b) the rates at each level are appropriately chosen. However, if the perfect knowledge of the previously decoded information is not available, [30] showed that such a mapper channel is suboptimal—for an AWGN channel with set-partitioning labeling the capacity loss is significant, whereas with Gray labeling the capacity loss is small.

Inspired by Rimoldi and Urbanke's work on "rate-splitting" on a Gaussian multiple access channel [31, 32], this paper deals with high spectral efficiency transmission over a point-to-point AWGN channel using a rate-splitting approach[1]. It is shown that the Shannon capacity limit of AWGN is practically achievable by means of creating "virtual sources", each of which employs a low-rate capacity-achieving code. The aggregate power is distributed among virtual sources, thus having an equivalent effect of splitting the aggregate rate. Each virtual source uses an independent code, called a component code in that level. One good property of the rate splitting multilevel coding scheme (RS–MLC) lies in the simplicity of successive cancelation decoding, like multistage decoding in the conventional MLC. The essence of RS–MLC can be regarded as a replacement of "Adder" channel for "Mapper" channel in the conventional MLC. The advantage of this scheme is twofold: first it admits a rigorous capacity achievability proof, allowing the Shannon capacity limit to be approached as closely in the high-SNR regime as it can be done in the low-SNR regime with powerful binary codes; second, it allows *arbitrary* power and rate allotment among individual component codes, thereby making system design more efficient and much easier.

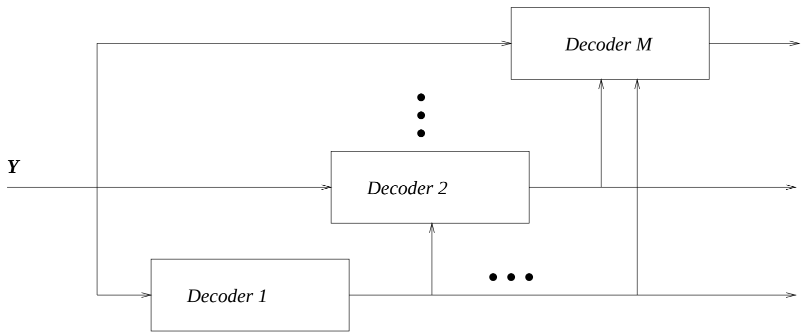

Figure 2: Multistage decoding.

RS–MLC is theoretically perfect in the sense that it successfully transforms a high-SNR channel into a set of low-SNR subchannels without any loss in terms of capacity, and in the meanwhile, admits a low-complexity MSD procedure. But in practice, one concern is the error propagation effect, which comes from the MSD procedure. The MSD works in a decision-directed manner, thus errors of previously decided levels propagate when the latter component codes are decoded in sequence. Analysis of error propagation will be treated in this paper, leading to simple rules to alleviate this deleterious effect. Design guidelines are illustrated by several examples. In addition, we describe a new synchronization scheme based on "rate splitting" concept that is particularly suitable for TC and LDPC.

[1] Actually this idea appeared first in the Ph.D thesis [32], though sketchily and without reference to the classic multilevel codes.

Although RS–MLC is capable of coming arbitrarily close to the Shannon capacity limit in the high-SNR regime, we show that this scheme is subject to greater delay by examining the random coding exponent. We find that the coding exponent of each virtual source (individual component code) is much smaller than that of a single-user case, which means a much larger coding latency[2] is unavoidable for RS–MLC to attain a fixed error probability if MSD is used. This observation indicates the insufficiency of traditional Shannon coding theorem, which focuses primarily on asymptotic performance of codes. It is more natural to ask, what is the maximal achievable rate with a fixed coding latency n and a target error probability p_w? Our results motivate further research on delay-constraint capacity and delay-constraint coding theory.

The paper is organized as follows: Section II elaborates the main idea of RS–MLC over an AWGN channel. A simple illustrative but rigorous proof from multiuser information theory is given. Section III is devoted to error propagation analysis and design critiques of practical systems. Power and rate allotment are considered. Design guidelines are illustrated by examples aiming to achieve large coding gains without sacrificing bandwidth. We examine the random error exponent of each equivalent subchannel in Section IV, which sheds light on inherent limitations of multilevel codes together with multistage decoding. Section V concludes the paper.

II. Approaching the Shannon Limit with RS–MLC

A. Basic Principle

Through orthogonal modulation described in [2], the point-to-point waveform channel can be accurately modeled as an equivalent ideal discrete-time AWGN channel whose capacity limit is [3]

$$C(P, \sigma^2) = \frac{1}{2}\log_2\left(1 + \frac{P}{\sigma^2}\right), \qquad (1)$$

where P denotes the transmit power, and σ^2 is the variance of the Gaussian noise. By means of the chain rule, one can easily verify that

$$C(P, \sigma^2) = C(P_1, \sigma^2) + C(P_2, \sigma^2 + P_1) + \cdots + C\left(P_M, \sigma^2 + \sum_{i=1}^{M-1} P_i\right) \qquad (2)$$

is valid for all nonnegative power allocation $P_1, P_2, \ldots, P_M$ such that $P = \sum_{i=1}^{M} P_i$. The chain rule says that the rate R of a single user transmitting at capacity can be regarded as an M-tuple $(R_1, R_2, \ldots, R_M)$ of M "virtual users" sharing the same AWGN channel but experiencing

[2]To construct good block codes, the major parameters of interest are the probability of a block decoding error (word error probability), denoted by p_w, the block length n, and the rate R. It should not be surprising that, if R is less than the capacity C of the channel, then one can hold R fixed and, by increasing n, find codes for which p_w becomes small exponentially with increasing n. This is the essence of the Shannon coding theorem. There are, of courses, prices to be paid by increasing the block length n. One of these is ***coding latency***. The first information bit in a block of incoming data must generally be delayed until a code word can be formed. The same case occurs at the receiver, where decoding also requires a complete code word. Note that this delay is irrelevant to processing power, and is hence called coding latency.

[3]In the paper we deal with only real-valued channels; however the results can easily be extended to complex channels, i.e., bandpass signals represented in the equivalent complex baseband. In this case, the factor of 1/2 in equation (1) does not appear.

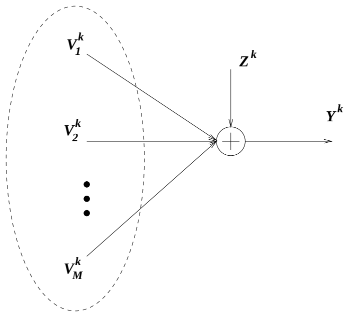

Figure 3: Equivalent Gaussian multiple access channel—creation of M virtual sources sharing the same AWGN point-to-point channel.

different noises, and finally the capacities of the M virtual users sum up to the point-to-point Shannon limit defined in Eq. (1). It is important to mention that the chain rule itself does not imply that multistage decoding achieves the capacity, whereas a rigorous proof will be established later.

By transforming point-to-point AWGN to an equivalent Gaussian multiple access channel (see Fig. 3), the signal received at time k is

$$Y^k = \sum_{i=1}^{M} V_i^k + Z^k, \tag{3}$$

where the noise samples Z^k are i.i.d. zero-mean Gaussian random variables with variance σ^2, and V_i^k is the symbol transmitted by virtual sender i with average power P_i. The following theorem establish the fact that the rate tuples defined in Eq. (2) are special "points" in the capacity region (thus achievable) and these points can be achieved via successive cancelation procedure.

Theorem (**Capacity Preservance**): With arbitrary power allocation, the capacity dictated by

$$C\left(\sum_{i=1}^{M} P_i, \sigma^2\right) = C(P_1, \sigma^2) + C(P_2, \sigma^2 + P_1) + \cdots + C\left(P_M, \sigma^2 + \sum_{i=1}^{M-1} P_i\right)$$

is achievable by M *individual* codes with a power distribution of $(P_1, P_2, \ldots, P_M)$.

Proof: Consider a discrete-time and frame-synchronous Gaussian multiple access channel with the power constraint $P = (P_1, \ldots, P_M)$, where P_i is the average power constraint for user i. Its capacity region [35, 36] is the subset of R^M containing the rate M-tuples $(R_1, \ldots, R_M)$ that satisfy

$$\sum_{j \in S} R_j \leq \frac{1}{2} \log_2 \left(1 + \frac{\sum_{j \in S} P_i}{\sigma^2}\right) \qquad \forall S \subseteq \{1, \ldots, M\}.$$

Note that the above in-equations define a capacity region over M-dimensional space. So to speak, any point in this capacity region is achievable using a joint ML decoder. However, there exist some special points, i.e., vertices in the capacity region, that allow a low-complexity decoding procedure known as successive cancelation. Selecting a consecutive subsets S such as $\{1\}, \{1, 2\}, \ldots, \{1, \ldots, M\}$, and applying the above inequality, we obtain

$$R_1 \leq \frac{1}{2} \log_2 \left(1 + \frac{P_1}{\sigma^2}\right)$$

$$
\begin{aligned}
R_1 + R_2 &\leq \frac{1}{2}\log_2\left(1 + \frac{P_1 + P_2}{\sigma^2}\right) \\
&\vdots \\
R_1 + \cdots + R_M &\leq \frac{1}{2}\log_2\left(1 + \frac{P_1 + \cdots + P_M}{\sigma^2}\right).
\end{aligned}
$$

Applying a series of subtractions of the inequality with their former counterparts yields rate tuples ("vertices") in the capacity region that are known to be achievable with successive cancelation, individually. These vertices can be represented as

$$
R_i \leq \frac{1}{2}\log_2\left(1 + \frac{P_i}{\sigma^2 + \sum_{j=1}^{i-1} P_j}\right), \qquad i = 1, \ldots, M.
$$

which are precisely the rate tuples defined in Eq. (2). This completes the proof.

The capacity achievability of the rate M-tuples of these vertices implies a low-complexity decoding scheme [35]. Rather than the complex joint decoding of M virtual sources, one can first decode the virtual source M, regarding other virtual sources as interference (virtual Gaussian noise); thus the corresponding noise variance becomes $\sigma^2 + \sum_{i=1}^{M-1} P_i$. When decoding the virtual source $M-1$, the interference effect of M can be subtracted because it is already known. Therefore, decoding the virtual source $M-1$ only observes a noise variance of $\sigma^2 + \sum_{i=1}^{M-2} P_i$. The same procedure continues until the last one experiencing noise σ^2 is decoded. This decoding mechanism is known as successive cancelation or multistage decoding [37].

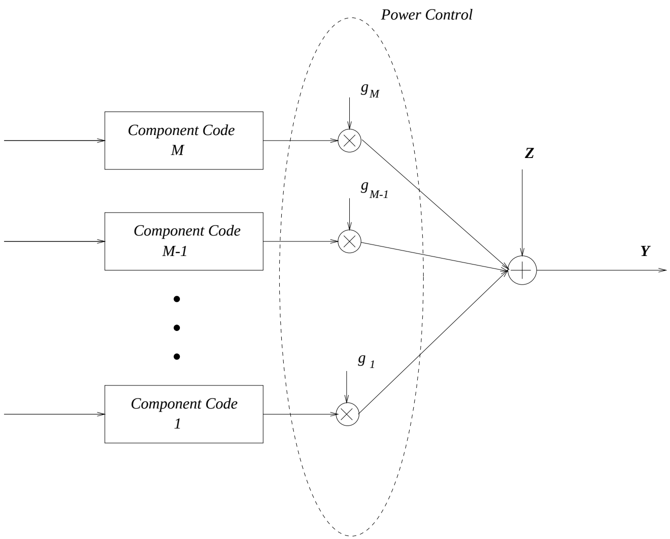

Figure 4: Rate-splitting multilevel coding scheme. Gain verctor $(g_M, g_{M-1}, \ldots, g_1)$ is employed to adjust the power of each level (virtual user) to satisfy the power distribution $(P_1, P_2, \ldots, P_M)$, where $P = \sum_{i=1}^{M} P_i$. The target capacity at level i is thus $\frac{1}{2}\log_2\left(1 + \frac{P_i}{\sigma^2 + \sum_{j<i} P_j}\right)$.

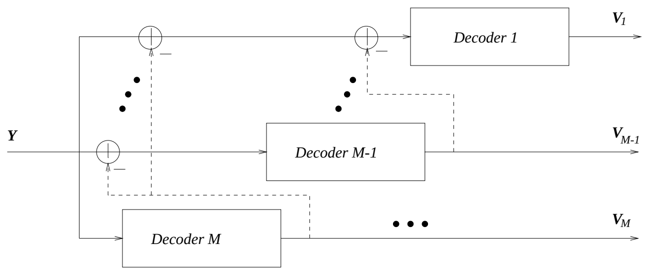

Figure 5: Multistage decoding procedure of the order of $M, M-1, \ldots, 1$.

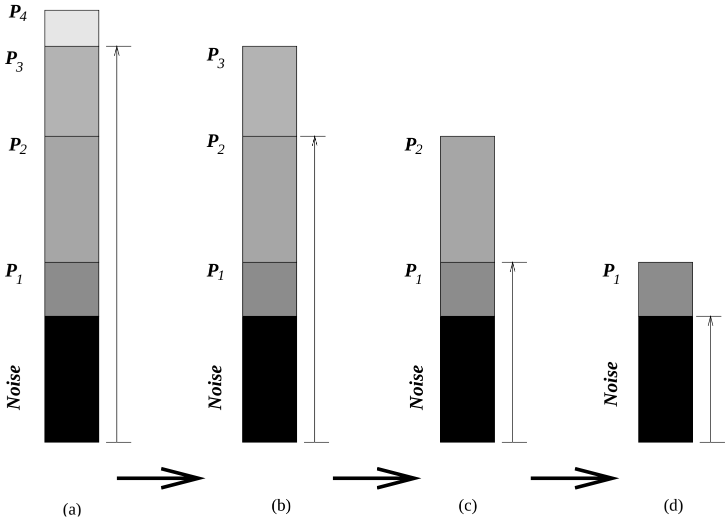

Figure 6: Noise variance in multistage decoding. (a) Decoding level 4 (i.e. the highest level of power P_4) sees background noise with variance $\sigma^2 + P_1 + P_2 + P_3$ because all transmit signals in level 3, 2, 1 are present as additional noise sources. (b) Decoding level 3 sees the variance of background noise as $\sigma^2 + P_1 + P_2$, since the signal in Level 4 has been subtracted out. (c) Likewise, noise variance in decoding Level 2 becomes $\sigma^2 + P_1$. (d) and noise variance in decoding Level 1 becomes σ^2.

Based on the capacity preservance property of virtual sources, a RS–MLC approach is immediate and straightforward. One can stack up several capacity-achieving low-rate codes (see Fig. 4), each code assuming a power allocation P_i and experiencing noise variance $\sigma^2 + \sum_{j=1}^{i-1} P_j$, respectively. The decoding procedure is much similar to the multistage decoding used in the conventional MLC, but with a prescribed decoding order from $M, M-1, \ldots$ to 1. The major difference between RS–MLC and MLC is that one can arbitrarily allocate power and rate among different levels, making system design much easier and allowing flexible error propagation control. Fig. 5 shows the multistage decoding procedure for RS–MLC, wherein previously decoded signals are subtracted out when decoding the current level. In Fig. 6 we show the effective noise variance at each decoding stage.

B. Connection to MLC

Connection of RS–MLC to the conventional MLC is straightforward by comparing Fig. 1 with Fig. 4. The major difference is that RS–MLC adopts an "Adder" channel whereas MLC a "Mapper" channel. In broad terms, an adder can be considered as a special mapper, thus

RS–MLC is a particular subset of MLC. Specifically, taking the output of component codes as binary labeling $\{0, 1\}$, and choosing the gain vector as $(2^{M-1}, \ldots, 1)$, RS–MLC can also be described as a conventional MLC with 2^M-PAM. Similarly, using two bits for each level at a time, a specific MLC with complex lattice $Z^2/2Z^2/4Z^2/\ldots/2^{M-1}Z^2$ considered in [38] turns out to be an RS–MLC scheme in which the mapper is precisely an adder; simulation results showed that this specific 4-ary partitioning schemes provide substantial savings in delay and complexity of multistage decoding compared to codes based on binary partitioning schemes in [26].

A general mapper channel imposes relatively strong constraint compared with a more flexible adder channel; Obviously, since RS–MLC assumes Gaussian signals at each level, no limitations are present. This advantage admits a rigorous capacity achievability proof form multiuser information theory. For the classic MLC, a similar argument was claimed in [24, 25] based on the chain rule of mutual information and assuming that perfect knowledge of previously decoded information is available. Without such an assumption, the classic MLC with multistage decoding can be modeled by a set of equivalent parallel bit-channel of BICM with ideal interleaving [30]. More specifically, it was shown that for an AWGN channel with a set-partitioning mapper, the capacity loss resulting from this mapper is significant, whereas with Gray mapper the capacity loss is quite small. Both mappers are suboptimal when the suboptimal multistage decoding is used. However, when an adder mapper is employed, there is no capacity loss as we have shown.

The most significant application of MLC (or RS–MLC) is to approach the Shannon limit closely in the high-SNR regime, provided that low-rate capacity-achieving codes are employed as component codes. With low-complexity multistage decoding, decoding is performed separately at each level; at the jth level the decisions at higher levels $(M, M-1, \ldots, j+1)$ are taken into account, whereas no knowledge of the lower levels $(j, \ldots, 1)$ is assumed. As the error probability at each level can be made arbitrarily small, the error propagation from upper to lower level is negligible. Empirically, by using TC of appropriate rate at each level, it has been shown that reliable transmission can be achieved within 1 dB of the Shannon limit [26]. Undoubtedly, if the more powerful LDPC codes are used as component codes, performance within tenths of a decibel of the Shannon limit is practical, albeit with greater delay.

III. Performance Analysis: Error Propagation

Practically it is desirable to design a coding system with a specified target error probability, which is usually determined by application scenarios. For example, in mobile communications, 10^{-3} is enough; however in data networks, 10^{-6} or lower is required. Therefore, proper power and rate allocation at each level, taking into account the effect of error propagation, is of great importance.

Assume p_i^t is the target error probability specified by applications at level i. Referring to Fig. 7, we know that if an error occurs at time index k, most likely all lower levels at time index k will have an error event due to wrong successive subtractions. Assume the errors coming from upper levels are not correctable and there are no overlaps of error events between different levels, then the correct probability at level i, p_i^c equals

$$\begin{aligned} p_i^c &= p_M^c \cdot p_{M-1}^c \cdots p_i^c \\ &= (1-p_M^t)(1-p_{M-1}^t)\cdots(1-p_i^t). \end{aligned} \tag{4}$$

As p_j^t is usually a small number, it is reasonable to drop the mutual multiplications of these small numbers, which yields

$$p_i^c = 1 - \sum_{j=i}^{M} p_j^t. \tag{5}$$

We can then denote p_i^{ep} as the error probability due to error propagation to obtain

$$p_i^c = 1 - p_i^t - p_i^{ep}. \tag{6}$$

Subtracting Eq. (5) from (6) yields

$$p_i^{ep} = \sum_{j=i+1}^{M} p_j^t. \tag{7}$$

It turns out that the decoding errors occurring in a higher level serve as an addition term to that of the lower levels. Owing to this "sinking behavior" of decoding errors, it is advisable to put the largest target error probability to the lowest level, and then the second largest to the second lowest level, and so on, thus alleviating the effect of error propagation to a maximum extent. It should be noted that the error propagation in Eq. (7) is actually an upper bound, because overlappings of errors between different levels are not taken into account.

Bearing in mind the sinking behavior described in Eq. (7), one can allocate suitable power and rate to each level, depending on the choice of component codes. A specific component code may be from block or convolutional codes, particularly capacity-approaching TC, LDPC and lattice codes. An appropriate choice might rely mainly on tradeoffs between code rate, error-correcting ability, coding latency and/or coding gain, etc. It is worth pointing out that the order of allocation of the available power should be from lower level to upper level in order to accommodate the chain effect of interference (i.e. each level acts like a noise source to other levels in one direction). An illustrative example is shown in Fig. 8.

We now consider several possible applications of RS–MLC. In all examples considered here, we present analytical results only, for the sake of simplicity, under the assumption that the

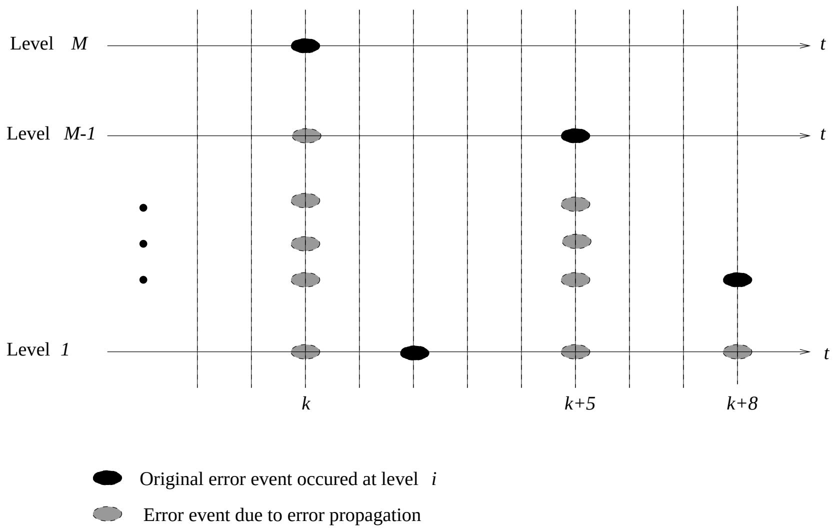

Figure 7: Error events in multistage decoding. At time k, an original error event at level M propagates to all lower levels; at time $k+2$, there is no error propagation because level 1 is the last one to be decoded; at time $k+5$, an original error event occurs at level $M-1$ and, correspondingly, error events occur at level $M-2, \ldots, 1$ due to error propagation.

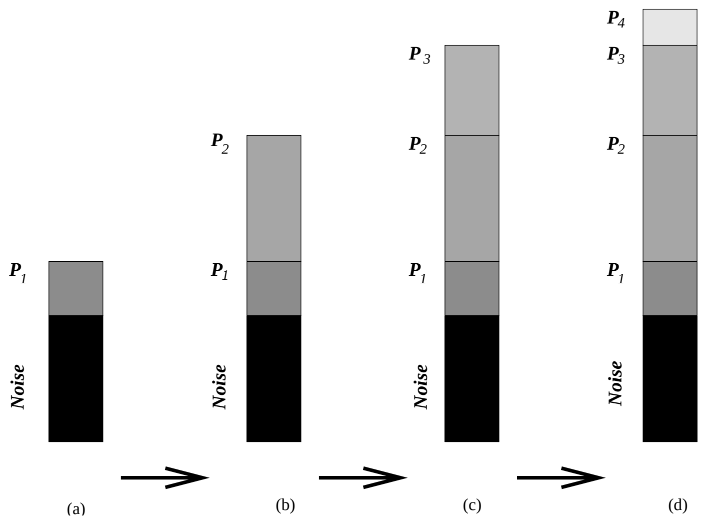

Figure 8: Procedure of allocating power and rate: (a) Allocate P_1 to level 1, then the capacity of this level is $C_1 = \frac{1}{2}\log_2\left(1+\frac{P_1}{\sigma^2}\right)$. (b) Allocate P_2 to level 2 with capacity $C_2 = \frac{1}{2}\log_2\left(1+\frac{P_2}{\sigma^2+P_1}\right)$. (c), (d) The same procedure continues up to $P = \sum P_i$. Note that the order of power allotment is merely inverse to the order of multistage decoding.

coding signals of each level is Gaussian, which is a prerequisite for the presented proof. These examples are primarily used to illustrate the design rules for RS–MLC. Assume that a component code has a coding gain of γ with respect to uncoded binary coherent BPSK, then we obtain the following approximate expression on error performance

$$p_e \approx Q\left(\sqrt{\gamma \mathrm{SNR}_{\mathrm{eff}}}\right),$$

where $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp(-y^2/2)dy$, $\mathrm{SNR}_{\mathrm{eff}}$ is the effective signal-to-noise ratio experienced in the corresponding level.

Application 1: Coding gain versus bandwidth

Example 1: Consider an application case of rate 1 bit/sample with an error probability of 10^{-6}, where the baseline uncoded binary modulation scheme operates about 5.76 dB away from the Shannon limit. Here we are concerned with coding gain without sacrificing bandwidth efficiency by means of RS–MLC. We split the total rate 1 bit/sample into two levels of rate 0.5 bit/sample, each employing an irregular rate-1/2 LDPC code [18] operating 0.3 dB away from Shannon capacity [4] at an error probability of 10^{-6}. Taking into account the error propagation effect of MSD, we select the target error probability of level 1 and level 2 to be 5×10^{-7}, thus the overall error probability is guaranteed to be below 10^{-6}. The error performances of level 2 (decoded first) and level 1 (decoded last) are plotted in Fig. 9, wherein the variance of background noise σ^2 is normalized to 1.0. It is demonstrated that a coding gain of 5.0 dB can be achieved by RS–MLC with multistage decoding.[5]

Example 2: Consider an RS–MLC scheme of rate 2 bits/sample. Similarly, we stack up four rate-1/2 irregular LDPC codes, each operating 0.3 dB away from the Shannon capacity limit

[4]This code is only 0.13 dB away from capacity of binary-input AWGN channel, and has a code blocklength of 10^6.

[5]This coding gain might be somewhat optimistic since we have assumed the signals in each level looks like Gaussian for the sake of simplicity.

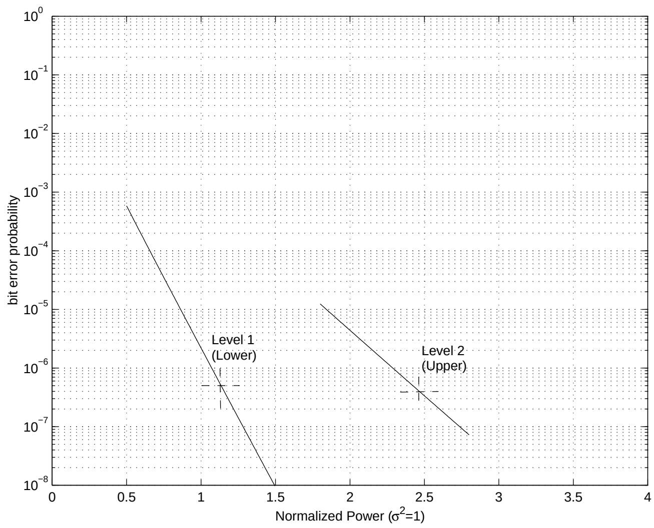

Figure 9: Analytical results: the AWGN channel variance σ^2 is normalized to 1. Upper level (i.e. level 2) is decoded first, seeing a background noise of $1.0 + 1.135 = 2.135$. The factor of 1.135 comes from the power of level 1, which can achieve an error probability of 5×10^{-7} when level 2 is subtracted. The total power allocated to levels 1 and 2 is $P_1 + P_2 = 1.135 + 2.423 = 3.558$. In the case of $\sigma^2 = 1$ and rate 1 bit/sample, the minimum power required by the Shannon capacity formula is 3.0, and then the scheme is only $10 \log_{10} \frac{3.558}{3.0} = 0.741$ (dB) away from the capacity limit. Recalling that the baseline uncoded binary modulation scheme operates about 5.76 dB away from the Shannon limit at 10^{-6}, a coding gain of 5.0 dB can be achieved.

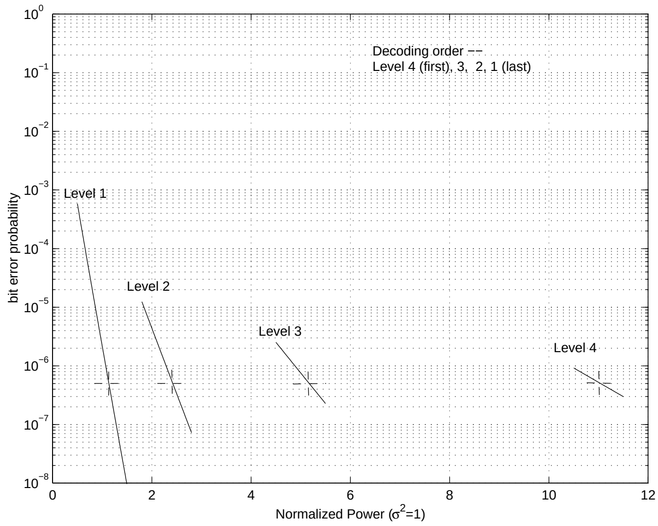

Figure 10: Analytical results: power allocation at each level is 1.135/2.423/5.173/11.045, respectively (from level 1 to level 4). The total energy is 19.776. Each level uses a rate-1/2 irregular LDPC code of [18]. This example operates only 1.2 dB away from the Shannon limit.

(the same code as in Example 1). The target error probability of each level is selected to be 5×10^{-7} so that the overall error probability keeps approximately 10^{-6}. Again multistage decoding is considered. In Fig. 10 we see that the total power allocated to four levels is $P = 19.776$. Recalling that a minimum power of 15.0 (with respect to the normalized noise variance 1) is required to reliably transmit 2 bits/sample, this example is operating $10 \log_{10} 19.776/15 = 1.2$ (dB) away from the Shannon limit.

Before ending this section, it is worth pointing out that the optimality of RS–MLC and its corresponding MSD holds only in the asymptotic case of capacity-approaching component

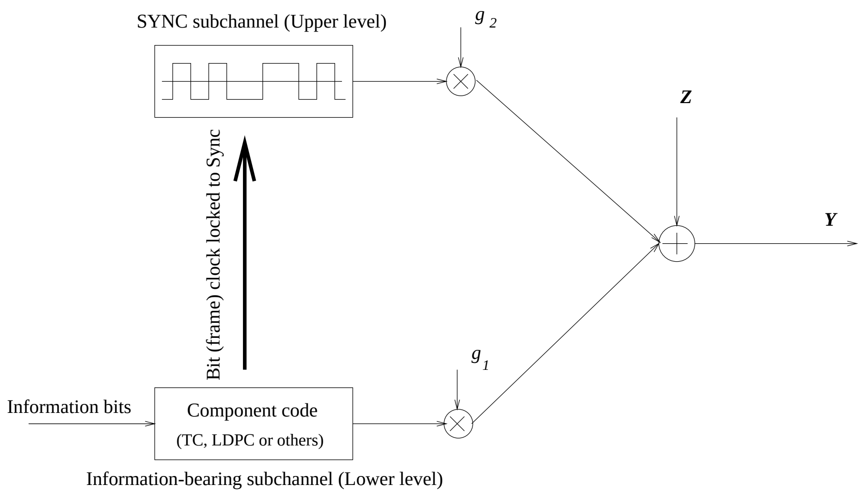

Figure 11: Rate-splitting synchronization scheme.

codes. Care must be taken if component codes operate somewhat further from the Shannon capacity. The reason for this is that each level acts as interference to its higher levels and hence the extra power needed at each level will magnify significantly from lower to higher levels. Surely, a full maximum likelihood decoding, other than the suboptimal multistage decoding, may alleviate this deleterious effect at the cost of implementation complexity. Instead of performing multistage decoding once per multilevel codeword, iterative decoding procedure has been suggested in [46].

Application 2: Synchronization of TC or LDPC

The invention of TC and LDPC codes can work properly at a very low SNR with iterative decoding. In such cases, carrier, symbol and/or frame synchronization turns out to be difficult because the information-bearing signal is quite noisy. The contradiction is that TC and LDPC may reduce the energy per transmitted symbol below the level needed for reliable tracking by the phased-locked loops in the coherent demodulator. Therefore, it is necessary to separate the synchronization channel (usually uncoded) from the information-bearing channel (coded).

It is natural to design a two-level system (as shown in Fig. 11): the upper level, which will be detected first, carries a pseudo-random (PN) sequence (uncoded), referred to as a "sync subchannel"; and the lower level carries a coded information sequence in which TC or LDPC may be employed to provide powerful error protection for information bits. The two subchannels use the same time clock, i.e. the information-bearing subchannels locked to the sync subchannel, synchronized on both bit clock and frame clock. The upper level, i.e. SYNC subchannel is "decoded" first, yielding correct carrier, symbol and frame synchronization. As the PN sequence in the upper level is a known pattern, its interference effect can be accurately subtracted when decoding the lower level, provided that the synchronization is achieved.

Optimum closed-loop phase estimators for modulated and unmodulated carriers based on maximum a posteriori (MAP) have been shown to take the form of feedback loops that attempt to continuously null the difference between the phase of the received signal and its estimate, produced by a controllable oscillator within the loop. It is well-known [47] that at a high loop SNR the variance of the phase estimate varies inversely as the loop SNR for unmodulated carriers, but that for modulated carriers such as BPSK, the variance of the phase error increases due to the inherent multiplication operations needed to generate the error signal; this additional degradation is generally referred to as "squaring loss" and becomes more severe in low SNR regime. In a rate-splitting synchronization scheme,[6] since the phase detector works on the sync subchannel that is uncoded, no squaring loss is induced and then such a scheme is appropriate for fairly noisy channels.

IV. Performance as a Function of Block Length

We have shown above that the Shannon capacity limit in the high-SNR regime can be approached by multiplexing low-SNR low-rate capacity-approaching codes without loss of optimality. Both rigorous proof and design examples validate this conclusion. Therefore it seems safe to say that reliable communication up to the Shannon capacity is practically feasible both in low-SNR and high-SNR regimes.

However, approaching the Shannon capacity limit is not the whole story of coding practice. The Shannon coding theorem states that, if R is less than the capacity C, no matter how close they are, surely there exist codes for which the word error probability p_w becomes small exponentially with increasing n. There are, of course, prices to be paid by increasing the block length, one of which is system delay. At the transmitter side, the first information bit in a block of incoming data stream generally must be delayed by n bits before a code word can be formed, and at the receiver it is the same case. Note that this system delay results inherently from the coding process and cannot be reduced by increasing processing power; it is thus referred to as *coding latency.*

The relation between block length, i.e., coding latency, and word error rate p_w can be reflected by the famous random coding exponent $E_r(R)$ [33]

$$E_r(R) \stackrel{def}{=} \max_{0\le\rho\le1} \{E_0(\rho) - \rho R\}, \tag{8}$$

where

$$E_0(\rho) \stackrel{def}{=} -\log_2 \left\{ \int_Y \left[\sum_{x\in X} \Pr\{x\} (f_Y(y|x))^{\frac{1}{1+\rho}} \right]^{1+\rho} dy \right\}. \tag{9}$$

The channel input and output are denoted by X and Y, respectively. It has been shown (see [40, 42]) that there exits at least one $(n,\ R)$ block code for which

$$p_w \le e^{-nE_r(R)}, \tag{10}$$

where n is the block length (also called coding latency) and R is the rate. In other words, we obtain an approximate estimate of the minimum coding latency, given p_w and $E_r(R)$, in the following form:

$$n \ge -\frac{1}{E_r(R)} \ln p_w. \tag{11}$$

Note that the random coding exponent $E_r(R)$ is a convex $\cup$, decreasing, positive function of R for $0 \le R < C$, and it is independent of blocklength n. From the point of view of system delay and complexity, a larger random coding exponent is quite desirable.

[6] here the term of power-splitting would be more appropriate since the sync subchannel does not transmit information bits.

As it is usually difficult to evaluate Eqs. (8) and (9) directly, we turn to Shannon's sphere packing lower bound and random coding upper bound [40], which bounded the word error probability for a block code with equal-energy codewords of a specific block size. The sphere packing bound can be written in the form [41]

$$\begin{aligned} p_w &\geq Q_{sp}(\theta) \\ &= \int_{\theta}^{\pi} \frac{(n-1)(\sin\phi)^{n-2}}{2^{n/2}\sqrt{\pi}\Gamma(\frac{n+1}{2})} \int_0^{\infty} r^{n-1} e^{-(r^2+nA^2-2r\sqrt{n}A\cos\phi)/2} dr\, d\phi \end{aligned} \tag{12}$$

where A is signal-to-noise "amplitude" ratio, i.e., $\sqrt{P/N}$, $\Gamma(\cdot)$ is the Gamma function, and θ is the root of the following equation:

$$\int_0^{\theta} \frac{n-1}{n} \frac{\Gamma(\frac{n}{2}+1)}{\Gamma(\frac{n+1}{2})\sqrt{\pi}} (\sin\phi)^{n-2} d\phi = 2^{-nR} \tag{13}$$

Shannon derived asymptotic approximation to these functions that are valid for large n

$$Q_{sp}(\theta) \approx \frac{[G(\theta)\sin\theta e^{-(A^2-AG(\theta)\cos\theta)/2}]^n}{\sqrt{n\pi}\sqrt{1+G^2(\theta)\sin\theta}[AG(\theta)\sin^2\theta - \cos\theta]} \tag{14}$$

where $G(\theta) = (1/2)[A\cos\theta + \sqrt{A^2\cos^2\theta + 4}]$, and Eq. (13) becomes, asymptotically,

$$\frac{\Gamma(\frac{n}{2}+1)(\sin\theta)^{n-1}}{n\Gamma(\frac{n+1}{2})\sqrt{\pi}\cos\theta} = 2^{-nR} \tag{15}$$

Shannon also computed an upper bound on performance by a "random coding" method. The random coding bound gives an expression for the ensemble average word error probability, averaged over the ensemble of all possible spherical codes, where each codeword is selected independently and completely at random, subject to an energy constraint. If the transmit rate is close to channel capacity, an asymptotic formula of the random coding upper bound turns out to be the sphere packing lower bound multiplied by a factor essentially independent of n, that is

$$\begin{aligned} p_w &\leq Q_{rc}(\theta) \\ &\approx Q_{sp}(\theta)\left(1 + \frac{AG(\theta)\sin^2\theta - \cos\theta}{2\cos\theta - AG(\theta)\sin^2\theta}\right) \end{aligned} \tag{16}$$

It is worth pointing out that, asymptotically, for rates near channel capacity, the multiplying factor in Eq. (16) is just a little over unity; thus the sphere packing lower bound and the random coding upper bound are close together – the asymptotic error exponents of the lower bound and the upper bound surprisingly turn out to be identical, i.e.,

$$\begin{aligned} E_{asy.}(\theta) &\stackrel{def}{=} \lim_{n\to\infty} -(1/n)\ln Q_{sp}(\theta) \\ &\stackrel{def}{=} \lim_{n\to\infty} -(1/n)\ln Q_{rc}(\theta) \qquad \text{(alternatively)} \\ &= \frac{1}{2}[A^2 - AG(\theta)\cos\theta] - \ln[G(\theta)\sin\theta] \end{aligned} \tag{17}$$

Table 1: Random coding exponents of RS–MLC: an example

Level i	Level 1	Level 2	Level 3	Level 4
E_r^i	1.4×10^{-3}	1.4×10^{-3}	1.4×10^{-3}	1.4×10^{-3}

where θ is the solution of a limiting expression of Eq. (15),

$$\lim_{n\to\infty} \frac{\Gamma(\frac{n}{2}+1)(\sin\theta)^{n-1}}{n\Gamma(\frac{n+1}{2})\sqrt{\pi}\cos\theta} = \lim_{n\to\infty} 2^{-nR} \tag{18}$$

which yields

$$\theta = \arcsin(2^{-R}) \tag{19}$$

Although the error exponent determined by Eq. 17 holds asymptotically (in the sense of blocklength n), it can act as a faithful surrogate for random coding exponent $E_r(R)$ because, by definition, $E_r(R)$ is independent of n. As it can be seen, the asymptotic assumption merely makes the calculation more tractable. Next, we employ Eq. 17 to weigh the performance of RS–MLC against that of single-user coding in terms of coding latency.

Different from an RS–MLC scheme in which several virtual users separately utilize partial power to transmit information, a single-user coding refers to using aggregate power by a single code without partitioning the aggregate rates into several codes. More specifically, consider the power allotment in example 2 (the four-level case). The power distribution is 1.135/2.423/5.173/11.045, respectively, from level 1 to level 4. Each level is supposed to transmit one-half bit per sample, and the background noise is normalized to $\sigma^2 = 1.0$. The multistage decoding starts at level 4, then goes to level 3, level 2, and finally level 1; thus the corresponding noise variance for each level is (1.0+1.135+2.423+5.173), (1.0+1.135+2.423), (1.0+1.135), and 1.0, respectively. Assuming each level uses a sphere-packing code—the best code we can imagine, by calculation of Eqs. (17) and (19) we obtain the random coding exponent [7] E_r^i for each level $i, i = 4, 3, 2, 1$ which are shown in Table 1.

To make comparison, consider the following three cases in which two or more lower levels are combined into a single level and others remain the same: the first case is to utilize the aggregate power of two lower levels and correspondingly the aggregate rate is 1 bit/sample. The aggregate power of this combined level is (1.135+2.423) and the resulting error exponent turns out to be 4.7×10^{-3}. The second case utilizes the aggregate power of three lower levels and then the aggregate rate is 1.5 bit/sample. The resulting error exponent is 1.04×10^{-2}. Finally, the third utilizes the aggregate power of the four levels and then the aggregate rate is 2 bit/sample. This is exactly the single-user coding case and the resulting error exponent is 1.87×10^{-2}. It is evident that the random coding exponent of single-user coding is significantly larger than that of RS–MLC. To attain the same error probability performance, an RS–MLC scheme requires a quite larger block length as compared to a non-splitting single user scheme. It can also be seen that the more the number of the levels, the smaller the corresponding random coding exponent.

Another way to view the coding latency inherent in RS–MLC is to investigate the word error probability performance versus block size. In Fig. 12, we compare the performance of

[7]The asymptotic error exponent in the sense of blocklength n can be used a good approximation of random coding exponent because, by definition, random coding exponent is independent of blocklength n.

RS–MLC with the three aforementioned cases which combine several levels into a single level and maintain the same aggregate rate; the performance curves are calculated assuming that the best code, i.e. Shannon sphere-packing code [8], is used as component code. Note that the last case in which combines all levels into a single level is precisely the non-rate-splitting case, or say single-user coding case. It is shown that single-user coding significantly outperforms RS–MLC in terms of coding latency.

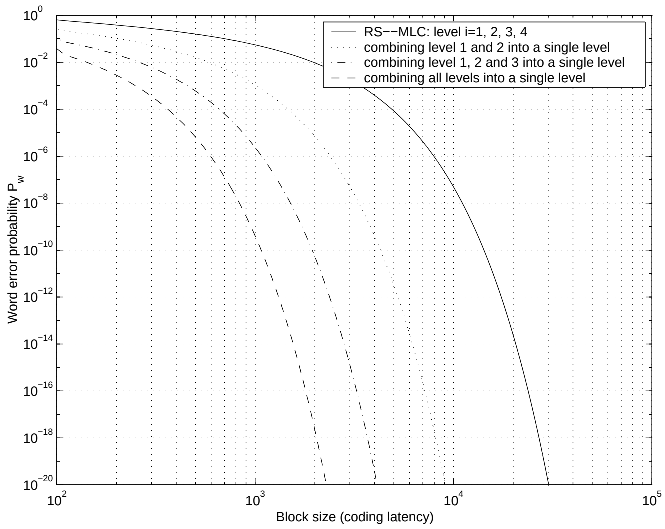

Figure 12: Single-user coding of combined lower levels assuming that a sphere-packing code is used.

The implication of this result is that, even though RS–MLC is capable of approaching the Shannon limit asymptotically, it turns out to be inefficient in terms of coding latency, which in turn determines the minimum system delay. Good codes should be able to approach a fixed error probability quickly without sacrificing throughput as the block size (coding latency) increases. Our results reveal that a capacity-achieving coding scheme like RS–MLC may not necessarily be a good one for delay sensitive applications. Further research on delay-constrained capacity and delay-constrained coding theory is thus desirable.

V. Concluding Remarks

We investigated a new multilevel coding technique for an additive white Gaussian noise channel, referred to as RS–MLC. By stacking up low-rate capacity-approaching codes, this technique allows the Shannon capacity limit in the high-SNR regime to be approached as closely as it can be done in the low-SNR regime with powerful binary codes. The significance of this work was enhanced by a recent advance in coding theory that a rate-1/2 LDPC code approaches Shannon limit within 0.0045 dB [19]. RS–MLC is also of practical interest because of its comparatively low-complexity successive cancelation decoding, or called multistage decoding. Contrary to the conventional multilevel codes whose optimality depends on appropriate rate choices on each level, RS–MLC is capable of arbitrarily allotting power and rate to each level by means of "rate splitting", which was originally proposed in a Gaussian multiple-access

[8]Such a code doesnot exist at all for $n > 2$, but it is the best code we can imagine.

channel by Rimoldi and Urbanke [31, 32]. It turns out that this treatment leads directly to a rigorous capacity achievability proof, which is already available from multiuser information theory.

One concern in RS–MLC is the error propagation effect, which comes from the low-complexity multistage decoding procedure. This MSD works in a decision-directed manner, thus decision errors propagate when decoding each component code in sequence. Error propagation is analyzed, and design guidelines are illustrated by several typical examples. In addition, we extend the rate splitting concept to a new synchronization scheme which is particularly suitable for low-SNR scenarios.

Implicit in the multistage decoding to achieve the Shannon capacity limit is the assumption of Gaussian behavior of each level. This requirement might be attained in principle by the shaping scheme [45] applied to each individual component code in RS–MLC, perhaps not because how large the shaping gain is, but because signals in lower levels serve as interference to others, thus they are required to be Gaussian-like.

Last, but by no means least importantly, we discussed the coding latency inherently in RS–MLC in order to attain a fixed error probability. RS–MLC exhibits poor performance in terms of coding latency. The implication is that an asymptotically optimal code may not necessarily be a good one. Future efforts towards information-theoretical understanding of delay-constrained capacity and related coding issues are motivated.

ACKNOWLEDGMENT

The author is greatly indebted to Prof. B. Rimoldi for the course on information theory and wireless communications at EPFL. His insights on Gaussian multiple access channel directly stimulated this work. I would also like to thank Dr. Evangelos Eleftheriou (IBM Zurich Research Lab) and Dr. Robert Fischer (Universitaet Erlangen-Nuernberg) for insightful comments and suggestions.

References

[1] C.E. Shannon, "A Mathematical Theory of Communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379-423 and pp. 623-656, July and Oct. 1948.

[2] G.D. Forney, Jr. and G. Ungerboeck, "Modulation and Coding for Linear Gaussian Channels," *IEEE Trans. on Information Theory*, vol. 44, pp. 2384-2415, Oct. 1998.

[3] A.R. Calderbank, "The Art of Signaling: Fifty Yeas of Coding Theory," *IEEE Trans. on Information Theory*, vol. 44, pp. 2561-2595, Oct. 1998.

[4] D.J. Costello, Jr, J. Hagenauer, H. Imai, and S.B. Wicker, "Applications of Error-Control Coding," *IEEE Trans. on Information Theory*, vol. 44, pp. 2531-2559, Oct. 1998.

[5] S. Verdu, "Fifty Years of Shannon Theory," *IEEE Trans. on Information Theory*, vol. 44, pp. 2057-2077, Oct. 1998

[6] E. Biglieri, J. Proakis, and S. Shamai, "Fading Channels: Information-Theoretic and Communications Aspects," *IEEE Trans. on Information Theory*, vol. 44, pp. 2619-2691, Oct. 1998.

[7] C. Berrou, A, Glavieux, and P. Thitimajshima, "Near Shannon Limit Error-correcting Coding and Decoding: Turbo Codes," *Proc. 1993 Int. Conf. Communication,* Geneva, Switzerland, pp. 1064-1070, May 1993.

[8] L.R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal Decoding of Linear Codes for Minimizing Symbol Error Rate," *IEEE Trans. on Information Theory*, vol. 20, pp. 284-287, March 1974.

[9] J. Hagenauer, E. Offer, and L. Papke, "Iterative Decoding of Binary Block and Convolutional Codes," *IEEE Trans. on Information Theory*, vol. 42, pp. 429-445, March 1996.

[10] S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, "Serial Concatenation of Interleaved Codes: Performance Analysis, Design, and Iterative Decoding," *IEEE Trans. on Information Theory*, vol. 44, pp. 909-926, May 1998.

[11] D. Divsalar and R.J. McEliece, "Effective Free Distance of Turbo Codes," *Electronics Letters,* vol. 32, pp. 445-446, Feb. 1996.

[12] S. Benedetto and G. Montorsi, "Unveiling Turbo Codes: Some Results on Parallel Concatenated Coding Schemes," *IEEE Trans. on Information Theory,* vol. 42, pp. 409-428, Nov. 1996.

[13] L.C. Perez, J. Seghers and D.J. Costello, "A Distance Spectrum Interpretation of Turbo Codes," *IEEE Trans. on Information Theory,* vol. 42, pp. 1698-1709, Nov. 1996.

[14] S. ten Brink, "Iterative Decoding Trajectories of Parallel Concatenated Codes," *3rd IEEE/ITG Conference on Source and Channel Coding,* (Munich, Germany), Jan. 2000.

[15] R.G. Gallager, *Low Density Parity Check Codes,* MIT press, Cambridge, Mass., 1963.

[16] D.J.C. MacKay and R.M. Neal, "Near Shannon Limit Performance of Low-density Parity-check Codes," *Electron. Lett.,* vol. 32, pp. 1645-1646, Aug. 1996.

[17] N. Wiberg, H.-A. Loeliger, and R. Kotter, "Codes and Iterative Decoding on General Graphs," *Euro. Trans. Telecommun.,* vol. 6, pp. 513-526, Sept. 1995.

[18] T. Richardson, A. Shokrollahi and R. Urbanke, "Design of Provably Good Low-density Parity Check Codes," submitted for publication.

[19] S.-Y. Chung, G. David Forney, T. Richardson, and R. Urbanke, "On the Design of Low-Density Parity-Check Codes within 0.0045 dB of the Shannon Limit," submitted to *IEEE Communin. Letters.*

[20] H. Imai and S. Hirakawa, "A New Multilevel Coding Method Using Error Correcting Codes," *IEEE Trans. on Information Theory*, vol. 23, pp. 371-377, May 1977.

[21] G. Ungerboeck, "Channel Coding with Multilevel/Phase Signals," *IEEE Trans. on Information Theory,* vol. 28, pp. 55-67, Jan. 1982.

[22] A.R. Calderbank, "Multilevel Codes and Multistage Decoding," *IEEE Trans. on Communications,* vol. 37, pp. 222-229, March 1989.

[23] G.J. Pottie and D.P. Taylor, "Multilevel Codes Based on Partitioning," *IEEE Trans. on Information Theory,* vol. 35, pp. 87-98, Jan. 1989.

[24] Y. Kofman, E. Zehavi, and S. Shamai (Shitz), "Performance Analysis of a Multi-level Coded Modulation System," *IEEE Trans. on Commun.,* vol. 42, pp. 299-312, Feb./Mar./Apr. 1994.

[25] J. Huber and U. Wachsmann, "Capacities of Equivalent Channels in Multilevel Coding Schemes," *Electron. Lett.,* vol. 30, pp. 557-558, Mar. 1994.

[26] 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.

[27] U. Wachsmann, R.F.H. Fischer and J.B. Huber, "Multilevel Codes: Theoretical Concepts and Practical Design Rules," *IEEE Trans. on Information Theory*, vol. 45, pp. 1361-1391, July 1999.

[28] G. D. Forney, M. D. Trott, S.-Y. Chung, "Sphere-Bound-Achieving Coset Codes and Multilevel Coset Codes," *IEEE Trans. on Information Theory*, vol. 46, pp. 820-850, May 2000.

[29] T. Wörz and J. Hagenauer, "Iterative Decoding for Multilevel Codes Using Reliability Information," *in Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Orlando, FL, Dec. 1992.

[30] G. Caire, G. Taricco, and E. Biglieri, "Bit-Interleaved Coded Modulation," *IEEE Trans. on Information Theory*, vol. 44, pp. 927-946, May 1998.

[31] B. Rimoldi and R. Urbanke, "A Rate-splitting Approach To The Gaussian Multiple-access Channel," *IEEE Trans. on Information Theory*, vol. 42, pp. 364-375, March 1996.

[32] R. Urbanke, "On Multiple Access Communication," Ph.D Thesis, Washington University, Saint Louis, 1995.

[33] R.G. Gallager, *Information Theory and Reliable Communication,* New York: Wiley, 1969.

[34] A. Ruiz, J.M. Cioffi, and T. Woerz, and K. Kasturia, "Discrete Multiple Tone Modulation with Coset Coding for the Spectrally Shaped Channel," *IEEE Trans. Commun.,* vol. 40, pp. 1012-1029, June 1992.

[35] T.M. Cover and J.A. Thomas, *Elements of Information Theory,* New York: Wiley, 1991.

[36] A.D. Wyner, "Recent Results in the Shannon Theory," *IEEE Trans. on Information Theroy,* vol. 20, pp.2-10, Jan. 1974.

[37] R.G. Gallager, "A Perspective on Multiaccess Channels," *IEEE Trans. on Information Theory,* vol. 31, pp. 124-142, March 1985.

[38] T. Mittelholzer, X. Lin, and J.L. Massy, "Multilevel Turbo Coding for M-ary Quadrature and Amplitude Modulation," *Proc. 1997 Inter. Symp. on Turbo Codes,* Brest, France, pp. 127-134, Sept. 1997.

[39] G. Ungerboeck, "Trellis-coded Modulation with Redundant Signal Sets—Part I and Part II," *IEEE Commun. Mag.*, Vol. 25, pp. 5-21, Feb. 1987.

[40] C.E. Shannon, "Probability of Error for Optimal Codes in a Gaussian Channel," *Bell Syst. Tech. J.,* vol. 38, pp. 611-656, May 1959.

[41] S. Dolinar, D. Divsalar, and F. Pollara, "Code Performance as a Function of Block Size," *TMO Progress Report 42-133,* Jet Propulsion Laboratory, Pasadena, California, pp. 1-23, May 1998.

[42] A.J. Viterbi, and J.K. Omura, *Principle of Digital Communication and Coding,* McGraw-Hill, 1979.

[43] R. De Buda, "Some Optimal Codes Have Structure," *IEEE J. Select. Areas Commun.,* vol. 7, pp. 893-899, Aug. 1989.

[44] R. Urbanke and B. Rimoldi, "Lattice Codes Can Achieve Capacity on the AWGN Channel," *IEEE Trans. on Information Theory,* vol. 44, pp. 273-278, Jan. 1998.

[45] G.D. Forney, "Trellis Shaping," *IEEE Trans. on Information Theory,* vol. 38, pp.281-300, March 1992.

[46] K.R. Narayanan and J. Li, "Bandwidth Efficient Low-Density Parity-Check Coding Using Multilevel Coding and Iterative Multistage Decoding," *2nd Int. Symp. on Turbo Codes & Related Topics,* Brest, France, 2000.

[47] J. G. Proakis, *Digital Communications,* McGraw-Hill, Inc., third edition, 1995.