NOVEL APPLICATIONS OF SPEECH PROCESSING IN AT&T NETWORK SYSTEMS PRODUCTS

Mahendra K. Verma, Dimitrios P. Prezas, Thomas L. Russell, Jr., Mostafa Hashem Sherif, and Reed Thorkildsen

*Mahendra K. Verma, Thomas L. Russell, Jr., and **Mostafa Hashem Sherif** are with AT&T Bell Laboratories at Holmdel, New Jersey. The late **Dimitrios P. Prezas** was with AT&T Bell Laboratories at Indian Hill Park, Naperville, Illinois. **Reed Thorkildsen** is with AT&T Bell Laboratories at Whippany, New Jersey. Mr. Verma is a member of technical staff in the Global Customer Networking Department. He works on systems engineering and applications of emerging technologies to global networks and services. He has a B.Sc. and M.Sc. in physics from the University of Delhi, India, and an M.Sc. and Ph.D. in electrical and electronic engineering from the University of Bradford, U.K. He joined AT&T in 1986. Mr. Prezas was responsible for development of speech coding and automatic speech recognition technologies for intelligent*

In this paper, we discuss speech processing applications in the integrated access and cross-connect system, digital cellular telephone system, operator services position system, and intelligent networks. These applications are used to show how the AT&T Network Systems product line allows telephone companies and business customers to implement solutions based on end-user characteristics, expectations, and requirements.

Introduction

This paper addresses four examples of speech technology applications. Integrated access and cross-connect and digital cellular telephone system applications enhance the use of existing network facilities, while the applications in operator services and in intelligent networks enhance operations efficiencies.

The integrated access and cross-connect system (IACS) is a wideband packet product that is being used in private and virtual private networking applications for efficient transport and multiplexing of multimedia traffic. The product integrates several technologies, including advanced techniques for coding speech and voiceband data signals, digital speech interpolation, echo cancellation, and packet protocols. This reduces the amount of facilities needed for multimedia traffic and thereby reduces cost to the network services providers.

Digital cellular telephone systems make available public mobile service that is integrated with the wired telephone network. Here the challenge is to reduce the channel bit rate, while retaining speech quality, to increase the number of channels in the radio bandwidth allocated to this service. Speech technology multiplies the number of telephone calls that can be made on crowded cellular telephone systems.

The operator services position system (OSPS) is a feature of the 5ESS® switch used by telephone operating companies for directory and toll assistance operator services. Speech recognition is being implemented for determining call type and whether a billed party will accept collect calls. Among the advantages of this technology are that it reduces operator fatigue and can supplement operators in emergencies.

77

Panel 1. Acronyms in This Paper

ADPCM	adaptive differential pulse-code modulation
AM	administrative module
ASR	automatic speech recognition
BRI	basic-rate interface
CCITT	International Telegraph and Telephone Consultative Committee
CELP	code-excited linear predictive
CM	communication module
DDS	digital data service
DICE	digital circuit emulation
DSP	digital signal processing
DTMF	dual-tone multifrequency
HDLC	high-level data link control; international standard developed by the International Standards Organization (ISO)
IAC	integrated access terminal
IACS	integrated access and cross-connect system
IAT	integrated access terminal
ISDN/SM	switching module supporting Integrated Services Digital Network
LPC	linear predictive coding
MIPS	million instructions per second
MPLPC	multipulse linear predictive coding
NSCX	network services complex
OSPS	operator services position system
OT	operator terminal
PC	personal computer
PCM	pulse-code modulation
PSM	position switching module
RPE	regular pulse-excited
SM	switching module
TDMA	time-division multiple-access
VSELP	vector-sum-excited linear predictive
TIA	Telecommunications Industry Association
TID	Telefónica Investigacion y Desarrollo
VDLC	virtual data link capability
VLSI	very large scale integration

Customer access to intelligent network services often requires the use of dual-tone multifrequency (DTMF) signaling. Automatic speech recognition (ASR) technology provides modernization of the user interface, especially where DTMF signaling is not available. In an international application, the 5ESS switch with ASR capability in integrated peripheral equipment will soon be used to give access to intelligent network features. These services will thus be made available to customers without DTMF service.

Integrated Access and Cross-Connect System

AT&T's IACS integrates speech, voiceband data, image data, digital data, and network management and control into one self-healing wideband packet network.[1-3] It cross-connects in both the circuit and packet domains, and offers easy load sharing and "graceful degradation" of voice traffic if overloads do occur. The IACS provides multiplexing, transport, and networking of analog and digital traffic. As shown in Figure 1, the IACS has two basic components: the integrated access terminal (IAT) and the integrated access controller (IAC). The IAT provides bandwidth on demand and offers uniform procedures for network management and support. The IAC is a management system for centralized control and maintenance of a network of IATs.

The IAT is of interest here: it uses speech technology in its multiple roles of traffic classifier, speech coder, packetizer, statistical multiplexer, and digital data interface. Digital data, signaling, image data, and network control information arrive from channel banks or other IATs at the primary rate of 1.544 Mb/s (megabits per second), or at 2.048 Mb/s, and from local concentrators via ISDN basic-rate interface lines.

The IAT differentiates between speech spurts and voiceband data. It further classifies voiceband data traffic on the basis of its speed. Echo cancellation within the IAT compensates for any echoes caused by transmission or packetization delays. However, speech detection, echo cancellation, and dynamic classification of voiceband data
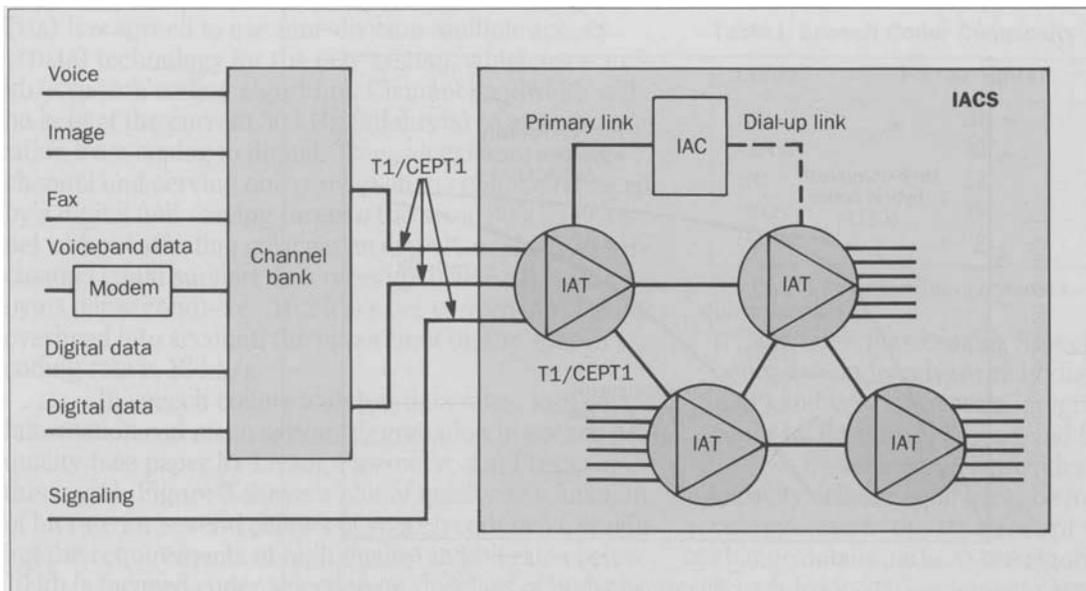
Figure 1. Architecture of integrated access and cross-connect system. A variety of analog and digital signals are processed, combined, and routed through a network of integrated access terminals. The integrated access controller is used for centralized management of several integrated access terminals. The local concentrator connects subrate data traffic to an integrated access terminal.

79

can be disabled if desired.

A variety of speech encoding algorithms are supported, such as fixed-rate adaptive differential pulse-code modulation (ADPCM) according to CCITT Recommendation G.721/G.723 (ANSI T1.303), and embedded ADPCM with 2 core bits and 2 enhancement bits according to CCITT Recommendation G.727 (ANSI T1.310). (CCITT is the International Telegraph and Telephone Consultative Committee.) Packetization of speech requires that voice packets be played out at regular intervals at the receiving end. This regular playing out is achieved through the procedures of the CCITT Packet Voice Protocol G.764 (ANSI T1.312).

Digital speech processing techniques are used to automatically adjust the amount of compression on data modem signals, depending on the bit rate. The IAT packetizes continuous-bit-rate digital data in one of two ways. Using digital circuit emulation (DICE), the IAT can interface to 64 kb/s (kilobit per second) digital data channels or digital data service (DDS) subrate channels of the DS0A

format. It can be provisioned to remove DDS idle codes and the redundant copies of the DS0A format. DDS subrate services (at 2.4 kb/s, 4.8 kb/s, and 9.6 kb/s) can thus be carried at rates close to the actual user bit rate; removing this unneeded information further reduces transmission cost.

The IAT's virtual data link capability (VDLC), extends DICE to protocols that use high-level data link control (HDLC) procedures at the link layer. It allows the IAT to detect and eliminate the HDLC flags in addition to eliminating DDS control, subrate framing bits, and DDS idle codes. Furthermore, VDLC allows partitioning and rejoining of HDLC frames, to accommodate layer frames.

The IAT interfaces directly with networks providing frame-relay services based on link-access procedures for D channel (LAPD). It therefore can network frame-relay services at the DS1 or fractional-DS1 level. Frame relay services are particularly suited in wide-area distributed computing networks and integrated networks at the primary rate of 1.544 Mb/s.

Figure 2. A comparison of the speech quality of several key voice coding technologies. A pulse-code modulation coder, as used in the public network, yields high quality but requires a high bit rate. A high-complexity hybrid coder, such as a CELP coder, yields high quality even at very low bit rates.
(See Panel 1 for definitions.)
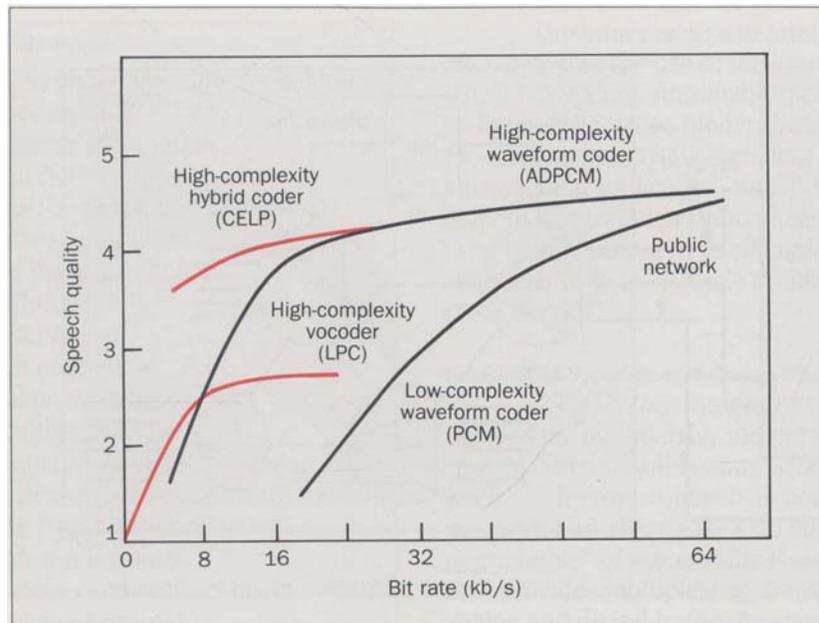
The IACS fits many traditional networking applications and opens avenues for new service concepts in point-to-point applications, in satellite networks, in virtual private networks, and in other services.[1,2,4-6] The ability to compress voice and data traffic while maintaining toll-quality service means that the IACS is ideally suited for facility relief applications where compression and dynamic bandwidth allocations are needed. Applications of the IACS in satellite networks rely on compression and cross-connection of packets; the IACS is readily suitable for multidestinational use because the address information is contained in the header of each packet. With IACS, new virtual networking services can be tailored to the specific needs of the traffic, whether voice, data, or image. These services are called "virtual" because there is no permanent assignment of bandwidth through the digital transmission facilities. Rather, bandwidth is allocated on demand and is statistically shared with other users of the same IACS.

Digital Cellular Telephone System

Today's cellular telephone system is based on repeated use of 832 frequency-modulation (FM) radio channels within the radio spectrum allocated by the Federal Communications Commission. To meet anticipated growth in the demand for cellular telephone service, from 2 million U.S. subscribers now to 18 million in 1995, a new generation of digital cellular equipment is needed, with digital signal processing techniques for low-bit-rate speech coding, spectrally efficient modems, and fast adaptive equalization. High-quality speech coders that operate at the lowest possible bit rate play a key role in a spectrally efficient digital cellular system. An 8-kb/s speech coding algorithm has been selected for use in digital cellular equipment. In addition to quality and bit rate, other factors that constrain the choice of a speech coder are robustness in the cellular environment and implementation complexity.

The Telecommunications Industry Association

(TIA) has agreed to use time-division multiple-access (TDMA) technology for the new system, which uses an 8-kb/s speech coding algorithm. Channel bandwidth will be kept at the current 30 kHz (kilohertz) to ease the transition from analog to digital. Thus, an existing analog channel unit serving one conversation could be replaced by a digital unit serving three in the same 30-kHz channel without affecting adjacent analog channels. A 30-kHz channel could support data rates up to 48.6 kB/s (kilobytes per second)–i.e., 16.2 kb/s per conversion. Taking overhead into account, the upper limit on the speech coding rate is 13 kb/s.

In speech coders with low data rates, loss of information can mean serious degradation in speech quality (see paper by Jayant, Lawrence, and Prezas in this issue). Figure 2 shows a plot of quality as a function of bit rate for several classes of speech coders.[7] Combining the requirements of high quality and bit rates below 10 kb/s focused coder selection on the class of high-complexity, hybrid coders.

Representative of hybrid coders are multipulse linear predictive[8] (used by British Telecom in Skyphone™ services), regular pulse-excited[9] (a variation of which was selected as the speech coder for use in the pan-European digital cellular system), and code-excited linear predictive (CELP).[10] (Skyphone is a trademark of British Telecom.) The coder selected for use in the U.S. digital cellular system is vector-sum-excited linear predictive (VSELP), a variation on CELP. With a bit-rate of 8-kb/s, this coder, combined with channel coding at a 5 kb/s rate, yields speech quality equal to or better than analog FM under similar conditions. The VSELP coder uses a special set of codebooks that are derived from vector sums of basic excitation vectors representative of speech contained in a speech database.

Table I shows a comparison of the number of millions of instructions per second (MIPS) for typical implementable speech coders. The number of MIPS that are required for CELP coders is at least an order of magnitude greater than required by a pulse-code modulation

Table I. Speech Coder Complexity

Coder	Bit rate (kb/s)	Complexity (MIPS)
PCM	64	<1
ADPCM	32	<1
RPE	13	6
MPLPC	10	10
CELP	8	15–30

See Panel 1 for definitions of acronyms.

(PCM) coder, for example. Speech coder technology is being driven by advances in digital signal processing (DSP) and very large scale integration (VLSI) devices (see paper by Purdue, Leikness, and Sharp in this issue).

Cellular service providers require that higher capacity cellular equipment be made available in 1991, yet revisions to the TIA standard will be issued that could change details of the VSELP algorithm. Thus, AT&T has scheduled VSELP implementation in a DSP in equipment for the Autoplex® cellular communications system scheduled for commercial availability in late 1991. A full-duplex CELP coder has been implemented in an AT&T WE® DSP-32C device[11] and has formed the basis for implementation of VSELP in an AT&T WE DSP-16A for the Autoplex digital cellular system.

81

Operator Services

The 5ESS switch OSPS product[12] is used by telephone companies for directory and toll assistance operator services. Automatic speech recognition is being implemented to automate some operator functions. The ASR feature can automate operator-handled toll and assistance calls, particularly for call-type recognition (e.g., collect and calling card calls) and for recognition of billing acceptance by called parties. Operators are always available when callers' requests are not recognized by the ASR equipment and for customer-requested operator handling (i.e., when a customer says "Operator").

Using existing OSPS capabilities, telephone companies can configure teams of automated positions that

Figure 3. Architecture for an operator services position system with automated toll and assistance. The host 5ESS switch gives access to the outside network for both operator terminals and automated terminals (those with automatic speech recognition).
(See Panel 1 for definitions.)
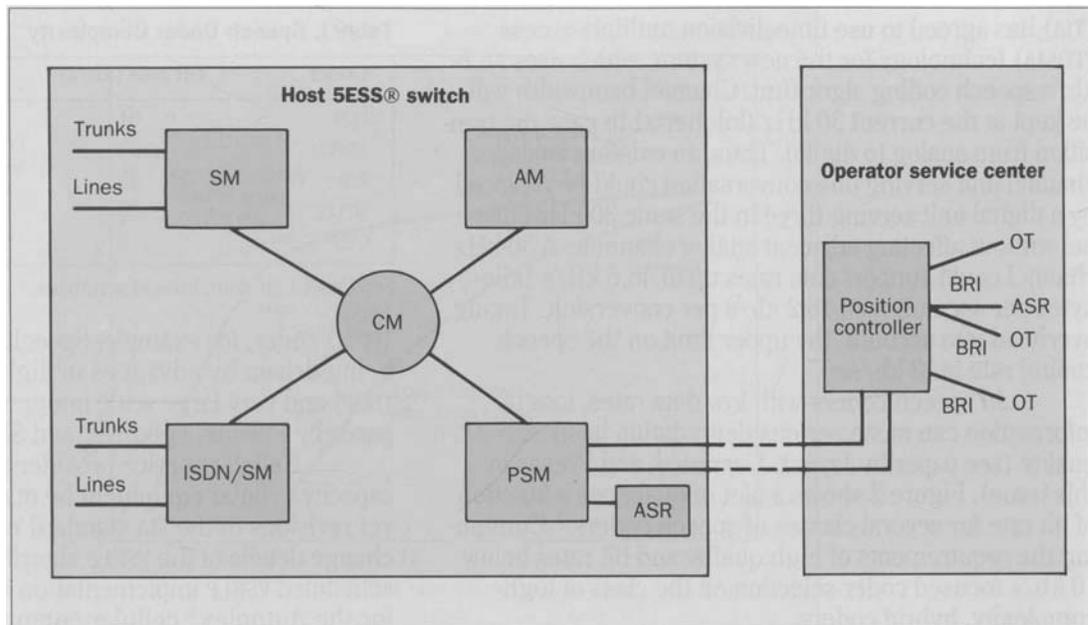
incorporate ASR, just as they configure human operator teams. An existing OSPS feature can be used to allow human operator teams to handle call volume traffic overloads experienced by the automated team. This preserves overall customer service at desired levels. The speech recognition algorithm proposed for this feature is a key-word-spotting algorithm, based on a hidden Markov model,[13] that can recognize words embedded in continuous speech with over 95 percent accuracy over typical telephone lines.

In a typical scenario, the caller will initiate the ASR service by dialing 0 plus the desired number. If the calling station is restricted to operator-handled treatment only, the call will be delivered to a human operator. If the dialed number fits any of the card service formats, the calling party will experience normal card service. Normally the caller will hear a tone prompt and can respond with the desired call type (e.g., collect, calling card, third number) either by speaking or using DTMF tones.

Collect. If the caller speaks "collect" or dials the DTMF digits for a collect call, the caller will be prompted for his or her name, which is recorded by the system. After speaking the name, the caller will hear normal network responses following outpulsing of the forward number—ringing, busy, or possibly intercept announcements (such as "all circuits busy"). The caller's line will be split or muted when the called party answers. Meanwhile, the called party hears a charge-acceptance prompting announcement including the recorded calling party name and a request to say "yes" to accept or "no" to reject the call (e.g., "You have a collect call from *recorded name*. Would you accept the charges? Please say 'yes' or 'no.'"). If the called party says "yes," he or she will be connected to the caller and both will hear an announcement that they can go ahead with the call. If the called party says "no," a denial announcement will be given to the caller (e.g., "Your call was not accepted. Please hang up."). Table II summarizes a typical collect call scenario.

Table II. Example of an Automated Collect Call

What the OSPS system says	What the customer says	What the OSPS system does
"At the tone, please say calling card, collect, third number, person, or operator"	"I want to call collect."	Spots the word *collect*
"Please say your name, now."	"Mary Roe."	Records speech *Mary Roe*; places call
"You have a collect call from *Mary Roe*. Will you accept the charges?"	"Huh? Oh, yes."	Spots the word *yes*
"Your collect call has been accepted."		Completes the call

Third Number. If the caller speaks "third number" at the prompt from the switch or dials DTMF digits requesting third number billing, the caller will be prompted to dial the third number. After dialing the third number, the caller will typically hear normal network responses (as above) and be connected to the called party when he or she answers. For third number calls that require confirmation of billing acceptance, the third party will hear an announcement upon answering the telephone. This prompting announcement includes a statement of the calling party's name, identification of the call as a third number call, and a request that the third party say "yes" to accept or "no" to reject the call billing.

Other Charging Types. Users of automated calling card service will receive the same treatment they do today; that is, they will be asked to dial a calling card number. Callers who request "person" will be transferred to operators for this premium service. Callers may explicitly request operator assistance by dialing 0 or saying "Operator."

Figure 3 depicts the basic 5ESS switch OSPS toll and assistance service architecture with ASR functionality. The automated positions will be traffic-engineered just as the present operator positions are. Because not all customers will want or be able to use the automated service, it is expected that there will be a slight increase in real-time utilization of the switch.

For billing purposes, OSPS will record that ASR was used to handle the call. This allows the telephone company to charge different rates for automated versus operator-handled calls, when desired or required.

Intelligent Networks

An emerging area of speech processing applications is that of intelligent networks, which offer network-based services via a combination of distributed intelligence and out-of-band common channel signaling. Intelligent network nodes like the network services complex (NSCX) provide interactive services such as toll-free 800 service calls.

Until now, the caller input to intelligent network services required a Touch-Tone or DTMF phone. Soon, however, ASR technology will be introduced to modernize the user interface, especially where rotary phones are used. Enhanced 800 services controlled through voice commands are the leading intelligent network application targets for ASR technology. Menu-driven 800 advanced call prompter or credit card services will respond to spoken digits in place of tones and automatically route calls or approve credit.

Intelligent network services have attracted the interest of international customers and have increased AT&T's potential for switching equipment sales. The first AT&T deployment with ASR capabilities will be in

83

Figure 4. Decision tree for a representative intelligent network service—news delivery for Telefónica. The caller need only say (and the system need only recognize) the spoken words "uno," "dos," and "tres."
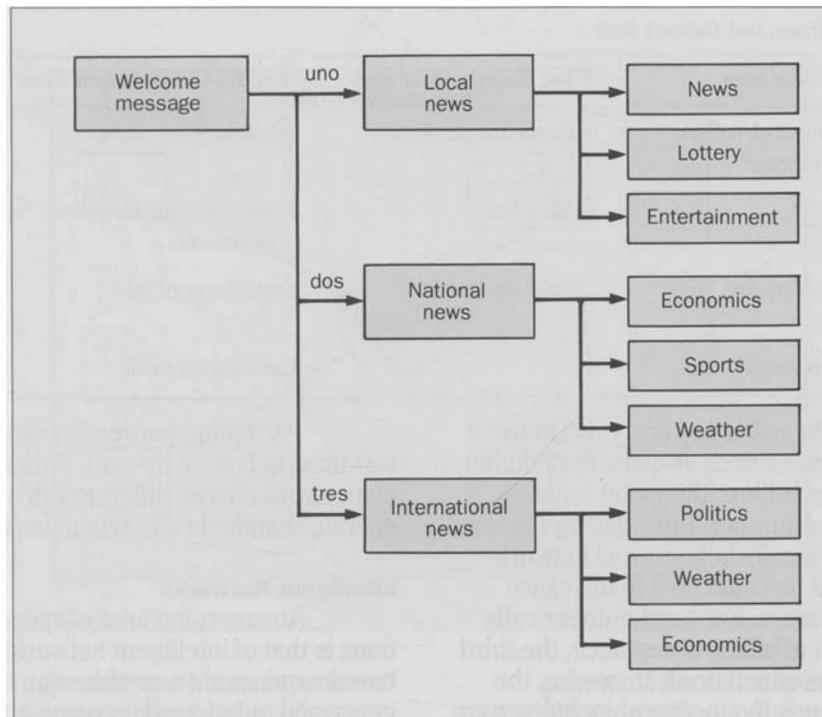
Spain in the first quarter of 1991. The small number of Touch-Tone phones in Spain (less than 5 percent) and the high-technology profile of Telefónica, the Spanish telephone administration, have been the primary factors leading to the new service. Telefónica will offer toll-free enhanced 900 services (equivalent to the U.S. 800 service) for the 1992 Summer Olympics in Barcelona. Such services may use decision trees like that in Figure 4.

The targets of the Spanish intelligent network application of ASR were conservatively set to accommodate the relatively old network and the pool of users who are unfamiliar with automated services. The ASR subsystem will support speaker-independent isolated-word recognition of the Spanish key words *uno, dos,* and *tres.* This provides menu variety and simplifies data collection for word modeling. The speech recognition

algorithm is based on hidden Markov models[13] and can spot a limited set of vocabulary words embedded in extraneous speech (see paper by Wilpon et al. in this issue). For isolated key words, the recognizer correctly identifies the word 98 percent of the time. However, our experience in the field indicates that often customers respond with extraneous speech in addition to the vocabulary word. For such vocabulary words embedded in typical customer responses containing extraneous speech, the recognizer rate is better than 95 percent. AT&T's ability to develop interorganizational synergies has resulted in swift technology transfer from research to development and a timely meeting of customer needs.

A flexible and cost-effective architecture was designed for this project. The ASR subsystem is an adjunct to the NSCX. Each subsystem is based on a per-

sonal computer (PC) with 12 dual-channel DSP-32C boards and can serve a total of 24 channels at one time. The DSP board is designed to accept speech through a coder-decoder (codec) from dedicated lines from the NSCX and to recognize the speech. A general-purpose interface bus feeds the ASR decision back the the NSCX. The entire ASR frame houses three subsystems, for a total capacity of 72 channels.

One of the key tasks of this project was collection and verification of spoken Spanish for language modeling. Fully automated PC-based speech collection stations—with optical disk drives for mass storage, telephone interfaces, announcement capabilities, and software tools for semiautomated speech editing—were installed at a Telefónica central office in Madrid. The stations were assigned toll-free 900 numbers. Personnel of Telefónica's research and development organization, Telefónica Investigacion y Desarrollo (TID), were trained to maintain the stations. A scenario resembling a menu-driven service in Spanish was employed to collect the spoken words "uno," "dos," and "tres." All digits, "cero" through "nueve" (0–9), and "si" and "no" were also collected in isolated form.

This data acquisition effort lasted over a period of approximately 8 months (April to November 1989) and recorded over 8000 speakers from a total of 9 Spanish provinces and 50 central offices. This database was initially verified by TID personnel. The *uno-dos-tres* portion of the database was then cross-checked by AT&T Bell Laboratories. Files containing inappropriate responses, signaling tones, or no speech at all were removed. About 2500 recordings of each of the words *uno, dos,* and *tres* were finally used as testing and training data.

Particular attention was paid to the performance issues of ASR technologies for recognizing key words embedded in unconstrained speech. The database described above included little representation of extraneous speech material. Thus, it was necessary to include additional material in the model formation process. Therefore a secondary collection effort was run by TID over a 2-month

period (December 1989 and January 1990), tailored specifically to gather extraneous material. A minitrial at the AT&T Network Systems International offices in Madrid with prototype ASR hardware in early February of 1990 provided more unconstrained data as well as preliminary field performance figures. Achievement of the desired speech recognition performance targets will be verified during a more extensive field minitrial run with TID on prototype hardware in November 1990.

Telefónica's efforts to upgrade its telephone network and AT&T's commitment to meet customer needs have been the driving forces in technology growth. Directions for the future may include increased vocabulary size (0 to 9, *si, no*) to accommodate larger menus and connected-digit recognition for advanced credit card services. Algorithmic advances that would allow such growth on the present system are feasible. In addition, the experience gained in this project will facilitate AT&T's development of ASR capabilities for other languages and applications, both within the U.S. market and abroad.

Conclusion

The applications discussed in this paper are examples of speech technology applications to AT&T Network Systems products. AT&T Network Systems is committed to incorporate these new technologies to enhance its product lines to better serve the needs of its customers.

References

1. W. J. Giguere, "New applications of wideband technology," *Proceedings of International Communications Conference*, Vol. 3, Atlanta, Georgia, April 15-19, 1990, pp. 997-999.
2. M. H. Sherif, F. D. Fite and M. C. Gruensfelder, "On the road to Universal Information Services (UIS): The Integrated Access Terminal," *Proceedings of ENTELEC'89*, New Orleans, Louisiana, March 19-22, 1989, pp. 42-48.
3. K. Sriram et al., "An integrated access terminal for wideband packet networking: Design and performance overview," *Proceedings of International Switching Symposium*, Stockholm, Sweden, June 1990, Vol. 6, pp. 17-74.
4. C. P. Ashoka and R. N. Paul, "Network consolidation via wideband

85

packet technology—Private network trial of IACS in AT&T IMS," *Proceedings of International Communications Conference*, Vol. 1, Atlanta, Georgia, April 15-19, 1990, pp. 154-158.

5. L. T. Corley, "BellSouth trial of wideband packet technology," *Proceedings of International Communications Conference*, Vol. 3, Atlanta, Georgia, April 15-19, 1990, pp. 1000-1002.
6. G. Forcina and T. Oishi, "An efficient satellite transmission scheme for the interconnection of remote wideband packet networks," *Proceedings of International Communications Conference*, Vol. 3, Atlanta, Georgia, April 15-19, 1990, pp. 1069-1074.
7. N. S. Jayant, "Coding Speech at Low Bit Rates," *IEEE Spectrum*, August 1986, pp. 58-63.
8. B. S. Atal and J. R. Rende, "A New Model of LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Paris, France, April 1982, pp. 614-617.
9. P. Kroon et al., "Regular Pulse Excitation—A Novel Approach to Effective and Efficient Multipulse Coding of Speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-34, No. 5, October 1986, pp. 1054-1063.
10. M. R. Schroeder and B. S. Atal, "Code-excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tampa, Florida, March 1985, pp. 937-940.
11. K. Swaminathan and R. V. Cox, "Design, Implementation, and Evaluation of a 8.0 Kbps CELP Coder on a Single AT&T DSP32C Digital Signal Processor," *Proceedings of IEEE Workshop on Speech Coding for Telecommunications*, September 1989, pp. 11-13.
12. Special Issue on Operator Services Position System, *AT&T Technical Journal*, Vol. 68, No. 6, November/December 1989.
13. J. G. Wilpon, C. H. Lee, and L. R. Rabiner, "Application of Hidden Markov Models for Recognition of a Limited Set of Vocabulary Words in Unconstrained Speech," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, Scotland, May 1989.

Biographies (continued)

network applications and secure voice equipment. He joined the company in 1979 and had a B.S. in physics from the University of Athens, Greece, and an M.S. and Ph.D. in electrical engineering from the Illinois Institute of Technology in Chicago. Mr. Russell is a supervisor in the Switching Services Concepts Planning Department. His group does systems engineering and develops prototypes for new AT&T Network Systems products, especially those involving automatic speech recognition and wireless technologies. He has a B.S. in physics and M.S. in mathematics from Stevens Institute of Technology, Hoboken, New Jersey. Mr. Sherif is a distinguished member of technical staff in the Advanced Digital Signal Processing Department. He works on systems engineering and standardization of wideband packet networks. He has a B.Sc. in electronics and communications and M.Sc. in circuits from Cairo University, Giza, Egypt, and a Ph.D. in bioengineering from the University of California at Los Angeles. Mr. Thorkildsen is a supervisor in the Radio Technology Department. His group incorporates speech processing technology, including low-bit-rate speech coding, into AT&T digital cellular systems. He has a Ph.D. from the University of Virginia, Charlottesville. He joined AT&T in 1981.