SPEECH PROCESSING: A PERSPECTIVE ON THE SCIENCE AND ITS APPLICATIONS

James L. Flanagan and Charles J. Del Riesgo

James L. Flanagan *was director of the Information Principles Research Laboratory at AT&T Bell Laboratories in Murray Hill until his retirement in August 1990, and* ***Charles J. Del Riesgo*** *is director of the Systems Peripherals Development Laboratory at the Columbus, Ohio facility. At AT&T Bell Laboratories, Mr. Flanagan has been responsible for research in digital speech processing, acoustics, linguistics, software engineering, image coding, and human-machine communications. He joined the company in 1957 and has a B.S. from Mississippi State University (Starkville) and both an M.S. and a Ph.D. from Massachusetts Institute of Technology (Cambridge), all in electrical engineering. Mr. Flanagan is now director of the Center for Computer Aids for Industrial Productivity at Rutgers University in New* (continued on page 13)

Though AT&T has broadened its scope to embrace information signals as varied as image, audio, data, facsimile, and medical diagnostics, voice communications remain central to the AT&T network. But even the applications of voice are shifting dramatically from the original "plain old telephone service" (POTS) into new services and capabilities for communication, computation, and information management. In particular, evolving techniques for human-machine communication by voice are opening new market opportunities, and are stimulating mass deployment of sophisticated systems that are easy to use and that will serve society's complex needs. This paper draws a perspective on the science of voice processing, and on some of AT&T's contributions to the fundamental understanding. It then discusses new applications that are now supported by this understanding, and by the remarkable advances that are being made in microelectronics. Finally, it speculates on scientific advances and the resulting technology expected in the coming decade.

Introduction

Voice is a preferred means for person-to-person communication. The acoustic signal conveys a sequence of audible sounds—speech—whose meanings (within cultural groups) have been agreed upon *a priori*. This acoustic "code" constitutes language.

But auditory signals propagate poorly over distances. And communications technology was spawned by the human desire to communicate between distant points. Its development was fueled by early understanding of electrical and acoustical phenomena, and by efforts to harness this physics to societal benefit. Significant inventions along the path of this evolution include the telegraph, telephone, electronic

amplification, radio, and—much later—the transistor, the stored-program digital computer, and fiber-optic transmission.

Not surprisingly, then, the early focus in communications was toward transporting voice signals faithfully over distances. The technologies for capturing, representing, conveying, making logical decisions upon, and reconstituting spoken information are traditionally termed *speech processing.*

The technical focus in speech processing has shifted and broadened over time. Originally the concept of facsimile reproduction of the acoustic waveform led to the telephone, where the scientific challenge was to extend the distance of transmission. Early transatlantic telegraph cables spanned great distances, but could not support the bandwidth required for speech waveform transmission. Thus, the need for transatlantic voice communication spurred the study of speech compression, which steered research efforts away from waveform concepts and toward the fundamental information properties of the speech signal. Work on "vocoders" (voice coders) was born of this interest, where the objective was voice communication over the least possible bandwidth. With the advent of transatlantic radio, with its relatively greater bandwidth, the emphasis changed again, and became directed less toward transmission economy and more toward encryption for privacy.

Though undersea electronic amplifiers made transatlantic telephone cables possible, the expense of such circuits returned the focus in speech processing to efficient use of bandwidth. The technology of TASI (time assignment speech interpolation) utilized the silent intervals in two-way conversation for a threefold increase in cable capacity. This technique exploits the statistics of speech energy bursts across a group of communication channels (originally 36 in the transatlantic cable).

Bandwidth efficiency for cable and radio remained a driving interest in speech processing until the advent of computers and digital information systems. As computers and information systems have become more sophisticated and capable of serving a great variety of needs, ease-of-use for the human has become central to their mass deployment and acceptance. Because speech is natural to human communication, great interest now attaches to giving complex machines the ability to interact conversationally with human users, and to representing spoken information efficiently for economical computer storage. This trend has been recognized for several years.[1]

Panel 1. Terms and Acronyms in This Paper

ADPCM	adaptive differential PCM
CELP	code-excited LP
DSP	digital signal processor
ISDN	Integrated Services Digital Network
MP-LPC	multipulse LPC
LP	linear prediction
LPC	linear prediction coefficient
PCM	pulse code modulation
POTS	plain old telephone service
SBC	subband coding
SP	signal processor
SPC	signal processor companion
TASI	time assignment speech interpolation
VLSI	very-large-scale integration

As twentieth century technology has matured, there has been a tendency for the sciences of communication, computation and information management to merge. Today's sophisticated information systems typically contain elements of each. Consequently, there is a strong drive toward integrating information modalities for voice, image, data, computation and conferencing into modern systems (see Berkley and Flanagan in this issue.)

Technology. The technologies embraced by speech processing include speech coding (or compression), synthesis, recognition, talker verification, and audio conferencing. The sciences that support these technologies are acoustics, digital signal processing, linguistics, and engineering. Practical implementations depend almost exclu-

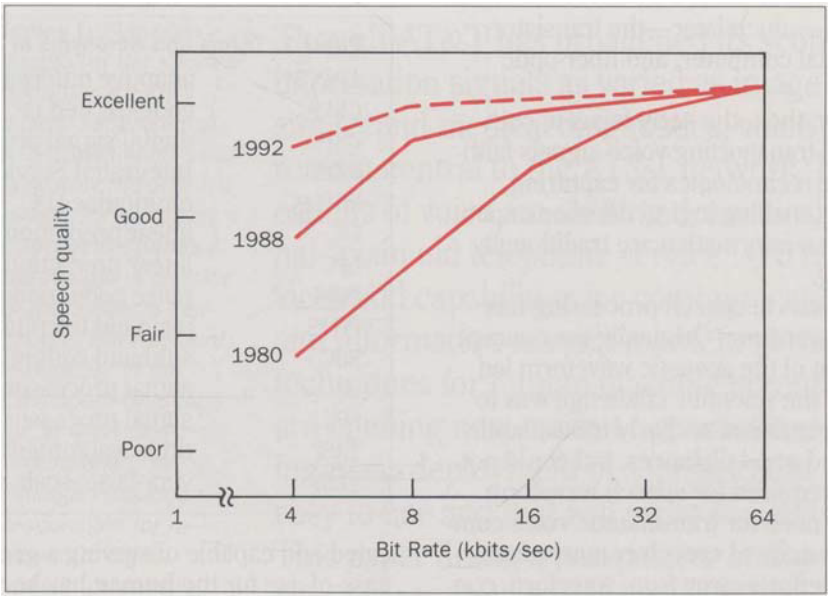


Figure 1. Qualitative representation of progress in speech signal compression (or low bit-rate coding) over recent years. Current challenges center on algorithms that provide high-quality speech coding at low transmission rates.

sively on the enormous advances in microelectronics.

Speech Coding. In the strict digital sense, speech coding originated with pulse code modulation (PCM). This digital representation—through Nyquist sampling and binary quantization—clings to the original telephone concept of preserving the acoustic waveform. There is little in the technique that is signal specific except bandwidth and dynamic range. Its great advantage over analog transmission is noise-free signal regeneration and the opportunity for digital encryption. The technique has served in telephone toll transmission for many years at the well-known 64 kilobits per second (kb/s) rate.

To gain greater efficiency, adaptive differential PCM (or ADPCM) uses characteristic correlation between time samples of the speech waveform (hence its short-time predictability) to code speech with high quality at 32 kb/s. This technique has been serving in the commercial telephone plant for over five years. Because it is slightly speech specific, voice-band data is automatically detected and treated specially by a fixed quantization.[2]

Coding methods for lower transmission bit-rates depart further from the concept of waveform preservation, and apply the perceptually adequate criterion of preserving the short-time amplitude spectrum. Unnecessary phase information is discarded in the process to achieve increased efficiency. Coding speech in contiguous subbands of the spectrum, and dynamic bit assignment according to the perceptibility of quantizing noise in the different frequency ranges, led to the concept of subband coding (SBC) for 16 kb/s.[3] This technique was key in providing at an early time the storage economy necessary for the successful deployment of AT&T's AUDIX voice mail system.[4] New algorithms now under development aim to increase still further the quality of this service.

Linear prediction (LP) technologies in various forms [such as LP coefficient vocoders (LPC), multipulse LPC (MP-LPC), and code-excited LP (CELP)] address good-quality coding in the transmission range 16 to 2 kb/s. Examples of current applications include digital

cellular radio and encryption for government communications. The technical challenge remains gaining the highest possible quality for the lowest possible bit-rate and hardware cost. Figure 1 illustrates the status of speech coding and the quality improvements realized over recent time.[5] Jayant, Lawrence, and Prezas in this issue offer a comprehensive discussion of current issues in speech and audio coding.

Speech Synthesis. In its most rudimentary form, speech synthesis—providing voice answerback for machines—is achieved by concatenation of human spoken, digitally-stored words and phrases. Extensive application of this technology — usually employing digital coding for efficient storage—has been made in the form of announcement machines. A typical example is the AT&T 14A Announcement System, which uses 9.6 kb/s MP-LPC coding to produce telephone intercept messages, coin-box announcements and voice prompts.[6] These systems give good speech quality at low hardware cost, but message versatility is limited in that only those human-spoken words and phrases that have been compressed and digitally stored can be used to assemble messages. Cost of storage limits vocabulary size, and lack of prosodic and contextual control limits contextual naturalness. Nevertheless, many thousands of announcement systems are in use today for tasks that do not require large vocabularies. A related and remarkable application of synthesis from low bit-rate stored parameters is the custom VLSI implementation of an LPC synthesizer offered earlier by Texas Instruments in its "Speak and Spell"™ toy.

Synthesis directly from unrestricted text aims for the ultimate versatility. The challenge is to achieve natural quality, while maintaining economies in the relatively complex hardware that is required. The ingredients typically include:

- A grapheme-to-phoneme (or phonetic) transformation, typically accomplished with the aid of a stored pronouncing dictionary and programmed letter-to-sound rules
- A syntactic (and eventually semantic) analysis to compute prosodic features of sound pitch, intensity, and duration
- A method for generating and interpolating (i.e., smoothing) the sound sequence. In an AT&T implementation, this is achieved from a stored library composed of fractions of syllables that are digitally coded as linear-prediction parameters.

The AT&T system runs on an 80386 computer with a single digital signal processor (DSP) chip as the synthesizer.[7] Synthesis quality is high in intelligibility, but machine-like in naturalness. A similar system is DecTalk™, trademarked and marketed by Digital Equipment Corporation and based upon research from the Massachusetts Institute of Technology.

As Hirschberg et al. in this issue show, the research frontier in text synthesis is in achieving human-like quality, varied voices, dialects and languages, and ultimately, in understanding the exquisite detail needed to duplicate the voice characteristics of an individual.

Speech and Speaker Recognition. The technology of automatic speech recognition has advanced explosively in the last several years because of new understanding in modeling speech sound sequences, characterization of large talker populations, and computation techniques for dealing with naturally connected words.[8] The progress has therefore been from early systems for speaker-trained template recognition of a few tens of isolated words (such as that used in the AT&T Liberty voice repertory dialer for cellular telephones) to emerging systems that are speaker independent and capable of dealing with connected speech in interactive conversation (including the use of "word-spotting" to deal with the inadvertent utterances humans typically make when talking with a machine). For high reliability, these systems are still limited to relatively small vocabularies (hundreds of words), but can readily serve large user populations. They also typically use finite state grammars that are designed for specific tasks. Speech input is therefore restricted to sentence constructions permitted in this

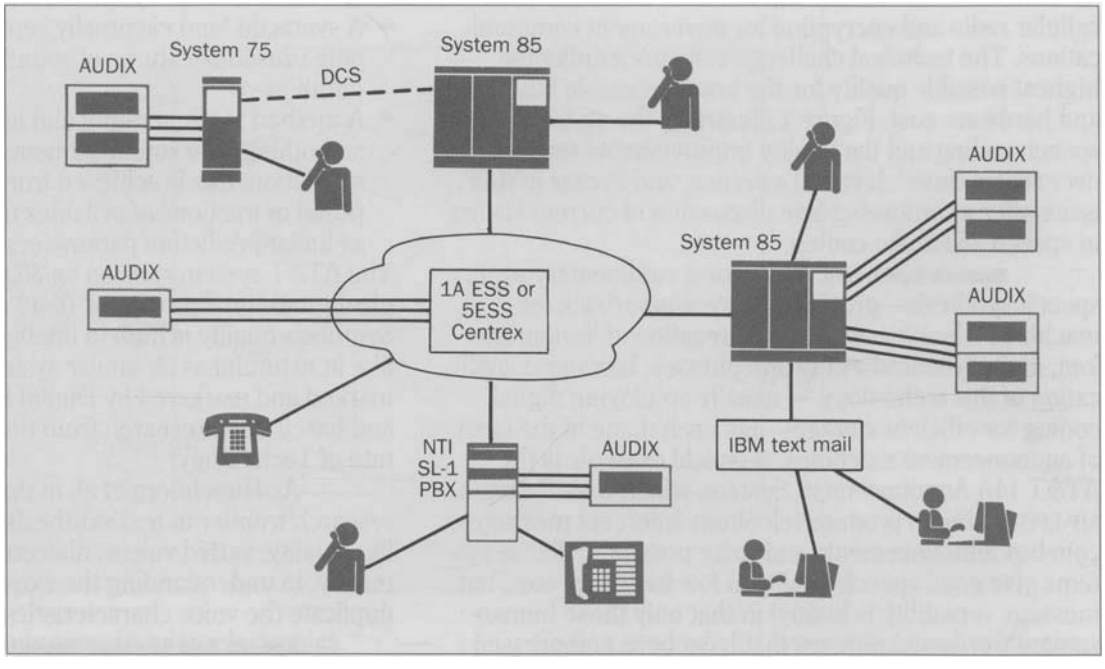


Figure 2. Through standardizing digital formats, AT&T's AUDIX system for voice messaging interfaces with a wide variety of telecommunications servers. Included are private branch exchanges (PBXs) such as AT&T's Definity® System 85 and System 75. Network connectivity is supported by central switches such as the 1A, ESS™ and 5ESS® switches.

delimited sub-set of natural language.

Research is now addressing vocabularies greater than 1,000 words, based on recognition units of subword length. Statistically-based subword models use hidden Markov techniques to estimate spoken word sequences. Improved models of language will aid the task by quantifying syntax, semantics and possibly even pragmatics; and by admitting input speech more representative of natural language. Substantial advances in economic high-speed computation will be needed to support large-vocabulary systems and natural language input. These advances can realistically be expected over the next decade. Wilpon et al. in this issue review current capabilities and the outlook for continued progress.

Talker recognition is closely related to speech recognition. But whereas speaker-independent *speech recognition* tries to ignore individual distinguishing characteristics, *talker recognition* systems measure these attributes to determine speaker identity. An important mode of talker recognition is *talker verification*, where a user wanting access to privileged data or restricted premises is required to make an identity claim. The machine must then authenticate the claim by measurements on the voice signal and comparison to previously obtained patterns. Current techniques use cepstral patterns to characterize talkers. They achieve accuracies in the high 90-percent range, nominally independent of user population size.[9] Experimental AT&T systems are now supported on a single DSP32C chip.

Teleconferencing. New understanding in acoustics, inexpensive electret transducers and sophisticated microelectronics combine to support autodirective speech-seeking beam-forming microphone arrays for hands-free audio conferencing. The benefit is high-quality sound pickup for large-group conferencing. Using multiple beam-forming and track-while-scan algorithms,

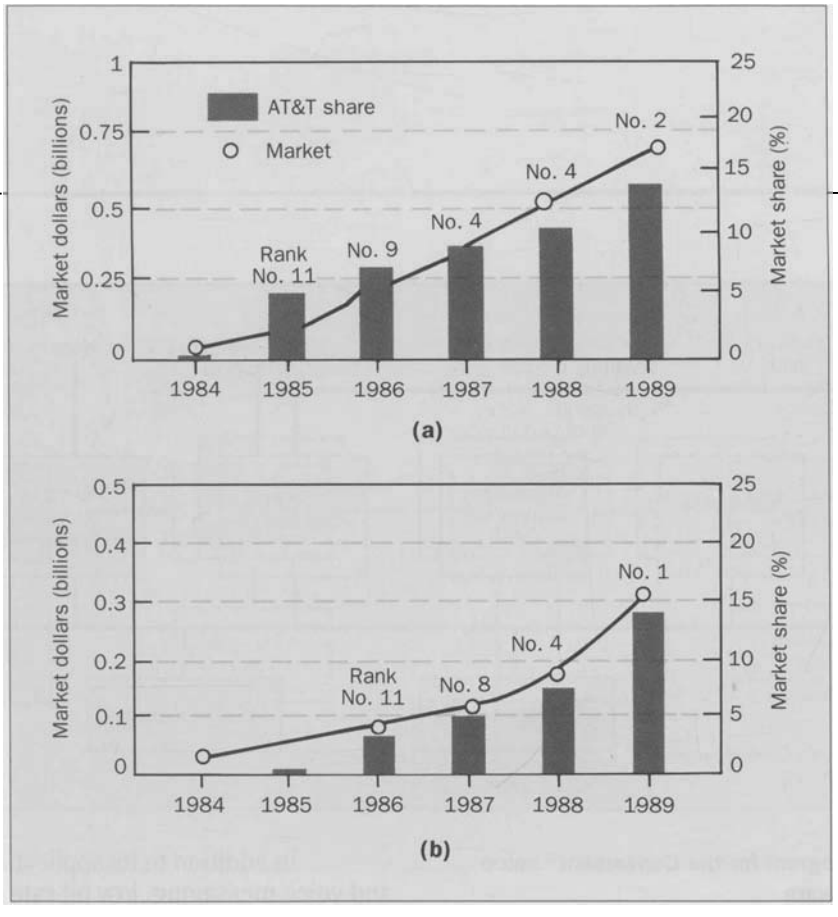


Figure 3. AT&T sales performance in (a) voice processing equipment; and (b) voice messaging equipment.

the systems can dynamically locate and lock onto active talkers to obtain a high quality signal in the presence of reverberation and noise.[10] Speech processing algorithms for distinguishing continuous speech from interfering noise permit the systems to point to and follow talkers as they move about the conference room.

Business

Until recently, commercial applications of speech processing have been largely in voice response, voice mail and transmission. Voice response and voice messaging (mail) systems use speech compression, or low bit-rate coding, for economic storage and high-quality readout. For example, AT&T's AUDIX system for voice mail uses 16 kb/s for its storage. This system has great versatility and interfaces with a great variety of communication servers[4] see Figure 2).

The domestic market for voice processing equipment (exclusive of transmission) is presently estimated to exceed $1 billion per year, and is growing at about 30 percent per year. This growth is due to the increasing ability of today's systems to successfully meet customers' business needs. Based on 1989 revenue, AT&T is a leading provider of voice processing equipment. The AUDIX and AUDIX Voice Power products have attained a 13 percent market share in the voice messaging segment (Figure 3a) and the Conversant® Voice Information System product has attained a 13 percent market share and the leadership position in the voice response segment (Figure 3b). The Conversant voice information system is based on an AT&T 6386 computer with multiple DSP32 signal processors that permit extensions to about 1.5 gigaflops of total processing power. The product includes telephone interface, Touch-Tone decoding, voice response and mail features.

Figure 4 depicts the Conversant® voice information system hardware architecture. The signal processor (SP) board contains a Motorola 68020 controller and two

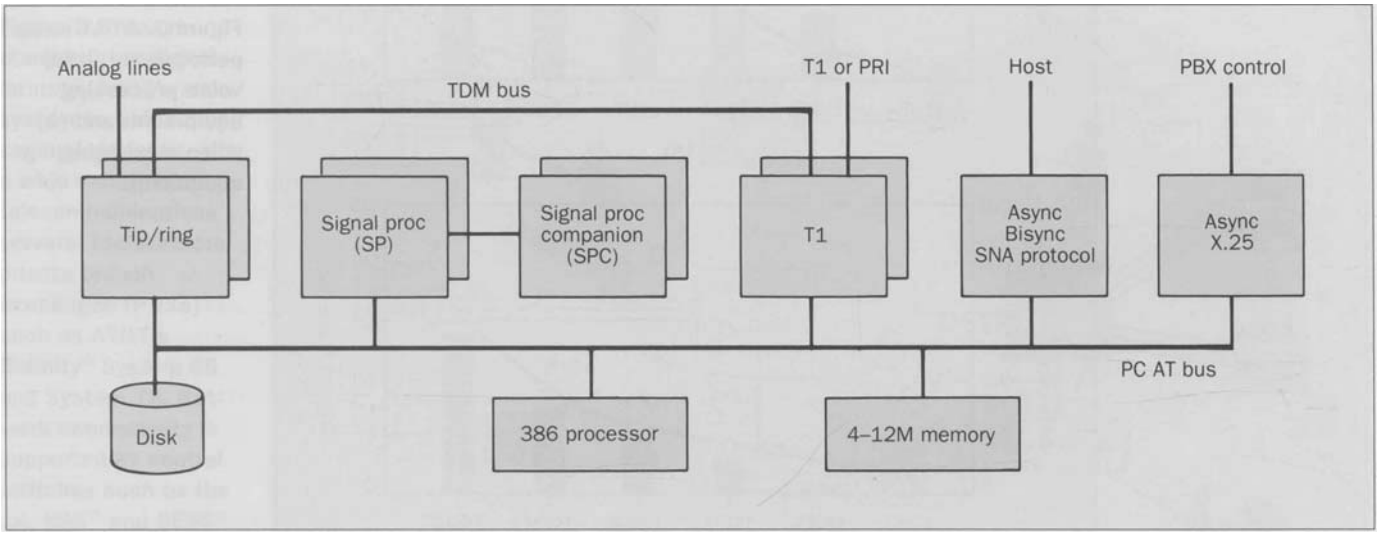


Figure 4. Architecture diagram for the Conversant® voice information system hardware.

DSP32C's for a total of 60 megaflops of computing power. The signal processor companion (SPC) board contains 12 DSP32C's and 4 mbytes of memory for a total of 300 megaflops of computing power. Up to 4 companion boards can be attached to a single SP board. One SP and 4 SPCs can be combined to perform 24 channels of speaker independent, connected digit speech recognition at economical per-channel cost.

The technologies of speech recognition and speaker verification are now emerging from the research laboratory in robust, practical forms. Their reliability and capability have been established through new research on speaker independence and techniques for recognizing key words in a surround of extraneous speech. These technologies, though computationally demanding, are economically supported by recent advances in microprocessors, such as those used in the Conversant voice information system equipment. Consequently, a number of significant new business applications for speech recognition and speaker verification are rapidly developing.

In addition to its applications in voice response and voice messaging, low bit-rate coding also has important applications in transmission. These applications have largely been focused on private line, experimental digital cellular telephone, specialized secure communications, and experimental wideband audio storage and transmission. With the advent of ISDN (Integrated Services Digital Network) and end-to-end public-switched digital connectivity, low bit-rate coding will provide a variety of new services for voice and audio information, including mobile wireless personal communication. Standards organizations—especially international groups—are currently active in all these areas. As with facsimile, standardization—along with ease of use and utility—will stimulate mass deployment of speech processing for both messaging and transmission. In this issue, Fischell et al., Verma et al., and Berkley and Flanagan discuss a number of these emerging applications.

The Future

Our projections for the future are based on several beliefs:

- Speech is a preferred means for human-machine

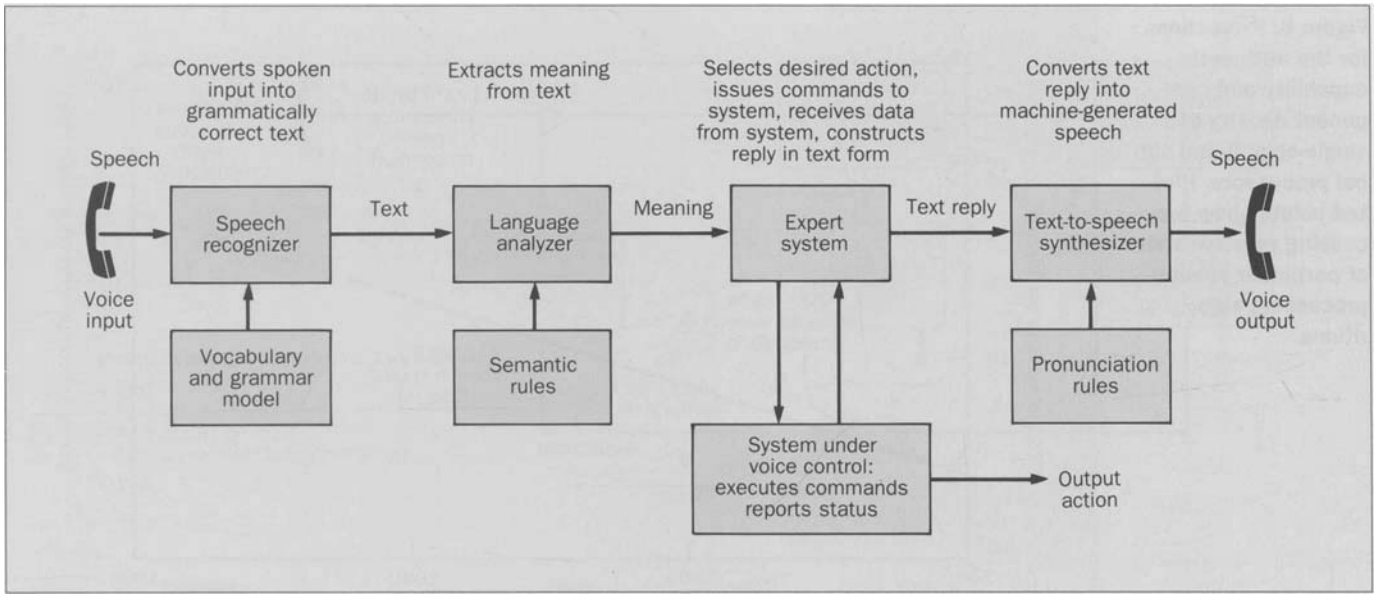


Figure 5. Generic use of speech recognition and text synthesis for task-specific applications. Speech recognizers typically operate with connected-word, speaker independent input using a vocabulary size of several hundred words, and a finite-state grammar and stored semantic rules that adequately span the specific task application. The expert system may be a database of great variety (ranging from airline flight information through the work-space model for a robot). Text synthesis permits great versatility for the machine to report its actions and conduct conversational exchanges with the human user.

communication.

- Pervasive switched digital connectivity (first in the form of basic-rate ISDN, and later as broader-band services) will permit immediate, ubiquitous access to communications, computation, and information.
- Personal (and personalized) information systems will become common, using integrated modalities for voice, image, data, and hands-free conferencing.
- Ease of use, low cost, and utility are crucial to mass deployment and acceptance of new systems and services.
- Digital signal processors will continue to expand in capability and will support speech-processing algorithms of enormous complexity.[11]

Our present office and home communications environments comprise telephones, speakerphones, personal computers (PCs), terminals, modems, TVs with cable, videocasette recorders (VCRs) and compact disc (CD) stereos. In the future these functions will be subsumed in workstations with optical stores, digital switched networking, and integrated capabilities for voice, image, audio, video, fax, data, graphics, and natural language communication. Spatial realism in sight and sound will be supported, as will terminal sensors that aid the human user in information capture and display, and provide hands-free operation.

In the coming era, as in the past, investing in superior knowledge will be good business. But the investment will be good business only if the knowledge is used expeditiously.

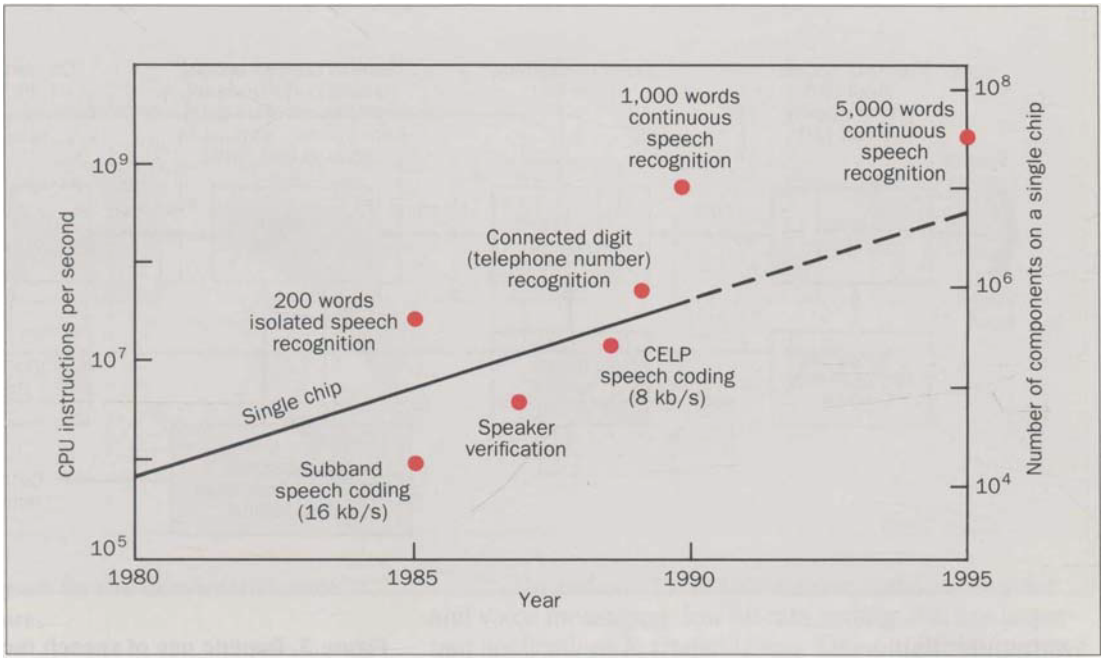


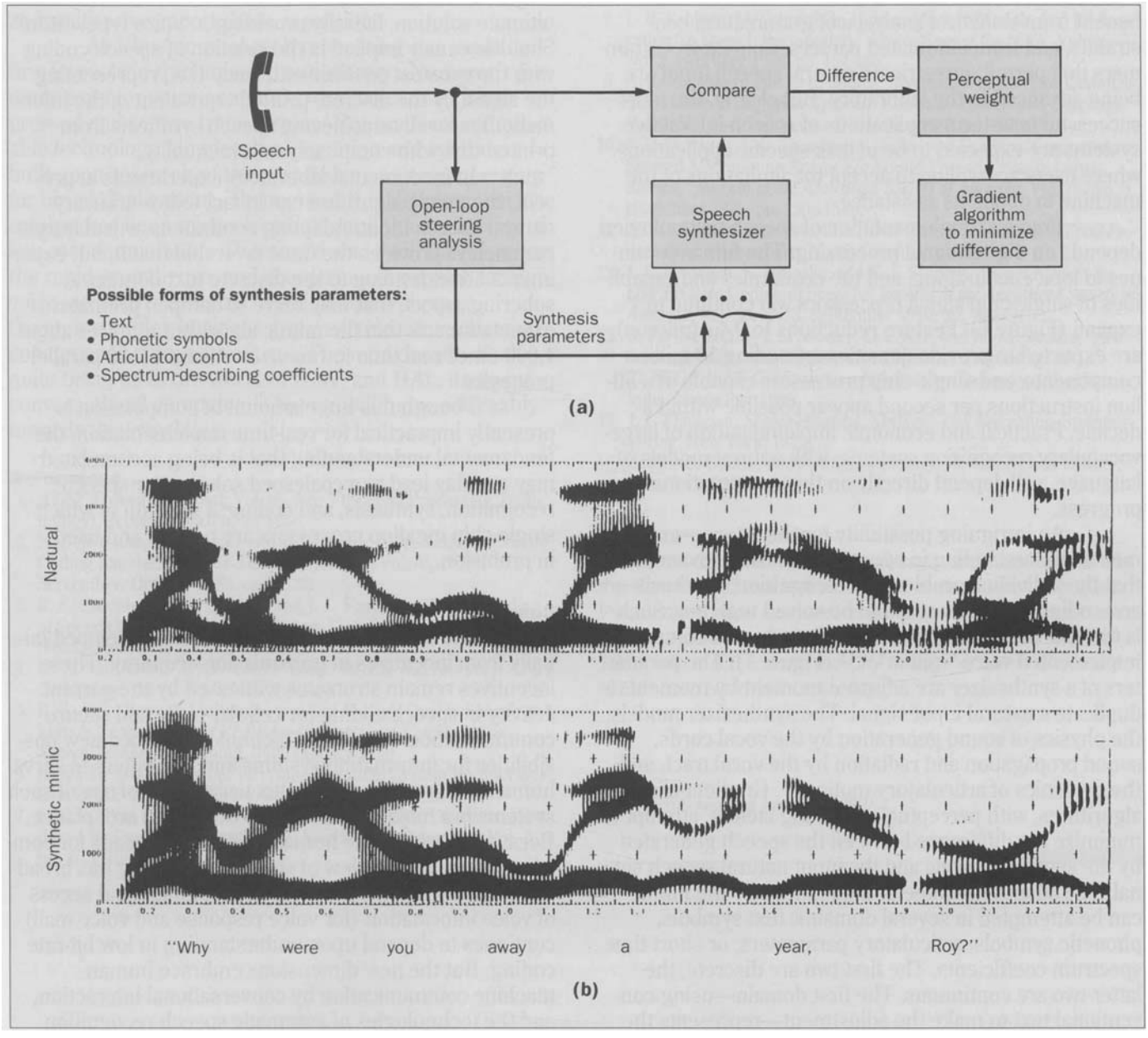
Figure 6. Projections for the arithmetic capability and component density of single-chip digital signal processors. Plotted points show processing requirements of particular speech-processing algorithms.

Fundamental understanding in coding of information signals, along with economic microelectronics, will provide high-quality speech, wideband audio, and image over commonplace digital capacities such as ISDN (see Berkley and Flanagan's paper on HuMaNet in this issue).

Conversational interaction with machines and computers—with natural connected speech, speaker independence, and high reliability—will for the foreseeable time remain restricted to task-specific applications (Figure 5). This implies word vocabularies in the order of several hundred, with models of grammar that are restricted finite-state subsets of natural language. Nevertheless, the astute user who understands and accepts the machine's non-human limitations will find the information system remarkably useful. Voice response through synthesis from unrestricted text completes the conversational capability. The challenge in the latter arena is to achieve natural quality as well as high intelligibility.

Vocabulary sizes for speech recognition will increase through research on sub-word units. Systems with vocabularies in excess of 1,000 words are already running in the laboratory, and the long-range goal of vocabularies exceeding 10,000 words seems realistic. Language models that span greater subsets of English are crucial to continued advances, and this work will

Figure 7. Concept of a computer-based speech mimic. (a) The mimic attempts to duplicate the natural speech signal through moment-by-moment adjustment of its control parameters. The parametric domain, in the most ambitious form, may be discrete text or phonetic characters, and in a continuous form, articulatory vocal-tract shape controls or spectral coefficients. (b) The sound spectrogram shows the computer mimic duplication of the input phrase *Why were you away a year, Roy?* Computation time is about 1,000 times real time on a parallel processor.


Speech input
Compare
Difference
Perceptual weight
Open-loop steering analysis
Speech synthesizer
Gradient algorithm to minimize difference
Synthesis parameters
Possible forms of synthesis parameters:
• Text
• Phonetic symbols
• Articulatory controls
• Spectrum-describing coefficients
(a)
Natural
Synthetic mimic
"Why were you away a year, Roy?"
(b)

benefit from statistical analyses of grammatical constraints, and from automated parsers. Context-free grammars that permit unrestricted natural speech input are being advanced in the laboratory. But clearly, the most successful near-term applications of speech-interactive systems are expected to be in task-specific applications where users are willing to accept the limitations of the machine to obtain its assistance.

Practical implementation of speech technologies depends on digital signal processing. The future continues to look encouraging, and the economies and capabilities of single-chip signal processors will continue to expand (Figure 6). Feature reductions to 0.4μ (micron) are expected to provide densities exceeding 10 million components, and single-chip processors capable of a billion instructions per second appear possible within a decade. Practical and economic implementation of large-vocabulary recognition systems, with natural models of language, will depend directly on this computational progress.

An intriguing possibility for the future—as fundamental understanding in speech processing advances—is that the individual problems of recognition, synthesis and coding may coalesce, and be solved together. Such is the goal of one ambitious AT&T study on a computer-implemented voice "mimic"[12,13] (Figure 7). The parameters of a synthesizer are adjusted moment by moment to duplicate a natural input signal. The synthesizer models the physics of sound generation by the vocal cords, sound propagation and radiation by the vocal tract, and the dynamics of articulatory motion.[14]. Gradient-climbing algorithms, with perceptual weighting factors, attempt to minimize the difference between the speech generated by the synthetic mimic and the input natural speech signal. The adaptive parameter adjustment for the mimic can be attempted in several domains: text symbols, phonetic symbols, articulatory parameters, or short-time spectrum coefficients. The first two are discrete, the latter two are continuous. The first domain—using conventional text to make the adjustment—represents the ultimate solution, literally providing a "voice typewriter." Simultaneously implied is the solution of speech coding with the greatest possible efficiency (i.e., representing the signal by the discrete printed equivalent of the information), as well as achieving speech synthesis from printed text with completely natural quality.

In fundamental laboratory experiments at present, the mimic algorithm can in fact follow arbitrary natural speech input. Adapting continuous articulatory parameters provides the most favorable result, but experiments are extending to the discrete text domain. A sobering aspect, that may serve to dampen premature expectations, is that the mimic algorithm requires about 1,000 times real time to run on an Alliant FX-80 parallel processor.

Though this huge amount of computation is presently impractical for real-time implementation, the fundamental understanding that is being accumulated may one day lead to a coalesced solution for speech recognition, synthesis, and coding: a solution in which single-chip gigaflop processors are needed and used in profusion.

Conclusion

The science of speech processing developed initially from incentives in transmission economy. These incentives remain strong, as witnessed by the current activity in speech coding for cellular radio and secure communication. But digital technology opened new possibilities for information systems and computers to serve humans in a variety of complex tasks. Ease-of-use of such systems is a major factor in their utility and acceptance. Because speech is the human's preferred means for communication, the purview of speech processing has broadened to these applications. Efficient storage and access of voice information (for voice response and voice mail) continues to depend upon understanding in low bit-rate coding. But the new dimensions embrace human-machine communication by conversational interaction, and the technologies of automatic speech recognition

and synthesis support these capabilities.

As a business, speech processing has made its first impacts in transmission and storage. Numerous products exist for voice announcement, voice mail, compressed and secure transmission. But practical and reliable technology for speech recognition and synthesis, built upon research of the past decade, is now emerging for broad deployment. Throughout this development, the catalyst has been microelectronics and the availability of low-cost computation. The next decade will therefore see the rapid growth, use, and acceptance of sophisticated voice-interactive systems. Applications will range through complex business systems, telecommunication, consumer and home products. The next decade won't quite bring us to the world of *2001*, and HAL, its fluent conversational computer,[15] but we will move noticeably towards this capability.

References

1. *AT&T Technical Journal*, Vol. 65, No. 5 (Speech Processing Technology), September/October 1986.
2. N. Benvenuto, G. Bertocci, and W. Daumer, "The 32 kb/s ADPCM Coding Standard," *AT&T Technical Journal*, Vol. 65, No. 5, September/October 1986, pp. 12-22.
3. R. E. Crochiere, S. A. Webber, and J. L. Flanagan, "Digital Coding of Speech in Sub-Bands," *Bell System Technical Journal*, Vol. 55, No. 8, October 1976, pp. 1069-1085.
4. M. A. Van Andel, "While You're Away, AUDIX Will Answer," *AT&T Technology*, Vol. 3, No. 3, 1988, pp. 34-41.
5. B. S. Atal, "Beyond Multipulse and CELP: High Quality Speech at 4 kbits/s," *Proceedings of VERBA—International Conference on Speech Technologies*, Rome, Italy, January 1990.
6. AT&T Network Systems, 14A Announcement System Product Bulletin 2664D PER-11/87, Morristown, New Jersey, 1987.
7. L. C. W. Pols and J. P. Olive, "Intelligibility of CVC Utterances Produced by Dyadic Rule Synthesis," *Speech Communication*, Vol. 2, pp. 3-13, 1983.
8. L. R. Rabiner, "A Tutorial On Hidden Markov Models And Its Application To Speech Recognition," *Proceedings of the IEEE*, Volume 77, No. 2, February 1989, pp. 157-186.
9. A. E. Rosenberg and F. K. Soong, "Evaluation of a Vector Quantization Talker Recognition System," *Computer Speech and Language*, Vol. 2, No. 3/4, September-December 1987, pp. 143-157.
10. J. L. Flanagan, J. D. Johnston, R. Zahn, and G. Elko, "Computer-Steered Microphone Arrays For Sound Transduction In Large Rooms," *Journal of the Acoustical Society of America*, Vol. 78, No. 5, November 1985, pp. 1508-1518.
11. AT&T Federal Systems, *AT&T DSP Parallel Processor, BT100 Product Brief*, Greensboro, North Carolina, 1988.
12. J. L. Flanagan, K. Ishizaka, and K. L. Shipley, "Signal Models for Low Bit-Rate Coding of Speech," *Journal of the Acoustical Society of America*, Vol. 68, No. 3, September 1980, pp. 780-791.
13. S. Parthasarathy, J. Schroeter, C. Coker, and M. M. Sondhi, "Articulatory Analysis and Synthesis of Speech," *Proceedings of TENCON-IEEE Conference on Information Technology*, Bombay, India, November 1989.
14. J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, Springer-Verlag, New York, 1972.
15. Arthur C. Clarke, *2001: A Space Odyssey*, New American Library, New York, 1972.

Biographies (continued)

Brunswick, New Jersey. Mr. Del Riesgo leads development of the Conversant® speech processor, and the AUDIX and AUDIX Voice Power voice systems, and call management, gateways, and data products PBX system adjuncts. He joined AT&T in 1961 with a B.E.E. from the City College of New York, and an M.E.E. from New York University.

(Manuscript received June 14, 1990)