FROM RESEARCH TO REWARD: A NATIONAL ACADEMY OF SCIENCES SERIES ABOUT SCIENTIFIC DISCOVERY AND HUMAN BENEFIT

Search for a Signal: The Role of Mathematical Codes in Fostering a Cell Phone Communication Revolution

Imagine trying to conduct a conversation with a friend across a busy city street. Not very easy, right? Yet, Americans do something like this a billion times a day,[1] every time they talk on their cell phones. Without realizing it, they are using mathematical codes, called error-correcting codes, to turn staticky radio signals into clear human speech.

Error-correcting codes have a different purpose from secret codes, which are more familiar to people thanks to spy movies and news about cryptography. Secret codes are intended to conceal information: for example, to prevent someone from eavesdropping on your phone call.[2] Error-correcting codes *reveal* information. They help your cell phone scrub unwanted noise from an incoming call. Both operations are essential for a cell phone, but recent years have seen major advances in error-correcting codes, with each new generation of phone incorporating vastly improved coding technology.

Why are error-correcting codes important? The decoding of incoming phone calls is one of the most difficult, power-intensive tasks your cell phone performs.[3] The more efficiently this operation can be done, the less power your phone will require. This means that your phone can travel farther from a cell tower and still get a usable signal.[4] Error-correcting codes enable the phone companies to deliver on their promise of "more bars in more places."

Ironically, today's most widely used error-correcting codes, called low-density parity check (LDPC) codes, were considered impractical when they were discovered. Robert Gallager, a young engineer at the Massachusetts Institute of Technology (MIT), developed LDPC codes in 1960 as a solution to a problem that few people even knew existed: how to transmit a message over a noisy communication line, such as a two-way radio, so the message comes through with near-perfect fidelity.[5] He wasn't expecting his discovery to become part of a multi-billion-dollar

[1] Smith, A. 2001. Trends in Cell Phone Usage and Ownership. Pew Research Center Internet & Technology. http://www.pewinternet.org/2011/04/28/trends-in-cell-phone-usage-and-ownership (accessed May 20, 2019).
[2] The technology used in many U.S. cell phones for concealment is called CDMA (code division multiple access). See, for example, Andrew Viterbi, an oral history conducted in 1999 by David Morton, IEEE History Center, Hoboken, NJ. http://ethw.org/Oral-History:Andrew_Viterbi (accessed May 20, 2019).
[3] Carlton, A. 2016. Surprise! Polar codes are coming in from the cold. *Computerworld*. https://www.computerworld.com/article/3151866/mobile-wireless/surprise-polar-codes-are-coming-in-from-the-cold.html (accessed May 20, 2019).
[4] Raymond Pickholtz, an oral history conducted in 1999 by David Hochfelder, IEEE History Center, Hoboken, NJ. https://ethw.org/Oral-History:Raymond_Pickholtz (accessed May 20, 2019). See especially section 4.11, "Multi-user detection."
[5] Hardesty, L. 2010. Explained: Gallager codes. *MIT News*. http://news.mit.edu/2010/gallager-codes-0121 (accessed May 20, 2019).

business. Back then you would have needed thousands of that era's computers to do the calculations necessary to implement his codes. No one anticipated the day would come when nearly everyone would own a device that couples a two-way radio to sophisticated signal processing and a built-in computer: a cell phone.

Gallager's invention has been called "a bit of 21st-century coding that happened to fall in the 20th century."[6] But it was no accident. It was possible because he worked in a research environment that encouraged him to tackle a fundamental problem, regardless of whether the solution was practical. "I was probably halfway between the ivory tower theorist, interested in only puzzle solving, and the engineer trying to do something useful," says Gallager today. "If the research climate then was anything like it is now, I would certainly have failed. The current emphasis on invention, starting a company, and making a billion would have ruled out the quest for understanding that is the basis of good science."

Gallager's research built on the game-changing work of mathematician and electrical engineer Claude Shannon, often described as the founder of information theory.[7] During World War II, Shannon, a research engineer at Bell Labs, worked on a variety of military projects, but he was bored by them. Late at night, while listening to jazz music in his apartment,[8] he worked on the problem that really captivated him: a theory of communication.

It was Shannon who coined the term "bit" to denote a unit of information (coded as a 1 or a 0).[9] In doing so, Shannon anticipated the digital revolution. The first cell phones were analog radios, but since the second generation (2G) they have been digital.[10] This means that they convert a spoken signal into a string of 1s and 0s and transmit them to the destination phone, which converts the 1s and 0s back to an audio signal. Digital technology is based on the work Shannon did 35 years before cell phones arrived on the market.[11]

During his research, Shannon realized that the conversions back to an audio signal can be messy: some 0s may have changed to 1s and vice versa, leading to noise or static. He wondered if it was possible to identify the mistaken bits and switch them back. Although his way of posing the question was novel, his solution was old: add redundancy.

---

[6] Costello, D., and G. D. Forney. 2007. Channel coding: The road to channel capacity. *Proceedings of the IEEE*,95(6):1150–1177. doi: 10.1109/JPROC.2007.895188.

[7] Collins, G. P. 2002. Claude E. Shannon: Founder of information theory. *Scientific American.* https://www.scientificamerican.com/article/claude-e-shannon-founder (accessed May 20, 2019).

[8] Soni, J., and R. Goodman. 2017. A Man in a hurry: Claude Shannon's New York years. *IEEE Spectrum*. http://spectrum.ieee.org/geek-life/history/a-man-in-a-hurry-claude-shannons-new-york-years (accessed May 20, 2019). (Excerpted from Soni, J., and R. Goodman. 2017. *A Mind at Play: How Claude Shannon Invented the Information Age*. New York: Simon and Schuster.)

[9] Soni, J., and R. Goodman, *op. cit.* Shannon's original paper is Shannon, C. A mathematical theory of communication. *Bell System Technical Journal* 27(3)379–423.

[10] Don Schilling, an oral history conducted in 1999 by David Hochfelder, IEEE History Center, Hoboken, NJ. http://ethw.org/Oral-History:Donald_Schilling (accessed May 20, 2019). Also see Perry, T. 2013. Captain cellular. *IEEE Spectrum* 50(5):52–55.

[11] Collins, G. P., *op. cit.*

People had been using redundancy to clarify messages for years. When soldiers recite the alphabet as Alpha, Bravo, Charlie, etc., they are adding information to help the receiver understand the intended letters. In a quiet room, the receiver could easily hear the difference between "B" and "C." But when yelling over the din of battle, the redundant syllables of "Bravo" and "Charlie" help a lot.

If you are working with bits, one way to add redundancy is to repeat each bit three times, so "1" becomes "111" and "0" becomes "000." If the receiving phone gets the message "101," it assumes that the middle digit "0" is an error and corrects it to a "1." This is more likely to be correct than "000," because it would take two errors in transmission to convert that signal to "101."

This simple scheme is called an error-correcting code. It has two valid code words, "000" and "111"; when any other word is received there has been an error. But this primitive "repeat three times" code is far from the most efficient. It slows every transmission down by a factor of three. Also, it fails to correct some mistakes, on the infrequent occasions when two out of three received bits are wrong.

Shannon realized that you could get more efficient error-correcting codes by combining two ideas. First, you can encode several bits at a time; for example, you might encode "0101" as "1101001." (Note that the code words have to be longer than the input blocks [the "0101"] because you are adding redundancy.) Second, the valid code words, such as "1101001," need to be easily distinguishable from each other; for instance, "1101101" would be too close to "1101001" leading to a high error rate during decoding.

Combining these two ideas, Shannon discovered that every communication channel has a fundamental speed limit: a rate of information transmission that cannot be exceeded without allowing errors to creep in. This speed limit (usually called the "Shannon limit" or the "Shannon capacity" of the channel) depends on two things: the signal-to-noise ratio and the bandwidth of the transmission.[12] Although it was originally expressed in terms of a bit rate, most engineers now think of the Shannon limit in terms of power usage. If your transmitter tries to exceed the data rate dictated by the Shannon limit for a given power, then the message will be received with errors. But if you transmit at a rate below the Shannon limit, and you have a sufficiently clever code, then you can get near error-free transmission.

But what is a "sufficiently clever" code? Shannon himself could not give an example of a near-capacity code. He could only prove that they must exist. But he did identify some clues. The longer the input blocks and code words the better. For instance, LDPC codes for most applications use block lengths of 1,000 to 10,000 bits. He also showed that a good way to make the valid code words distinguishable *on average* is to choose them randomly.

[12] Hardesty, L. 2010. Explained: The Shannon limit. *MIT News*. http://news.mit.edu/2010/explained-shannon-0115 (accessed May 20, 2019). Also see Wikipedia. Shannon–Hartley theorem. https://en.wikipedia.org/wiki/Shannon–Hartley_theorem (accessed May 20, 2019).

Unfortunately, random codes are hard to decipher. The kinds of codes that engineers knew how to decipher were exactly the opposite: highly structured. “The common saying was that all codes were good except the ones we could think of,” says Gallager, who used Shannon’s theories in his own work many years later.

For a long time, even the best codes still required more than twice as much power as a hypothetical code achieving the Shannon limit, and the conventional wisdom was that the limit would never be reached. But in 1993, French engineers Claude Berrou and Alain Glavieux stunned the coding community with a new method called turbo codes that requires only 10 percent more power to transmit than the theoretical minimum.[13] A turbo encoder uses two separate encoders—one encoding the message directly, the other scrambling it first and then encoding. The scrambling provides the dose of randomness that Shannon’s theorem required. On the decoding end, Berrou and Glavieux ingeniously applied a technique called belief propagation, which emulates the “message-passing” behavior used to solve a crossword or Sudoku puzzle.

Three years after Berrou and Glavieux’s groundbreaking work, British scientist David MacKay rediscovered Gallager’s long forgotten work on LDPC codes.[14] LDPC decoders resemble crossword puzzle solvers even more strongly than Berrou and Glavieux’s decoders, because they use multiple processors (instead of just two). (Thus, for an LDPC code with 4,000-bit block lengths and 8,000-bit codewords, you have 12,000 clues.) Like a crossword puzzle solver, the decoders piece together a solution from many separate but overlapping pieces of information. When you solve an across clue, you gain information about the solution to the overlapping down clues, and vice versa. Successful solution therefore depends on a complex process of combining information between the across and down clues. An especially important ingredient is the use of “soft decisions.” For example, you may think that a certain letter is an “A,” but you write it in pencil instead of pen because you’re not sure. The LDPC decoder does this by assigning each bit a probability of being a “0” or “1” and updating the probability as more information is available.[15]

Such complexity requires tremendous computing power. In the 1960s you couldn’t even fit thousands of computer processors into a large warehouse. But by the late 1990s you could fit them onto a small silicon chip. The world was finally ready for Gallager’s LDPC codes.

Thanks to their head start, Berrou and Glavieux’s turbo codes were the predominant error-correcting technique used in 3G phones.[16] But Gallager’s LDPC codes operated even closer to the Shannon limit, and they also worked better for high-fidelity applications like digital video. For all of these reasons, LDPC codes played a much larger role in 4G wireless standards.

---

[13] Costello, D., and G. D. Forney, *op. cit.* Engineers usually talk in terms of decibels, but because this meaning of “decibel” is not widely understood by the public, the decibel measure has been converted to a power ratio.

[14] MacKay, D., and R. Neal. 1996. Near Shannon limit performance of low-density parity-check codes. *Electronics Letters* 23:1645–1646.

[15] For an explanation with more mathematical details, see Shokrollahi, A. 2003. LDPC Codes: An Introduction. https://www.ics.uci.edu/~welling/teaching/ICS279/LPCD.pdf (accessed May 20, 2019).

[16] A. Carlton, *op. cit.*

Of course, technology never stops, and cell phone companies are starting to roll out the next generation of wireless, 5G. Just like its predecessors 3G and 4G, 5G features a new coding technology. In 2009, a Turkish-born, MIT-educated engineer named Erdal Arikan (a former doctoral student of Gallager) designed "polar codes" that also approach the Shannon limit, yet are more structured than turbo or LDPC codes.[17] Like the LDPC codes, they were intended at first as a theoretical project, and even Arikan was surprised when they turned out to be so useful. The mathematical structure means that their efficiency can be calculated and guaranteed. In practical applications polar codes still perform worse than LDPC codes, but they are closing the gap, and 5G phones will use both kinds of codes for different tasks.[18]

Throughout the many generations of mobile communication, one factor has remained constant. Theoretical ideas, often developed well in advance of any practical application, have been crucial to ever-improving communications. Without the basic research done by Shannon, Gallager, and others, millions of Americans might still be searching for a cell phone signal.[19]

This article was written by Dana Mackenzie for *From Research to Reward*, a series produced by the National Academy of Sciences. This and other articles in the series can be found at www.nasonline.org/r2r. The Academy, located in Washington, DC, is a society of distinguished scholars dedicated to the use of science and technology for the public welfare. For more than 150 years, it has provided independent, objective scientific advice to the nation.

[17] Actually, LDPC codes can be either structured or unstructured, but in accordance with Shannon's theorem the less structured ones tend to work better.

[18] Carlton, A. *op. cit.*

[19] In addition to the published articles cited, the author benefited from the following interviews: David MacKay (now deceased), February 2, 2005; Michael Tanner, February 3, 2005; Dan Costello, February 8, 2005, and July 19, 2017; Keith Chugg, February 9, 2005.