Research in speech communication

JAMES FLANAGAN

CAIP Center, Rutgers University, New Brunswick, NJ 08903

ABSTRACT Advances in digital speech processing are now supporting application and deployment of a variety of speech technologies for human/machine communication. In fact, new businesses are rapidly forming about these technologies. But these capabilities are of little use unless society can afford them. Happily, explosive advances in microelectronics over the past two decades have assured affordable access to this sophistication as well as to the underlying computing technology. The research challenges in speech processing remain in the traditionally identified areas of recognition, synthesis, and coding. These three areas have typically been addressed individually, often with significant isolation among the efforts. But they are all facets of the same fundamental issue—how to represent and quantify the information in the speech signal. This implies deeper understanding of the physics of speech production, the constraints that the conventions of language impose, and the mechanism for information processing in the auditory system. In ongoing research, therefore, we seek more accurate models of speech generation, better computational formulations of language, and realistic perceptual guides for speech processing—along with ways to coalesce the fundamental issues of recognition, synthesis, and coding. Successful solution will yield the long-sought dictation machine, high-quality synthesis from text, and the ultimate in low-bit-rate transmission of speech. It will also open the door to language-translating telephony, where the synthetic foreign translation can be in the voice of the originating talker.

Speech is a preferred means for communication among humans. It is beginning to be a preferred means for communication between machines and humans. Increasingly, for well-delimited tasks, machines are able to emulate many of the capabilities of conversational exchange. The power of complex computers can therefore be harnessed to societal needs without burdening the user beyond knowledge of natural spoken language.

Because humans are designed to live in an air atmosphere, it is inevitable that they learn to convey information in the form of longitudinal waves supported by displacement of air molecules. But of the myriad acoustic information signals, speech is a very special kind. It is constrained in three important ways:

- by the physics of sound generation in the vocal system,
- by the properties of human hearing and perception, and
- by the conventions of language.

These constraints have been central to research in speech and remain of paramount importance today.

This paper proposes to comment on the field of speech communication in three veins:

- first, in drawing a brief perspective on the science;
- second, in suggesting critical directions of research; and
- third, in hazarding some technology projections.

FOUNDATIONS OF SPEECH TECHNOLOGY

Speech processing, as a science, might be considered to have been born from the evolution of electrical communication. Invention of the telephone, and the beginning of telecommunications as a business to serve society, stimulated work in network theory, transducer research, filter design, spectral analysis, psychoacoustics, modulation methods, and radio and cable transmission techniques. Early on, the acoustics and physiology of speech generation were identified as critical issues for understanding. They remain so today, even though much knowledge has been acquired. Alexander Graham Bell was among those who probed the principles of speech generation in experiments with mechanical speaking machines. (He even attempted to teach his Skye terrier to articulate while sustaining a growl!) Also, it was recognized early that properties of audition and perception needed to be quantified, in that human hearing typically provides the fidelity criterion for receiving speech information. Psychoacoustic behavior for thresholds of hearing, dynamic range, loudness, pitch, and spectral distribution of speech were quantified and used in the design of early telecommunication systems. But only recently, with advances in computing power, have efforts been made to incorporate other subtleties of hearing—such as masking in time and frequency—into speech-processing algorithms. Also, only recently has adequate attention been turned to analytical modeling of language, and this has become increasingly important as the techniques for text-to-speech synthesis and automatic recognition of continuous speech have advanced.

About the middle of this century, sampled-data theory and digital computation simultaneously emerged, opening new vistas for high-quality long-distance communication and for simulating the engineering design of complex systems rapidly and economically. But computing technology soon grew beyond data sorting for business and algorithm simulation for science. Inexpensive arithmetic and economical storage, along with expanding knowledge of information signals, permitted computers to take on functions more related to decision making—understanding subtle intents of the user and initiating ways to meet user needs. Speech processing—which gives machines conversational capability—has been central to this development. Image processing and, more recently, tactile interaction have received similar emphases. But all these capabilities are of little use unless society can afford them. Explosive advances in microelectronics over the past two decades have assured affordable access to this sophistication as well as to the underlying computing technology. All indications are that computing advances will continue and that economical computation to support speech technology will be in place when it is needed.

INCENTIVES IN SPEECH RESEARCH

Acoustic waves spread spherically and do not propagate well over distances. But communication over distances has long
been a need in human society. As understanding of electrical phenomena progressed, the electrical telegraph emerged in the mid-nineteenth century. Following this success with dots and dashes, much attention turned to the prospect of sending voice signals over electrical wires. Invention of the telephone is history.

In the early part of the twentieth century, the incentive remained voice communication over still greater distances. Amplification of analog signals, which attenuate with distance and accumulate noise, was needed. In 1915 transcontinental telephone was achieved with marginal fidelity by electromechanical "repeaters." Transatlantic telegraph cables could not support the bandwidth needed for voice, and research efforts turned to "vocoders" for bandwidth compression. In 1927, as electronics technology emerged, transatlantic radio telephone became a reality. Understanding of bandwidth compression was then applied to privacy and encryption. Transatlantic voice on wire cable had to await the development of reliable submersible amplifiers in 1956. With these expensive high-quality voice circuits, the interest in bandwidth conservation again arose and stimulated new developments, such as Time Assignment Speech Interpolation, which provided nearly a three-fold increase in cable capacity.

From the mid-twentieth century, understanding emerged in sampled-data techniques, digital computing, and microelectronics. Stimulated by these advances, a strong interest developed in human/machine communication and interaction. The desire for ease of use in complex machines that serve human needs focused interest on spoken language communication (1-3). Significant advances in speech recognition and synthesis resulted. Bandwidth conservation and low bit-rate coding received emphasis as much for economy of storage (in applications such as voice mail) as for savings in transmission capacity. The more recent developments of mobile cellular, personal, and cordless telecommunications have brought renewed interest in bandwidth conservation and, concomitantly, a heightened incentive for privacy and encryption.

As we approach the threshold of the twenty-first century, fledging systems are being demonstrated for translating telephony (Fig. 1). These systems require automatic recognition of large fluent vocabularies in one language by a great variety of talkers; transmission of the inherent speech information; and natural-quality synthesis in a foreign language—preferably with the exact voice quality of the original talker (Fig. 2). At the present time, only "phrase book" type of translation is accomplished, with limited grammars and modest vocabularies, and the synthesized voice does not duplicate the quality of individual talkers. Translating telephony and dictation machines require major advances in computational models of language that can accommodate natural conversational grammars and large vocabularies. Recognition systems using models for subword units of speech are envisioned, with linguistic rules forming (i) acceptable word candidates from the estimated strings of phonetic units, (ii) sentence candidates from the word strings, and (iii) semantic candidates from the sentences. Casual informal conversational speech, with all its vagaries and nongrammatical structure, poses special challenges in devising tractable models of grammar, syntax, and semantics.

**TECHNOLOGY STATUS**

A fundamental challenge in speech processing is how to represent, quantify, and interpret information in the speech signal. Traditionally, research focuses on the sectors of coding, speech and speaker recognition, and synthesis.

**Coding.** High-quality digital speech coding has been used for many years in telecommunications in the form of Pulse Code Modulation (PCM), using a typical transmission rate of 64k bits/sec. In recent years, capacity-expanding Adaptive Differential PCM (ADPCM) at 32k bits/sec has served in the telephone plant, particularly for private lines. Economical systems for voice mail have derived from compression algorithms for 16k bits/sec Sub-Band Coding and low-delay Code Excited Linear Prediction (CELP), and this technology—implemented for 8k bits/sec—is currently being tested in digital mobile cellular telephones.

Signal quality typically diminishes with coding rate, with a notable "knee" at about 8k bits/sec. Nevertheless, vocoder rates of 4k and 2k bits/sec are finding use for digital encryption over voice bandwidth channels. The challenge in coding is to elevate quality at low transmission rates. Progress is being made through incorporation of perceptual factors and through improved representation of spectral and excitation parameters (4).

There are experimental reasons to believe that high quality can be achieved at rates down to the range of 2000 bits/sec. Improvements at these rates may come from two directions: (i) dynamic adaptation of perceptual criteria in coding, and (ii) articulatory modeling of the speech signal.

In coding wideband audio signals the overt use of auditory perception factors within the coding algorithm ("hearing-
specific" coders) has been remarkably successful, allowing wideband signal representation with an average of less than two bits per sample. The implication of this is that FM stereo broadcast quality can be transmitted over the public switched digital telephone channels provided by the basic-rate ISDN (Integrated Services Digital Network) (Fig. 3). Alternatively, one can store up to eight times more signal on a high-fidelity compact disc recording than is conventionally done.

For stereo coding, the left-plus-right and left-minus-right signals are transform-coded separately (typically by 2048-point FFTs). For each spectrum at each moment, a masking threshold is computed, based on the distribution of spectral energy and on critical-band masking in the ear. Any signal components having spectral amplitudes less than this threshold will not be heard at that moment in the presence of stronger neighbors; hence, these components need not be allocated any bits for transmission. Similarly, if bits are assigned to the stronger components so that the quantizing noise spectrum is maintained below this masking threshold, the quantizing noise will not be audible. The computation to accomplish the coding, while substantial, is not inordinate in terms of presently available DSP chips.

This and related techniques are strongly influencing international standards for speech and music coding. And it appears that continued economies can be won through perceptual factors such as masking in the time dimension. (See subsequent discussion of temporal masking.)

Recognition and Synthesis. Unhappily, advances in recognition and in synthesis, particularly in text-to-speech synthesis, have not been strongly coupled and have not significantly cross-nurtured one another. This seems to be largely because recognition has taken a pattern-matching direction, with the immensely successful hidden Markov models (HMMs), while synthesis has relied heavily on acoustic phonetics, with formant models and fractional-syllable libraries contributing to the success. Nevertheless, the techniques are destined to be used hand in hand in voice-interactive systems. Both can benefit from improved computational models of language.

Present capabilities for machine dialogue permit intelligent fluent interaction by a wide variety of talkers provided the vocabulary is limited and the application domain is rigorously constrained (ref. 5; Fig. 4). Typically, a finite-state grammar is used to provide enough coverage for useful conversational exchange. Vocabularies of a couple hundred words and a grammar that permits billions of sentences about a specific task—say, obtaining airline flight information—are typical. Word recognition accuracy is above 90 percent for vocabularies of several hundred words spoken in connected form by a wide variety of talkers. For smaller vocabularies, such as the digits, recognition accuracies are also in the high 90s for digit strings (e.g., seven-digit telephone numbers) spoken in connected form. With currently available signal processor chips the hardware to support connected-digit recognition is relatively modest.

Again, a significant frontier is in developing computational models of language that span more natural language and permit unfettered interaction. Computational linguistics can make strong contributions in this sector.

**Talker Verification.** Using cepstrum, delta cepstrum, and HMM techniques, the ability to authenticate "enrolled" talkers over clean channels is relatively well established (6). The computation needed is easily supported, but not much commercial deployment has yet been seen. This results not so much from any lack of desire to have and use the capability but to an apparently low willingness to pay for it. Because speech recognition and talker verification share common processes, combining the features in an interface is natural. The investment in recognition can thereby provide verification for a minimal increment in cost. New applications of this type are emerging in the banking sector where personal verification is needed for services such as cash-dispensing automatic teller machines.

**Autodirective Microphone Arrays.** In many speech communication environments, particularly in teleconferencing and in the use of voice-interactive terminals, it is more natural to communicate without hand-held or body-worn microphones. The freedom to move about the work place, without tether or encumbrance, and to speak as in face-to-face conversation is frequently an advantage. Autodirective microphone arrays, especially beam-forming systems, permit good-quality sound pickup and mitigate the effects of room reverberation and interfering acoustic noise (7).

High-performance, low-cost electret microphones (Fig. 5), in combination with economical distributed signal processors, make large speech-seeking arrays practical. Each sensor can have a dedicated processor to implement beam forming and steering. A host controller issues appropriate beam-forming and beam-pointing values to each sensor while supporting algorithms provide sound source location and speech/non-speech identification. The array is typically used with multiple beams in a "track-while-scan" mode. New research on three-dimensional arrays and multiple beam forming is leading to high-quality signal capture from designated spatial volumes.

**CRITICAL DIRECTIONS IN SPEECH RESEARCH**

**Physics of Speech Generation: Fluid-Dynamic Principles.** The aforementioned lack of naturalness in speech generated from compact specifications stems possibly from two sources. One is the synthesizer's crude approximation to the acoustic properties of the vocal system. The other is the shortcomings in control data that do not adequately reflect natural articu-

---

**Fig. 3.** Influence of digital representation on audio signal quality. Increasingly complex algorithms are used as representation bits per sample diminish. Hearing-specific coders incorporate human perceptual factors, such as masking in frequency.

**Fig. 4.** Recognition and synthesis systems permit task-specific conversational interaction. Expansions of vocabulary size, talker independence, and language models that more nearly approach natural spoken language, together with high-quality synthesis, are research targets (5).

---

**Table 1.**

<table>
<thead>
<tr>
<th>Audio Quality</th>
<th>Wideband</th>
<th>Telephone Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Fair</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Bits/ Nyquist Sample</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech Specific Coders</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Hearing Specific Coders</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Overall</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

FIG. 5. Large two-dimensional array of 408 electret microphones. Each microphone has a dedicated chip for beam forming.

lation and prosody. Both of these aspects affect speech quality and certainly affect the ability to duplicate individual voice characteristics.

Traditional synthesis takes as its point of departure a source-filter approximation to the vocal system, wherein source and filter do not interact (Fig. 6). Typically, the filter function is approximated in terms of a hard-walled tube, supporting only linear one-dimensional wave propagation. Neither is realistic.

Advances in parallel computation open the possibility for implementing speech synthesis from first principles of fluid dynamics. Given the three-dimensional, time-varying, soft-wall vocal tract, excited by periodically valved flow at the vocal cords and by turbulent flow at constrictions, the Navier-Stokes equation can be solved numerically on a fine space-time grid to produce a remarkably realistic description of radiated sound pressure. Nonlinearities of excitation, generation of turbulence, cross-modes of the system, and acoustic interaction between sources and resonators are taken into account. The formulation requires enormous computation, but the current initiatives in high-performance computing promise the necessary capability.

Computational Models of Language. Already mentioned is the criticality of language models for fluent, large-vocabulary speech recognition. Tractable models that account for grammatical behavior (in spoken language), syntax, and semantics are needed for synthesis from text as urgently as for recognition (Fig. 7). Statistical constraints in spoken language are as powerful as those in text and can be used to complement substantially the traditional approaches to parsing and determining parts of speech.

Information Processing in the Auditory System; Auditory Behavior. Mechanics and operation of the peripheral ear are relatively well understood. Psychoacoustic behavior is extensively quantified. Details of neural processing, and the mechanism for interpreting neural information, are not well established. But this does not preclude beneficially utilizing behavioral factors in speech processing. Over the past, telecommunications and audio technology have exploited major

FIG. 6. Traditional representation of sound generation and propagation in the vocal tract. One-dimensional approximation of sound propagation permits computation of pressure and velocity distributions along tract and at radiating ports. Turbulent excitation is computed from the Reynolds number at each location along the tract. Vocal cord simulation permits source-filter interaction.
TABLE III

TRIGRAM PROBABILITIES (%)

<table>
<thead>
<tr>
<th>TRIGRAM</th>
<th>ITALIAN</th>
<th>JAPANESE</th>
<th>GREEK</th>
<th>FRENCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>igh</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>ett</td>
<td>70</td>
<td>0</td>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>cci</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>feu</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>oco</td>
<td>0</td>
<td>61</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>mus</td>
<td>0</td>
<td>86</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>los</td>
<td>4</td>
<td>0</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td>dis</td>
<td>3</td>
<td>0</td>
<td>74</td>
<td>5</td>
</tr>
<tr>
<td>bia</td>
<td>0</td>
<td>6</td>
<td>75</td>
<td>0</td>
</tr>
<tr>
<td>eeu</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>nte</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>dis</td>
<td>10</td>
<td>6</td>
<td>0</td>
<td>61</td>
</tr>
<tr>
<td>gea</td>
<td>0</td>
<td>0</td>
<td>38</td>
<td>14</td>
</tr>
<tr>
<td>eil</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>

FIG. 7. Illustrative probabilities for selected text trigrams across several languages (10 in total). While the number of possible trigrams is on the order of 20,000, the number of trigrams that actually occur in the language is typically fewer by an order of magnitude—constituting great leverage in estimating allowed symbol sequences within a language and providing a tool for estimating etymology from the individual probabilities. (Data from K. Church, AT&T Bell Laboratories.)

Aspects of human hearing such as ranges of frequency, amplitude, and signal-to-noise ratio. But now, with inexpensive computation, additional subtleties can be incorporated into the representation of audio signals. Already high-fidelity audio coding incorporates some constraints of simultaneous masking in frequency. Additionally, masking in time is an obvious target of opportunity (Fig. 8). Relatively untouched, so far, is the esoteric behavior of binaural release from masking, wherein interaural phase markedly controls perceptibility.

Coalescing Speech Coding, Synthesis, and Recognition. The issues of coding, recognition, and synthesis are not disjoint—they are facets of the same underlying process of speech and hearing. We might strive therefore for research that unifies the issues from the different sectors. Better still, we might seek an approach that coalesces the problems into a common understanding. One such effort is the “voice mimic.”

The voice mimic attempts to generate a synthetic speech signal that, within perceptual accuracy, duplicates an input of arbitrary natural speech. Central to the effort is a computer model of the vocal cords and vocal tract (to provide the

FIG. 8. Illustration of the time-frequency region surrounding intense, punctuate signals where masking in both time and frequency is effective.

FIG. 9. Computer voice mimic system. Natural continuous input speech is approximated by a computed synthetic estimate. Spectral differences between real and synthetic signals are perceptually weighted and used in a closed loop to adjust iteratively the parameters of the synthesis, driving the difference to a minimum.
acoustic synthesis) (Fig. 9), a dynamic model of articulation described by nearly orthogonal vocal-tract shape parameters (to generate the cross-sectional area function), and, ideally, a discrete phonetic symbol-to-shape mapping. A perceptually weighted error, measured in the spectral domain for natural and synthetic signals, drives the synthesis parameters so as to minimize the mimicking error, moment by moment. Open-loop analysis of the input natural speech is useful in steering the closed-loop optimization.

Ideally, one would like to close the loop at the text level, in which case the problems of recognition, coding, and synthesis coalesce and are simultaneously solved—the result producing as one, a voice typewriter, the ultimate low bit-rate coder, and high-quality text synthesis. Present realities are removed from this, but good success is being achieved on connected input speech at the level of articulatory parameter adjustment.

Lest enthusiasm run too high, it should be quickly mentioned that the required computation is enormous—about 1000 times real time on a parallel computer. Or, for real-time operation, about 100 billion floating-point operations are required per second (100 GFlops). This amount of computation is not as intimidating or deterring as it once was. Through highly parallel architectures, one can now foresee teraflop capability (though it is presently less clear how to organize algorithms and software to utilize this power).

**Robust** Techniques for Speech Analysis. Most algorithms for coding and recognition can be made to perform well with "clean" input; that is, with high-quality signal having negligible interference or distortion. Performance diminishes significantly with degraded input. And machine performance diminishes more precipitously than human performance. For example, given a specific level of recognition accuracy, the human listener can typically achieve this level with input signal-to-noise ratios that are 10 to 15 dB lower than that required by typical automatic systems.

A part of this problem appears to be the linear analysis used for most processing. Linear predictive coding, to estimate short-time spectra, is representative. Sizeable durations of the signal contribute to computation of covariance values, so that extensive amounts of noise-contaminated samples are averaged into the analysis. One alternate procedure of interest at present is to eliminate the worst noise-contaminated samples and reconstitute the discarded samples by a nonlinear interpolation algorithm. Another is the use of auditory models of basilar membrane filtering and neural transduction for characterizing signal features.

**Three Dimensional Sound Capture and Projection.** High-quality, low-cost electret microphones and economical digital signal processors permit the use of large microphone arrays for hands-free sound capture in hostile acoustic environments. Moreover, three-dimensional arrays with beam steering to the sound source and multiple significant images in a reverberant enclosure provide significant improvements in pickup quality. Spatial selectivity in three dimensions is a by-product (Fig. 10). Computer simulations are providing designs that are being digitally implemented and tested in real environments.

Design of receiving arrays is similar to that for transmitting (or projecting) sound—though the costs of transducers for receiving and transmitting differ greatly. Increased spatial realism in sound projection will result from this new understanding.

**Integration of Sensory Modalities for Sight, Sound, and Touch.** The human's ability to assimilate information, perceive it, and react is typically more limited in rate than the transmission capacities that convey information to the user terminal. The evolution of global end-to-end digital transport will heighten this disparity and will emphasize the need to seek optimal ways to match information displays to human processing capacity.

Simultaneous displays for multiple sensory modalities provide benefits if they can be appropriately orchestrated. The

![Fig. 10](image-url)  
*Fig. 10.* Signal-to-noise ratios measured on two octaves of speech for a cubical microphone array composed of $7 \times 7 \times 7$ sensors and positioned as a "chandelier" at the ceiling center in a computer-simulated hard-walled room of dimensions $7 \times 5 \times 3$ meters. Source images through third order are computed, and multiple beams are steered to the source and its images. The off-focus response reflects spatial selectivity.*
sensory modalities of immediate interest are sight, sound, and touch. Our understanding of the first two is more advanced than for the latter, but new methods for force feedback transducers on data gloves and “smart skin” implementations aspire to advance tactile technology (9) (Figs. 11 and 12).

Ease of use is directly correlated with successful integration of multiple sensory channels. On the speech technology side, this means integration into the information system of the piece parts for speech recognition, synthesis, verification, low bit-rate coding, and hands-free sound pickup. Initial efforts in this direction are designed for conferencing over digital telephone channels (10). The speech features allow call setup, information retrieval, speaker verification, and conferencing—all totally under voice control. Additionally, low bit-rate coding of color images enables high-quality video transmission over modest capacity (Fig. 13).

SPEECH TECHNOLOGY PROJECTIONS—2000

How good are we at forecasting technology? In my experience, not so good. But not so bad either. I recently got out a set of vugraphs on coding, synthesis, recognition, and audio conferencing that I prepared in 1980. These were made for 5-year and 10-year forecasts as part of a planning exercise. To my surprise about half of the projections were accurate.

Notable were subband coding for initial voice-mail products (called AUDIX) and 32-kbps ADPCM for transmission economies on private line. But there were some stellar oversights. My 1980 vugraphs of course did not predict CELP, though I was in intimate contact with the fundamental work that led to it.

Despite the intense hazard in anticipating events, several advances seem likely by the year 2000:

- **Speech signal representation of good perceptual quality at <0.5 bit/sample.** This will depend on continued advances in microelectronics and, especially, the incorporation of psychoacoustic factors into coding algorithms.

- **Multilingual text-to-speech synthesis with generic voice qualities.** Multilingual systems are emerging now. The outlook for duplication of individual voice characteristics by rule is not yet supported by fundamental understanding. But generic qualities, such as voice characteristics for man, woman, and child, will be possible.

- **Large-vocabulary (100K-word) conversational interaction with machines, with task-specific models of language.** Recognition of unrestricted vocabulary, by any talker on any subject, will still be on the far horizon. But task-specific systems will function reliably and be deployed broadly. A strong emphasis will continue on computational models that approximate natural language.

- **Expanded task-specific language translation.** Systems that go substantially beyond the “phrase-book” category are possible, but still with the task-specific limitation and generic qualities of voice synthesis.

- **Automated signal enhancement, approaching perceptual acuity.** This is among the more problematic estimates, but improved models of hearing and nonlinear signal processing for automatic recognition will narrow the gap between human and machine performance on noisy signals. Comparable recognition performance by human and machine seems achievable for limited vocabularies and noisy inputs. Interference-susceptible communications, such as air-to-ground and personal cellular radio, will benefit.

- **Three-dimensional sound capture and projection.** Inexpensive high-quality electret transducers, along with economical single-chip processors, open possibilities for combating multipath distortion (room reverberation) to obtain high-quality sound capture from designated spatial volumes. Spatial realism in projection and natural hands-free communication are added benefits. Current research suggests that these advances are supportable.

- **Synergistic integration of image, voice, and tactile modalities.** Although the constituent technologies for sight, sound, and touch will have imperfect aspects for the foreseeable time, proper
design of application scenarios will enable productive use of these modalities in interactive workstations. Human factors engineering is central to success. Expanded utility of tactile displays depends on new transducer developments—for example, the design of transducer arrays capable of representing texture in its many subtleties.

* Requisite economical computing. Indications are that microelectronic advances will continue. Presently deployed on a wide basis is technology that provides computations on the order of 50 Mflops on a single chip and costs less than a dollar per Mflop. Before the year 2000, the expectation is for wide deployment of 0.35 μm (and smaller) technology, with commensurate gate densities. Computation on the order of 1 Gflop will be available on a single chip. This availability of computing will continually challenge speech researchers to devise algorithms of enormous sophistication. If the challenge is in fact met, the year 2001 may actually see a HAL-like conversational machine.

In addition to current university research, this paper draws liberally from material familiar to me over a number of years while at AT&T Bell Laboratories, for whom I continue as consultant. I am indebted to Bell Laboratories for use of the material and for kind assistance in preparing this paper. I am further indebted to the Eighteenth Marconi International Fellowship for generous support of this and related technical writings. Components of research reported here were supported in the CAIP Center by the National Science Foundation and the Advanced Research Projects Agency under contracts MIP-9314625, IRI-933214946, DABT 63-93-C-0037, and DAST 63-93-C-0064.