~Judith A. Markowitz

VOICE BIOMETRICS

Who are you? Your voice alone can be used to verify your personal identity —unobtrusively and invisibly.

"IT'S ME!"

This pronouncement is usually made over the telephone or at an entryway out of sight of the intended hearer. It embodies the expectation that the sound of one's voice is sufficient for the hearer to recognize the speaker. In short, "It's me!" is the original real-world, speaker-recognition challenge.

It is possible today to automate a growing number of speaker-recognition tasks with such technologies as speaker verification and speaker identification. Like human listeners, voice biometrics[1] use the features of a person's voice to ascertain the speaker's identity. Systems performing this function have been applied to real-world security applications for more than a decade. Their use is increasing rapidly in a broad spectrum of industries, including financial services, retail, corrections, even entertainment. Here, I provide an overview of speaker verification and speaker identification, focusing on deployed, real-world technologies and the types of applications being used today.

Voice-biometrics systems can be categorized as belonging in two industries: speech processing and biometric security (see Figure 1). This dual parentage has strongly influenced how voice-biometrics tools operate in the real world.

Speech processing. Like other speech-processing tools, voice biometrics extract information from the stream of speech to accomplish their work. They can be configured to operate on many of the same acoustic parameters as their closest speech-processing relative—speech recognition. And like speech recognition, they benefit from lots of data, good microphones, and noise cancellation software. Voice biometrics are vulnerable to some of the same conditions that cause speech-recognition systems to perform poorly: background and channel noise; variable and inferior microphones and telephones; and extreme hoarseness, fatigue, or vocal stress.

There are also important differences between voice-biometrics systems and other speech-processing technologies, including speech recognition. The most significant is that voice biometrics technologies do not know what a person is saying, relying on speech recognition to do that. Moreover, the trend toward speaker independence that characterizes speech recognition cannot exist for voice biometrics. By definition, voice biometrics are always linked to a particular speaker. As

[1]Speech-processing researchers prefer the term "speaker recognition." People outside the speech-processing industry often confuse it with "speech recognition," which refers to a speech-processing technology that recognizes what a person is saying. This confusion has led to some use of speech-recognition tools for security applications beyond the abilities of speech-recognition technology. The result is a weak form of password or passcode security. "Voice recognition," another confusing term, is often used to refer to speech recognition but misleadingly suggests some speaker identification is involved.

a result, they require some type of enrollment for each user. The need for enrollment is an attribute voice biometrics shares with its relatives in the biometric-security industry.

Biometric security. Membership in the biometrics industry influences how voice-biometrics systems are used. Biometrics-based technologies are applied most often in security, monitoring, and fraud prevention where they positively identify individuals and distinguish one person from another. These abilities differentiate biometrics from all other forms of automated security. A card system can, at best, determine only whether a person has a viable access card, and password security can determine only whether the person knows the proper password. None of them verify that the person presenting the card or entering the password is the individual authorized to do so.

Biometric systems determine whether a biometric sample, such as a fingerprint or spoken password, comes from a specific individual by comparing that sample with a reference biometric—a sample of the same type of biometric provided by the individual in question. Developers of voice biometrics called this a "reference voiceprint." As with reference templates for other biometrics, reference voiceprints are evaluated in terms of the number of times they mistakenly accept a false claim of identity as a legitimate claim and the number of times they reject a legitimate speaker as an impostor.

The most significant difference between voice biometrics and other biometrics is that voice biometrics are the only commercial biometrics that process acoustic information. Most other biometrics are image-based. Another important difference is that most commercial voice biometrics systems are designed for use with virtually any standard telephone on public telephone networks. The ability to work with standard telephone equipment makes it possible to support broad-based deployments of voice biometrics applications in a variety of settings. In contrast, most other biometrics require proprietary hardware, such as the vendor's fingerprint sensor or iris-scanning equipment. This distinction—standard versus proprietary input device—is beginning to disappear. The recent development of inexpensive, high-quality cameras, for example, now enables wider deployment of face-recognition applications.

RICK STROMOSKI

Types of Voice Biometrics

The following sections outline the best-known commercialized forms of voice biometrics: speaker verification and speaker identification.

Speaker verification. Speaker-verification systems authenticate that a person is who she or he claims to be. If, for example, the person speaking at the beginning of this article had shouted, "It's Julie," rather than, "It's me!," the intended hearer would have to perform speaker verification based on that identity claim.

Figure 2 shows a typical speaker-verification process. It begins with a claim of identity, such as entering an account number or ID code, presenting a credit card or ATM card, or allowing the system to access the ID of a user-linked input device, such as a specific cellular telephone. Typically, a person enters the ID manually using a telephone keypad or keyboard. If the system uses speech recognition in conjunction with speaker verification, it may request verbal input of an ID code.

The system uses the ID to retrieve the reference voiceprint for the person authorized to use the ID, as in Figure 2. It then requests a sample of speech from

Figure 1. The voice-biometrics family tree.

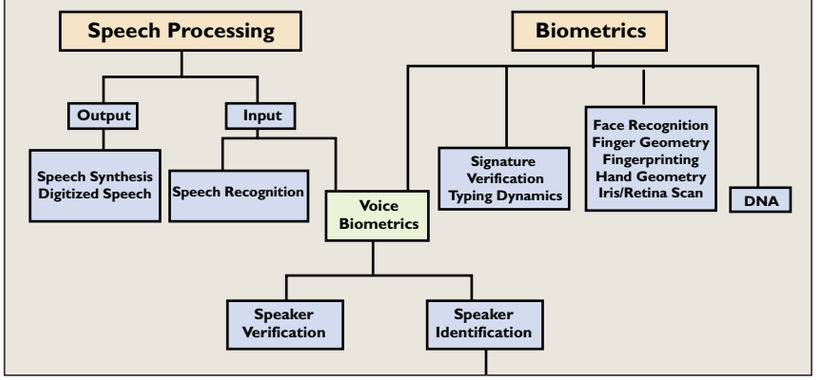
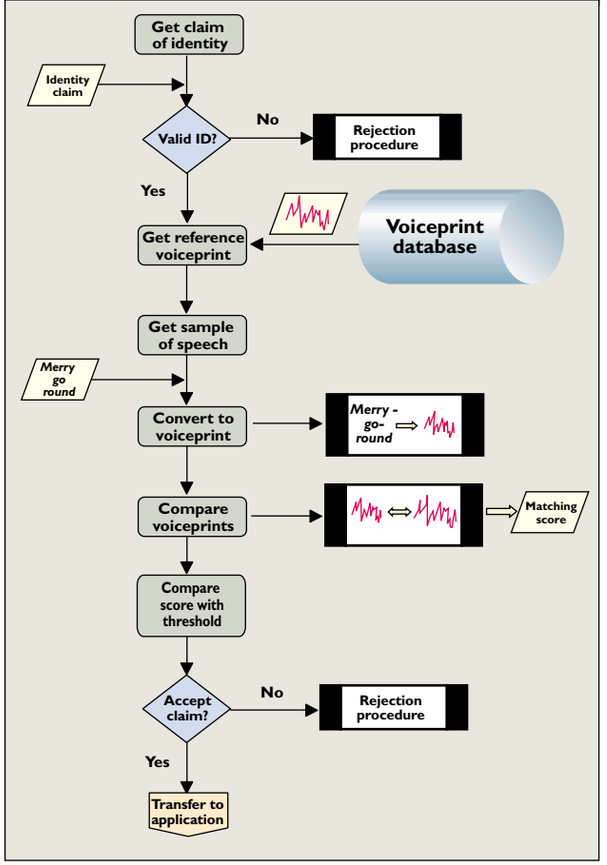


Figure 2. Speaker verification.



the claimant. The newly input speech is converted into a voiceprint and compared to the reference voiceprint. The results of the comparison are quantified and compared to an acceptance/rejection threshold to determine whether the two voiceprints are similar enough for the system to accept the identity claim.

Figure 3 shows several ways of interacting with speaker-verification systems. Most commercial systems are text-dependent. They request a password, account number, or some other prearranged code. Because it requests a password, the system in Figure 3a is text-dependent. Text-dependent systems provide what the data-security industry calls "strong authentication."[2] Strong authentication requires the use of at least two different kinds of security. In the case of text-dependent speaker verification, the person must have the correct voice (an example of "Who you are" security) and also know the proper password (an example of "What you know" security).

The system in Figure 3b displays a text-dependent, voice-only approach that uses the account number as both identity claim and password. Speech recognition decodes the input, and speaker verification uses the same input as the biometric sample it compares to the reference voiceprint.

Figure 3c shows an example of "text-prompted" technology.[3] Text-prompted systems ask the speaker to repeat a series of randomly selected digit strings, word sequences, or phrases. Text prompting requires longer enrollment than text-dependent technology, because the reference voiceprint it generates must contain all the components that will be used to construct challenge-response variants. As Figure 3c indicates, verification also takes longer.

Text prompted verification is well-suited to high-security and high-risk systems, such as those used to monitor felons in home-incarceration and community-release programs. Text prompting is also useful when there is legitimate concern about the use of sophisticated recordings of impostors, because the responses requested from the user are selected randomly, making it difficult to create and play a tape recording with the requested items in the proper sequence.

Relatively few applications require text prompting, because it's fairly difficult to defeat a good commercial speaker-verification system with a recording. The voice signal input into a microphone or telephone held close to the speaker's mouth differs markedly from a signal captured even as close as a foot away from the speaker. Moreover, many commercial speaker-verification systems look for telltale auditory signals, distortions, exact matches, and other indications that a recording has been used. As a result, creating a recording that can fool

[2]The three basic classes of security are: what you have (such as a key, a card, or a token); what you know (such as a password or a PIN); and who you are (biometrics).
[3]Vendors have begun using the term "challenge-response" to refer to these systems.

Figure 3. Types of speaker verification.

Step 1: System: *Please enter your account number using your touch-tone keypad*
Caller: **235167**
Step 2: System: *Please say your password*
Caller: ***Merry go round***
System: *Thank you.*

(a) Text-dependent verification.

System: *Please say your account number*
Caller: ***235167***
System: *Thank you.*

(b) Text-dependent verification with speech recognition.

Step 1: System: *Please enter your account number using your touch-tone keypad*
Caller: ***235167***
Step 2: System: *Please say 42-69*
Caller: ***42-69***
System: *Please say 83-24*
Caller: ***83-24***
System: *Please say 99-48*
Caller: ***99-48***
System: *Thank you.*

(c) Text-prompted verification.

Call Center Agent with caller ID: *What can I do for you, Ms. Jones?*
Caller: ***I want to transfer $100,000 from my savings account with you to an offshore account in Bimini that I have just opened.***
Agent: *Let me look into that. Please wait while I am getting the information*
PAUSE while the system is verifying the identity of the caller
Agent: *Thank you for your patience. I will process that request now.*

(d) Text-independent verification.

these systems is a difficult and costly challenge.

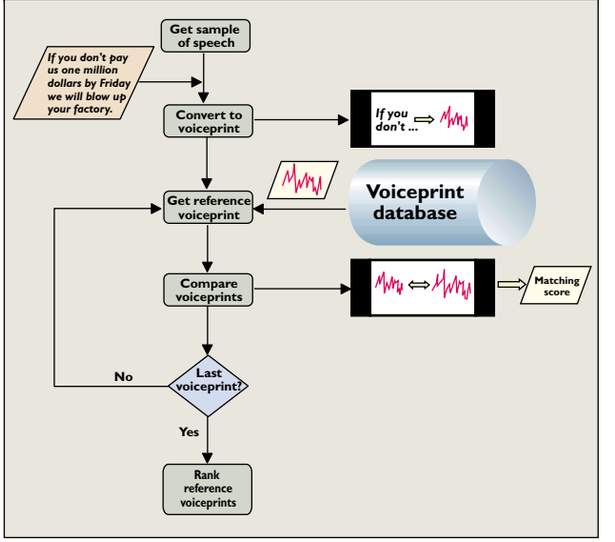
Text-independent verification accepts any spoken input, making it possible to design unobtrusive, even invisible, verification applications that examine the ongoing speech of an individual. The ability of text-independent technology to operate unobtrusively and in the background makes it attractive for customer-related applications, because customers need not pause for a security check before moving on to their primary business objective. For example, the text-independent speaker verification in Figure 3d is invoked by an agent in a bank's customer-service call center. The system verifies that the caller—who has requested a large transaction—is indeed the authorized customer. Another application of this technology, now under development, is continuous speaker verification over wireless mobile telephones [1].

Text-independent technology is much more difficult to implement than text-dependent or text-prompted technology. It requires longer samples of speech and is more sensitive to the acoustic quality of the input. Another potential concern centers on privacy. The application's ability to operate in the background gives text-independent technology the potential for being used without the subject's knowledge.

Speaker identification. Speaker identification[4] assigns an identity to the voice of an unknown speaker. The assertion "Its me!" requires speaker identification, because the intended hearer is expected to assign the proper identity to the speaker based on the voice alone. In most cases, speaker identification is more difficult than speaker verification, because it involves multiple comparisons of utterances that are likely to be different from each other and may not have been recorded with comparable equipment. In forensic and intelligence applications, for example, the stored samples may have been acquired by hidden equipment at a distance from the speaker in a noisy environment. When a suspect is incarcerated, forensic speaker identification may be easier to perform if the court permits multiple recordings to be made of the suspect's voice using equipment comparable to that used when the crime was committed.

Unlike speaker verification, speaker identification does not expect to receive a claim of identity. Processing begins when a sample of speech from of an unknown speaker is presented to the system (see Figure 4); the sample may be live or recorded. It is almost always text-independent and in some applications, such as those in law-enforcement, the speaker may not know the sample was even taken. The system then converts the sample into a voiceprint and systematically compares the new voiceprint with all or with a specified subset of the system's reference voiceprints.

Figure 4. Speaker identification.



[4]Speaker identification is sometimes called "speaker recognition." This lack of precision is unfortunate, because the term also refers to the entire class of "voice-biometrics." The resulting ambiguity is another reason I prefer the term "voice biometrics" for referring to the class of speaker-identity technologies.

Table 1. Example deployed applications.

Function	Application Type	Example
Security	Data and data networks	**BMC Software.** Password reset (over the telephone) using virtual help desk.
		Illinois Department of Revenue. Off-site access to secure data networks.
		INTRUST Bank. Internal wire transfers
	Physical/site access	**U.S. Immigration and Naturalization Service.** Entry to U.S. and Canada during off hours; port of entry at Scobey, Mont.
		Girl Tech. Door access control system and locked box for children.
		City of Baltimore. Evening and weekend access to the five main city buildings.
Fraud Prevention	Telephone network security (toll fraud)	**University of Maryland, College Park.** Toll-free long-distance lines for faculty and staff.
		GTE TSI. Integration of speaker verification into wireless security package offered to carriers.
	Transaction security	**Home Shopping Network.** Automated product-ordering over the telephone.
		Glenview State Bank. Transfer of money between accounts of a bank customer.
Monitoring	Time and attendance monitoring	**SOC Credit Union.** Time and attendance of part-time employees.
		Salvation Army. Time and attendance of Salvation Army workers.
	Corrections monitoring	**New York City Department of Probation.** Tracking of juvenile and adult probationers.
		Dane County Jail, Madison, Wisc. Monitoring of home-incarcerated offenders.

The system in Figure 4 was configured to rank the reference voiceprints in terms of how likely they are to contain the voice of the person generating the sample. Other systems select one or two potential identities from among the reference voiceprints. Some implementations allow the system to report that the voice does not match any of the reference voiceprints in its database. This strategy has been applied to controlling cellular toll fraud when more than one person may be authorized to place calls from a specified phone. If the voice sample supplied by the caller fails to match the set of reference voiceprints associated with a mobile phone, the caller is not allowed to place calls with that phone.

Enhancing performance. As in speech recognition, the performance of voice-biometrics systems is adversely affected by noise in the telephone channel and by other acoustic variability. Developers of speech-recognition tools can build models of network noise into complex, data-rich speaker-independent word models. By contrast, voice biometrics work with speaker-dependent models created from a limited amount of data spoken on a particular telephone.

Cross-channel and cross-device mismatches arise when a person enrolls on one device or network (such as a high-quality wireline telephone in an office) and attempts to voice-verify on a device or network with markedly different acoustic properties, such as an inexpensive wireless telephone.

The most common approaches to attenuating noise, as well as channel and device mismatches, are voiceprint adaptation, cohort modeling, and world models. Voiceprint adaptation involves modification of the original enrolled voiceprint to incorporate data from successful verifications into the statistics of the original model. Adaptation makes it possible to include acoustic information about the range of devices a person uses for verification into the voiceprint. Adaptation can also update the voiceprint with regard to variations in the person's voice. For example, a person may speak quite differently when under stress or when tired.

Cohort modeling and world models are performance-enhancement techniques used in speaker verification. Systems that use cohort modeling identify individuals whose voices are similar to the voice of a newly enrolled user. During verification, the system compares the new input to each of the cohorts, as well as to the voiceprint of the person whose identity is being claimed. A world model is a group model containing a spectrum of voices. Systems using world models perform only two comparisons: one with the voiceprint for the claimed identity and one with the world model. The basic assumption underlying the use of cohorts and world models is that new input from authorized users will match their reference voiceprints better than they match the voiceprints of other people—even in adverse conditions.

Commercial Applications and Trends

Most commercial applications of voice biometrics provide security, fraud prevention, or monitoring; see Table 1 for a partial list of deployed applications in these three areas. The applications described in the following sections reflect the dominant trends in commercial deployment; the majority are speaker-verification applications, most of which use text-dependent technology. They also reflect the diversity and creativity being applied to real-world implementations of voice biometrics.

Data security (Illinois Department of Revenue). The

Illinois Department of Revenue (IDOR) is the taxing body for the state of Illinois, collecting tax and other financial information from individuals and businesses and storing it in databases protected by several levels of security. One of IDOR's responsibilities is to perform tax audits of Illinois businesses and out-of-state companies doing business in Illinois. Audits are done on the premises of the audited business, which can be as far away as California and Florida. In the course of performing an audit, an auditor may need information from the IDOR databases. Before speaker verification was installed, the auditor called a supervisor who had an authorized person transfer the data to a disk and send it to the auditor. By the time the auditor received it, the information could be as much as 10 days old.

In 1994, IDOR installed a text-dependent speaker-verification system, configuring it to require anyone trying to access its computer network to voice-verify over a telephone line before being connected to a data line [3, 7]. The system has been in continuous operation since that time and expanded to 664 users, including IDOR managers, programmers, and selected individuals from other government agencies, such as the Illinois comptroller's office, the Illinois secretary of state, and the U.S. Internal Revenue Service. It allows managers to take laptops to meetings while maintaining email contact with people at the agency. It also enables programmers on call to work from home during emergencies after work hours or on weekends. Prior to the installation of speaker-verification, some programmers would have to drive up to 30 miles to IDOR facilities to handle emergencies.

IDOR officials report they have seen no evidence that any unauthorized person has gotten into its databases since speaker verification was installed.

Physical access (Girl Tech, Inc.). Girl Tech develops and sells products and services reflecting the play preferences of pre-teen girls [6] and is committed to making technology more accessible to them. The main constraints on integration of advanced technology into Girl Tech products are cost, size, and ease-of-use.

Privacy is an important issue for pre-teen and early-teen girls concerned about siblings entering their rooms to read their diaries or borrow their things. Addressing the need for privacy, Girl Tech incorporated chip-based text-dependent speaker verification into its Door Pass and Password Journal products. Door Pass is a brightly colored plastic device that attaches to a bedroom door. A child activates its motion sensor by pressing the on/off button; whenever the door moves, Door Pass demands the password. If the proper password is not supplied in the correct voice, Door Pass registers an intruder and sounds an alarm. When the child returns, Door Pass welcomes her and reports the number of intruders it foiled during her absence.

Future applications will be text-independent and combine voice biometrics with other speech and biometric technologies.

Password Journal is a password-protected plastic box that stores a diary or other personal items. Anyone seeking to open Password Journal must say the correct password in the proper voice. Like Door Pass, Password Journal reports the number of intruders attempting to open it.

Transaction security (Home Shopping Network). Home Shopping Network (HSN), a division of USA Networks, Inc., pioneered the electronic retailing industry in 1977. Its 24-hour-a-day programming reaches 74 million households through broadcasts, cable, and satellite dishes. It has 3.8 million active customers (called "members") and enrolls approximately 7,000 new members a day [3, 5]. Members buy items by calling one of HSN's toll-free telephone numbers. Orders can be placed with an agent or through an interactive voice-response system using touch-tone input. The company's call volume and 24-hour-a-day programming make automated transactions a financial and customer-service necessity. However, HSN has two main reservations about the touch-tone system. Almost 30% of its members cannot use it, because they do not have touch-tone telephones. It is also vulnerable to fraud perpetrated by anyone who knows a member's identification codes.

Starting in 1999, HSN began deploying a hands-free, member-authentication system on its 800-num-

Home Shopping Network is converting its touch-tone order-entry operation to speech recognition.

bers. Now, when a member calls, the system answers with "Welcome to Home Shopping Network. Please speak your telephone number, starting with the area code." Speech recognition decodes the spoken input. If the telephone number is not registered already, the member is transferred to the enrollment procedure. If more than one person is enrolled for that number, the system uses speaker identification to determine which, if any, of the members associated with the number is calling. Otherwise, it uses speaker verification. Callers who are voice-verified successfully are transferred to the interactive voice-response product-ordering module. When verification fails, the caller is transferred to an operator.

By the end of June 2000, HSN had installed the system on 9 of its 10 800-numbers and expected to complete work on the main 800-number soon after. More than 450,000 members were enrolled, and verification was reported to be performing at a 98% rate.

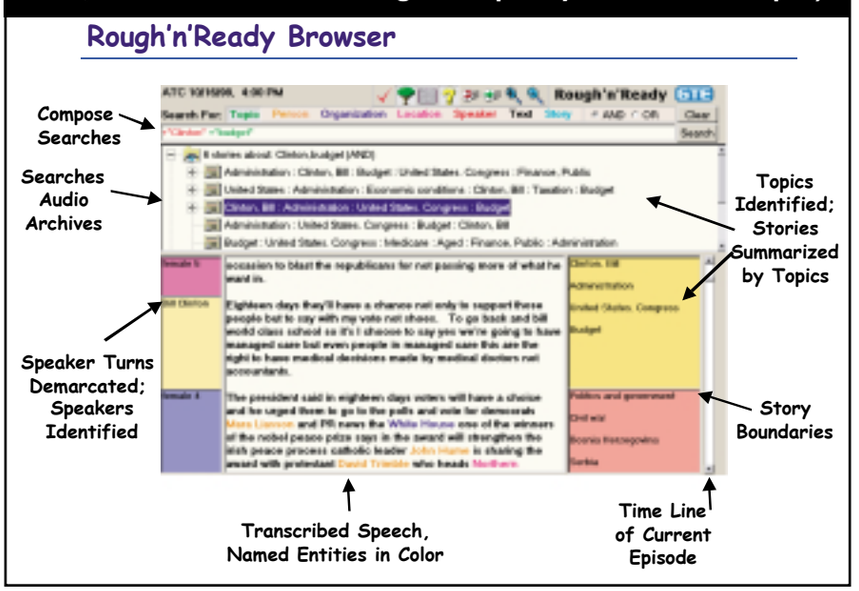
Corrections monitoring (New York City Department of Probation). The New York City Department of Probation is the second largest probationary agency in the U.S., annually supervising 90,000 adults (about 60,000 at any one time) and 4,000 juveniles. Short Term Alternative to Remand Treatment (START) is a program the Department established in 1993 as an alternative sentencing option for defendants with split sentences [4]. A split sentence consists of up to six months incarceration followed by five years probation. The probationers serve the jail portion of their sentence under domicile restriction.

START combines electronic monitoring with intensive supervision—office contact twice a week supplemented by at least one field contact per week. The program uses radio-frequency-controlled devices, primarily bracelets, for individuals confined to a single location. Bracelets are not suitable for other types of probationers, such as juveniles who attend school and adults who work.

In 1999, START implemented text-prompted voice-based tracking for the first 30 days of every START sentence and began using the system as an alternative to electronic bracelets when a city court imposes curfew on a probationer lacking long-distance telephone service.

Before probationers enter the program, START officers discuss all conditions of probation with them, including voice tracking. A description of the voice tracking system appears in an addendum to The Conditions of Probation document the probationer is required to

Figure 5. User interface for Rough'n'Ready, a broadcast news system. (The Rough'n'Ready system was developed at BBN/GTE Internetworking [2] and acquired by Lernout & Hauspie Speech Products in 1999; the interface here is being developed by Lernout & Hauspie.)



sign. The probationer is then enrolled in the voice-tracking system under the supervision of a START officer. Most probationers are given pagers. The system randomly pages these probationers who must return the page within 10 minutes. Some probationers, including juveniles, are required to call the voice-monitoring center at a given time and from a pre-specified location. Failure to return a page or make the scheduled call can result in a violation of probation.

For probationers who stay in compliance for six months, START reduces the level of monitoring by eliminating voice tracking. For probationers who violate probation, voice tracking is reinstituted.

START reports the voice system has worked so well that Family Court now uses the same technology to monitor its curfews.

Outlook

Current research and market trends indicate that future applications of voice-biometrics will be text-independent and incorporate other speech-processing and biometric technologies. Such applications are already in demand in several markets. For example, health-care, financial services, and other industries that handle large numbers of sensitive documents have begun to incorporate multiple biometrics into their security strategies. The use of products for multiple and layered biometrics is further supported by declining prices on biometric sensors and development of standards, facilitating the development of multibiometric applications. In April 2000, the BioAPI Consortium[5] released version 1.0 of its BioAPI specification, and the following month, Microsoft announced its intention to develop its own application programming interface standard.

The wireless industry, Internet security providers, and telecom services providers all support development of unobtrusive, text-independent speaker verification and identification to secure the communications environments of the future. Some approaches focus on chip-based security embedded in wireless telephones and PDAs. Other solutions require more powerful technology in communications networks, such as the work on continuous verification being done at the University of Wales [1] and elsewhere. Other approaches involve integration of speaker verification and other biometrics with public key infrastructure encryption and digital certificates for securing e-commerce applications.

Deployment of speech-recognition applications also spurs demand for voice-based security. Companies with existing speech-recognition applications are adding speaker verification as a way of extending these applications to secured transactions or as replacements for PIN-based security. Moving in the other direction, HSN, for example, is converting its touch-tone order-entry operation to speech recognition, so members who voice-verify successfully can order using just their voices. HSN is also weighing whether to incorporate a second level of speaker verification for large orders.

Developers are also applying the combined power of text-independent speaker identification, speech recognition, and other technologies to the automation of entirely new types of tasks. One such application is the automatic indexing, search, and retrieval of information in audio sources, such as tape recordings and news broadcasts [8]. Figure 5 shows the user interface for Rough'n'Ready, one of the systems described in [8]. Developed by BBN technologies, Rough'n'Ready applies a battery of speech-processing technologies to the task of analyzing and indexing audio and video recordings, such as news broadcasts.

These trends indicate acceptance of speaker verification and identification and that voice biometrics technologies are increasingly viewed as components in larger, more complex solutions. C

[5]The BioAPI Consortium was formed in 1998 by Compaq Computer, IBM, Identicator Technology, Microsoft, Miros, and Novell (representing the Speaker Verification API standard workgroup) for the purpose of developing a specification of a standardized API compatible with a wide range of biometrics application programs and biometrics technologies. Consortium members now also include biometrics vendors and consultants (Identicator, IriScan, ITT Industries, J. Markowitz Consultants, Keyware, Mytec, National Biometric Test Center, and Visionics) and biometrics users (Barclays Bank, Intel, Kaiser Permanente, U.S. National Institute of Standards in Technology, and the U.S. National Security Agency.

References

1. Auckenthaler, R., Carey, M., and Mason, J. Speaker-centric score normalisation and time pattern analysis for continuous speaker verification. In *Proceedings of ICASSP 2000, the IEEE International Conference on Acoustics, Speech, and Signal Processing* (Istanbul, Turkey, June 5–9). IEEE Press, Piscataway, N.J., 2000, 48–50.
2. Kubala, F., Colbath, S., Liu, D., Srivastava, A., and Makhoul, J. Integrated technologies for indexing spoken language. *Commun. ACM 43,* 2 (Feb. 2000), 48–56.
3. Markowitz, J. Ieri, oggi, domani: Speaker recognition yesterday, today, and tomorrow. In *Proceedings of COST250 Workshop on Speacker Recognition in Telephony* (Rome, Italy, Nov. 10). European Co-operation in the Field of Scientific and Technical Research, 1999.
4. Markowitz, J. Hands on with ... The New York Department of Probation. *Voice ID Quart. 4,* 2 (Apr. 2000).
5. Markowitz, J. Hands on with ... Home Shopping Network. *Voice ID Quart. 3,* 4 (Oct. 1999).
6. Markowitz, J. Hands on with ... Girl Tech. *Voice ID Quart. 3,* 2 (Apr. 1999).
7. Markowitz, J. Hands on with ... Illinois Department of Revenue. *Voice ID Quart. 2,* 3 (Oct. 1998).
8. Maybury, M. News on demand, special section. *Commun. ACM 43, 2* (Feb. 2000), 32–79.

Judith Markowitz (JMarkowitz@POBox.com) is president of J. Markowitz Consultants in Evanston, IL and Chicago.