Robert M. Gray

The 1974 Origins of VoIP

History often unfolds in unforeseen ways. The story told here provides an example in which developments in digital signal processing (DSP)—speech coding, in particular—had a profound impact on the early development of the ARPANET, the ancestor of the Internet.

I discovered the story while I was preparing a historical talk for the Special Workshop in Maui (SWIM). SWIM was intended to bring together a collection of pioneers in digital speech, especially in the development of linear predictive coding (LPC). Unfortunately for the workshop (but fortunately for me), two of the pioneers, John D. Markel and my brother A.H. "Steen" Gray, Jr., were not able to attend, so I was invited to tell the tale of their contributions to the early development of LPC. My connections were both filial and technical: I had begun working on LPC in the mid-1970s with John and Steen, so I knew many of the contributions and players. As I conducted interviews and tracked down sources for my Maui talk (see the link at the end of the article for more detail, references, and related material), the subject matter expanded from speech and DSP to the first successful attempts to transmit real-time packet speech. The story shows how packet speech, recently rediscovered and made popular as voice over IP (VoIP), was first successfully demonstrated in 1974 on the ARPANET and how the Internet protocol (IP) emerged largely as a result of that effort.

EDITORS' INTRODUCTION

Robert M. Gray was born in November 1943 at North Island Naval Air Station, San Diego, California. He obtained his B.S. and M.S. degrees (1966) from the Massachusetts Institute of Technology and a Ph.D. degree (1969) from the University of Southern California. Since 1969, he has been with Stanford University, where he is the Lucent Technologies Professor of Engineering. Dr. Gray's work focused on digital signal processing for speech coding and recognition, image coding, and segmentation applications. He authored the books *Probability, Random Processes, and Ergodic Properties* (1988) and *Entropy and Information Theory* (1990) and coauthored the books *Random Processes* (1986), *Vector Quantization and Signal Compression* (1992), *Fourier Transforms: An Introduction for Engineers* (1995), *Image Segmentation and Compression Using Hidden Markov Models* (2000), and *Stochastic Image Processing* (2004). Dr. Gray received numerous awards, among which are the IEEE Centennial Medal (1984), IEEE Signal Processing Society's Society Award (1993), IEEE Information Theory Society Golden Jubilee Award for Technological Innovation (1998), IEEE Third Millennium Medal (2000), and a 2002 Presidential Award for Excellence in Science, Mathematics, and Engineering Mentoring (PAESMEM). One of Dr. Gray's happiest professional moments was receiving the latter award at the White House with his family and the former student who organized his nomination. Dr. Gray has had an over three-decade-long collaboration with Lee Davisson and an over one-decade collaboration with Richard Olshen. In all of his collaborators, reliability and wit are the qualities that he appreciates the most. An open and straightforward person, Dr. Gray has a quick-wit and no-nonsense approach that melts barriers and simplifies interactions.

One of Dr. Gray's nonprofessional passions is listening to music: "you can't always get what you want, but if you try sometime, you just might find, you get what you need" sings Mick Jagger, a favorite of our guest. Other passions are gilded age history, maritime history, reading, and hiking.

In the story told next, Dr. Gray brings to light the digital signal processing roots of a modern concept, voice over IP. He uncovers the historical facts and voices the opinion of a speech and network research and industry community that contributed directly to the emergence of successful packet speech transmission. He fills in the details and tells the story modestly as a remote observer. We insist that he share his close relationship with major players in, and proximity to, the events recollected. As always, we would love to hear from you . . .

—Adriana Dumitras and George Moschytz
"DSP History" column editors
adrianad@ieee.org,
moschytz@isi.ee.ethz.ch

HOW IT ALL BEGAN

Two threads of the story began in 1966. In December 1966, Shuzo Saito of NTT and a young Nagoya University doctoral student, Fumitada Itakura, published [1] as a report of the NTT Electrical Communication Laboratory. The report described a statistical approach to speech coding, wherein short segments of speech were modeled using Gaussian autoregressive processes. The basic idea was to form a maximum likelihood selection of the underlying probabilistic model based on observed speech data, where the model was described by regression or linear prediction (LP)

1053-5888/05/$20.00©2005IEEE

coefficients. These coefficients characterize the optimal linear predictor of a data sample given a finite collection of earlier values. The coefficients, combined with voicing and pitch information, were communicated to a receiver to permit local synthesis of an approximation to the original speech. The system was an example of a *vocoder,* in contrast to a *waveform coder,* since the encoder explicitly transmits descriptions of a model (corresponding to the vocal tract parameters) and an excitation rather than attempting to directly reproduce a waveform. This approach to speech vocoding, which would later be named linear predictive coding, became the most significant and widely used low-bit-rate compression scheme for digital speech. A wide literature now exists on linear prediction in a speech context, with acknowledged classics including [2] and [3]. The underlying mathematical theory, however, is decades older. The published history of linear prediction methods specifically applied to speech coding and recognition began with Saito and Itakura's report.

Essentially the same algorithm, but for a different application and with a completely different derivation based on a maximization of a Shannon differential entropy, was presented at a conference in October 1967 by John Burg as a method for spectrum estimation [4]. In November 1967, Bishnu S. Atal and Manfred R. Schroeder applied linear prediction ideas to speech in a waveform speech coder that used LP coefficients to form a prediction residual, which was coded and transmitted along with the LP coefficients [5]. The technique was a form of adaptive predictive coding, which did not involve explicit modeling; it was a waveform coder and not a vocoder. In November 1969, Atal presented an LPC speech coder at the Annual Meeting of the Acoustical Society of America, and an abstract was published in 1970 [8]. The widely read complete paper was published with Hanauer in 1971 and gave LPC its name [9]. Today, Itakura and Atal are universally recognized as the "fathers" of LPC speech because of their independent early contributions of linear prediction methods to speech processing and their subsequent outstanding contributions to its development.

Returning to 1966, another thread of the story began at the University of California at Santa Barbara (UCSB), where Glen Culler introduced his On-Line system. The system allowed real-time signal processing at individual student terminals, arguably the first real-time DSP in a classroom. For the story at hand, however, the key point is that Culler was becoming renowned for building fast and effective computer systems. In 1968, Culler joined ARPANET pioneers Elmer Shapiro, Lenny Kleinrock, and Larry Roberts to complete the specification of the Interface Message Processor (IMP), which was the basic "node" of the ARPANET. Also in 1968, John Markel began Ph.D. work at UCSB and took a job at the Speech Communications Research Lab (SCRL). Markel allegedly moved from Arizona to California to escape taking a Ph.D. language exam in French and took a Fortran exam at UCSB to fulfill the requirement. In Santa Barbara, Markel also turned his considerable z-transform and programming skills toward developing software implementations of the Itakura and Saito LPC algorithms. In August 1968, Burg presented a method (now called the Burg algorithm) that found an equivalent set of LP parameters, i.e., the reflection coefficients [6].

ARPANET GROWS

In January 1969, Bolt Beranek & Newman received a contract sponsored by the Advanced Research Projects Agency (ARPA) to build the first four IMPs. Also in 1969, Culler cofounded Culler-Harrison, Inc., which developed some of the best early array processors. His pioneering contributions to computing and networking were recognized years later when he received the National Medal of Technology from President Clinton in 1999. Thanks largely to Culler, in 1969 UCSB became the third node on the ARPANET, joining the University of California in Los Angeles (node one), the Stanford Research Institute (node two), and the University of Utah (node four). Every ARPANET node had different equipment and software; this fact not only set the pattern for the future but also focused the early work on interface and standardization issues. In July of the same year, Itakura and Saito applied ideas from the classical statistical literature on partial correlations to develop the partial correlation (PARCOR) variation on the autocorrelation method [7]. The idea and implementation were similar to those of Burg's algorithm (reflection coefficients and partial correlation coefficients are the same except for their sign), but PARCOR had lower computational complexity.

The first efforts toward developing packet speech transmission on the ARPANET were initiated in 1972 by Bob Kahn (shortly after his joining ARPA) along with Jim Forgie of Lincoln Labs and Dave Walden of Bolt Beranek & Newman. Early experiments to simulate the transmission of portions of digital speech at 64 kb/s indicated that a major change in packet handling was required, along with serious data compression of the speech. Kahn, speaking to his friend Danny Cohen, who was then working at Harvard on real-time visual flight simulation, described an ARPA project that he was supporting at the Information Sciences Institute (ISI) of the University of Southern California in Marina del Rey to bring real-time speech and video to the ARPANET. In 1973, Cohen moved to ISI, where he worked with Steve Casner, Randy Cole (and others), and SCRL on real-time operating systems and eventually real-time signal processing of both speech and video.

To explore the possibilities of packet speech on the ARPANET, Kahn formed the Network Secure Communications (NSC) group and became its "éminence grise." The acronym NSC arose because of ARPA's interest in supporting encrypted speech over the net, but it soon changed meaning to Network Speech Compression. Eventually, it became known unofficially as the Network Skiing Club, reflecting the fact that winter meetings were often held at Alta, Utah. The original NSC members were the Information Sciences Institute, the University of Utah, Bolt

Beranek & Newman, MIT Lincoln Laboratory, and the Stanford Research Institute. This group was soon joined by SCRL and Culler-Harrison, Inc. In addition to the core group, many other institutions participated in occasional meetings, including Texas Instruments, the U.S. Naval Research Lab, Harris, Inc., the National Security Agency, and Bell Telephone Laboratories.

Many speech compression techniques were studied in the NSC project. As Cohen became more involved, he learned about LPC as a promising candidate for speech compression from his SCRL colleagues. The version of LPC ultimately chosen for use in the NSC was the software implementation by Markel and Gray at SCRL and UCSB. At that time, Markel and Gray, along with Hisashi Wakita, were publishing reports and papers on the Itakura and Saito algorithms and providing Fortran code for the algorithms and associated signal processing, pitch detection and coding, and parameter conversion. The software development for the packet implementation was divided among the Information Sciences Institute, Bolt Beranek & Newman, Lincoln Lab, and the Stanford Research Institute.

NEW DEVELOPMENTS: TCP AND NVP

In 1974, two major developments in the history of the Internet occurred. The first was the specification of the Transmission Control Protocol (TCP) by Bob Kahn and Vint Cerf. The second was the development and description of the Network Voice Protocol (NVP) by Danny Cohen and his colleagues, who spelled out the details of how real-time speech could be communicated on the ARPANET. The NVP used only the basic ARPANET message headers and not the TCP protocol because Cohen had realized early on that the packet and reliability constraints of the available TCP implementation made it unsuitable for real-time communication. He had also argued, in his discussions with Vint Cerf in early 1974, that the extraction from the original TCP of a simpler protocol more amenable to real-time processing was required. Cohen characterized the difference between real-time traffic and reliable data transmission as the difference between milk and wine: you had to deliver the milk quickly before it spoiled even if you spilled some on the way, but you could deliver wine a lot more slowly.

In August 1974, the NVP was successfully tested using continuous variable slope delta (CVSD) modulation at 16 kb/s in real-time but with poor quality of speech between the Information Sciences Institute and Lincoln Lab. In December of the same year, the first real-time, documented experiment of two-way LPC packet speech communication took place at 3.5 kb/s over the ARPANET between Culler-Harrison, Inc., and Lincoln Lab using the basic Markel and Gray LPC algorithms coupled with NVP. LPC packet speech conferencing followed in 1976 between Culler-Harrison, Inc., the Information Sciences Institute, and Lincoln Lab at 3.5 kb/s. (A video of one of these 1976 conferences, produced by the Information Sciences Institute participants and converted from film to MPEG video by Billy Brackenridge of Microsoft, Inc., may be found following the link at the end of this article. The video demonstrates the quality of the speech and provides a fascinating peek at the computers, clothes, and haircuts of the time.) The NVP was formally published in 1976 [10] with Cohen as lead author. The following quotation testifies to both the variety of hardware and the remarkable friendly cooperation among the participants:

> The Network Voice Protocol (NVP), implemented first in December 1973, and has been in use since then for local and transnet real-time voice communication over the ARPANET at the following sites:
>
> - Information Sciences Institute, for LPC and CVSD, with a PDP-11/45 and an SPS-41
> - Lincoln Laboratory, for LPC and CVSD, with a TX2 and the Lincoln FDP, and with a PDP-11/45 and the LDVT
> - Culler-Harrison, Inc., for LPC, with the Culler-Harrison MP32A and AP-90
> - Stanford Research Institute, for LPC, with a PDP-11/40 and an SPS-41.
>
> The NVP's success in bridging the differences between the above systems is due mainly to the cooperation of many people in the ARPA-NSC community, including Jim Forgie (Lincoln Laboratory), Mike McCammon (Culler-Harrison), Steve Casner (Information Sciences Institute), and Paul Raveling (Information Sciences Institute), who participated heavily in the definition of the control protocol; and John Markel (Speech Communications Research Laboratory), John Makhoul (Bolt Beranek & Newman, Inc.) and Randy Cole (Information Sciences Institute), who participated in the definition of the data protocol. Many other people have contributed to the NVP-based effort, in both software and hardware support.

In April 1977, James Flanagan from Bell Laboratories, Inc. applied for a patent for "packet transmission of speech" and U.S. Patent 4,100,377 was subsequently granted in 1978. The granting of this patent was perceived by members of the NSC as being inconsistent with the open development process that led to the successful transmission of packet speech over the ARPANET. However, the patent office did grant this patent, and its term of applicability has long since expired; thus, its role in the development of packet speech (as noted in this article) was not significant over the long run.

IP AND TCP SEPARATE

In August 1977, Cohen, Cerf, and Internet legend Jon Postel agreed to explicitly separate IP from TCP to allow for real-time applications. As the first step, they created the User Datagram Protocol (UDP). The separation of IP and TCP became official in January 1978 in TCP version 3, and it stabilized in TCP/IP version 4, which is still in use today. The obvious irony is that the current popular view is of VoIP as novel, when in fact IP was specifically designed to handle real-

time signal transmission such as that of digital speech!

Perhaps the NSC project is best described in hindsight by Randy Cole: "It's hard to overstate the influence that the NSC work had on networking . . . the NSC effort was the first real exploration into packet-switched media, and we all know the effect that's having on our lives 30 years later."

Acknowledgments

Many thanks to J.D. Markel, A.H. "Steen" Gray, Jr., John Burg, Charlie Davis, Larry Rabiner, Mike McCammon, Danny Cohen, Steve Casner, Richard Wiggins, Vishu Viswanathan, Jim Murphy, Cliff Weinstein, Joseph P. Campbell, Randy Cole, Rich Dean, Andreu Veà, Vint Cerf, and Bob Kahn for conversations, interviews, oral histories, and documents. Thanks to Adriana Dumitras, George Moschytz, John Gill, and Michael Godfrey for editorial suggestions and corrections.

REFERENCES

[1] S. Saito and F. Itakura, "The theoretical consideration of statistically optimum methods for speech spectral density," Electrical Communication Laboratory, NTT, Tokyo, Rep. 3107, Dec. 1966.

[2] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, Apr. 1975.

[3] J.D. Markel and A.H. Gray, Jr., *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.

[4] J.P. Burg, "Maximum entropy spectral analysis," presented at the 37th Meeting of the Society of Exploration Geophysisists, Oklahoma City, OK, Oct. 1967.

[5] B.S. Atal and M.R. Schroeder, "Predictive coding of speech signals," in *Proc. 1967 AFCRL/IEEE Conf. Speech Communication and Processing*, Cambridge, MS, 1967, pp. 360–361.

[6] J.P. Burg, "A new analysis technique for time series data," presented at the NATO Advanced Study Institute on Signal Processing with Emphasis on Underwater Acoustics, Enschede, The Netherlands, Aug. 1968.

[7] F. Itakura and S. Saito, "Analysis synthesis telephony based on the partial autocorrelation coefficient," in *Proc. Acoust. Soc. of Japan Meeting*, Jul. 1969.

[8] B.S. Atal, "Speech analysis and synthesis by linear prediction of the speech wave," presented at the 78th Meeting of the Acoustical Society of America, San Diego, Nov. 1969. (Abstract in *J. Acoust. Soc. Am.*, vol. 47, p. 65, 1970.)

[9] B.S. Atal and S.J. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, vol. 50, pp. 637–655, Aug. 1971.

[10] D. Cohen, "Specifications for the Network Voice Protocol," USC/Information Sciences Institute, ISI/RR-75–39, Mar. 1976.

For more detail, references and related material see http://ee.stanford.edu/~gray/lpcip/. SP