Policing Digits

New keys for keeping digital data straight

By IVARS PETERSON

A little bit can make a big difference. That's the danger in transmitting or storing data as strings of ones and zeros — whether encoded in radio waves broadcast by a distant spacecraft or in microscopic pits on the gleaming surface of a compact disk.

Suppose that the sequence 00101011 means "plus." Static picked up by a radio receiver's antenna or a speck of dust marring a disk's surface can readily flip a bit from 0 to 1 or 1 to 0, perhaps changing the original sequence to 00101101, which stands for "minus." Addition suddenly becomes subtraction.

To circumvent such problems, researchers have over the years invented a variety of ingenious strategies to ensure that a computer or some other digital device receiving information can automatically detect and correct random errors. These mathematically based "error-correcting codes" now permit clean sound reproduction even from a scratched or dirty compact disk, accurate storage of data on a computer's hard drive, and reliable data communication at low power over long distances.

"Almost every time digital information is transmitted or stored in a real-world application, some form of error coding is required," says P. Vijay Kumar of the Communication Sciences Institute at the University of Southern California in Los Angeles. It's an invisible but pervasive technology critical to the reliable functioning of many types of information systems.

The trouble is that these error-correcting codes require the addition of extra digits to a message or signal. This additional information allows a received message or block of stored data to be checked for errors, which can then be corrected. But it also reduces the rate at which information can be transferred. And it takes time to encode and decode the raw data.

Hence, users of these codes must balance maximizing the reliability of data transmission against keeping the rate of information transfer as high as possible. They also need to consider the overhead imposed by the time-consuming encoding and decoding operations required before and after data transmission.

Researchers are constantly on the lookout for practical error-correcting codes that are more efficient and compact than those currently in use. Two groups have now uncovered a remarkable, previously unsuspected mathematical link between two types of error-correcting codes originally thought to be quite distinct. This discovery, which allows certain complicated but highly efficient error-correcting codes to be expressed in terms of simpler, easier-to-use types, opens up the possibility of achieving quicker error-free communication.

"It's a very pretty result," remarks Neil J.A. Sloane of AT&T Bell Laboratories in Murray Hill, N.J., who was one of the discoverers. Sloane's Bell Labs colleague A. Robert Calderbank and Patrick Solé of the National Center for Scientific Research (CNRS) in Valbonne, France, share credit for the finding. Kumar and A. Roger Hammons Jr., now at Hughes Aircraft Co. in Canoga Park, Calif., independently worked out the same result.

A person speaking over a crackly telephone line has several options for making himself or herself intelligible at the other end. For example, the person can talk louder or more slowly. If the noise persists or proves excessive, the person can repeat words several times or even spell them out.

The key to digital error detection and correction also lies in redundancy. For example, one can simply repeat each of the digits of a message a certain number of times. Thus, the message 101 could be transmitted as 111000111, and the computer at the receiving end would decode this sequence to get 101. If one bit had changed during transmission, two others would still be correct, and the computer would select the majority digit as the correct entry.

But this method requires the transmission of a large number of extra digits, making it slow and inefficient. Another, more practical approach requires adding a small number of "check" digits to the end of each segment, or block, of a message. This procedure resembles the verbal ploy of saying "a as in alpha, r as in Romeo," and so on, to help a listener correctly identify the letters of a spelled-out word.

For example, suppose a message can be conveyed in blocks of four digits each, ranging from 0000 to 1111. According to an error-correcting scheme known as the Hamming code, the addition of a carefully defined sequence of three digits to each of these blocks makes it possible to detect and correct errors that corrupt a transmitted message.

Upon encoding, 0000 becomes 0000000, 0001 becomes 0001110, 0010 becomes 0010101, and so on, right up to 1111, which becomes 1111111. These extended blocks are known as codewords. Although the pattern in the codeword list may not be evident to the eye, mathematical techniques readily pick it out.

The trick is to create such a strong pattern in the set of codewords that random flips of one or two bits in any block will stand out like a sore digit. In fact, mathematicians and communications experts have developed a number of different mathematical schemes that provide useful sets of strongly patterned codewords.

Consider the message 1000 0101. Encoded, it reads: 1000111 0101101. After transmission, the message may look like this: 1000011 0001101. The first block doesn't exist among the codewords in the Hamming scheme, but it differs in only one place from the codeword 1000111. No other codeword comes this close. So the decoding computer selects this codeword as the most likely possibility and corrects the flawed block accordingly.

Of course, the computer may still make the wrong selection, depending on the severity of the digit scrambling that occurred during transmission. But by choosing the extra digits in each codeword carefully, communications experts can significantly reduce the probability of such cases of mistaken identity.

To be practical, error-correcting codes must be designed so that encoded messages may be easily generated and recognized by simple computer routines. So-called linear codes, which exhibit strong patterns, meet this criterion. Nearly all error-correcting codes now in commercial use are linear.

However, researchers have long known of nonlinear codes that are far more efficient—requiring fewer extra digits per block—than equivalent linear codes. But decoding nonlinear codes has generally proved so cumbersome and complicated it hasn't been worthwhile to use them.

"It has been very hard to encode and decode them," Sloane says. "But they are great codes. In fact, you can prove that they're better than any linear code can possibly be."

A few years ago, Kumar started studying a cellular telephone technology called "code division multiple access" (CDMA), which enables many users to broadcast simultaneously over the same communications channel. He brought to this project both considerable practical expertise and a high degree of mathematical sophistication, a combination not often found among researchers.

In the CDMA approach, signals are kept straight by assigning a separate codeword or sequence of digits — like those used in error-correcting codes — to each user as an identifying tag. Because having a greater number of codewords would permit more users to obtain access to the system, Kumar was interested in finding alternative schemes for producing more codewords.

He and his coworkers succeeded in creating a potentially useful, efficient linear code based not on binary digits (0,1) but on quaternary digits (0,1,2,3). The surprise came when they expressed their new linear quaternary code in binary form. That is, they replaced each 0 with 00, 1 with 01, 2 with 11, and 3 with 10 to get binary strings twice the length of the original quaternary strings.

Kumar and Hammons recognized in the binary strings the Kerdock code, one of the well-known, highly efficient nonlinear codes that had previously proved so difficult to use. Similar relationships linked quaternary sequences with other nonlinear codes.

"You get the [nonlinear] codes in a really simple and beautiful manner," Sloane declares.

Meanwhile, unaware of the work that Kumar and Hammons had done, Sloane, Calderbank, and Solé started a little later and followed a different mathematical track to come to the same conclusion. Calderbank found out about the USC results when he noticed the title of a talk by Kumar and Hammons scheduled for presentation at an information theory symposium. He telephoned Kumar.

"We talked briefly, but it was clear there was a great deal of overlap in our results," Kumar says. The two groups had worked out essentially the same thing, though each had chosen to elucidate different details.

"We decided everyone would be better served by writing one paper rather than two," Calderbank says. The joint paper will appear later this year in the IEEE TRANSACTIONS ON INFORMATION THEORY.

New mathematical work is quickly unraveling the relationship between the novel quaternary codes and familiar binary ones. "After we made the basic observation, we covered a lot of ground very fast," Calderbank says. And this ground-breaking effort has continued at a rapid pace.

"There have been lots of nice spinoffs," Sloane notes.

Sloane, Kumar, Calderbank, and others have followed up intriguing links between these codes and a variety of mathematical topics, including aspects of number theory, group theory, and geometry. The new codes even suggest simple, alternative strategies that statisticians can use to design experiments for testing the validity of hypotheses.

This improved understanding of the mathematical underpinnings of codes also means that several families of highly efficient error-correcting codes could soon be available for general use in communications. The Hughes Network Systems division of Hughes Aircraft, for example, has filed a patent on an application of quaternary codes to cellular telephone systems. One product already incorporates the new technology.

Kumar has heard from colleagues that the nicest thing about the new result is that it provides a very simple answer to a complicated question. "If something is simple, you tend to use it more," he notes. "There's no reason now not to use quaternary codes."

"There are lots of potential applications here, and some very beautiful mathematics," Sloane adds.

Inspired by the new results, several groups are now looking for even better codes, going beyond the quaternary digits to other sets of numbers. They are also looking for improved decoding techniques that take advantage of the newly discovered relationship between binary and quaternary codes.

Indeed, just working out for quaternary codes what is already known about binary codes would be a major undertaking. "Basically, the theory of quaternary codes is wide open," Kumar says.

"I think it's really going to change coding theory," Sloane remarks. And consumers may soon be picking up pocket communicators that use the new codes to keep their airborne messages straight.

"What we hoped for when we wrote the first paper, and what seems to be coming true, is that there's an ... iceberg out there," Calderbank says. "The tip of the iceberg is the binary field, but the rest of the iceberg is interesting, too. There's a lot of it that's under water. We think there's a lot of exploration to do." □

Nabbing errors at the grocery store

All kinds of businesses use identification numbers, from the bar codes on food packages to the long strings of numbers on credit cards and airline tickets. These numbers often incorporate an extra "check" digit as a means of detecting forgery or error.

Typically, the mathematical schemes that underlie the assignment of check digits permit computer detection of incorrectly entered or scanned identification numbers. Unlike error-correcting codes, however, they can't automatically fix the mistake.

Mathematician Joseph A. Gallian of the University of Minnesota at Duluth has studied the various methods used to generate check digits in commercial situations. For example, the last digit of an airline ticket number should equal the remainder left after dividing the rest of the digits by 7. Thus, the ticket number 170004595703 is presumably correct, because dividing 17000459570 by 7 leaves 3 as the remainder.

This is the simplest but least effective method of assigning a check digit, Gallian says. Applied to a garbled version of the original number, this method catches some, but not all, possible errors.

A more complicated scheme used for bar codes detects a larger proportion of errors. For example, a box of Kellogg's Corn Flakes may have the following number: 0 38000 00127 7. Suppose the scanner at the local supermarket reads this number as 0 58000 00127 7. How does the computer connected to the scanner detect the error?

The computer is programmed to add together the digits in positions 1, 3, 5, 7, 9, 11 and triple the result, then add this tally to the sum of the remaining digits. If the result doesn't end with a zero, the computer knows the entered number is incorrect. Try it on the Cheerios bar code displayed on the previous page.

"This simple scheme will detect over 90 percent of all possible errors," Gallian says. Many credit card issuers use a slightly different method, which catches 98 percent of the most common errors. — *I. Peterson*