Toward Vision 2001: Voice and Audio Processing Considerations

Lawrence R. Rabiner

The broad goal of *Vision 2001* is to provide seamless, easy-to-use, high-quality, and affordable communications between people and machines—anywhere and any time. To achieve this goal, the fields of computing, communications, and networking must converge in a variety of information terminals and network services. To make Vision 2001 a reality, there must be a significant number of advances in signal processing, computing, processing, networking, memory, and communications. This introductory paper examines the voice and audio processing implications of Vision 2001. It also discusses both current AT&T capabilities and areas in which these enablers are not yet sufficient to realize Vision 2001.

Introduction

In late 1991, AT&T created a concept for communications in the 21st century, which is generally referred to as *Vision 2001*. The essence of this concept is a communications environment in which access to people, machines, and information is both easy and convenient, and where every conceivable type of communication and message service is ubiquitous and readily accessed.

This dream of providing seamless, easy-to-use, high-quality communications between individuals, groups of people, and machines—anywhere, any time, and at an affordable price—presupposes knowledge of the following terms and definitions:

- *Seamless communication.* A user selects any of several communication methods (voice and voice messaging, video and video messaging, e-mail, facsimile, and exchange of handwritten material or data files) and integrates it—in what appears to be an effortless process—with any other form of communication. During a telephone call, for example, a caller should be able to jot a note on a message pad and send it, so that the other party receives it simultaneously, without affecting the call's quality or protocol in any way. Seamless communication implies having the capability of converting one form of communication to another more convenient type, as required. For example, an e-mail or facsimile message can be converted to a voice message if the receiving party does not have access to a display terminal. Seamless communication also implies ease-of-use, namely, a well-thought-out and efficiently designed user interface.
- *High-quality communication.* This term means that a user notices no significant signal degradation, regardless of the environment or transmission medium. For example, a telephone call and video message should sound and look the same, respectively, over both wired and wireless networks. High-quality communication also implies both ease-of-use and convenience. In many instances, voice commands used to control the communications flow are more convenient and simpler than touch-tone pads, pen input, and keyboards.
- *Anywhere communication.* This term implies the existence of a worldwide infrastructure that couples directly into the long-distance and local networks, allowing personal communicators to be used both indoors (either wired or wireless) and outdoors (wireless). Anywhere communication also implies having a roving capability (the party one might want to reach could

be anywhere), along with the availability of an up-to-date, network database containing the current location of every potential system user. This capability allows user-selected features and user-specific knowledge, such as speech patterns, to be accessed readily while traveling.

- *Any-time communication.* A user can communicate whenever desired, even if the person to whom the communication is being sent is unavailable or does not want to accept it at the time. The key concept surrounding any-time communication is *integrated messaging*, which provides each user with a single access number for an integrated, attached mailbox that receives all types of communication. Thus, integrated messaging can be used for telephone calls, video messaging, combination voice and video messaging, e-mail, facsimile transmissions, handwritten notes, and data exchanges. In such an environment, telephone calls and video messages are combined when a recipient is not available or doesn't want to receive calls directly. E-mail, facsimile transmissions, handwritten notes, and data messages are routed automatically to a recipient's integrated mailbox.
- *Communication at reasonable cost.* This term encompasses both moderately priced, high-performance terminals (including small size and low power requirements for wireless terminals) and a flexible, intelligent, high-capacity worldwide network that can effectively handle a large volume of audio, video, and data traffic at low unit cost.
- *Habitable environment.* This term means that the communication products and services are useful. That is, there is a utility associated with each of them from a customer's perspective, and they are easy to use (easy to learn, install, access, and operate). Communication products must also be aesthetically pleasing or attractive, at least in the sense that they do not annoy or distract a user. The challenge of creating such habitable human/machine interfaces is the focus of the paper by Cosky et al.[1]

To make Vision 2001 a reality, advances are required in signal processing (speech and image), computer science, processor design, memory and battery technology, networking, and communications. This paper focuses on one key Vision 2001 technology: *voice and audio processing.* It describes current research capabilities for each segment of voice and audio processing, as well as areas in which additional exploration is required to realize the dream of universal voice communication.

Panel 1. Abbreviations, Acronyms and Terms
MOS—mean opinion score
TTS—text to speech
VEST—Voice English-Spanish Translater

Segments of Voice and Audio Processing

Voice and audio processing can be separated conveniently into the following segments:

- *Speech and audio coding*, which compresses information in a speech or audio signal for efficient transmission or storage;
- *Text-to-speech synthesis*, which converts normal text to spoken language to transmit a message from a machine to an individual;
- *Speech recognition*, which extracts message information from spoken language to control the actions of a machine by means of voice commands;
- *Speaker verification*, which verifies a person's identity to control access to information, networks, or physical premises;
- *Language identification*, which identifies an individual's spoken language to provide customized services in the correct language;
- *Language translation*, which provides two-way spoken communication between individuals who do not speak the same language; and
- *Electroacoustics*, which provides appropriate microphone, loudspeaker, and echo-control platforms so that high-quality voice communication can be maintained in any environment.

This paper examines, in each of these segments, Vision 2001 challenges, discusses current technology, and estimates the communication capabilities foreseen in the next five years.

Speech and Audio Coding. Almost all speech transmitted over the existing public-switched telephone network is band-limited to the 200- to 3,400-hertz (Hz) range. This band-limited signal is often referred to as *telephone-bandwidth speech.* It can be represented accurately by a digitized signal having a 64-kbits/s data rate. An important speech-coding concern is devising efficient algorithms for

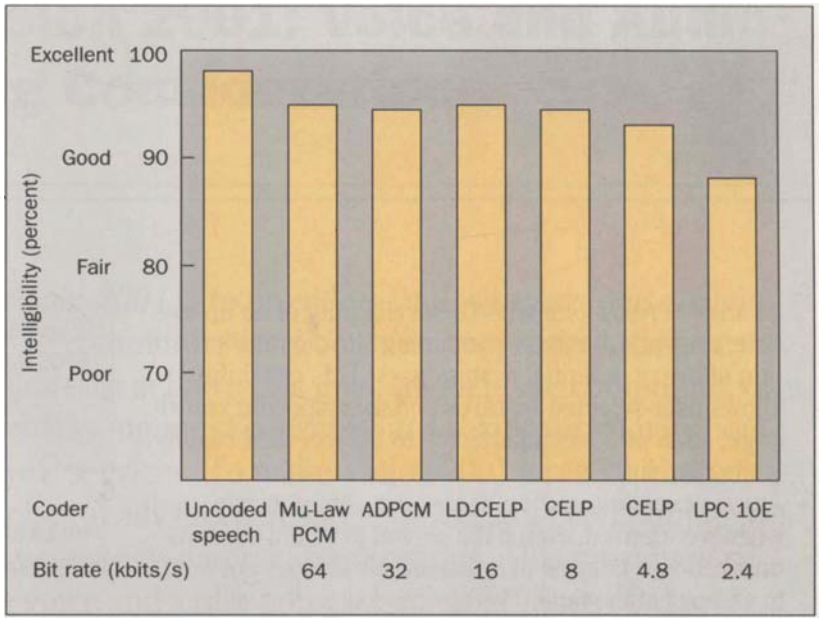

Figure 1. This bar chart shows the speech-intelligibility scores of several telephone-bandwidth coders, as a function of bit rate. Results are shown for uncoded speech and for bit rates of 64, 32, 16, 8, 4.8 and 2.4 kbits/s. The results show that intelligibility is robust to decreasing bit rate.

reducing the bit rate needed to transmit and store telephone-bandwidth speech while maintaining encoded-speech quality.

In recent years, teleconferencing has created a need for transmitting and storing broader-bandwidth speech (covering the range of 50 to 7,000 Hz) in digitized form. As such, it is often referred to as wideband speech, which can be represented accurately by a digitized signal having a 128-kbits/s data rate. Speech coding also involves creating efficient algorithms for reducing the bit rate of wideband speech, again while maintaining speech quality. In addition, multimedia applications on the information superhighway have created a need for efficient transmission and storage of full CD-quality digital audio. Fulfilling this need is a challenge, because the CD bandwidth range is 10 to 20,000 Hz and CD storage rates are about 1.4 Mbits/s. Furthermore, target data-transmission facilities are designed with 64-kbits/s and 128-kbits/s rates. Audio-coding algorithms attempt to bridge the gap in bit rates while maintaining full (or close to full) CD quality.

The goal of speech and audio coding is to exploit signal redundancies, signal structure (such as periodicity and correlations), and knowledge of human perception (masking, for example) to code the signal to various bit rates while maintaining quality as high as possible. To understand the current capability in speech and audio coding, it is worthwhile examining curves of signal quality versus bit rate. Measuring speech quality is a challenge itself. Signal quality, for speech signals, has two key dimensions: *intelligibility* and *subjective quality*, which are measured in terms of a subjective rating scale. Intelligibility is usually measured by asking listeners to identify one of a pair of rhyming words. Such intelligibility tests are called *diagnostic rhyme tests*. Subjective quality is usually measured in terms of a mean opinion score (MOS) rating, by which a listener rates speech (or audio) quality on a five-point scale—five is excellent, four is good, three is fair, two is poor, and one is unacceptable.

Figures 1 and 2 illustrate the intelligibility and MOS for telephone-bandwidth coders. Results are shown for uncoded speech and bit rates of 64, 32, 16, 8, 4.8 and 2.4 kbits/s. The results show that intelligibility is more robust to decreasing bit rate than is subjective quality.

For wideband speech (with its broader bandwidth), MOS scores of four or higher have been measured for bit rates of 16 kbits/s and up, while scores fall below four for lower bit rates. The major application of wideband speech coding is teleconferencing, in which 16-kbits/s transmission is almost always feasible. This bit rate, therefore, is the one having current interest.

For digital audio (for example, music) coding, a two-channel CD-bandwidth digital audio signal can be coded at a rate of 128 kbits/s and have a MOS score of 4.5. The demand for full CD quality is projected to be strong and growing. CD audio coding facilitates listening to "live concerts," previewing new CDs, providing CD-audio for movies on demand and other interactive TV services—all over broadband networks, such as ISDN.

AT&T's efforts in voice and audio coding have resulted in a wide range of products and services, including:

- Telephone-bandwidth coding standards at several bit rates;
- Digital telephone-answering machines; and
- Secure telephony devices.

Over the next several years, significant improvements in the ability to code speech and audio are envisioned. Furthermore, several important milestones are now being focused on, including:

- Development of a 13-kbits/s cellular-telephone bandwidth coder that provides network-quality coding (MOS of 4.0 or higher) over the wireless network.

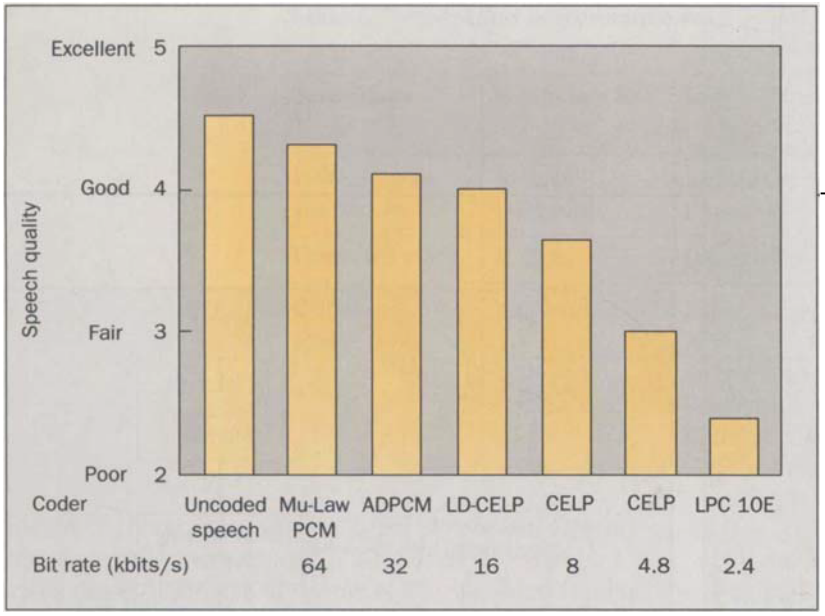

Figure 2. Speech-quality mean opinion scores of several telephone-bandwidth coders, as a function of bit rate, are shown in this bar graph. MOS tests use a five-point rating scale. The attribute 5 denotes excellent quality (no noticeable impairments); 4, good (only very slight impairments); 3, fair (noticeable but acceptable impairments); 2, poor (strong impairments); and 1 bad (highly degraded speech). The results show that subjective quality is not robust to decreasing bit rate.

- Development of a 16-kbits/s wideband coder for audio/video teleconferencing systems (integrated with acoustic echo cancelers) providing high-quality, full-duplex communication for groups of people.
- Development of a 2.4-kbits/s telephone-bandwidth coder providing cellular quality (MOS on the order of 3.5 to 4.0), and enabling the use of satellites for voice communication.
- The capability of speeding up or slowing down coded speech messages (without seriously affecting sound quality) for the purpose of enabling systems to provide an advanced messaging feature. The latest release of the AT&T AUDIX® voice-messaging system has this capability, made possible by using an appropriate 16-kbits/s coder.

Two papers in this issue discuss speech and audio coding in more depth. The paper by Cox et al. expands on current issues in wireless coding of speech.[2] The paper by Jayant and Chen details the state of the art of audio coding and its potential applications.[3]

Text-to-Speech Synthesis. AT&T 's goal in speech-synthesis research is to develop a machine having an intelligible, natural-sounding voice for conveying information to a user (generally in the form of a text file) in a desired accent, language, and voice (male, female, or child). This development would enable such a machine to converse with a user in response to either touch-tone (keyboard) or spoken queries for information (ostensibly from a database accessible by the machine, either directly or by means of a network connection).

Figure 3 shows a block diagram of a system used for converting text to speech (TTS). System input is an arbitrary text message (usually, but not always having appropriate punctuation). The first task of the TTS system is to convert the text string into a sequence of phonetic symbols (indicating the sounds to be spoken), along with a set of prosody markers (indicating the intonation and emphasis on certain words as inferred from a linguistic analysis of the text). This text-to-sound and prosody conversion involves a combination of linguistic analyses, including dictionary look-up of word pronunciations, rules for exceptions and unusual cases, and algorithms for generating appropriate word durations, pitch, and loudness contour for the speech. Once the phonetic symbols and prosody markers have been determined, the next step in the TTS process is to assemble the appropriate speech units, computing both the pitch and duration contours. A store of elemental sound units is required to create intelligible speech. The size of the sound inventories varies from as few as 700 sound units to as many as 4,000, depending on the language being synthesized. The final two steps in the TTS process are speech synthesis from the sequence of sound units, and digital-to-analog conversion to create an analog speech signal.

TTS systems have been used in a variety of telecommunications and desk-top applications, such as:

- *Network servers*, which convert text-based e-mail (or facsimiles) to voice-based messages available over standard voice networks;
- *Voice previewers*, which speak draft text material, and can be used to spot errors in the text and to generate feedback (using a different sensory modality) about the effectiveness of the text-based message; and
- *Feedback mediums*, for various telephone information services (for example, stock price quotations, sports scores, directory assistance, and banking services).

To approach the Vision 2001 goal of high-quality TTS synthesis, the following developments must occur during the next several years:

- Significant enhancement of the system's prosodic com-

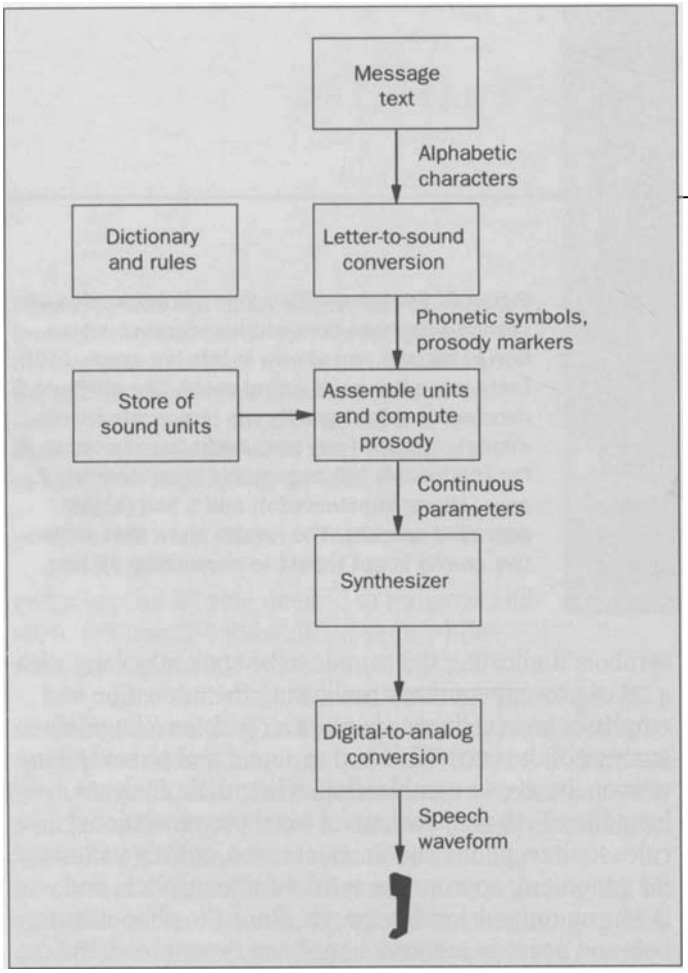

Figure 3. This block diagram illustrates a full text-to-speech (TTS) synthesis system. The first task of the TTS system is to convert the text string into a sequence of phonetic symbols and a set of prosody markers. The next step is to assemble the appropriate speech units. The final two steps are synthesis of speech from the sequence of sound units and digital-to-analog conversion, which creates an analog speech signal.

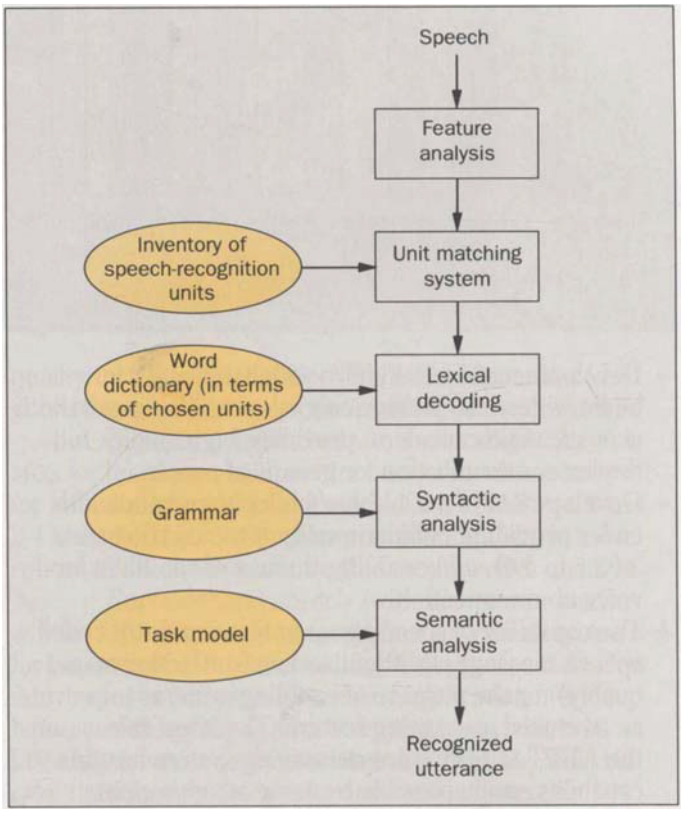

Figure 4. A complete speech-recognition system is shown, which incorporates the syntactic and semantic analysis modules. This system employs a pattern-recognition approach. The first step is a spectral analysis of the signal, called *feature analysis*. The output of feature analysis is a set of parameters that characterizes the time-varying spectrum of the speech signal. These parameters are matched against a set of stored patterns to provide a lattice of possible speech matches.

ponents (duration, pitch, and emphasis) to improve the overall synthesis quality;
- Production of female and children's voices and the investigation of regional accents;
- More effective text-semantics exploitation (over the course of several sentences or even paragraphs) so the synthesizer can speak more clearly;
- Exploration of alternate forms of synthesis, such as waveform methods (for example, stored sounds) and articulatory methods (stored vocal-tract configurations), to take advantage of natural constraints in speech production and perception; and
- Development of a unified software architecture for TTS, so that each new language (or talker or regional accent) can create new synthesis units automatically (and rapidly), devising new duration, prosody, and discourse rules appropriate to the language (and talker and accent).

The paper by Sproat and Olive provides a comprehensive overview of current strengths and weaknesses in speech synthesis.[4]

Speech Recognition. The goal of AT&T speech recognition research is to provide a human/machine user interface that can recognize (or "understand") and respond to spoken input, in the form of voice commands. Approaching this goal, speech recognition technology

Table I. Recognition performance for several technologies and tasks

Technology	Vocabulary Size	Task	Word-Error Rate (Percent)
Isolated words and phrases	10 digits	Isolated digits	0.1
	1,000 words	Basic English	4.3
Connected words	11 digits	Connected digits	0.2
Continuous speech	1,000 words	Database management	4.5
	2,000 words	Airline reservations	2.3
	20,000 words	Reading from the *Wall Street Journal*	11.3

increasingly provides greater convenience and ease of use, beginning with control of advanced information services through the use of simple voice commands and progressing toward more futuristic applications.

Speech recognition technology can be separated conveniently into the following three categories:

- *Isolated word and phrase recognition*, in which a system is trained to recognize a discrete set of command words (or phrases) and to respond appropriately.
- *Connected word recognition*, in which a system is trained on ("learns") a discrete set of vocabulary words (for example, digits), but it is required to recognize fluent sequences of these words (for example, telephone numbers and credit card numbers).
- *Continuous speech recognition*, in which a system is trained on a discrete set of subword vocabulary units (for example, phonemes and syllables), but it is required to recognize fluent speech, such as grammar-constrained sequences of words. For more advanced applications, the vocabulary can be unlimited and the job of the recognizer is to understand the meaning of the spoken input.

Each of these technology categories has led to a range of current and anticipated applications in telecommunications.

Figure 4 is a block diagram of the pattern-recognition approach to speech recognition. This approach has been applied to each of the previously discussed technology categories. The first step in pattern recognition is a spectral analysis of the signal (called *feature analysis* in Figure 4), because the information in the speech signal is carried, over time, in the distribution of energy in the audio frequencies. The output of the feature analysis is a set of parameters that characterizes the time-varying spectrum of the speech signal. These parameters are matched against a set of stored patterns (which might be whole words, phrases, or subword speech units) to provide a lattice of possible matches to the speech. (Scores for the "goodness" of each match are computed for this lattice.)

When training is done on whole words (or phrases), the unit-matching and lexical-decoding blocks of Figure 4 coalesce into a single, word-matching (or phrase-matching) block. When the training patterns are subword units, the unit-matching block provides a lattice of subword-unit match scores, and the lexical-decoding block deciphers the lattice into a word lattice based on a *word dictionary*. The final two blocks of the recognizer, *syntactic analysis* (based on a word grammar) and *semantic analysis* (based on a task specification) are used to decode the word lattice into the best scoring sentence that is both syntactically correct and semantically meaningful.

Table I illustrates the current (research-laboratory) capability of recognizing speech. For isolated word recognition, word-error rates are as low as 0.1 percent for the ten isolated digits and 4.3 percent for a select vocabulary of 1,000 isolated words. For connected word recognition, word-error rates as low as 0.2 percent have been obtained for recognition of fluent digit strings composed of combinations of 11 digits (including both zero and "oh"). For continuous speech recognition, the reported error rates range from 2.3 percent to 4.5 percent for vocabularies of 1,000 to 2,000 words and highly constrained tasks, to 11.3 percent for a 20,000-word vocabulary for read speech, excerpted verbatim from the *Wall Street Journal*.

Current speech recognition technology supports a wide range of applications in telecommunications, including those that automate attendant functions (for example, operator services, order taking, and directory-listing retrieval), and those that provide new ways to control information services (for example, voice banking and access to databases). Some services, used every day by millions of subscribers, have been created within the AT&T network. They include *voice recognition call processing* for automation of "operator-plus" (O+) calls, 800 speech-recognition service for call distribution, and the WorldPlus™ service for multilingual access to information. Speech recognition has also appeared in telecommunications products, most notably the hands-free, voice-

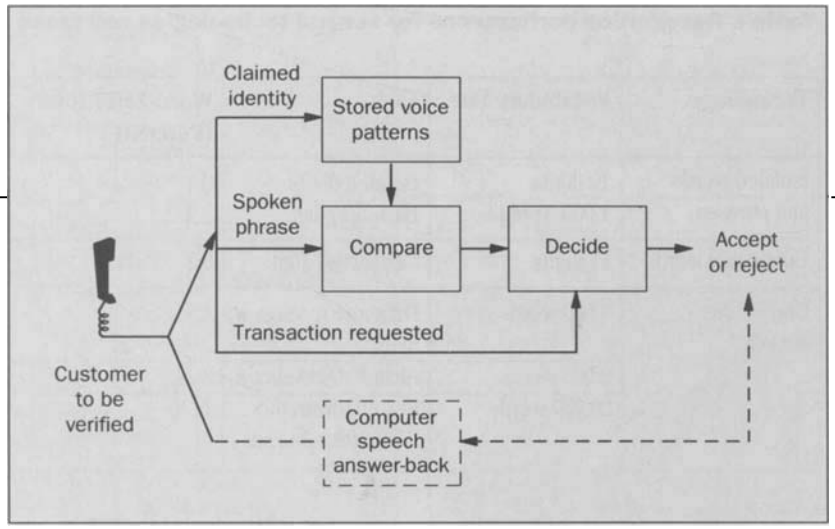

Figure 5. This illustration represents an integrated speaker-verification system. A customer wishing verification provides a claimed identity, which is the spoken phrase acceptable by the verification system, and the transaction requested. A comparison of the spoken phrase with the appropriate, stored voice pattern provides a comparison score. Depending on the transaction requested, the decision to accept or reject an identity claim is made and sent back to the customer.

dialing cellular speakerphone introduced by AT&T early in 1993.

The following, significant goals in speech recognition technology must be achieved to meet the challenges of Vision 2001:

- *Improved overall robustness.* Often, speech recognizers that perform extremely well in the laboratory have error rates ten times higher in field use due to such factors as sensitivity to background noise, differences in telephone hand sets, variations in transmission channels, and talker idiosyncrasies.
- *Rejection of extraneous speech.* Reliable rejection of utterances that do not contain valid command words or sentences (for example, extraneous speech and background sounds) is needed.
- *Recognition of commands embedded in carrier speech.* The so-called word-spotting problem must be solved; for example, extracting the phone number from such utterances as, "My phone number is 647-1234." When the recognizer seeks a seven-digit telephone number, it must treat the speech at the beginning of the sentence as an extraneous carrier phrase and ignore it.
- *Real-time implementation of speech-recognition algorithms on low-cost processors.* Such implementation enables speech recognizers to be embedded in almost any product.

AT&T anticipates significant progress toward meeting the aforementioned goals during the next several years. In this issue, an extensive discussion of these goals is provided by Juang, Perdue, and Thomson.[5] Additionally, the factors affecting the implementation of voice processing solutions on digital-signal-processing chips are discussed in the paper by Crochiere and Boddie.[6]

The long-term challenge in speech recognition is to provide a natural-language, spoken-dialogue capability for any AT&T product or service. This means researchers and developers must learn how to create speech-recognition systems (and those capable of understanding speech) for specific applications that are easy to use, accurate, and robust. Such applications must have the flexibility to accommodate any language, almost unlimited vocabularies, and unconstrained syntax. To achieve these goals, AT&T must bring together—on a single platform—the extensive research programs on acoustics, linguistics, natural language, and semantics. AT&T has already begun research on developing speech-understanding systems for accessing both the Official Airline Guide and *Wall Street Journal* articles by means of natural-language voice queries.

Speaker Verification. As computers, networks, and telephone systems become more widely interconnected, the need for security to prevent break-in, theft of service, or mayhem becomes more and more important. For Vision 2001, security is a matter of claiming a speaker identity and verifying the claim. The identity claim is often done by identifying the terminal, by means of automatic number identification, from which the communication is initiated. Verification is accomplished either in a text-independent manner or by using key phrases ("This is John Doe speaking"), both of which are based on speech-pattern talker models. A user can access whatever services are offered after being identified and verified as a valid customer.

Figure 5 shows a block diagram of an integrated, speaker-verification system. A customer wishing verification provides a claimed identity (to access the appropriate, stored voice pattern), the spoken phrase suitable to the

verification system (for example, the key phrase or a digit sequence), and the transaction requested. A comparison of the spoken phrase with the appropriate, stored voice pattern provides a comparison score. Depending on the transaction requested, the decision to accept or reject an identity claim is made and sent back to the customer. The *accept* or *reject* message usually appears on some type of video screen, but sometimes it is transmitted through a computerized, speech answer-back system.

Decision criteria can be adjusted according to the task. For example, in banking transactions, a simple match could be required to check an account balance, while an authorization to withdraw funds could require a much higher level of match.

A speaker-verification system can make two types of errors: it can either reject a true customer (false reject) or accept an impostor (false accept). The goal of most speaker-verification systems is to constrain the rate of false-reject errors to less than one percent while minimizing instances of false-accepts.

Currently, under controlled laboratory conditions, speaker verification by voice can be both very accurate and highly reliable. Assuming availability of a cooperative user uttering a machine-specified digit string, a speaker-verification system can achieve false-accept and false-reject rates as low as 0.3 percent using a fixed transducer and transmission system in a minimum-background-noise environment. It is still a formidable challenge, however, to develop a practical speaker-verification system that is immune to disparate background noises, as well as tolerant of different transmission systems, telephone hand sets, and talkers.

One challenge is to create a real-world system in the next five years. Such a system will be based on both technology improvements and new discoveries. Its performance will be highly accurate; that is, it will accept true speakers reliably, reject impostors, and robustly handle the variability inherent in all typical situations. Additional challenges are to develop a system that minimizes the need for training, and one that can effectively adapt to changes in a talker's voice patterns over time.

Language Identification and Translation. An important aspect of Vision 2001 is the need for machine-assisted communication between those who do not speak the same language. This requirement leads naturally to the need for the following capabilities:

- Identification of a talker's language from unconstrained, spoken input (optimally, a short phrase or even a word or two);
- Speaking to a talker in the language he or she is speaking (for example, a speech-synthesis capability in a wide variety of languages);
- Recognition of voice commands spoken in a wide range of languages; and
- Translating the message content of spoken commands from one language to another.

The goal of the AT&T language-identification research program is to ascertain the language of a talker accurately. Two key methods are used in reaching this goal:

- Spotting commonly spoken keywords that occur only in the target language; and
- Estimating which language is being spoken based on statistics of occurrence of various sounds and sound combinations.

Current laboratory capability is about 91-percent language accuracy on a standard National Institute of Standards and Technology test of the following three languages: English, Mandarin Chinese, and Spanish. The test is based on about 50 seconds of unconstrained speech. A near-term goal is to develop a system capable of identifying speech from a set of 11 languages—English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin Chinese, Spanish, Tamil, and Vietnamese—based on two seconds of real-time spoken input. This must be done for any talker and utterance, with greater than 95-percent (laboratory) accuracy. This system would have applications in such global systems as AT&T Language Line® or WorldPlus™ services.

AT&T researchers have demonstrated—in the laboratory—a highly accurate language-translation system having a vocabulary of about 500 words with a highly constrained task syntax. This system was called Voice English-Spanish Translater (VEST), and it provided language translation for a banking and currency-transfer task. The system operated in near real time on an array of 128 AT&T DSP32C digital-signal-processing chips.

The long-term goal of AT&T language-translation research is to provide both speech-understanding and translation capabilities for the following five languages in real time: English, Spanish, French, Mandarin Chinese, and Japanese. In addition, vocabularies of about

2,000 words per language would be provided for prescribed tasks, such as hotel reservations or game playing over a communications link.

Electroacoustics. Three essential requirements of Vision 2001 are the ability to provide:

- "Anywhere" communication;
- High-quality, person-to-person communication in open-field environments (for example, group offices); and
- High-quality communication between groups of people (conferencing).

To satisfy these requirements, electroacoustic transducers (microphones and loudspeakers) are needed. Such transducers produce high-quality signals (both voice and video) in any acoustic environment, and they facilitate development of teleconferencing systems having high-performance, acoustic echo control for full-duplex, eyes-free and hands-free communication. The ideal communication system, within the Vision 2001 concept, would contain a mix of microphones, loudspeakers, noise-cancellation equipment, and echo-control systems, all of which would be "smart." That is, they would use information about the sound field to adapt to almost any acoustic environment, achieving optimum signal-to-noise ratios.

The current focus of AT&T research in electroacoustics is to implement designs for:

- A "smart" microphone (array) that can find the correct (or desired) acoustic source, track its motion, and effectively ignore all other extraneous sounds. Such a smart microphone works with either a telephone (3,200-Hz) or wide-bandwidth (7-kHz) speech signal.
- A "smart" loudspeaker (array) that can focus sound over a field that is either as narrow as a cone directed to a single listener or as wide as the entire room in which the loudspeaker is placed.
- An echo-cancellation and suppression system that adapts rapidly to the acoustic environment.
- A noise-control system that seeks to reduce the ambient noise in the communication equipment (generally achieved through proper design of axial cooling fans), and one that minimizes the noise entering the communication system by using sophisticated signal-processing methods (for example, anti-noise technology and active noise cancellation).

A range of products and services is evolving within AT&T, based on advances in electroacoustic research. These products include specialized microphones for multimedia PCs, consumer products, and video terminals, as well as acoustic echo control in speakerphones, and noise-control systems in switches and PBXs. The interplay between acoustics and signal processing in designing advanced, electroacoustic systems is discussed in the paper by Baumhauer et al.[7]

The long-range challenge of electroacoustics research is to design specialized systems of microphones, loudspeakers, echo-control devices, and noise-control equipment that provide a high-quality communications environment anywhere, any time.

Open Architectures

Over time, voice-processing functionality is becoming a commodity, because an understanding of how to design and implement a range of reliable voice services—on standard processors—is becoming widespread. Concurrently, the platforms on which the processing is implemented should be open, so that multiple users (and vendors) can contribute readily and easily to the overall functionality of the system. Open-architecture approaches to voice processing are discussed in the paper by FitzGerald and Moosmiller.[8]

Summary

Rapid technological advances are leading to the communication and information systems of the 21st century. Moreover, AT&T is progressing toward realizing the needs of Vision 2001—seamless, easy-to-use, high-quality communications—anywhere, any time, and at an affordable price. In both voice and audio processing, the path along which to proceed is clear.

Recent advances in speech and audio coding, speech synthesis, speech and speaker recognition, language identification and translation, and electroacoustics have helped in defining the user interface to terminals and network services. Advances in voice and audio processing, over the next several years, will continue to move technology closer to achieving the goals of Vision 2001.

Acknowledgment

The author expresses appreciation to David Isenberg and Judy Tschirgi of AT&T Bell Laboratories for valuable feedback and consultation on this paper.

References

1. M. J. Cosky, B. L. Lively, L. A. Roberts, and B. L. Wattenbarger, "Talking to Machines Today and Tomorrow: Designing for the User," *AT&T Technical Journal*, March/April 1995, Vol. 74, No. 2, pp. 81-91.
2. R. V. Cox, P. Kroon, J-H Chen, R. Thorkildsen, K. M. O'Dell, and D. S. Isenberg, "Speech Coders: From Idea to Product," *AT&T Technical Journal*, March/April 1995, Vol. 74, No. 2, pp. 14-22.
3. N. S. Jayant and E. Y. Chen, "Audio Compression: Technology and Applications," *AT&T Technical Journal*, March/April 1995, Vol. 74, No. 2, pp. 23-34.
4. R. W. Sproat and J. P. Olive, "Text-to-Speech Synthesis," *AT&T Technical Journal*, March/April 1995, Vol. 74, No. 2, pp. 35-44.
5. B-H. Juang, R. J. Perdue Jr., and D. L. Thomson, "Deployable Automatic Speech Recognition Systems: Advances and Challenges," *AT&T Technical Journal*, March/April 1995, Vol. 74, No. 2, pp. 45-56.
6. R. E. Crochiere and J. R. Boddie, "Digital Signal Processors: Toward Vision 2001," *AT&T Technical Journal*, March/April 1995, Vol. 74, No. 2, pp. 71-80.
7. J. C. Baumhauer Jr., S. H. Early, J. H. Fikus, S. L. Gay, and M. A. Zuniga, "Audio Technology Used in AT&T's Terminal Equipment," *AT&T Technical Journal*, March/April 1995, Vol. 74, No. 2, pp. 57-70.
8. C. W. FitzGerald and J. P. Moosmiller, "Architecture of the Intuity™ Response Application Programming Interface (IRAPI)," *AT&T Technical Journal*, March/April 1995, Vol. 74, No. 2, pp. 92-101.

(Manuscript approved February 1995)

Lawrence R. Rabiner *is director of the Information Principles Research Laboratory at AT&T Bell Laboratories in Murray Hill, New Jersey. He leads research efforts in several key areas of information sciences: speech and image processing, interactive systems (including handwriting recognition), digital signal processing, and communications. Mr. Rabiner received B.S., M.S., and Ph.D. degrees in electrical engineering, all from the Massachusetts Institute of Technology in Cambridge. He joined AT&T in 1962.*