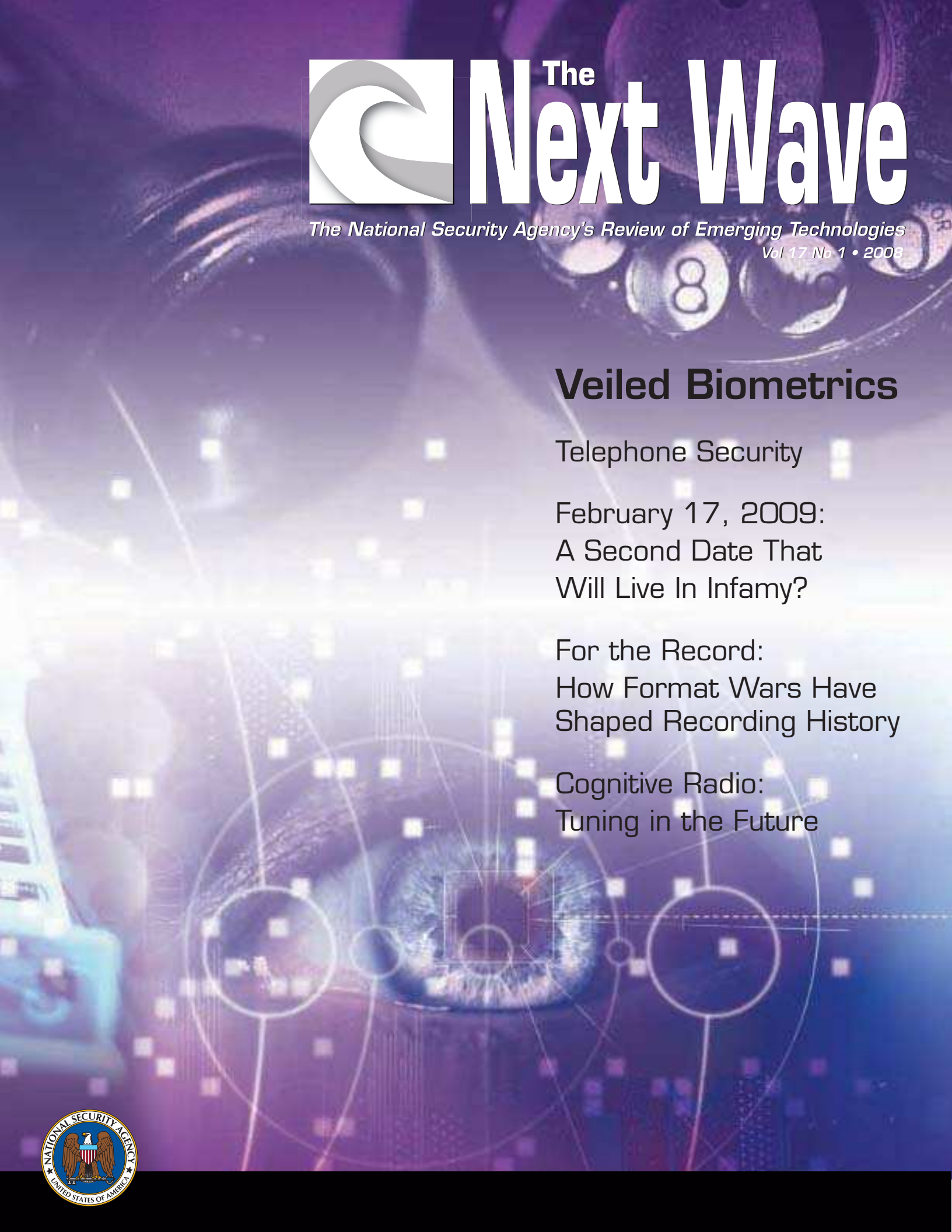


The Next Wave

The National Security Agency's Review of Emerging Technologies
Vol 17 No 1 • 2008

Veiled Biometrics

Telephone Security

February 17, 2009:
A Second Date That
Will Live In Infamy?

For the Record:
How Format Wars Have
Shaped Recording History

Cognitive Radio:
Tuning in the Future

The Next Wave

Letter from the Editor

For the first issue of 2008, The Next Wave (TNW) presents a selection of articles covering a wide range of technologies from storage to telephony to digital broadcasting.

Within this issue are two constants: security and change. Security is the theme of our two feature articles. These two articles show how we adapt security in an ever changing world. Our first feature article, Veiled Biometrics, describes the techniques of fuzzy vaults and fuzzy extractors, both used to make the storage and retrieval of biometric data more secure and reliable. Our second feature comes from the Center for Cryptologic History and spans telephone secrecy through the development of SIGSALY, the secure telephone terminal developed during World War II.

Our three focus articles all concentrate on the digital revolution and some of the changes it brings. We introduce three technologies that are in different stages of adoption: The first is an overview of cognitive radio—while still in the research stage, it is posed to dramatically change current radio operations. Next is an article on the ongoing struggle for supremacy in the DVD storage wars—three competing technologies all vying to be the market winner. Finally, we have an article that briefly discusses the upcoming demise of analog TV in 2009—a legacy technology making way for digital technology.

We hope you enjoy the diversity of technologies in this first issue of 2008. Also, we hope that our presentation of these articles helps you with technological changes you encounter.

The Next Wave is published to disseminate significant technical advancements and research activities in telecommunications and information technologies. Mentions of company names or commercial products do not imply endorsement by the U.S. Government. Articles present views of the authors and not necessarily those of NSA or the TNW staff.

For more information, please contact us at
TNW@tycho.ncsc.mil

CONTENTS

FEATURES

FOCUS

Veiled Biometrics

Who are you? How do you prove your identity if it is challenged? The process of providing proof of one's identity is known as authentication. Although there are numerous ways to authenticate oneself, they all fall into three categories: what you have, what you know, and what you are. Table 1 below provides examples of factors that fall into each of the three categories. One of these factors alone, or a combination of factors, can serve as establishment of one's identity.

The authentication process involves two stages: enrollment and verification. The enrollment phase entails collecting and storing information related to authentication. For example, suppose that you wish to access your bank account from an ATM. To do so, you first must enroll in the ATM system. In order to enroll, you provide the bank with a PIN and the numbers from the magnetic stripe on your ATM card are recorded. This data is stored, hopefully securely, and made available for retrieval at the time of verification. During the verification stage, you will be required to reproduce your authentication information, in this case an ATM card and PIN. The system will then compare it with the data stored from the enrollment process, and determine if a match has occurred. If the match is not exact, access to your account will be denied.

In the ATM example, the forms of authentication used, i.e., the ATM card and PIN, are examples of what you have and what you know, respectively. The question arises whether these factors truly provide proof of identity. Couldn't someone with your ATM card and PIN impersonate you and gain access to your account? It was in answer to these types of questions that the use of biometrics, or what you are, was first proposed for authentication purposes.

Biometrics, literally "to measure life", is the science of developing methods of effectively measuring and analyzing an individual's biological (and behavioral) characteristics for identification and/or authentication purposes. In the developing science of biometrics, biological characteristics that have been, and continue to be, explored include fingerprints, irises, retinas, hand geometry, vascular patterns, and facial characteristics.

The concept and practice of identifying and/or authenticating an individual's identity based upon inherent biological features has been around for centuries. In 500 B.C., there is evidence that the Babylonians recorded their business transactions on clay tablets that included the involved parties' fingerprints. Beginning in the 14th century, Chinese merchants stamped the fingerprints and footprints of children and used them for identification purposes. In the 19th century, via the direction of Richard Edward Henry of Scotland Yard, police began the practice of fingerprinting criminals for identification purposes. In 1936, ophthalmologist Frank Burch formally proposed identifying individuals based upon their iris patterns.

As biometrics technology and methodology advances, scientists are able to more effectively extract various biological characteristics from an individual for use in identification and/or authentication algorithms. Moreover, as computing systems grow more powerful and plentiful, biometrics recognition systems can be im-

Form	Example
What we know	Passwords, PINs, pass phrase, mother's maiden name
What we have	Tokens, CAC card, RSA token, Smart card
What we are	Biometrics: fingerprint, iris, speech pattern, walking gait

Table 1: Forms of Authentication

plemented in an efficient and cost effective manner. As such, systems relying upon biometric identification and/or authentication algorithms are being employed for use in both the business and security sectors. For example, allowing an individual to pay for their groceries via the presentation of their fingerprints (thereby accessing their financial information) or allowing an individual to bypass airport security checkpoints via presenting their iris in order to verify their identity and criminal history.

As the use of biometric identification/authentication algorithms in various systems proliferates, so does the concern for the security and privacy of an individual's biometric information. The biometric data that is collected, transported, and stored must be secured during all parts of the authentication process. A compromise of this data can have a great impact on the person since a body part cannot be revoked.

Generally, this biometric information is stored in the clear in the form of a template, a conglomeration of feature vectors, within a central repository or database. In order to protect an individual's stored biometric information, it is necessary and desirable to develop a method of securing the biometric data in a veiled manner while being able to efficiently and effectively utilize and access it for identification/authentication purposes. As such, an individual's biometric information would be transformed and stored in a non-recoverable form, while still allowing for the successful identification/authentication of the individual upon their presentation of suitable biometric information.

This concept is similar to many password authentication systems: In order to authenticate the user, the system requires that the user input a password. This password is then hashed and the hash is compared with a database consisting of the hashes of the passwords of authorized users. If a match is found, then the user is identified as an authorized user. Thus, ideally, a sort of hashing approach would be suitable for biometric data. However, biometric information is generally extremely noisy, or "fuzzy", data (both a product of the individual and of the biometric processing technology employed). Therefore, simply hashing an individual's biometric information in the same manner as one would hash a password is ineffective: It would not produce a repeatable result. Thus, merging biometrics with cryptography is a tricky proposition (see Table 2).

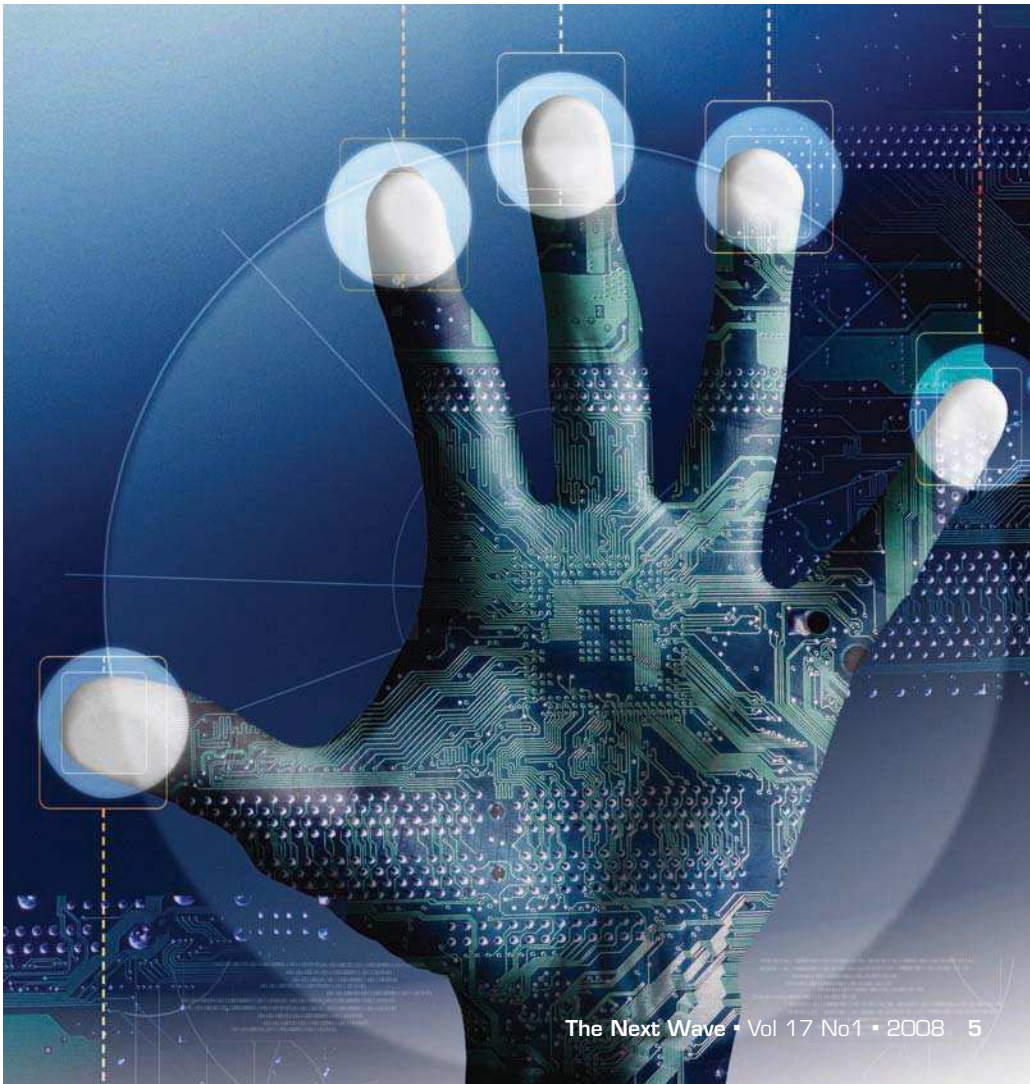
Biometrics	Cryptography
Non-uniform in structure	Uniform in structure
Generated through biometric reading	Randomly generated
Variable data length	Consistent fixed key length
Fuzzy – not exactly reproducible	Exactly reproducible
What you are	What you have

Table 2: Property Comparison of Biometrics and Crytography

Where hashing a user's biometric information fails in authentication/identification systems, veiled biometrics is able to succeed. Veiled Biometrics is a powerful concept, from which several algorithms have been developed, involving transforming a user's biometric information into a secure form that allows for repeatedly successful identification/authentication. Research into this area has been advanced within the National Information Assurance Research Laboratory (NIARL). Authentication researchers, teaming with mathematicians, have investigated and developed a proof of concept for securing data such as biometrics that can be revoked. Two research papers [1] and [2] have been used as a basis for the work done in this area. They both address securing noisy or fuzzy data.

Fuzzy Vaults and Fuzzy Extractors

Two specific types of tools that may be employed in developing Veiled Biometric algorithms, namely, Fuzzy Vaults and Fuzzy Extractors, are described in this sec-



tion. Both of these tools allow the fundamental requirements of a Veiled Biometric algorithm to be satisfied: They facilitate the secure storage of the users' biometric data and they have the ability to cope with the fact that each new reading that the biometric system takes will be different from the last.

Both Fuzzy Vaults and Fuzzy Extractors make use of error-correcting codes. The reader should note that this is a novel use of error correction. Traditionally, error-correcting codes are used to ensure accurate data transmission over a noisy channel. However, in the context of Veiled Biometrics, error-correcting codes are employed in order to "correct" errors that occur between biometric readings.

FUZZY VAULTS

The concept of a Fuzzy Vault was first introduced by Juels and Sudan in 2002 [1]. The premise of their Fuzzy Vault construction is to allow a set of fuzzy data (in this case, the biometric reading) to serve as a "locking set" for a secret. Only the user with the correct "key" can "unlock" the Vault and retrieve the secret. The "key" is allowed to be fuzzy in that any key which is "very close" to the correct key will successfully unlock the Vault. This simple fuzzy locking and unlocking idea is the essence of the Fuzzy Vault construction.

The Fuzzy Vault works as follows: Upon enrollment into a system, a reading B is taken of the user's biometric data. The data from this reading is used to form the locking set L, which is used in turn to lock some secret s. The set L is then stored on the server, sorted together with enough random noise that L becomes indistinguishable from the noise. After enrollment, the user can attempt to gain access to the system by presenting another biometric reading, from which an unlocking set U will be created. If this reading is very similar to the original B, the error-correcting code will allow the set U to unlock the Vault and recover the secret s. Recovery of s allows the user to gain access to the system.

It is important to note two things about the Fuzzy Vault. First, to an intruder who gains access to the server, biometric data is indistinguishable from random noise. Second, the biometric reading of an unauthorized person attempting to gain access to the system will be very different from the original B that was used to enroll. Thus, the error-correcting code will be unable to construct an unlocking set and the secret s will remain locked inside the Vault.

FUZZY EXTRACTORS

The idea of a Fuzzy Extractor was proposed in 2004 by Dodis and Reyzin [2]. A Fuzzy Extractor can be viewed as a "fuzzy" hashing algorithm, in that similar inputs will hash to the same value. Upon enrollment, a reading B is taken of the user's biometric data. A hash H(B) is stored on the server, along with a helper string h. The helper string h holds a small clue to what the original B was. After enrollment, the user can attempt to gain access to the system by presenting a new biometric reading C. If C is very close to B, the error-correcting code will be able to use the information in h to correct the errors in C and reproduce B. Therefore, H(B) can be computed and matched with what is stored on the server and the user is authenticated.

Once again, there are two important points to note. First, although h leaks some amount of information about B, this amount is so small that it would still be computationally infeasible to recover B given h. Second, the biometric reading of an intruder attempting to gain access to the system will be so different from B that the error-correcting code will not be able to correct enough errors. The error-correcting code will use h to correct as many errors as possible and will produce B', its best attempt at reproducing B. However, in this case, H(B') will not match H(B) and the intruder will be unable to authenticate

Securing Fingerprint Authentication

In the two preceding sections, we described a set of mathematical algorithms and principles that can be applied to any biometric technology. Biometric technologies vary widely in how they represent and compare readings. Because of this diversity, each biometric presents unique challenges to the developer of an authentication system.

The fingerprint is one of the oldest and most widely accepted biometrics and, thus, has been one of the first to be researched for use in veiled biometric systems. The mathematical algorithms employed in authentication systems, whether the Fuzzy Vault, Fuzzy Extractor, or otherwise, all require fingerprint readings to be converted to an abstract mathematical format. This section describes how to accomplish this goal.

Traditional Fingerprint Representation

In any fingerprint system, whether for authentication or forensic use, we can think of the fingerprint as undergoing a series of transformations. Each transformation converts the data to a form that is simpler to use, but also causes the loss of some information. Consider, for example, a fingerprint recovered at a crime scene. The fingerprint data begins as the actual ridges and valleys on the individual's finger. This pattern of ridges is left on a surface at the crime scene in the form of a latent fingerprint, containing some of the information from the original finger. Crime scene investigators then dust and photograph the

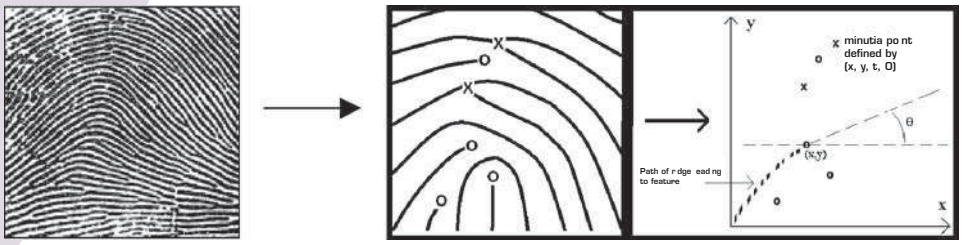


Figure 1 – Transforming a fingerprint image and locating minutiae

latent print to produce a representation of the fingerprint that they can use in further analysis.

A similar result is achieved by a digital fingerprint scanner, which captures an image of the fingerprint electronically. The quality of the data will vary greatly between the two schemes, but both cases produce a simple 2-dimensional image of the fingerprint ridge structure. This representation might be sufficient, if it was only necessary to visually compare a handful of fingerprints. However, in both law-enforcement and digital authentication, it is desirable to automate the comparison process and have a level of accuracy that transcends visual comparison. Both of these goals are achieved through a process called "feature extraction."

Feature extraction is the process where complex data is reduced to a simpler set of key identifying characteristics. For fingerprints, this has traditionally involved extracting the fingerprint's "minutiae" from a fingerprint image.

Fingerprint minutiae are points on the print that describe the key features of the ridge pattern. The two main types of minutiae are the "ridge ending" (where a ridge terminates) and the "ridge bifurcation" (where a ridge splits into two ridges). The set of these minutiae for a given fingerprint is sufficient to accurately identify and compare fingerprints. Fingerprint minutiae have been used successfully for many years in both law enforcement and digital authentication. Minutiae can be described with varying levels of complexity, but the most important qualities of a minutia point can be described by the following four quantities.

1. x-coordinate (horizontal position);
2. y-coordinate (vertical position);
3. Angle (angular direction of the fingerprint ridge);
4. Minutia type (ridge ending, ridge bifurcation, etc.).

Alignment-Free Representation

Comparison is the cornerstone of biometric systems. Any usable fingerprint

FUZZY VAULT DETAILS:

Enrollment:

Step 1

Obtain an initial reading B of the user's biometric data. For this application, it is assumed that B can be represented as a set of features, each of which is an n-long bit string that has been naturally identified with an element of $GF(2^n)$. That is, $B= \{b_0,b_1,\ldots,b_n\}$, where $b_i \in GF(2^n)$.

Step 2

Generate a random secret $s=\{a_0, a_1,\ldots a_k\}$ with $a_i \in GF(2^n)$, and embed it into the coefficients of a polynomial $p(x)=a_0 + a_1x +\ldots+ a_kx^k \in GF(2^n)[x]$. Note that k is a carefully chosen parameter that balances error tolerance and security.

Step 3

Form the locking set, $L=\{(x_i,y_i) \mid x_i=b_i \ \& \ y_i = p(x_i)\}$, and store this on the server sorted together with random noise of the form (x_j , y_j), where $x_j \notin B$ and $y_j \neq p(x_j)$. Since the polynomial p is secret, the points from L that correspond to actual biometric data are indistinguishable from the noise. Also, store a hash of s, $H(s)$, for verification.

Verification:

Step 1

The user presents a new biometric reading $B' = \{\beta_0,\beta_1,\ldots,\beta_m\}$. The new set of features, B', is used to form an unlocking set U as follows: For each $\beta_j \in B'$, if $\beta_j = x_j$ for some (x_j,y_j) on the server, include the point (x_j,y_j) in U. The set U now consists of some points that are on the polynomial p and some points that are part of the random noise, but the two types of points are indistinguishable.

Step 2

Use Reed-Solomon decoding to attempt to reconstruct p from the points in U. If, and only if, B' is very similar to B, most of the points in U will be points on p and the decoding will be successful. Otherwise,the decoding process will produce some polynomial $q \neq p$.

Step 3

If the secret polynomial p (and hence the secret s) was successfully recovered in Step 2, then $H(s)$ is computed correctly and the user is authenticated. Otherwise, access is denied.

FUZZY EXTRACTOR DETAILS:

Enrollment:

Step 1

Obtain an initial reading B of the user's biometric data. In this application, it is assumed that that B can be represented as an n-long bit string.

Step 2

Construct a (n,k,d) BCH code, where k (and hence d) is carefully chosen to balance error tolerance and security, and compute the syndrome $syn(B)$. We will use $syn(B)$ as the helper string h.

Step 3

Store $H(B \oplus x)$ on the server, where H is some hash function and x is an n-long bit string chosen uniformly at random. The use of x ensures that the input into the hash function is uniform. Also store x and h on the server.

Verification:

Step 1

The user presents a new biometric reading B'.

Step 2

Compute $h \oplus syn(B') = syn(B) \oplus syn(B')$ and use BCH decoding to obtain the corresponding error vector v. Note that if, and only if, B' is very similar to B, the BCH decoding will work correctly and yield $B' \oplus v = B$. Otherwise, the decoding process will yield $B' \oplus v = C \neq B$.

Step 3

Compute $H(B' \oplus v \oplus x)$. If this matches the hash value $H(B \oplus x)$ stored on the server, the user is authenticated. Otherwise, access is denied.

system must provide a way to efficiently compare one fingerprint to another and return a useful answer. In a minutiae-based system, comparison is straightforward, provided that the fingerprints being compared can first be aligned. Because a fingerprint will never be read in exactly the same orientation (the position of the finger on the scanner will be slightly different from reading to reading), the position of the minutiae from two readings must be aligned before they can be compared for similarity.

In a conventional biometric authentication system, alignment is possible because the minutiae for the enrolled fingerprint are stored in the clear. However, in a veiled biometric system the enrolled fingerprint is stored with added "noise" for the sake of security; this makes alignment infeasible. It is therefore necessary to represent the fingerprint data in a format that does not need to be aligned. One way to achieve an alignment-free representation for fingerprints is to use minutiae pairs. Instead of considering single minutiae points, points are considered in pairs, with each pair being described by the relationship of the two points to each other. Angles and position are no longer measured with respect to the (x,y) plane, but for each point they are measured with respect to the other point. With this scheme, each pair of points can be described by the following five quantities.

1. **Distance between the two points;**
2. **Adjusted angle of the first point;**
3. **Adjusted angle of the second point;**
4. **Type of the first point;**
5. **Type of the second point.**

Because these values do not depend on the orientation of the fingerprint when it is scanned, the "minutiae pair" representation does not require alignment. The measurements for distance and angles must then be rounded for the sake of error tolerance. The result is a set of pairs that represent the fingerprint and can be compared efficiently with future readings without the need to align the two templates. This result provides a representation of the data in a mathematical format that can be used by Veiled Biometric algorithms.

Beyond Biometrics

The concept of an error-tolerant key extends to fields other than biometrics. In fact, any data set that is "fuzzy" can be used as input to a Fuzzy Vault or Fuzzy Extractor. One example of such a data set is described below.

A current project under development involves adding tamper detection capabilities to a circuit board. This board contains a series of emitters and detectors, and is covered by a diffractive coating. The goal is to secure a key based upon the unpredictable but reliable effect that this coating has upon the detectors.

To create an enrollment template, each emitter is activated in succession, and the detector readings are recorded. This yields a very reliable data structure. Although the values of each successive reading will be different, the structure of the data obtained will be the same. Thus, we can use this enrollment data to construct a Fuzzy Extractor.

As mentioned earlier, a Fuzzy Extractor requires two components to be stored. The first is some minimal amount of information about the enrollment template, which allows a presented template to be corrected so long as it is 'close enough' to the original. The second component is a hash of the enrollment template itself, which can also be thought of as part of a secure key.

For the tamper detection system, the readings of the detectors will vary somewhat with temperature and voltage. However, with the proper choice of an underlying Error Correcting Code, a Fuzzy Extractor will allow the enrollment template to be recovered even if small errors occur in every reading. Once the template has been reconstructed, it is a simple process to compute the key.

However, if anything has damaged the protective coating on the board, then the diffraction properties will change, resulting in significantly different readings. This difference will in turn exceed the error correction capacity of the Extractor, an incorrect template will result, and the key' obtained will be worthless.

Thus, while Fuzzy storage algorithms may have been created to support biometrics, they can readily be extended to any application that needs to recover an exact value from volatile data.

References

[1] Juels, A. & Sudan, M. (2002) A fuzzy vault scheme. IEEE International Symposium on Information Theory, 408.

[2] Dodis, Y., Reyzin, L. & Smith, A. (2004) Fuzzy extractors: how to generate strong keys from biometrics and other noisy data. Lecture Notes in Computer Science 3027: Advances in Cryptology, 523-540.

Telephone Security

This article presents a history of significant milestones in the development and deployment of high-level telephone security for the US and its UK allies during World War II. As a backdrop it briefly covers the methods to provide telephone privacy from shortly after the telephone was invented continuing through the introduction of the radiotelephone. Using these as precursors, what follows are the measures the US government took to assure telephone secrecy. The challenges enciphered telephony (ciphony) has placed on science and engineering are emphasized over operational history, which is better dealt with by others referenced herein.

When telecommunications came on the scene in the 1840s, the "dots" and "dashes" of Morse messages, though readily adaptable to confidentiality, were mainly coded for brevity. Before filing at the telegraph office, commercial users coded their messages privately, not much different than for sensitive mail. During the Civil War both the Union and Confederate armies used telegraphy as the prime source of command and control. Nomenclator tables, a combination of codes and ciphers, were the dominant method to provide message confidentiality.

Operational US telephone privacy, with the exception of jargon codes, had to wait over fifty years after Alexander Graham Bell first transmitted speech electrically (1876). Unlike the telegraph, which could be encrypted offline by either manual or mechanical methods, telephone scram-

bling had to be done in real time. Therefore, in its early days the telephone operating company had no other technical option but to transmit in the clear. Customers accepted the minimal risks from wiretappers or operator monitoring. However, to thwart eavesdropping on their overseas radiotelephone circuits, AT&T introduced telephone privacy in the nineteen twenties by frequency transposition. Operationally effective in its time, it offered only technical challenges to all but the most concerted interloper, a far cry from the online cryptographic security available for telegraphy circa 1920. At the onset of World War II, telephone secrecy became a high priority "cost be damned program" drawing attention at the presidential level. It remained, however, beyond the realm of economic reality until the Internet.

This article covers the major milestones of strategic telephone secrecy from its World War II genesis at the Bell Telephone Laboratories (BTL).

Bell Telephone Labs: The Crucible of Telephone Secrecy

Communication security has been a major concern of governments since time immemorial. The advent of telecommunications raised the specter within both government and the private sector of how to protect signals outside the control of the parties involved. The Bell Telephone Laboratories pioneered research in U.S. communication innovation. Telephony security could vary from jargon codes, physical security of the medium (e.g., protected distribution systems), to noise masking or cryptography (transposition or substitution under the control of a code or a key). Except for jargon codes, Bell Lab engineers filed for patents on the others starting before 1920.

A patent for noise masking was filed in 1919 by R.D. Parker, which claimed that "superimposing...a current of continuously varying frequency" derived from a phonograph record on the speech was a means of insuring secrecy. The recipient subtracted a synchronized replica of the masking noise thereby recovering the speech. This was a novel idea, but uncorrectable distortion over wireline or radio media made it operationally impractical at that time. BTL engineers continued to experiment with scrambling analog speech in the frequency and/or in the time domain to provide radiotelephone privacy. In the 1920s AT&T introduced the A-3 system to deny the casual listener intelligible speech. The A-3 "diced" the speech spectrum into five bands transposing and inverting them (of the 3,600 possible combinations, only six were operationally usable).

Breakthrough

Shortly before the Japanese attack on Pearl Harbor, President Roosevelt established the National Defense Research Committee (NDRC). Chaired by Vannevar Bush of Massachusetts Institute of Technology (MIT), it was premised on civilian control of military research. Bush brought together 6,000 of America's brightest academics and private sector engineers and scientists to promote and organize military research. One group in the NDRC, recognizing the importance and urgency of planning for a worldwide communications network, enlisted BTL to assist the Army Signal Corps with its systems engineering tasks including communications security. Message traffic was readily securable, but voice transmissions were not, especially radiotelephone where interception was easy and privacy methods primitive.

Dr. O.E. Buckley, who became president of BTL in 1940, was charged with contacting the military and others concerned with speech security, ciphony. In his study of military communications, R.K. Potter, Buckley's alternate representative, identified two distinct areas of need: 1) short-term mobile privacy and 2) long-term, high-echelon secrecy, both suitable for telephone circuits. Buckley, a strong ciphony advocate, undertook this work at the Bell Labs without a written contract under the auspices of the Chief Signal Officer (the NDRC eventually accepted BTL's proposal).

Development

A very tightly held program, designated Project X (SIGSALY) for the high-echelon strategic system, was initiated in October 1940. BTL's task was to expeditiously develop, produce, and deploy fixed-plant highly secure telephone terminals to be operated and maintained by Signal Corps personnel. A small group of Bell Lab researchers under A.B. Clark, notably R.K. Potter, Harry Nyquist, R.C. Mathes, and D.K. Gannett, investigated a suitable speech processor for SIGSALY. The team expanded to conduct research on encryption algorithms and modems for transmitting the signal over voice frequency channels.

The speech processor design capitalized on Homer Dudley's work circa 1935 on a voice coder (vocoder) for commercial privacy and channel derivation (i.e., deriving several channels in place of one) applications. The underlying principle of a vocoder was one of analysis and synthesis. The analyzer measures the voice energy from multiple filters across the audio frequency spectrum and also measures the fundamental pitch of the speaker. Variations in a speaker's delivery are nominally limited to 25Hz. The synthesizer creates harmonics of the speaker's pitch, which are modulated by the slowly varying spectrum energies. In the case of unvoiced sounds (i.e., "s" or "sh"), noise serves as the "carrier." (Figure 1 shows a block diagram of the vocoder.) The resulting output is synthetic speech, which, though intelligible, leaves much to be desired for speaker recognition (positive identification). Research on the cryptographic component proved to be a more daunting challenge.

James Harris Rodgers received a patent in 1881 on a circuit-hopping system, which under control of relays, transmitted over two or more circuits in rapid succession.
– The Codebreakers, David Kahn, 1967

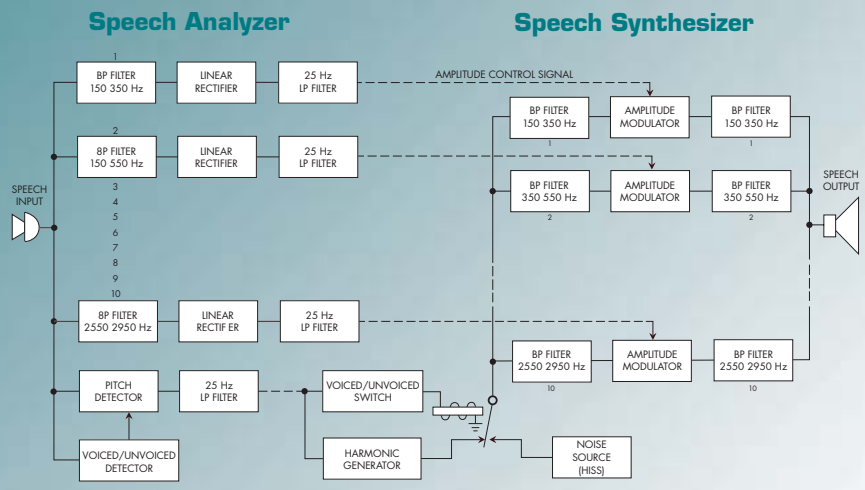


Figure 1: Vocoder (From A History of Engineering and Science in the Bell System, 1925-1975, M.D. Fagen, editor, 1978)

Potter's survey of eighty speech "secrecy" patents found a common fault in all. Like the A-3 they provided only technological surprise not cryptographic security – a determined and resourceful interloper could undo them. Rejecting these approaches, Potter pursued a different course in early 1941: noise masking the analyzer output. The results were similar to those of R. D. Parker. Next, Potter proposed digital substitution, using the method patented in 1919 by G. S. Vernam of AT&T for encrypting Teletype on-line: modulo2 addition of a five-level plain text tape with a random five-level key tape. (Table 1 shows Vernam's encryption modulo2.) Potter's experiments of quantizing vocoder channels to on-off signals added modulo2 to binary keys, though secure, produced badly mutilated synthesized speech, unacceptable to the listener.

K \ S	0	1
0	0	1
1	1	0

Table 1: Vernam Algorithm Modulo 2

Subsequently M.E. Mohr constructed a quantizer for up to ten levels. After experimenting with it, the team decided to encode the vocoder channels into six nonlinear amplitude steps (senary). The adoption of senary steps at the syllabic rate (25Hz) was a compromise between received voice quality and expected radiotelephone transmission margins, i.e., fading, noise and linear distortion.

In May 1941 Potter and Nyquist concluded that, mathematically, modulo6 addition of a nonpredictable senary key (where all six levels were equally probable) to senary plaintext would produce a cryptographically secure senary cipher. (See Table 2 for the Potter-Nyquist Modulo6 Encryption.) R. C. Mathes invented an electronic "re-entry" circuit for modulo6. (Though not told it was for SIGSALY, Claude Shannon, the father of Information Theory, was consulted early on about the modulo6 encryption.) The remaining elements of the system were the modem and source(s) of key.

K \ S	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

Table 2: Modified Vernam Algorithm Modulo 2

The modem team, having had considerable experience with Teletype transmission over radio, was faced with the problem of designing a modem for a six-level signal vice the customary binary FSK. Amplitude modulation was discarded since selective fades could be as high as 20db on transatlantic radio. They adopted a scheme of frequency shift keying six frequencies in each channel every 20ms (the equivalent of 129bits/sec per channel for a 600 baud senary signal). The transmit modem consisted of a twelve senary FM signals (170 Hertz spacing) covering the audio spectrum, which could be transmitted over ordinary voice frequency telephone lines to an independent sideband HF radio transmitter.

To take maximum advantage of off-the-shelf Teletype components, the engineering design was based on a parallel architecture throughout. Figure 2 shows the transmitter, composed of twelve separately filtered channels from the speech processor (codec) through the encryptor to the modem. The codec analyzer measured the energy in ten channels across the audio spectrum (150 to 2,950Hz); two channels (a main and vernier) measured the fundamental pitch or no pitch of the speaker. The analyzer outputs were quantized to six discrete levels via "steppers," RCA 2051 gas thyratrons, one stepper for each level, firing at twenty millisecond intervals; the pitch frequency (main and venier) was similarly quantized.

The receiving radio translated and sent the encrypted signal over telephone lines to the distant SIGSALY terminal. The receive terminal separated the twelve enciphered channels, demodulated each channel, and synchronously decrypted with matching keys. The decrypted spectrum channels drove the vocoder synthesizer. Figure 3 shows a logical block diagram of the SIGSALY receiver.

A sixteen-inch record stored prerecorded one-time encryption key (SIGGRUV) that when added modulo6 to each of the codec steppers produced twelve cipher streams. Three additional tones, the first for turntable changeover and the other two for synchronization, were also recorded.

As an alternate senary key source, BTL developed the "thrashing machine" (SIGBUSE: SIGSALY alternate key generation system). It consisted of an array of clattering relays and telephone selector switches controlled by pseudorandom key from M-228 rotor machines (SIGCUM: teletype encipherment system). The M-228 machines were developed by the Signal Corps for on-line Teletype encryption. A full duplex SIGBUSE system, housed in five bays, produced senary key on-line at 600 baud. Though not as secure or reliable as the SIGGRUV, SIGBUSE did not pose

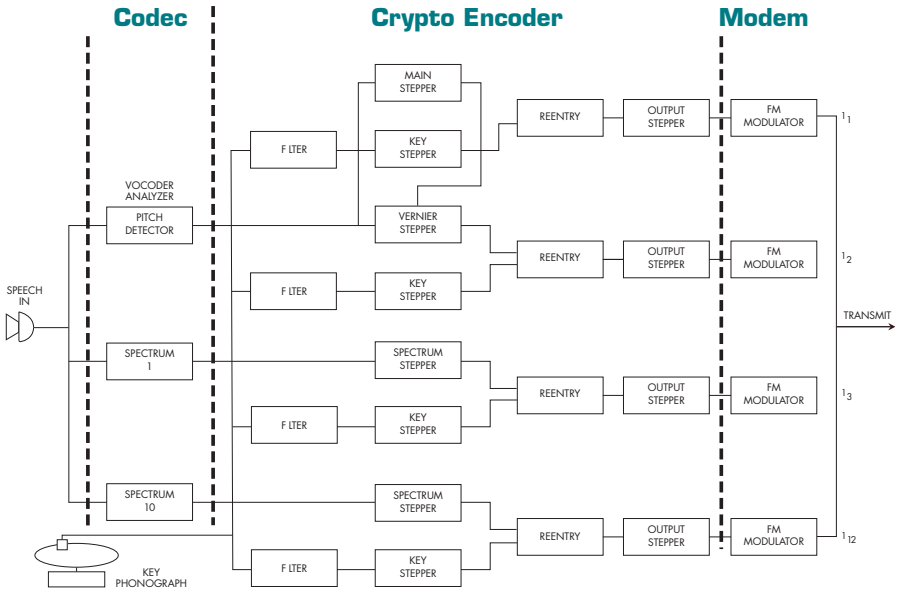


Figure 2: X System transmitter (From A History of Engineering and Science in the Bell System, 1925-1975, M.D. Fagen, editor, 1978)

the physical security concerns of distributing twelve minutes of key per one sixteen-inch record. SIGBUSE handled operational traffic up to secret, whereas the one-time key was used for top secret voice conferences.

Development of One-time Key System

Digitizing Gaussian noise produced the onetime key records described above. Noise outputs of twelve RCA 2051 thyratrons each sampled fifty times per second were quantized to six uniformly distributed levels via steppers similar to those used in the codec. The stepper outputs amplitude modulated twelve 170Hz spaced tones from 595 to 2,295Hz, which were combined and recorded on vinyl phonograph records at 33 1/3 rpm. Key production initially done in New York City by Bell Lab personnel was eventually taken over by ten officers and twenty-five enlisted members of the 805th Signal Service Company at the Pentagon in December 1944. By incorporating the BTL modifications (SIGSOBS: SIGSALY primary key generation system), the Signal Corps was able to manufacture two acetate recordings (SIGJING: acetate key records) at once, lowering the cost. Two playback terminals were associated with every SIGSALY terminal, each providing of unique key for a full duplex top secret conference. See Figure 4 for the patent diagram on which SIGSOBS was based.

In March 1942 one channel of the system was tested on an HF simulator to determine its performance under artificial fading conditions. It passed. The completed experimental model was quickly tested for operation and overall stability, and was continually being used as a test bed for design refinements and for training Signal Corps personnel. By April 1942 a complete set of drawings was ready to be turned over to Western Electric.

Deployment

In early 1943 Alan Turing, the UK's premier cryptologist, visited Bell Labs to accredit the system for the British government. The assistant chief signal officer had bestowed jurisdiction for ciphony to the Signal Intelligence Service (SIS) in February 1942. However, this author could not find correspondence from the National Archives and Records Administration files where a Signals Intelligence Service (SIS) or an Army Communications Service (ACS) official had accredited SIGSALY.

During the first official SIGSALY conference, inaugurated on July 15, 1943, between Washington and London, Dr. Buckley said, "...it must be counted among the major advances in the art of telephony."

From 1943 to 1946, twelve SIGSALY terminals provided secure teleconferencing intratheater, for the White House staff and the General Staff in Washington to Theater Commanders and our British allies. In the case of the Pacific Theater, the Pentagon terminal was connected to an HF radio terminal in Oakland, California, by full-period AT&T telephone lines.

SIGSALY was initially operated and administered by the Signal Corps. The General Staff assumed the responsibilities starting in March 1944 by the order of the secretary of war. Colonel Humelsine's Staff Communications Branch at the Pentagon handled the classification, priority, reproduction, and distribution of SIGSALY transcripts and secure (SIGTOT: teletype encipherment system) message traffic. Captain Dorothy Madsen wrote a General Staff Circular for eligible users, set up the administrative procedures, and personally edited all transcripts.

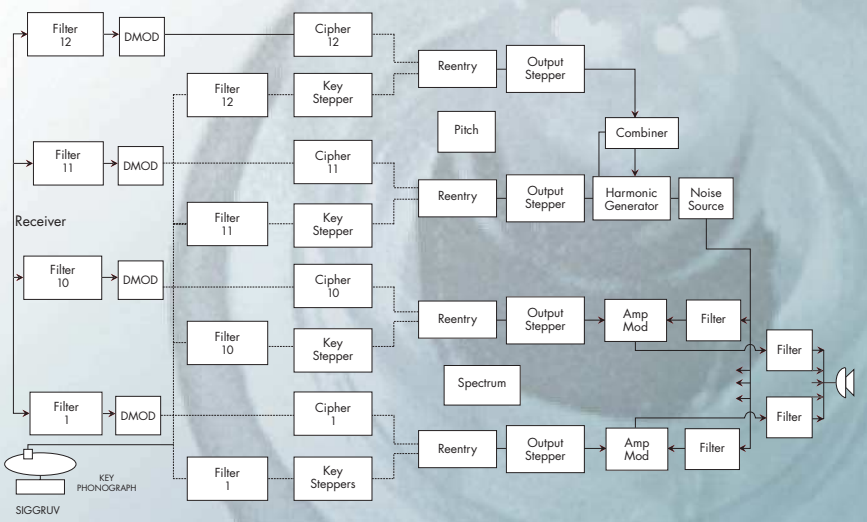


Figure 3: X System receiver (From A History of Engineering and Science in the Bell Sstem, 1925-1975, M.D. Fagen, editor, 1978)

Prime Minister Churchill spoke frequently to many senior officials including President Truman on SIGSALY but ironically FDR never used it.

The 805th Signal Service Company was in charge of the overseas terminals, and to the extent possible followed the above procedures. The Signal Corps retained technical responsibility for transmission and encryption. The Army Communication Service couriers distributed SIGSALY key records worldwide and in conjunction with AT&T Long Lines supported the 805th with radiotelephone and Teletype transmission facilities. One SIGSALY terminal occupied thirty seven-foot relay racks and required over 30kw of power.

Until SIGSALY was decommissioned, the terminals and key production facilities were operated and maintained by the 81 officers and 275 enlisted men of 805th Signal Service Company with a small complement of Bell Labs personnel. The dedication and know-how of the 805th Signal Service Company kept SIGSALY availability extremely high under difficult wartime conditions. In his book The Green Hornet, Donald Mehl describes the travails of SIGSALY on the OL-31 barge that followed General MacArthur on his island-hopping campaign from Australia to Manila to the Japanese surrender on Tokyo Bay. The total program cost over its service life—R/D, procurement, training and Operation/ Maintenance (O/M)—was estimated to be $28M.

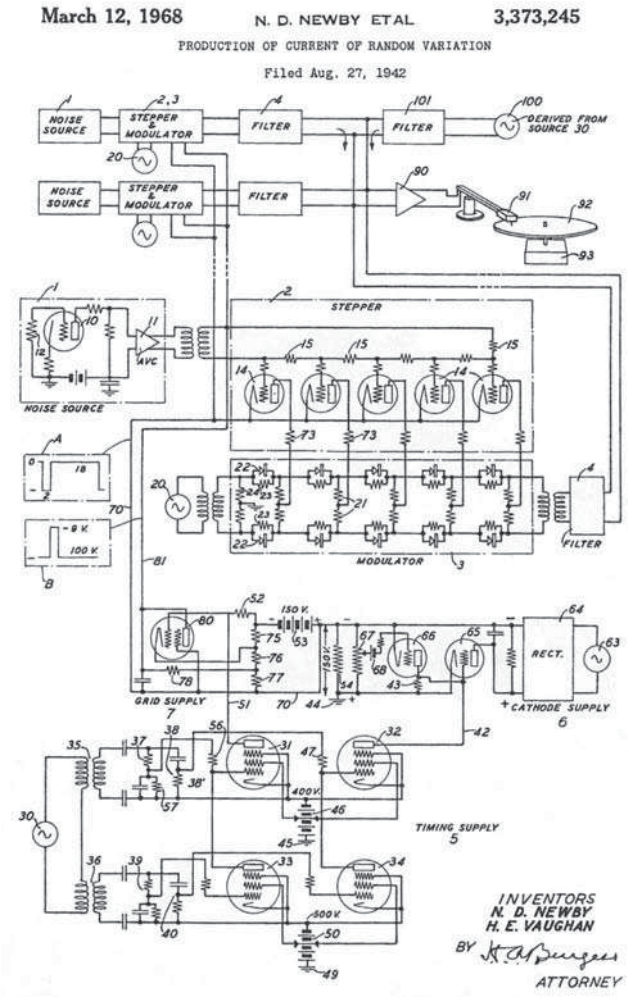


Figure 4: Patent diagram on which SIGSOBS was based.

SIGSALY Decommissioning/ Disposition

In February 1946 Major Luichinger submitted the results of his study and recommendations concerning the discontinuance of Overseas Secure Telephone Service to General Stoner, chief of Army Communication Service. In it he reported that in the last three months of 1945 operational SIGSALY traffic showed a continuing downward trend—Frankfurt averaging less than one call per day, which represented about 50 percent of the total. He recommended that all ETO terminals except Frankfurt and Berlin be terminated; only the Tokyo terminal on OL-31 barge was to remain operational.

On 13 August 1946 General Stoner, now the Assistant Chief Signal officer, in a memorandum to the Director of Intelligence, addressed Major Luichinger's report regarding the storage and destruction of SIGSALY and associated equipment. In summary it directed that

a) All equipment be returned to the Zone of Interior (ZI)

b) Six overseas terminals, one key production facility (SIGSOBS), be destroyed

c) Six be stored as war reserves in the ZI with two SIGSOBS and Off Premises Systems

The report stated that "Upon their return in the fall 1946 SIGSALY terminals were to be transferred to the Army Security Agency until the state of the art permitted a replacement system."

Other World War II Ciphony Systems

As the first SIGSALY equipments were rolling off the Western Electric production line in 1943, the Bell Lab researchers were redesigning it. They subsequently developed "Junior X" (AN/GSQ-2,3), which occupied six five-foot bays. It used miniature vacuum tubes, serial vice a parallel architecture, and a key generator in lieu of a one-time key. In the fall of 1944, the Signal Corps contracted for GSQ-2,3 production with delivery set for March of 1946, too late for WWII service. (See Figure 5 for a block diagram of the AN/GSQ-2,3.)

Also during the later stages of the war, BTL built and tested a multichannel Line-of-Sight (LOS) radiotelephone system (AN/TRC-6) for the Signal Corps. It saw only limited service in Europe as the first binary coded speech transmission system (analog Pulse Position Modulation (PPM)).

Prime Minister Churchill spoke frequently to many senior officials including President Truman on SIGSALY but ironically FDR never used it.

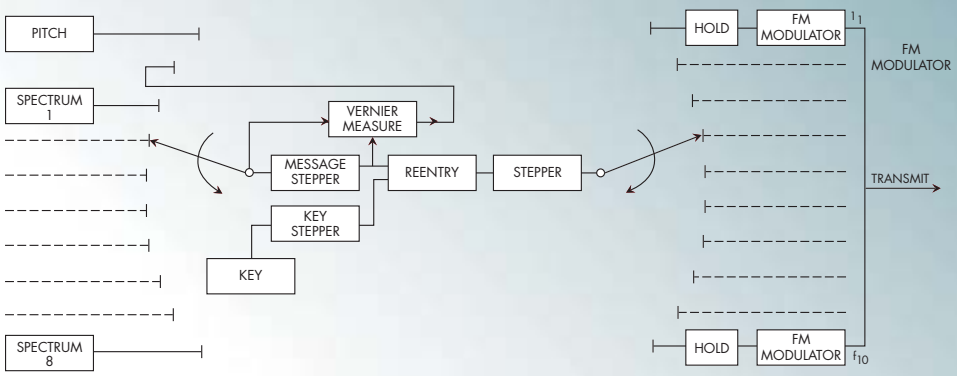


Figure 5: AN/GSQ-3 (From A istiry of Engineering and Science in teh Bell System, 1925-1975, M.D. Fagen, editor, 1978)

Conclusions

As an engineering accomplishment, SIGSALY was in a class by itself, especially if one considers the sheer magnitude of BTL/Western Electric starting from scratch to deliver the first operational terminals in thirty months. From a technology standpoint, SIGSALY had many operational "firsts":

- The first to use digital speech compression
- The first modem to use digital FM rather than binary (FSK)
- The first to extract and record digital (senary) key from a noise source
- The first to use a nonbinary Vernam encryption algorithm
- The first to store and distribute digital key on phonograph records
- The first to use Protected Wireline Distribution (OPEPS)

A few days after the war ended, the War Department renamed the Signals Security Agency (nee Signal Security Service nee Signal Intelligence Service), the Army Security Agency, placing it under Army Intelligence Staff instead of being subordinate to the Office of the Chief Signal Officer .

Army Security Agency's mission remained the same as the SIS: codebreaking and codemaking, COMINT and COMSEC. Under the latter a small contingent was established to conduct R/D on future voice and data systems while an operational group continued to produce and distribute keying material, certify system security, issue operating doctrine, and handle equipment procurement.

References

Bell Telephone Publications

"AN/TRC-6 – A Microwave Relay System," H.S. Black, Bell Laboratories Record, Dec. 1945

"Spectra of Quantized Signals," W. R. Bennett, The Bell System Technical Journal, July 1974

Books

A Brief History of Cryptology, J.V. Boone, 2005

A History of Engineering and Science in the Bell System 1925-1975, M. D. Fagen, Editor, 1978

Alan Turing, The Enigma, Andrew Hodges, Simon & Schuster, 1983

A World War II WAC's Memoir: My Journey to the Pentagon's Top Secret Command Center, Dorothy Madsen (Lt. Col., USAR, Ret) to be published

Frequency Analysis Modulation and Noise, Goldman, 1948

Information and Secrecy, Colin Burke, 1994

The Codebreakers, David Kahn, 1967

The Green Hornet, Donald Mehl, 1997

Top Secret Communications of WWII—Sigtot, Donald Mehl

AIEE

"Certain Topics in Telegraph Transmission Theory," H. Nyquist, Transactions , AIEE, 1927

National Archives and Records Administration

RG 111, Office of the Chief Signal Officer

RG 227, National Security Agency

RG 457, National Defense Research Committee

SIGSALY Speech Encipherment System RC-220-T1 Technical Manual Vol. A, Bell Telephone Laboratories School for War Training, NARA DECLASSIFIED 5/11/96

National Security Agency

"Speech and Facsimile Scrambling and Decoding," Monograph No. 17, 1969

"The ABC of Ciphony," Fred E. Buck, NSA Technical Journal, July 1956

"The SIGSALY Story," Patrick Weadon, NSA, 2000

The Start of the Digital Revolution, R.R. Peterson & J. V. Boone, NSA, 2000

The Quest for Cryptologic Centralization and the Establishment of NSA, 1940-1952, NSA, Thomas L. Burns, 2005

Unpublished Notes

"A World War II Wac's Memoir: My Journey to the Pentagon's Top Secret Command Center," Preface, [Internet], Dorothy (Meg) Madsen, available at http://www.ww2wac.com/page11 html

"History of SIGGRUV – The SIGSALY Recording Project," David Kemper, Sept. 1995

February 17, 2009

A Second Date that Will Live in Infamy

Time is ticking away toward the biggest technology challenge since Y2K. February 17, 2009 will mark the end of over-the-air analog television and the beginning of digital-only broadcasting.

For the majority of Americans, February 17th will be just another day with no difference in their television viewing. Those that use cable or satellite for their television programming will be unaffected by the switchover to digital-only over the air broadcast. Those that have already purchased a digital TV (DTV) will still be able to receive the over-the-air digital broadcast that many stations already transmit. The problem lies with the 20 percent of U.S. households that receive analog broadcast as their primary television signals. This translates to 45 million TV sets plus 28 million additional sets receiving analog broadcast in homes that have at least one other TV hooked up to cable or satellite services.

There are several options to successfully transition these 73 million analog television sets into the digital age. One option is to tie the sets into either a cable or satellite system that converts the digital signal into analog. A second option is to buy a digital-to-analog converter box. A third option would be to replace the analog set with a digital–ready TV set. Of course, the analog sets will continue to work with gaming consoles, VCRs, DVD players, and similar products that you currently use.

What could make things worse for the transition is that 61 percent of Americans have no idea what is due to happen on February 17, 2009. The two federal agencies responsible for a successful transition are the Federal Communication Commission (FCC) and the Commerce Department. To raise awareness and "to keep the heat on" these agencies, Congress held two hearings in October 2007 and have more scheduled. In 2008, the National Association of Broadcasters plan calls for extensively broadcasting public-service announcements on the digital TV transition.

To assist the DTV transition, the National Telecommunications and Information Administration (NTIA), a part of the Department of Commerce, will administer the "Digital-to-Analog Converter Box Coupon" program. Beginning in January 2008, each U.S. household is eligible to receive up to two $40 coupons to purchase digital-to-analog converter boxes. The coupons are good at participating consumer electronic retailers for eligible converter boxes only. These boxes are estimated to sell for $50 to $70.

The move towards digital broadcast began back in 1987 when the FCC began work on a new High Definition television (HDTV) system. The FCC adopted the HDTV standard in 1996 and most television stations were given a free second transmission channel for digital broadcast. FCC rules were set to encourage simultaneous broadcast of the analog and digital channels. The broadcasters were supposed to complete the DTV transition by the end of 2006 and return some of their analog spectrum to the government. Because broadcasters were slow at making the transition, Congress enacted and the President signed the DTV Transition and Public Safety Act of 2005. In addition to setting the February 17, 2009 deadline for the end of analog television broadcast, the law set forth plans to auction the spectrum that is saved by going to digital broadcast. It also set aside 24 MHz of spectrum for public safety use.

There are several advantages to DTV. Digital transmissions are less susceptible to interference from transmissions on other channels and providers can allocate closer spaced channels. Broadcasters can transmit multiple streams per channel. For example, channel 2 offers multiple streams. On channel 2.1 they transmit a stream of DTV, on 2.2 they transmit a stream of HDTV, and on 2.3 they transmit a stream of widescreen. (It should be noted that DTV and HDTV are not the same. HDTV has a better audio and video quality.) Instead of offering different format on the multiple streams, some stations will offer different programming. For example, 2.1 could be the normal programming and 2.2 could be a 24-hour view of the weather radar. Another advantage of DTV is no fuzzy reception. Either you will have a perfect picture or you will have no reception.

The final advantage of digital broadcast over analog is the savings in spectrum. Each of the 66 analog channels has a 6 MHz lane and takes up a total of 396 MHz of bandwidth. DTV does not need as much bandwidth and all of the TV stations would need only 288MHz. DTV would free up 108MHz. The spectrum that would be given back to the government is currently the channels 52-69, 698 MHz through 806 MHz, which is referred to as the 700 MHz band.

The government is auctioning 62MHz of the 700MHz band starting January 24, 2008 in what is known as Auction 73. The auction will include licenses for Economic Areas (176 licenses for 698-704/728-734 MHz and 176 licenses for 722-728MHz), Cellular Areas (734 licenses for 704-710/734-740 MHz), Regional Economic Area Groupings (12 licenses for 746-757/776-787 MHz), and one nationwide license (758-763/788-793 MHz). The government expects to sell the spectrum for a minimum of $12.5 billion. The frequency characteristics make this spectrum so valuable. The signals will propagate further and provide better coverage than the current cellular frequencies and will penetrate walls and buildings. The FCC has described the 700MHz band as the last beachfront property.

There is much speculation about what the winners of the 700MHz band auction will do with their newly obtained spectrum. It could be the third pipe into the home (with the phone line and cable line being the first two pipes). A nationwide wireless network using the 700MHz frequencies would cost about $2 billion to build versus $4 billion for the 1900MHz band. Additionally, there is speculation that WiMax may be deployed in the 700MHz band giving broadband network connectivity to the masses.

In summary, avoid being surprised when the analog over-the-air broadcast ends by either having your televisions hooked up to a cable or satellite service, purchasing a digital television, or purchasing an analog-to-digital converter box. As February 17, 2009 marks the end of one era, stay tuned to see what new technology innovations are over the horizon as a result of the 700 MHz auction.

References:

"FCC Releases Full Text of 700 MHz Second Report and Order; Auction to Begin by January 28, 2008." http://www.bingham.com/Media.aspx?MedialID=4592.

Ted Hearn, "Don't Panic. Yet.", http://www.multichannel.com/article/CA6417227.html.

Auction 73 700 MHz Band, http://wireless.fcc.gov/auctions/default.htm?job=auction_factsheet&id=73.

Brian Peters, "High Tech Coalition Renews Call for Successful DTV Transition," http://www.dtvcoalition.com/News/Read.aspx?ID=71.

www.hdtv.gov

Bryan Gardiner, "FAQ: Inside the High-Stakes 700-MHz-Spectrum Auction," www.wired.com/techbiz/it/news/2007/09/auction_faq.

Andrew Kantor, "Coming TV change won't make your set obsolete," www.freepress/net/news/12083.

Charles Pope, "Digital TV: Plot Thickens Feb. 17 2009," The Oregonian, October 28, 2007. (OregonLive.com)

Om Malik, "700MHz Explained in 10 Steps," http://gigaom.com/2007/03/14/700mhz-explained.

For The Record

How Format Wars Have Shaped Recording History

Modern technology and culture may find themselves in closest harmony revolving around the phonograph record. Reflections on the twentieth century conjure images that include the spotted dog Nipper peering quizzically into a Victrola, ecstatic teenagers at sock hops, and psychedelic album covers. Our cultural icons include Elvis, the Beatles, and Madonna. We live in an age when we are encouraged to "spin up", "get on track", and "stay in the groove"—language from the age of the gramophone.

The recording industry has supplied the lifeblood of popular entertainment for over a century, opening a window on the World's dreams, thoughts, and memories and preserving them for future generations. Technologies invented in the 1800s have made popular entertainment easily accessible from every corner of the planet. The phonograph record lies at the heart of those technologies, in one format or another. With so much power to shape culture, it is no wonder that the recording industry has been a battlefield for entrepreneurs who seek to profit from it.

Only a few years ago, "records" were thought of as the black, vinyl platters that spun out hit songs by the Bee Gees and symphonies by Beethoven. Vinyl records have since yielded to cassette tapes, Compact Discs, and Flash drives. Their ancestors include piano rolls, organettes, and music boxes. Recording formats have come and gone with regularity over the years—some slipping quietly into obscurity, others managing to define entire generations. The days of the gramophone record may be past, but the importance of recording popular culture and the entertainment that shapes it will never be lost.

The public's insatiable desire to be informed and entertained has inspired countless technological advances. Whether the recording technology of the day is papyrus or the iPod, it reflects from every facet of society. Accountants, doctors, engineers, teachers, military generals, priests, and bureaucrats inevitably find practical applications for recorded information. Practicality might drive the research necessary to pioneer new technologies, but entertainment can determine the direction those technologies take.

The quest to supply an incessant demand for richer, more rewarding media experiences continues to motivate entrepreneurs to come up with the next revolutionary media format. The cycle of invention, adoption, and obsolescence has repeated itself countless times throughout history, in ever tightening gyres. Today's pace of invention is measured by the stopwatch of Moore's Law, with cutting-edge technologies making their way to the recycle bin in a matter of months.

Recording media have been among the most visible technologies to pass quickly into obsolescence. Victrolas, cassette players, and video tape recorders clutter attics and garages worldwide. These modern antiques merge with the generations of popular recording formats that preceded them, and they are just as sure to be joined by many formats to come.

Winning the title of "best format" often comes with more than a little bit of luck and the willingness to put up a hard fight. Despite the difficulty of dispatching the prevailing technological format from its throne, and the inevitability of some day becoming equally archaic, each contending format is zealously defended by its inventors, investors, and early adopters. Few challengers can be winners in the Format Wars. Fewer still are remembered as champions.

Inventors who survive the Format Wars will go to great lengths to protect their secrets—and their profits. In modern times, patents are used to lay claim to the latest technological trend, and lawsuits are filed to challenge infringements. In millennia past, priesthoods preserved the mysteries of the oracle, and death could befall someone who might expose them.

Recording in High Definition

A format war seems to flare up every decade or so, and it can be a matter of years before an uncontested winner is declared. The new millennium ushered in one especially significant contest, to determine who will reign over the coming era of high-definition (HD) media.

The DVD Forum, caretakers of the prevailing laser recording format, had appointed HD DVD as the rightful successor to DVD—the current media king. Sony, who had been relegated to the sidelines in past battles, saw an opportunity for a coupe by getting to market first with Blu-ray, the company's laser disc contender.

Five years of intense fighting pitted Sony's Blu-ray camp against Toshiba and its allies championing HD DVD. To the surprise of many, on 12 February 2008, Toshiba raised the white flag and conceded defeat. As the dust settles on the laser disc battlefield, the outlook for a new generation of high-definition media storage has become clearer. Blu-ray stands as heir apparent in a long line of popular recording formats.

The current format champion may not wear the crown for long, if history is any indicator. New challengers are already lining up in the wings for a shot at the title, and they, in turn, will take up their positions on the battlefront of the Format Wars. Before we look to the conflicts that lie ahead, it can be enlightening to peer back into the past.

Checking the Records

Entertainment plays a major role in determining the prevailing format for recording information. Humans possess an innate passion for amusement, which has spawned technologies for creating music albums, movies, and video games.

The influence of the entertainment industry has reached extraordinary proportions in modern times. MP3s, video players, and game systems are considered essential equipment among a generation accustomed to immediate and unlimited access to wireless jukeboxes, cinemas, and arcades. But the obsession for songs, plays, and games did not recently evolve—it is tightly woven through human experience, from before history can account for it.

Ancient Artifacts

Music has been at the heart of human entertainment, perhaps for tens of thousands of years. Some of the earliest human technologies were developed for the sole purpose of making music. Prehistoric artisans crafted flutes from reeds and lutes from gourds. As technologies for creating music evolved, so did the desire to record what was played.

Musicians enjoyed a place of honor in early civilizations. Melodies performed during religious rituals were handed down orally as part of the priestly mysteries. By the time of the Golden Age of Greece, prizes were heaped on winning musicians at the Ancient Olympic Games. Some of the songs performed by choruses 2,500 years ago can still be sung today, reconstructed from fragments of ancient sheet music. Greek composers used a special notation system to record their melodies on papyrus. The sheet music they created may represent the first recording format for music.

The Greek engineering skills were preserved by Roman and Ottoman conquerors. Clockworks, invented by the Greeks, were combined with mechanical and hydraulic engines to create increasingly complex and even programmable machines. Some of these devices were designed to play musical instruments. Courtesans from Baghdad to Bologna were routinely treated to elaborately orchestrated displays that included boatloads of robot musicians and singing mechanical knights—all playing recorded music.

Skilled engineers crafted mechanical devices to mystify and entertain audiences for several centuries following the fall of the Greek Empire. Such automata became popular amusements, especially during

10,000 BC and Before
Memorization: In prehistoric times, melodies for songs and instruments were passed from one generation to the next by memorization.

100 BC
Clockworks: The Antikythera mechanism, crafted in the second century BC, was a programmable clockwork computer.

1200
Automata: Gears rigged to pull cables and trip levers were used to animate a variety of different contraptions, since at least the first century AD.

450 BC
Sheet music: Musicians in ancient Greece created the first copies of sheet music by adding musical notation to lyrics written on papyrus.

the Renaissance. Leonardo da Vinci was among the many artisans commissioned to create the cable-and-pulley-driven contraptions that would play back the music and actions recorded into them. The concept behind da Vinci wiring up an automaton's performance and a computer programmer writing code are essentially the same—only the materials and scale differ.

The Birth of the Modern Recording Industry

By the end of the Renaissance, recorded music had found its way into mechanical clocks, musical toys, and snuff boxes. The automata that had entertained audiences during much of Western history evolved into the barrel organ, ushering in an era of popular recorded music.

Barrel Organs

Barrel organs brought musical performances once reserved for wealthy urbanites to fairgrounds and street corners of small towns across Europe. Scores for elaborate tunes would be encoded with metal pins hammered onto the surface of a wooden drum, called the barrel. An organ grinder then turned a crank that was fastened to the barrel, causing the pins to pass over levers attached to pneumatic valves. When a valve was open, forced air would pass though a corresponding organ pipe to produce the desired note.

Several different tunes could be recorded on a single barrel by staggering the arrangements of the pins. To select another tune, the musician simply shifted the barrel forward or back to align a set of pins with the corresponding set of valves. Operating a barrel organ was primarily an engineering exercise. The music had already been created by the recording artist who wielded the hammer to set the pins in the barrel.

Player Pianos

By the late nineteenth century, the barrel organ had evolved into the player piano. Instead of using pins to open and close pneumatic valves, holes in a paper scroll allowed air to pass directly through corresponding valves. Mechanical amplifiers lifted the relatively heavy hammers sufficiently to strike the piano keys, without requiring so much force that the paper roll would be ripped apart by the forced air.

Successive improvements in the design of the player piano led to sophisticated recording processes that not only sped up production time for making piano rolls, but also provided subtly accurate recordings of individual performances. Accomplished concert pianists eventually agreed to have their performances recorded for posterity—and for profit. Competing piano companies and independent publishers began recording hundreds of classical tunes, popular ballads, and commissioned pieces for the piano.

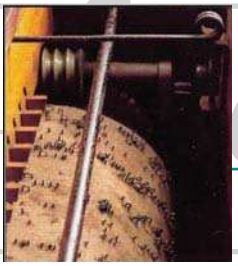
By the early 1900s, piano rolls in a variety of formats and materials flooded the music market. The availability of inexpensive piano rolls brought professional-quality musical performances into homes and social centers around the world. Player pianos were turning up in the parlors of back-country farm houses and small town saloons as well as in fashionable sitting rooms and concert halls.

Perforated rolls of paper were not the only medium for recording music at the turn of the nineteenth century, however. A revolutionary technology patented by Thomas Edison would soon replace the player piano as the family entertainment center and usher in the modern era of recorded entertainment.

Recording Cylinders

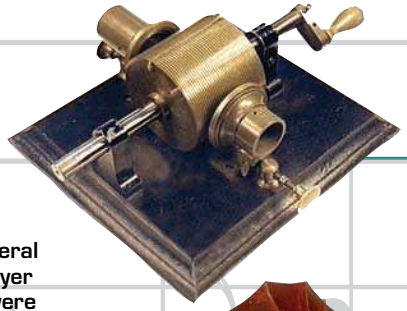
Edison had made his now-famous recording of "Mary had a little lamb" on a tinfoil covered cylinder in 1877. Early models of Edison's "talking machine" were sold mostly as office dictation machines under the Ediphone label. A home version of the talking machine, along with a limited number of prerecorded two- and later four-minute cylinders, was trademarked as the Phonograph.

Edison's talking machines attracted only a small audience. Attempts to miniaturize the recorded cylinders to make talking dolls also failed, due partly to the fragility of the tinfoil recording medium. Edison soon lost interest in his talking machine, turning his attention to the task of inventing the electric light. He even allowed the British



1400
Barrel organ: By the fifteenth century, the roller used in music boxes to pluck notes when metal pins passed by was adapted to create the barrel organ. The barrels used in cathedrals could be as large as 10 feet in diameter.

1876
Piano roll: Several pneumatic player instruments were exhibited at the Centennial Exposition of 1876 in Philadelphia, launching the earliest versions of the player piano



1877
Tinfoil cylinder: Thomas Edison introduced the phonograph in 1877. The first working talking machine was indented primarily for taking dictation.

1886
Wax cylinder: The Graphophone used engraved recordings on wax coated cylinders, making a significant improvement to the embossed cylinders used in the phonograph.

patent on the phonograph to expire in 1885, thinking the invention was a failure.

Although Edison had given up on promoting the phonograph, some of his rivals foresaw the potential for a machine designed to play recorded music. Charles Sumner Tainter partnered with Alexander Graham Bell and his cousin Chichester Bell to market an improved version of the phonograph. They jokingly called their invention the "Graphophone"—a play on the Edison trademarked name. Using prize money A. G. Bell had won for his invention of the photophone, the three young inventors established Volta Laboratory in Washington, D.C., and set to work building an improved talking machine.

Tainter and the Bells made numerous revisions to Edison's design, which they carefully documented to protect the several patents they had been awarded. The three men were so concerned that Edison would take credit for their inventions that, in 1881, they sealed a detailed account of their research in a box they deposited in the confidential archives of the Smithsonian Institution.

After several years of refining the graphophone, Tainter had managed to overcome many of the phonograph's shortcomings. Most significantly, Tainter developed a better recording medium by hardening the coating on the cylinder with carnauba wax and refining the etching process.

Against Tainter's better judgment, the American Graphophone Company approached Edison's representatives in May 1887, to propose the possible merger of the two companies. Tainter's associates demonstrated the advantages of the graphophone, including the improved wax cylinder. The merger offer was rebuffed, and, as Tainter had predicted, Edison immediately set out to develop a wax replacement for the phonograph's tinfoil cylinder.

By November of the same year, Edison filed a competing patent for a wax cylinder. Due to Tainter's carefully documented research, the courts upheld his patent claim. But by then, the wax cylinder was already destined to be a victim of the Format Wars.

Recording Discs

While the graphophone and the phonograph competed to dominate the emerging record industry of the 1890s, Emile Berliner was attempting to invent a better recording format. Berliner was already a pioneer in voice technologies, having invented the microphone in 1876, which he sold to the Bell Telephone Company for $50,000. He used that capital to fund his efforts to create audio reproductions that could be commercially marketed.

Berliner resorted to using an "indestructible" flat disc with lateral-cut groves for recording, instead of the fragile, vertically grooved cylinders Edison and Tainter were using. The lateral recording format had been tried in 1857 for linguistic analysis, but it was never patented for recording. Berliner patented his design in May 1887 and set out to launch The Gramophone Company.

Berliner's talking machine design was superior in that the hard vulcanized rubber discs were more durable than wax cylinders, and gravity held the stylus needle securely in the recording track. The platter format was also easier to ship and store, and it provided space for attaching a label. Most important, Berliner's sound discs, as he called them, could be reproduced from a master—something the wax cylinder could not do—opening the door to mass produced records.

Berliner later abandoned using rubber for pressing sound discs in favor of Durinoid, a shellac compound used for making buttons and electrical parts. By 1895 shellac records had become the new recording standard.

Edison scoffed at the scratchy sounding gramophones, and he persisted in believing that the public would reject a talking machine that could not also record voices. Edison was wrong. The gramophone was immediately popular. Within a decade of the introduction of sound discs—or gramophone records, as they grew to be called—the new recording format had nearly replaced wax cylinders. Even Edison had to concede that times had

changed, and in 1912 his company started producing gramophone records, as well.

The first major Format War had ended.

New Technologies Usher in New Recording Formats

Music was recorded on shellac records for half a century. But an experimental product developed by an Ohio tire company was destined to usher in the next generation of record technology—vinyl.

Vinyl Records

In 1926, Dr. Waldo Semon, a chemist at the BFGoodrich Company, in Akron, Ohio, developed a kind of plasticized rubber. The tire manufacturer could not immediately find a commercial use for the curious substance, but polyvinyl chloride (PVC) soon proved to have advantages for pressing records.

RCA Victor released the first vinyl record in 1930. Introduced as Program Transcription discs, the new discs were designed to play back at 33 1/3 rpm—half the speed of popular 78 rpm records. Despite the longer playing time and superior sound quality compared to shellac records, the new recording format was a commercial failure. Most potential customers could not afford the expensive playback equipment just when the Great Depression was getting underway. Those who could were disappointed by how the heavy pick-ups that were available at that time quickly wore through the soft vinyl discs.

Radio had also come of age by the 1930s, making records virtually obsolete. Vinyl was used at that time primarily for the recordings that were sent to disc jockeys for radio commercials and prerecorded programs, largely because the vinyl material was less likely to break on the way to the radio station. DJ copies of recorded music soon followed, for the same reason. Transcription services, which typically recorded at 33 1/3 rpm on 16-inch and 12-inch discs, also used vinyl, because their records were typically played only once.

World War II was ultimately responsible for the transition to vinyl records. Shellac had become scarce during and after the war, so record companies turned to vinyl to press some of their 78 rpm records. Vinyl records were lasting longer by then, because lighter tone arms prevented the needle from eating into them as quickly.

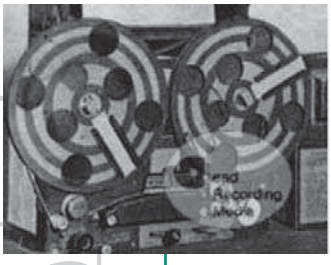
Columbia Records saw a brighter future for vinyl, however. The company had spent 10 years developing its microgroove technology and the equipment to record and play back music at the slower 33 1/3 rpm speed. During that time, Columbia made sure that every 78 rpm recording they made was also recorded at 33 1/3 rpm on 16-inch discs or on tape, another new recording medium. Columbia's library was already stocked with high-fidelity recordings when the new, long-playing format was introduced, giving the company a major competitive edge.

Columbia released the first Long Play (LP) record in 1948, packing 22 minutes of music on each side of a 12-inch disc. The company's focus at that time was on classical music, which required the longer playing time. Columbia had decided to continue recording its shorter pop tunes at 78 rpm. Less than one year later, RCA moved in on the single-recording market and started producing seven-inch singles at 45 rpm. The new "LPs" and "45s" were immediately successful. These popular recording formats helped define the 1950s and revive the flagging record industry.

Audiotape

While Columbia was developing the LP, the company was also experimenting with another recording medium—magnetic tape. Audio recordings on a magnetic medium had been around for almost as long as the gramophone. Valdemar Poulsen, a Danish engineer, demonstrated the first magnetic audio recording in 1898. For Poulsen's Telegraphone, a length of wire

1930
Vinyl record: RCA Victor launched "Program Transcription" discs, made from vinyl instead of shellac. The12-inch records could be recorded at 33 1/3 rpm, and they were not a fragile as 78 rpm gramophone discs.



1935
Audiotape: The Magnetophone was introduced at the 1935 Berlin Radio Fair. The accidental discovery in 1939 of high-frequency bias greatly improved recording sound fidelity.

1951
Video Tape Recorder (VTR): Bing Crosby Enterprises demonstrated the first black and white magnetic tape recording, spurring international efforts to develop a practical video recorder.

1963
Compact Cassette: Philips introduced the audio cassette medium that would make home recording popular for the first time since the phonograph was introduced.

was magnetized by running it across a recording head while sounds were being introduced. By running the same wire across a playback head, the magnetic pattern would reproduce the original sounds.

In 1928, German engineer Fritz Pfleumer adapted Poulsen's invention by substituting the recording wire with a long strip of paper coated with iron oxide powder. Pfleumer's audiotape was soon to be manufactured by German chemical giant BASF, and the Magnetophone it played on was in full production by 1935.

Although Columbia Records had been recording masters on audiotape since the 1940s, company executives did not foresee audiotape as a threat to their newly released vinyl LPs. The conditions were right for the next round of the Format Wars.

While the 1950s belonged to LPs and 45s, records gave way to tape recordings in the 1960s. Reel-to-reel tape recorders had been used on sound stages and in recording studios ever since 1948, when Bing Crosby premiered an Ampex Model 200 tape recorder to record his radio show. A loyal following of music enthusiasts brought Ampex and Wollensak tape recorders into their homes, but the format was too bulky and cumbersome for most casual listeners.

Ten years later, RCA developed a compact recorder that used ¼-inch magnetic tape that was conveniently preloaded in a plastic cartridge. The reversible, four-track tape provided a typical playtime of 30 minutes per side of stereo sound. Even at this smaller size, compared with the recently introduced pocket transistor radio, the tape recorder was still too large for general use.

In wasn't until Philips launched the Compact Cassette, in 1964, that audiotape became a serious challenger to records. The low cost and high fidelity of portable cassette tapes and the convenient recording option helped to convince music lovers to turn in their turntables for cassette recorders. The 1979 introduction of the Sony Walkman all but sealed the fate of the record player. After reigning over the recording industry for three-quarters of a century, the gramophone-style record became just another victim of the Format Wars.

Videotape

Magnetic tape was not limited to making audio recordings. Soon after Ampex introduced the Model 200 tape recorder, the company modified the machine to record video as well. Once again, Bing Crosby Enterprises (BCE) gave the world's first demonstration of a revolutionary recording technology. On 11 November 1951, the first practical video tape recorder (VTR) played back blurry black-and-white images of what it had recorded.

RCA followed two years later with a color VTR. Although the picture quality was somewhat better, the machine still was not practical. With the tape traveling at 360 inches per second, a 15-minute recording would require a tape reel over 10 feet in diameter.

It wasn't until 1956 that a commercial VTR with high enough quality for broadcast television was marketed. The desk-size Ampex Quadruplex machines used a four-head system that recorded on two-inch tape. The introduction of videotape spelled the end for the film-based kinescopes, which had been used in studios since the inception of television.

Videotape found its way into households in the 1970s. Sony introduced its half-inch Betamax video cassette recorder (VCR) in 1975, setting the stage for what was perhaps the most notorious format war of all time. The next year, rival Japanese companies, led by JVC, launched the Video Home System (VHS), effectively declaring war on Sony. After a decade-long battle, VHS triumphed to control the home entertainment industry. Sony did not

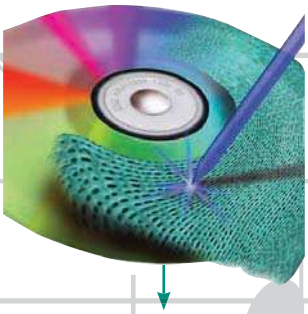
1971
Video Cassette Recorder (VCR): The Sony U-matic VCR, introduced in Tokyo in September 1971, was the world's first commercial videocassette format.

1978
LaserDisc (LD): *Jaws* was released in 1978 as the first movie available on LD. Although Philip's LD format experienced limited success, the technology paved the way for future generations of recording formats.

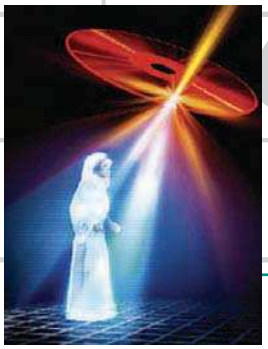
1982
Compact Disc (CD): Philips and Sony rolled out the first CD players in 1982. The first CD recording was ABBA's *The Visitors*

1976
Video Home System (VHS): JVC introduced the VHS in 1976—one year after Sony launched Betamax. The two companies battled for over a decade to decide which format would dominate the home recording market.

1991
DVD: The DVD Consortium avoided the type of conflict that slowed the adoption video cassettes in the 1980s, reaching a compromise between the Philips/Sony partnership and the Toshiba alliance to set the DVD standard.



2006
Blu-ray: Sony's Blu-ray format relied heavily on Hollywood backing to eventually convince Toshiba to drop support for its HD-DVD format.



2010 and beyond
The next generation recording format could be based on florescence, holographs, or 3D optical storage solutions.

surrender until 2002, when the company officially retired Betamax.

Laser Optical Discs

The recording industry endured numerous format changes during the twentieth century—most of them minor, a few revolutionary. But the greatest transition came at the close of the century, when digital technology overtook its analog predecessors. Digital technology had its roots in computer science. The proliferation of personal computers by the 1990s drove demand for a multipurpose storage format that could record audio, video, and data. A new recording medium was needed, and it needed to be digital.

Digital recording technology was not all that new, however. In 1958, David Paul Gregg invented the first laser disc. The new optical storage format was a major departure from the magnetic tape media used at that time for recording audio, video, and computer data.

Nearly twenty years passed before the laser disc made its first commercial appearance. MCA Discovision released the Videodisc player to consumers in 1976. *Jaws* debuted in 1978 as the first movie released in Laserdisc (LD) format.

A decade later, Philips teamed up with Sony to develop a compact disc (CD) format more convenient than the cumbersome 12-inch Laserdisc platter. The partnership ushered in the era of the ubiquitous CD, one of the most successful recording formats of all time.

Sony and Philips set out in the early 1990s to improve the CD by developing a high density version. Their efforts resulted in the release of the MultiMedia Compact Disc (MMCD) format. Meanwhile Toshiba and its allies were developing a Super Density (SD) compact disc. A deal was brokered in 1995 between the competing factions, yielding a new optical disc format—the DVD. That agreement bestowed the lion's share of the standard—and the royalties—on the Toshiba alliance, setting the stage for the coming battle over the high definition optical disc market.

Sony had been stung before, with the release of the Betamax videocassette. So as not to be shut out again, Sony teamed up with Philips to develop the technology for the next generation of media players. Instead of recording and reading data with red wavelength lasers, Sony's researchers turned to shorter, violet wavelengths to create the Professional Disc for DATA system, also known as PDD or ProDATA. The new optical disc system evolved into Blu-ray. Toshiba countered by developing its own "blue"-wavelength technology, the Advanced Optical Disc, which eventually became HD DVD. This time around the two competing standards were not to be compromised.

A surprise challenger climbed into the HD format ring with the Versatile Multilayer Disc (VMD). American startup company New Medium Enterprises (NME) arrived late to the market, but their VMD system offered a low-cost alternative HD format.

Two small European companies, MultiDisc Ltd. in London and TriGM International in Belgium, were working on a different wavelength—literally. Their solution for extending disc storage capacity was to stack more layers of data—possibly as many as 20 of them—instead of shortening the optical wavelength. The innovation led to the development of VMD. In 2004, NME acquired all of the intellectual property assets for VMD. NME launched its line of VMD systems and discs in US markets in September 2007.

NME has carved out its own niche market for HD video, with a targeted audience numbering in the billions. The New York based company secured agreements with major Indian film producers to distribute box office hits from that country's movie capital, commonly referred to as Bollywood. Indian films are particularly popular in much of Asia and the Middle East, accounting for billions of dollars in ticket sales each year. NME entered the HD home entertainment market with a library of VMD movies less than half the size of its competitors, but the company hoped to draw from an extensive archive of Bollywood films and expand on its small but growing number of Hollywood titles.

Although Toshiba withdrew HD DVD from competition in 2008, VMD is not the only alternative optical storage medium to Blu-ray. China announced in 2003 that Beijing E-world Technology, a multicompany partnership, had developed its own optical medium-based digital audio/video format, Enhanced Versatile Disc (EVD). The CD-size medium is physically a DVD disc that can store data in up to three layers. This approach makes it possible for an EVD to hold up to 15 GB on a single

disc, about three times the storage capacity of a DVD.

In December 2006, twenty Chinese firms unveiled 54 prototype EVD players, announcing that they intended to switch to the home-grown format by 2008. Support for the Chinese format fell off sharply when sales failed to materialize. Only a few movies were made available for EVD, and the low-cost players have not been enabled for high definition decoding.

E-World uses a similar approach to optical storage as NME, so EVD movies are compatible with both players. The two companies have been working together since mid-2007 to breathe life into the EVD format. A proposed partnership between E-World and NME could open China's doors to a red-laser HD format, if the deal can be consummated.

The Transition to HD

The world's transition to HD is rapidly gaining momentum. Come 17 February 2009, the Federal Communications Commission (FCC) will bring an end to all analog television transmissions in the United States (more on this topic can be found in "February 17, 2009: A Second Date that Will Live in Infamy?" also in this issue). After that day, over-the-air analog signals will no longer be transmitted.

Over the past decade, broadcast networks have been steadily adding to their fare of digital television (DTV) broadcasts, enticing some discriminating viewers early to the new standard. More HDTV programming, coupled with falling set prices and increased screen sizes, has accelerated HD conversion. By the end of 2007, approximately 43 percent of US households had at least one digital television set, and nearly 75 percent of all digital televisions currently sold are capable of displaying in high definition.

Comparing Optical Disc Technologies

Optical discs are at the heart of the devices that drive two thriving consumer markets: home entertainment and personal computing. As the resolution of the movies we watch and the video games we play increases, so does the demand for higher capacity media discs. Computing power also continues to grow exponentially, prompting the need for more room to store data. The manufacturers of these devices will reign over the next generation of digital entertainment.

The first generation of HD players saw Blu-ray using two layers per side on their discs and VMD four. Blu-ray had the early lead, storing 50 GB per side, compared to 20 GB for VMD. VMD has since doubled its storage capacity—the HD VMD50 stores 48 GB per side. NME is promoting a 10-layer standard, which is planned for release sometime in 2008. With Sony already working on a three-layer specification and NME claiming that a 20-layer disc is possible, 100 GB storage might not be far off. VMD is not limited to a red laser format, however. Applying VMD's multi-layer technology to blue laser technology could yield a 200 GB disc.

Perhaps the biggest advantage for NME to competitively challenge Sony is the low cost of production. VMDs and VMD players can be manufactured using the same equipment and processes that produce inexpensive DVDs. In an attempt to capture market share, Blu-ray started selling at below-cost prices that are dipping under $400. When NME launched VMD in the US, the basic player sold for less than $200. Red-laser discs are also cheaper to manufacture, costing less than half the price of a Blu-ray disc. Additionally, VMD production can take advantage of using existing manufacturing infrastructure. VMD's biggest disadvantage may be that the system is not interactive, while Blu-ray is capable of supporting videogame play and Internet access.

The Future of HD Formats

The question remains whether NME will be able to mount a serious challenge to Sony's Blu-ray. NME may lack the clout to overcome being shutout by the Hollywood movie industry. The company's strategy of targeting markets in India and much of Asia could pay off, however, with the cost advantages of VMD potentially overwhelming Blu-ray, especially in Asian markets. Another possible outcome is for Sony to acquire the advantages of VMD's multi-layer disc technology. This could be done through a partnership with NME, a direct buyout of the company, or Sony independently developing a competing technology.

Another possible scenario is that viewers will prefer to download HD content directly to their media devices, bypassing the home theater player altogether. The role of high-capacity optical discs would then shift to that of a data transfer and storage medium.

There will always be a need for recording information, and the format that provides the most storage is likely to attract the largest market. The HD generation of optical disc formats has multiplied the storage capacity of DVDs, accommodating 50 gigabytes of information and more. But work has already begun on the generation of data storage that will follow. Florescence, holographs, and various 3D optical storage solutions are under development, promising as much as a terabyte of storage on a thumbnail-size chip. The first fruits of these efforts are expected by the end of this decade.

The next round of Format Wars will likely introduce consumers to radical shifts in technology, prefaced by "nano" and "quantum" and "tera". The change may be imperceptible, however—songs and movies and games are already being downloaded from the ether, a trend that is sure to escalate. Recordings stored on the racks of personal libraries will become nostalgic vestiges of a pre-ambient age. But no matter what technological advantages for recording data, the winner of the next round of Format Wars might offer, its success ultimately will depend on its ability to capture our memories, our thoughts, and our dreams.

Cognitive Radio: Tuning In The Future

When you think of radio, you might picture a silver plastic box sprouting an antenna and blaring the latest tune from Nickelback, or maybe it's the stereo mounted in the dashboard of your SUV, or even the pocket transistor you got for your fourteenth birthday. And you would be right. But in the not-too-distant future, the radio you think about could be thinking about you.

Cognitive radio (CR) does more than play music—CR functions as the "brains" behind a host of wireless devices. It knows from moment to moment the most efficient broadcast mode, which devices to communicate with, and what information is needed. The artificial intelligence that makes CR "smart" could provide a solution to the problem of dwindling bandwidth the world faces today and a model for interactive wireless communications that will be part of our everyday lives tomorrow.

Shrinking Bandwidth

Radio frequencies have been used for decades to convey communications—radio and television broadcasts, walkie-talkies, ham radio transmissions. In the early days of radio, only a handful of broadcasts competed for an exclusive frequency within the vast spectrum of radio waves. Today, radio bandwidth is crowded, carrying the signals of countless new technologies—radar, cellular phones, the Internet, navigational systems, security alarms, garage door openers, and even remote-controlled toys.

Theoretically, the spectrum of radio frequencies is unlimited, with wavelengths ranging from just a millimeter short—EHF (extremely high frequency) to those infinitely long—extra low frequency (ELF). But a relatively narrow band of radio frequencies carries the preponderance of communications traffic—traffic volume that is projected to burgeon globally for years to come.

Most nations have set up agencies to assign radio frequencies to various users, parceling out spectrum for public services, military networks, and commercial use. In the United States, the Federal Communications Commission (FCC) has controlled the allocation of radio frequencies since 1934.

But room is running out for allotted radio frequencies to coexist without interference. The shrinking amount of available bandwidth is constraining the growth of some popular technologies and stifling the emergence of new ones. The eminent shortage of radio bandwidth is, in part, what inspired the vision of cognitive radio.

The Evolution of Radio

Radio helped to usher in the Golden Age of Invention at the turn of the 20th century. Late-nineteenth century physicists had been conducting experiments with

wireless signals. The successful outcomes from those experiments led Nikola Tesla to forecast in 1897 the immanent creation of an apparatus that would "…transmit intelligible messages to great distances." Since the time of the first wireless message, radio has evolved from controlled bursts of static, emitted by spark-gap transmitters, to digital signals relaying images from satellites exploring the outer reaches of our solar system and beyond.

Over the ensuing century, methods for radio wave propagation have progressed from damped wave to continuous wave to multiplex transmissions, while radio hardware advanced from vacuum tubes to transistors to microchips. But by the 1990s, further development in signal and hardware technologies confronted new obstacles, when bandwidth seemed to be running out and microchip size was reaching its physical limits. Different strategies for radio were needed that could accommodate new generations of communications technologies. The early vision for radio had been to provide a way to quickly deliver messages over great distances—a sort of wireless "Pony Express." But in the age of World of Warcraft, OnStar, and Facebook, electronic communications are becoming more immersive, interactive, and intimate.

It was nearly 100 years after Tesla provided the first demonstration of wireless communications that visionaries turned their attention from improving radio signals and hardware to better managing them. Radio in the 21st century needed to be more than just a messenger service, carrying information over a set path and delivering it at a specified destination. For radio to become the medium that would manage the conversations, inquiries, and electronic resources for an entire world, it would need to be aware, adaptive, and intelligent. To sustain the complex web of future communications, radios needed to "think," and not just react. The next generation of radios would need to manage their own environments, anticipate network demands, and satisfy the needs of their users. They would be truly "cognitive radios."

The Vision of Cognitive Radio

The phrase cognitive radio was coined in 1998, by Joseph Mitola III, as part of his doctoral research at Stockholm's Royal Institute of Technology. By then, Mitola had spent nearly a decade championing software radio, also named by him.

Software radio had been paving the path to CR, through the introduction of new generations of smaller and more powerful wireless devices that had become software capable and reprogrammable. Wireless devices quickly evolved to the point that their functions were fully controlled by software, introducing an era of software defined radio (SDR). The emergence of SDR marked a first step on the path to cognitive radio.

Mitola proposed his vision for a new kind of SDR—an "ideal cognitive radio (iCR)"— as part of his doctoral thesis. He described this CR in an interview for The Great Minds, Great Ideas Project as "really smart radio that would be self-aware, RF-aware, user-aware, and that would include language technology and machine vision along with a lot of high-fidelity knowledge of the radio environment."

Mitola proposed his vision for cognitive radio as how he imagined CR might evolve, perhaps 20 years in the future, to include autonomous machine learning, vision that exceeds that of a camera, and spoken or written language perception. To

Mitola attributed CR with seven capabilities

Sensing:
Reckoning RF, audio, video, temperature, acceleration, location, and more;

Perceiving:
Determining what is in the "scene" monitored by the sensor domains;

Orienting:
Assessing the situation to determine if it is familiar and reacting immediately if necessary;

Planning:
Identifying alternative actions to take on a deliberative time line;

Making Decisions:
Deciding among the candidate actions to choose the best action;

Taking Action:
Exerting effects in the environment, including RF and human-machine communications; and

Learning Autonomously:
Learning from experience gained from the first six capabilities.

Mitola's surprise, not only did the FCC adopt the proposed technology before he completed his dissertation, but "cognitive radio" came to be called alternatively "Mitola radio."

Cognitive radio, then, represents "intelligent" radio. But intelligence is measured across a continuum, exhibiting differing expressions and levels of intelligence. For cognitive radio, the path to "ideal" intelligence is demarked by three plateaus: (1) awareness, (2) adaptiveness, and (3) cognition. The starting point on this path is software defined radio.

Software Defined Radio

Cognitive radio rests on a foundation of the software that controls it. Over the decades since the introduction of software radio, programs controlling wireless devices evolved from basic frequency hopping and automatic linking to include networking capabilities, multiple waveform interoperability, and field upgradeability. The development of devices fully controlled by software marked the emergence of SDR.

An SDR typically has a radio frequency (RF) front end with a software-controlled tuner. The concept behind SDR, which Mitola proposed in 1991, constitutes a shift in emphasis for radio performance—whereas digital radios of the 1990s derived about 80 percent of their functionality from hardware, Mitola envisioned multiband, multimode radios with at least 80 percent of their functionality provided in software. Such radios would be able to switch to available RF spectrum, implement different waveforms, or comply with pre-defined policies while running on general-purpose hardware.

SDR is more versatile than standard radio, because it is not constrained to a small band of radio spectrum. Instead of using a hardware tuner to hone in on a specifically allocated frequency among a small licensed band, software controls programmable hardware to tune an SDR to any frequency across a broad spectrum. Software can also be encoded to enable an SDR device to implement any waveform, further extending the agility of SDR. An SDR device can also be programmed to adhere to policies that define protocols for prioritizing access to controlled bandwidth.

An SDR device has the potential to serve multiple purposes and access various parts of the spectrum, simply by loading new software. Software for an SDR can be used to change transmitter, receiver, and antenna characteristics as well as transmission protocols. Well optimized software can also implement forward error correcting codes and source coding of voice, video, or data. Programming modifications to the software can be made remotely and rapidly, providing a single SDR with the flexibility to handle a variety of communications uses.

Although each software-defined radio needs an individual set of instructions to work from, a network of SDRs can still be more efficient than traditional radios, over the long term. After the Department of Defense turned to the Joint Tactical Radio System (JTRS) program to replace existing tactical radios with SDRs, the military was able to reduce the number of radios in its network by more than 75 percent. Through "Jitters," as the program is colloquially called, radio inventory is scheduled to be reduced from 750,000 legacy units to 180,000 SDRs.

SDR has led to the creation of numerous commercial communications products, as well. "Universal radio" devices, designed to integrate various media functions, have steadily grown in popularity since the early 1990s, following SDR's introduction. Today, SDR gadgets abound, competing to serve as cell phone, pager, PDA, multimedia player, and game machine—often all within the same unit.

Radio Awareness

Whereas conventional radio passively responds to signals, cognitive radio is actively aware—it is self aware, user aware, RF aware, and situation aware. A CR device can discern its own position, the locations and actions of its users, the radio frequencies that are available to it, and the general surrounding conditions.

Self aware: A cognitive radio must know its position in space and time to be able to establish its network relationships. Geolocation devices are commonplace for automotive and marine navigation, and location-based services can also be used to locate cell phone users. As satellite global positioning systems such as GPS grow more sophisticated, they are able to inform CR devices of their geographic position anywhere in the world, to within several feet.

Home- and office-based communications as well as point-of-service solutions depend on more accurate units of measurement than even GPS can provide. Indoor communications also require tracking even when a device is not in the satellite's line-of-site. For product messages that "find" passing customers or misplaced keys that announce their whereabouts, Bluetooth and RFID (radio frequency identification) signals can pinpoint locations to within inches.

A cognitive radio is also aware of its own specifications, providing "plug-and-play" capability. A radio can be built from a variety of components. Each module in a radio assembly is automatically configured to function properly in the CR device.

Self-awareness extends to knowing how to operate, as well. For a cognitive radio to comply with regulations that apply for a certain location or situation, policies specifying constraints on radio operation can be built into the radio or accessed over the network. By conferring with configuration databases that carry radio protocols, the CR device would remain within physical and regulatory limits.

User aware: A cognitive radio needs to be aware of more than just the user's location—it needs to associate specific locations and situations with user preferences. A cognitive radio could rely on biometric input to identify a user and assign appropriate authorizations for access. Based on behavioral models learned over time for that particular user, the device would be able to associate concepts such as "home" or "bank" and even anticipate the user's needs as they related to the location, time, or current circumstances.

RF aware: As an outgrowth of SDR, cognitive radio is aware of available radio frequencies—which frequencies are in use and when they are free, where transmitters and receivers are located, and what signal modulation is employed.

A cognitive radio is also on the lookout for other cognitive radios. CRs accessible within the immediate area can collaborate across their wireless network to share tasks and information.

Situation aware: Cognitive radios are aware, moment to moment, of any possible situation of time, position, signal, propagation, and other relevant conditions. By programming in self-assessment capabilities, a CR device does not have to be deconflicted in advance.

Adaptive Radio

Awareness is only the first plateau for intelligent communications. For radio to evolve as cognitive, it must also qualify as adaptive. Software Defined Radio is able to perform multiple roles, based on the purpose of the software. But SDR is still dependent on the set of instructions it is given; it is not able to adapt to changing conditions. The proliferation of communication technologies has given rise to environments teaming with electronic messages. Coordinating the operations of a fluctuating number of RF devices within a dynamic RF environment poses a formidable, if not impossible challenge.

Understanding its environment is only the starting point for a cognitive radio. As technologies mature, communications devices are becoming able to adapt to their environments as well. Once a mobile phone recognizes that its current radio frequency is fading or no longer available, it must automatically seek out an alternate frequency that can sustain a signal, and then switch to it. Radio that is adaptive as well as aware defines a plateau on the ascent to ideal cognitive radio.

Three main approaches have been taken to untangle the web of RF signals spun by competing devices: frequency coexistence, time coexistence, and interference coexistence. For an RF device to adapt to its environment, it can (1) move to a different frequency, (2) wait its turn on a given frequency, or (3) limit interference by constraining its power or reducing proximity with other devices.

When a frequency is no longer available, as when the licensed user reappropriates it, a CR can be assigned a different frequency or wait to fill gaps in the transmission. Controlling interference among signals from various RF devices poses a trickier challenge, however. The two options are to limit the signal strength of the competing devices or increase the distances between the components.

Dynamic Spectrum Access

Software defined radio exhibits a kind of static intelligence—its adaptive behavior is hard coded and, as such, inflexible. The next development towards more intelligent radio calls for devices that adapt dynamically to their environment. A relatively new application stands at the forefront among the latest radio technologies to deliver this higher level of radio intelligence—dynamic spectrum access (DSA). DSA is viewed by many as a potential solution for the immediate challenges

	Software Capable	Software Programmed	Software Defined	Aware	Adaptive	Cognitive
Experiment with different settings						
Learn about the environment						
Modify radio parameters in response to inputs						
Measure quality of service (QoS)						
Full software control of functions						
Upgradeable in the field						
Multiple waveform interoperability						
Networking capabilities						
Programmable cryptography						
Automatic linking						
Frequency hopping						

Figure 1: Six Plateaus to iCR

that face network spectrum resources. Although DSA is actually a static application, it provides for dynamic, flexible, and autonomous spectrum access. The wireless networking architectures and technologies underlying DSA enable devices to dynamically adapt their spectrum access according to policy constraints, spectrum availability, propagation environment, application performance requirements, and other criteria.

The spectrum-dependent devices in a DSA system can dynamically change their parameters to access multiple dimensions of the spectrum resource including frequency, space, time, and signal codes. The agility and enhanced distribution of spectrum data provided by DSA should enable spectrum-dependent systems to share spectrum resources in near-real time among a large number of users.

Thomas Taylor, representing the Office of the Assistant Secretary of Defense, likens the transformation from current static spectrum allocation to DSA with the transition in the 1980s from circuit-switched to packet-switched networking. The gains in efficiency and improvements in interoperability should also be comparable.

"NeXt Generation" Radio and Beyond

Cognitive radio is slowly moving from theory to reality. To stimulate research that will lead to adaptive radio, the Defense Advance Research Projects Agency (DARPA) sponsored the neXt Generation (XG) communications program. XG builds on the SDR capabilities developed for JTRS by adding DSA capability.

DSA can be classified as either (1) coordinated—requiring a spectrum control and management infrastructure, or (2) opportunistic—where a group of devices autonomously sense the environment and access spectrum according to established policies. It is the opportunistic approach to DSA that has driven the XG program.

The XG program, which was launched in 2002, is funded by DARPA and managed by the Air Force Research Laboratory (AFRL). Equipped with a neXt Generation system of radios, the DoD planned to provide future warfighters with assured military communications, no matter where in the world they were deployed. The expressed goal of the XG program was to develop the technology and system concepts to access 10 times more spectrum with near-zero setup time; simplify RF spectrum planning, management, and coordination; and automatically deconflict operational spectrum usage. The XG radio—a computer with a wideband RF front-end—is smart enough to sense its signal environment, understand the spectrum rules that apply at its location, and take action based on those rules.

Shared Spectrum Company (SSC), currently the prime contractor under the XG program, successfully carried out the world's first successful test of DSA in a live demonstration, held at Fort A.P. Hill, Virginia, in April 2006. The successful demonstration of a DSA-capable system has set the stage for propelling the XG program to the next level.

For XG radio to be fully deployed, the system must be robust as well as affordable. In pursuit of these goals, DARPA embarked on the Wireless Network after Next (WNaN) program in 2005, with the intention of transitioning WNaN technology to the Army in 2010.

The goal of fully implementing WNaN depends on the success of the Adaptive Network Development (WAND) effort. WAND was set up by DARPA to develop the technology for establishing densely connected networks of inexpensive wireless nodes. To meet the DoD's requirements, WAND-enabled networks will need to be ultra-large (tens of thousands of nodes), highly-scalable, and highly adaptive. Ultimately, this richly interconnected fabric will provide superior battlefield communications at lower system cost.

Approaching Cognitive Radio

Although radio communications technology is becoming truly adaptive, cognitive radio remains a vision for the future. CR integrates the performance of software radio with the intuitiveness of machine intelligence. Through CR, many single-purpose gadgets connected to the Internet converge into a "personal digital assistant" that anticipates the user's needs and adapts to them.

Designers of next-generation mobile devices are already taking advantage of the processing resources in SDR platforms. Early adaptations to CR have produced mobile devices with a broad range of capabilities that target diverse world markets. Commercial applications of DSA are certain to soon follow. This piecemeal implementation of CR has led to some confusion in characterizing cognitive radio, with some functions narrowly defined for application-specific devices.

Improvements in network speed are propelling the development of cognitive radio. Third-generation (3G) communication systems are currently capable of high-speed Internet access, video downloads, and location-based services. Broadband networks such as Code Division Multiple Access–Evolution-Data Optimized (CDMA-EVDO) and High-Speed Downlink Packet Access (HSDPA), commonly referred to as 3.5G networks, are transmitting data at rates four-to-five times faster than current 3G networks. And trials for 4G networks are already underway in some countries, promising a further ten-fold increase in transmission speeds by the end of this decade.

The scheduled release and reassignment of RF spectrum has also spurred efforts to redefine radio communications, and hastening the transition to cognitive radio. The FCC is scheduled to reallocate radio bandwidth repossessed from VHF/UHF television bands no longer used after the mandatory conversion to digital television in February 2009. (More on this topic can be found in "February 17, 2009: A Second Date that Will Live in Infamy?" also in this issue.)

To prepare for the transition of some of these channels, to support dynamically assigned frequencies the Institute of Electrical and Electronics Engineers (IEEE) is preparing standard 802.22. The working group on Wireless Regional Area Networks (WRANs) is charged with develop-

ing a standard for "a cognitive radio-based PHY/MAC/air_interface for use by license-exempt devices on a non-interfering basis in spectrum that is allocated to the TV Broadcast Service." Protocols in the standard are intended to protect licensed incumbent services in the TV broadcast bands from harmful interference the new service might cause. The stated goal of IEEE is "to equal or exceed the quality of DSL or cable modem services, and to be able to provide that service in areas where wireline service is economically infeasible, due to the distance between potential users."

With the looming potential for blistering network speeds and nearly limitless bandwidth, industry is lining up to provide the hardware and software that will drive what could become a CR revolution. The first commercial cognitive radio certified by the FCC was introduced in 2006 by Florida-based Adapt4. Adapt4's XG1 cognitive radios can use up to 45 radio channels simultaneously, with each channel able to change frequency instantly to avoid interference. A host of other XG devices are sure to quickly follow.

Embracing Cognitive Radio

Cognitive radio is poised to be the platform for next-generation communications. SDR's ability to tap radio spectrum that is only temporarily available makes the architecture's adoption especially enticing. However, not everyone is eager for opening up the airways to SDR. Companies that already have secured expensive licenses for assigned bandwidth would, naturally, prefer to retain exclusive rights to those frequencies. The emergence of cognitive radio might be stalled by regulatory wrangling, but it is not likely to be halted.

Cognitive radio could more easily be applied first for non-commercial uses. Military organizations and emergency response teams are not subject to licensing agreements. Early versions of CR networks are already being established in these limited environments of unquestionable priority, with little opposition.

First responders are likely to be among the early adopters of CR. Disasters can easily disable communication systems by severing cables and downing cell towers. CR would ensure uninterrupted communications despite the flood of urgent mobile transmissions necessary for managing a crisis. Public safety teams could be given precedence on congested frequencies, with critical-care providers receiving top priority.

CR systems for emergency response teams are already becoming available. In August 2007, Spectrum Signal Processing by Vecima, which also contributed research for the JTRS program, began shipping an integrated wireless communications subsystem designed for the quick deployment of communications services for first responders. Spectrum's SDR architecture is designed to support full CR functions.

Only radio that is both aware and adaptive has the potential to be truly cognitive. Ultimately, iCR devices will anticipate changes in location, user preferences, and RF availability. And they learn to do so through their own experiences.

The speed and convenience of a growing number of radio devices coupled with steadily declining prices has begun to overwhelm crowded airways. The additional services cognitive radio can provide are sure to add to the problem of bandwidth congestion. But CR also promises a solution for that problem.

Like many disruptive technologies before it, cognitive radio might advance haltingly. Some people are sure to greet the new opportunities with enthusiasm. Others will fail to adapt and be left behind, wondering what happened. Only a few among the previous generation could envision how personal computers and the Internet would change everyday life. That experience may have prepared this generation to think seriously about the coming of cognitive radio.

Resources

T. Charles Clancy (2006). "Dynamic Spectrum Access in Cognitive Radio Networks," [Dissertation, submitted to the Faculty of the Graduate School of the University of Maryland, College Park in partial fulfillment of the requirements for the degree of Doctor of Philosophy].

J. Mitola III (2006), Cognitive Radio Architecture: The Engineering Foundations of Radio XML. John Wiley & Sons, Inc., Hoboken, New Jersey.

Thomas J. Taylor, Office of the Assistant Secretary of Defense.(2007) Managing the Air Waves: Dynamic Spectrum Access and the Transformation of DoD Spectrum Management . Crosstalk: The Journal of Defense Software Engineering, July 2007 issue.

i039211