THE FUTURE OF SPEECH PROCESSING TECHNOLOGY

by

Alan L. Bridges
VeXP Research/Systems, Ltd.
2754 Pine Hill Dr. N.W.
Kennesaw, Georgia 30144

ABSTRACT

Many "New Technologies" will ease the "Man-Machine Interface" through a variety of Input/Output (I/O) techniques. These include: sonic digitizers (bit pads), the optical "mouse", object recognition (machine vision), hand print character recognition, touch technology for video display terminals (VDTs), interactive computer graphics, Artificial Intelligence (Natural Language), and speech processing technology. Here, speech processing technology is used as somewhat of a "catchall" term, referring both to human speech as input to computers and to computer speech as output to humans. It includes speaker and speech recognition, speaker dependent/independent systems, isolated word recognition, connected speech recognition/understanding, speech synthesis and voice response systems. Advances in computer technology, mainly in algorithms and VLSI technology, have fostered new abilities to provide speech input and output. The major applications areas for speech processing technology include consumer, commercial, government/military, industrial and rehablitative systems. Speech recognition and synthesis devices are available for microcomputers, as in Texas Instrument's Voice Management System for their Professional Computer, as aids to the physically handicapped, for "hands-busy, eyes-busy" voice data entry applications, for limited vocabulary voice activated word processing, as voice response systems, and for utilization within Voice Mail or Store-and-Forward Messaging Systems. Although the possibilities for speech processing applications are numerous, the real-life applications for end-users are still somewhat restricted in number. However, during the 1990's advances in technology will bring about new developments, such as the as the Voice Activated Typewriter (connected speech recognition) and the Intelligent Interface Machine (IIM) from the Japanese. The Intelligent Interface Machine (IIM) or Fifth Generation Computer (as proposed by the Japanese) will offer: natural language; the ability to learn, associate and infer; the ability to understand the contents of its database, as well store, retrieve and pass along information; applied speech understanding; and applied picture and image understanding. Thus, future directions in computer technology include the integration of many compatible technologies into one system, i.e., Artificial Intelligence, Speech Processing Technology, Parallel Architecture, and Interactive Graphics.

KEYWORDS: Man-Machine Interface, Artificial Intelligence, Speech Processing Technology, Speaker and Speech Recognition, Speaker Dependent/Independent Recognition, Isolated Word Recognition, Connected Speech Recognition/Understanding, Speech Synthesis, Voice Response, Voice Store-and-Forward Messaging, Voice Management System, Voice Activated Typewriter (VAT), Intelligent Interface Machine (IIM), and VLSI Technology.

I. INTRODUCTION

Humanizing the man-machine interface has been the subject of many papers and new computer designs over the past couple of years or so. Techniques being utilized include sonic digitizers, the optical "mouse", object recognition (machine vision), hand print character recognition, touch technology for video display terminals, interactive computer graphics, Artificial Intelligence (Natural Language), and speech processing technology. Future developments, such as the Intelligent Interface Machine (IIM) or Fifth Generation Computer (as proposed by the Japanese), will integrate many of these technologies into one system. Speech processing technology, that is, automatic speech generation (response and synthesis) and automatic speech recognition (speaker and speech recognition/understanding) will be a must for man-machine communications.

II. SPEECH PROCESSING TECHNOLOGY FOR MAN-MACHINE COMMUNICATIONS

Speech is clearly the easiest and most natural means of comnmunication among people. Extending speech to machine interactions has or is being brought about by a variety of technological advancements, including new input methods and interactive software (natural languages). At the present time, there are many new products on the market and others on the way, designed around user needs as opposed to attributes of the machine. Word recognition units are being used at many sites within the United States and Japan where people give simple spoken commands to a computer-controlled device or enter spoken data into a computer. Although still in its infancy, the currently available technology can give a machine the ability to "talk" or generate "humanlike" speech and/or recognize or understand speech.

A. SPEECH GENERATION

Speech generation is far more common at the present time. Speech response technology as such has been around for over twenty years, but recent advancements in speech synthesis have led to substantial improvements. Toys, games, appliances, automobiles, clocks and microcomputers have the ability to "talk."

There are two basic types of automatic speech generation systems. The traditional audio or voice response unit, utilizes pre-recorded syllables, words or phrases and then a series of pre-recorded words that are sequenced yo respond to number codes. The basic elements of a voice response system include: provisions for storage of vocabulary, rules for forming messages from elements of the vocabulary, and a program for composing voice response messages.

Speech synthesis systems are based on digital storage (magnetic RAM or ROM) of phonetic structures based on phonemes or elements of speech sounds. The digital approach to speech synthesis permits the creation of more than 300,000 words of working vocabulary, which is in excess of the 50,000-word working vocabulary of everday English speech.

There are at least two digital means of synthesizing speech: voice encoding and waveform encoding (modelling of the vocal tract). Almost all of the technical problems of speech communication, including speech generation and recognition, can trace their fundamental solutions to speech analysis. For the most part, speech analysis techniques assume that the parameters of the underlying speech model change slowly with time. This leads to a number of "short-time" analysis techniques, such as Fourier analysis and analysis-by-synthesis.

Direct speech synthesis forms speech signals or responses directly from phonemes, the elements of speech. The basic physical structures that this type of synthesizer must accurately reflect include: an electrical analog of the human vocal tract, a program to specify the desired sound of the vocal tract parameters, and the control interface of the vocal tract. Thus, the major problem associated with synthesis-by-rule or constructive synthesis is one of proper encoding of phonemes. With the possible combination of phoneme and emphasis commands, almost any phrase can be formed by carefully assembling these elements into a speech program. The set of rules must consider factors like garbling effects of wrod boundaries, stress variations and pitch- and timing-contour problems. This method does enjoy the advantage of an unlimited vocabulary and low memory requirements. It does not require speech input, but rather utilizes direct text-to-phonemic conversion or direct phoneme encoding (as in Votrax Type-'n'-Talk Speech Synthesizer). The major disadvantage of stringing phonemes together is that the sound quality is poor and somewhat mechanical sounding. Allophones, diphones, demisyllables and morphs can be used to approach higher quality speech, but at a considerable increase in memory size.

Speech analysis-synthesis techniques include a variety of methods: Linear Predictive Coding (LPC), Partial Autocorrelation (PARCOR), Pulse Code Modulation (PCM), and Linear Delta Modulation. Data rates of 600 to 800 bits/second characterize formant synthesis, which follows peaks in the speech spectrum. Formant synthesis can compress speech data dramatically, but implementing them requires extensive processing power with algorithms that are far from perfected.

LPC-type synthesizers are characterized by data rates of 1,200 to 10,000 bits/second. About half of the current speech synthesis chip manufacturers use

the relatively simple LPC technique for estimating the parameters of the speech signal. LPC is used to estimate the basic parameters of speech, i.e., pitch, formants, etc., and to represent speech with low bit rate transmission and storage. The basis for LPC is that a speech sample can be approximated as a linear combination of past speech samples. Thus, very accurate estimates of speech parameters can be provided with relatively fast speed of computation. With LPC fewer bits are required to produce each word. Consequently, a compression of data occurs (100-to-1). The importance of linear prediction lies in the accuracy with which the basic model applies to speech. The parameters of this model are the voiced/unvoiced classification, the pitch period for voiced speech, the gain parameter, and the coefficients of the digital filter (all vary slowly with time). Unlike other parametric encoding methods LPC permits the implementation of an effective speech synthesizer with available LSI on a single chip. Rather than creating speech from synthetic phonemes, LPC-based speech systems are based upon the conversion to digital format of speech from an actual human voice. Hence, a practical support structure is necessary and includes evaluation boards (National Semiconductor and TI), development systems (such as TI's MULTIAMPL Development System), speech data development services (Hitachi) and training workshops.

The current market is dominated by speech synthesizers, vocabulary read-only memories and encoding/modulation chips. Voice response systems now being sold utilize several different techniques, but the general trend is toward all digital techniques (such as LPC) to process speech. Voice response systems are beginning to be interactive--with the computer not only asking questions, but recognizing or understanding limited vocabularies as well. The market for speech synthesizing devices is expected to grow rapidly through mid 1980's and growing at a slower rate thereafter. By 1990, annual shipments should top $900 million.

B. SPEECH RECOGNITION

On the other hand, speech recognition is a much more difficult technological problem to solve and requires significant capital investment. It will, however, become more pervasive within the next decade as technological breakthroughs are made at the semiconductor level. Improvements in software models or algorithms and placement of these algorithms in VLSI (Very Large Scale Integration) technology will improve the present, somewhat limited, capabilities of speech recognition systems, which represent a large, rapidly emerging market which is, for the most part, untapped.

For speaker/speech recognition systems the primary task is to either verify a speaker's identity (either a yes or no answer as to whether the speaker is who he says he is) or to identify the speaker from some known group of speakers (speaker verification and identification). Accordingly, applications include controlled access to information, data or secured areas and automatic credit transaction systems.

Speech recognition systems, which include Automatic Speaker/Speech Recognition (ASR) and Voice Data Entry (VDE) systems, convert the acoustic information (waveform) into a written equivalent of the information content in the spoken message. The very nature of speech recognition is dependent upon the constraints placed on the speaker, speaking situation and message content. Speech input to a computer system appears to be most practical when:

* a specific vocabulary is used
* an operator must handle equipment or documents with both hands or eyes busy
* wearing a headset microphone does not interfere with work tasks
* there is a cost benefit if the informationoriginatesby voice data entryrather than other means.

1. SPEAKER DEPENDENT AND SPEAKER INDEPENDENT RECOGNITION

Speaker/speech recognition devices can be either speaker-dependent or speaker-independent. Speaker-dependent devices are designed to understand a particular individual's speech patterns and are used widely, although they have limitations compared to speaker-independent systems that recognize voices of many operators. The basic concept of

speech recognition is rather simple--it is essentially a pattern recognition problem. The basic speech recognition system must digitize an analog signal or voice waveform and compare it to a stored reference pattern or vocabulary. Word templates can be formed by a variety of techniques, including bandpass filtering, analog-to-digital conversion, zero-crossing detection and Fast Fourier Transform (FFT) analysis. The input word template is processed and then compared to a series of templates or patterns stored in memory. Comparison algorithms are necessary and some sort of decision logic must be used to make a choice between available stored templates.

Speaker independent systems do not store individual voice patterns and thus do not have to be trained to the individual operator; rather, a speaker's voice inputs are compared with averaged voice print patterns. The difficulties of adapting speaker independent are apparent: wide variations in human speech--in terms of pitch, intensity and duration--permits only isolated word recognition or limited vocabulary connected speech. Thus, speaker independent systems are not as accurate or as flexible as speaker dependent systems. At the same time, speaker independent systems offer a larger future market, i.e., telephone directory recognition of names and automated information systems.

2. ISOLATED WORD RECOGNITION

Isolated word recognition systems do not have the problems associated with continuous word recognition systems since all words are separated by pauses. Typical vocabularies vary from 50 to 300 words. These systems can recognize a word spoken in isolation with an accuracy approaching 99 percent. The vocabulary and/or speaker can be changed, but this usually requires a "re-training" session. The general paradigm is one of comparing the parameter or feature representation of the incoming utterance with the prototype reference patterns of each word in the established vocabulary. Decision that must be made in this process include how to normalize for variations in speech, what parametric representations to use, how to adapt new speakers or vocabularies, how to measure or distinguish between two similar utterances, and how to speed up the matching process.

A typical word recognition system divides the acoustic signal into separate phonetic and spectral signals, where each feature is detected and converted separately into a bit pattern. Since some bit patterns can be longer, a time normalizer is used. As a consequence, the combination of phonetic and spectral features versus a number of time units forms a matrix representing the word, which is then compared with matrices for vocabulary words stored in memory. A match is made under a correlation processor. As long as the cost/performance requirements for isolated word recognition systems do not demand an order of magnitude of improvement, the present systems offerings will continue to be practical and cost effective. The principle avenues of improvement will be in basic algorithms, i.e., reference pattern representation and search strategies.

Speech recognition systems have been described as "user friendly," making speech recognition a strong contender for executive workstation control and operation, as in TI's new Voice Management System for The Professional Computer which can recognize up to 50 user-trained utterances and Votan's V5000 with a maximum logical vocabulary of 256 words (both are speaker dependent word recognition). Speech input can also simplify data base retrieval/inquiry systems. Order entry simply becomes a matter of speaking into a microphone or into a telephone to a remote voice entry system (such as the VOTAN VX Series). Low end systems are most noted for isolated word recognition capability, while high end systems have the ability to recognize connected words and hence larger vocabularies.

3. CONNECTED SPEECH RECOGNITION AND UNDERSTANDING

Continuous speech recognition (1,000-word vocabularies or larger) is still in the laboratories. Most continuous recognition systems have constraints that relate to the normalcy, age, sex and/or identity of the talker; the environmental conditions in which the user is speaking; the syntax and lexicon (vocabulary or systematic arrangement of words) of admissable phrases; and/or the number of different lexical entities. In connected or continuous speech recognition it is difficult to tell where one word ends and the other begins. This is basically due to the fact that the characteristic acoustic patterns of words exhibit a great variability

depending on the context. As the vocabulary size increases, the "confusability" increases because the simple emthods of representation and matching become too expensive and unacceptable. Continuous Speech Recognition (CSR) is more difficult than continuous speech understanding as error tolerance must be minimal, whereas for a correct interpretation the system would only have to understand about half of the words. Thus, techniques for compacting the representations and for reducing search efforts by constraining the number of possible words that can occur at a given point are of interest.

CSR systems must keep their program structure simple by using only some task-specific knowledge and by requiring that the speaker speak clearly and use a quiet room. Connected speech recognition uses a technique of analysis and description rather than classification, that is, moving away from pattern recognition towards hierarchical systems where subparts are recognized and grouped (concatenated) together to form larger and larger units. In terms of mathematics, we have moved from a signal space to a symbol space representation. Since CSR systems do not have the advantage of isolated word recognition systems, of knowing the beginning and ending of the words, the beginning must be specified prior to the match. Necessary evils are thus the rror and uncertainty in segmentation, labeling and matching--this means that algorithms, such as tree searching, must be used to select an optimum match of words. An exact match cannot be found until the next word in the sequence or the ending context is found. Most all CSR systems contain a set of alternative word sequences arranged in descending order of their likelihoods to represent the partial sentence completed so far. Given the word sequence with the highest likelihood, the task-specific knowledge (phonological rules, lexicon and syntax) generates all the words that can follow the sequence. Each of these words is matched against the unmatched symbol string to estimate the conditional likelihoods of occurence. this process is repeated until the whole utterance is analyzed and an acceptable word sequence is determined. If CSR systems are to achieve the higher accuracies necessary for their success (99%), than better search, matching, and segmentation/labeling techniques are essential. The most crucial aspect of CSR involves the environmental conditions under which the system must operate, i.e., quiet rooms versus real-life situations.

When we change from connected speech recognition to connected speech understanding (CSU), there is a change in perspective from that of matching acoustic patterns to one of interpretation of acoustic signals in light of knowledge about syllables, words, and sentences; about the rules of conversation; and about the subject under discussion. Reducing the problems of error and ambiguity that result from large vocabularies and connected speech is one objective of restricted speech understanding research. CSU must recognize the utterance even when it is not quite grammatical or well formed and even in the presence of noise. The requirement is actually somewhat relaxed as it is the intent of the message that matters rather than every phoneme or word in the message.

C. VOICE STORE-AND-FORWARD

Voice store-and-forward involves the real time encoding, compression and storage of a speech message for later retrieval. While voice response primarily involves static messages, voice store-and-forward involves continuously changing messages. Spoken words are digitized, compressed and stored, then are immediately available for playback or can be transmitteed to a host processor for later retrieval. No advance programming is necessary as most systems are user programmable, allowing the user to record or update his chosen messages or phrases on-line, without delay. All digital voice mail systems have the capability to record and play back the sounds of any language, since they are not based on a particular linguistic model of speech, as are phoneme-based systems. Software allows the selection of telephone, microphone, headset, or speaker voice input/output. When the digitized voice is played back, the speaker's identity and intonations are clearly identifiable.

Users of voice mail system have "Mail Boxes," which store voice messages from other users. The user can retrieve a message and then dictate his answer. The Voice Mail system can automatically deliver the message. The same message can be sent to hundreds of people as simple as sending it to one. Furthermore, messages can be created during the daytime and queued up for subsequent transmission at

night. The system can be unattended when the transfer takes place. These voice mail systems have been promoted as white-collar productivity tools that attack real-time problems of business communications, or what is called "telephone tag." Many companies have entered this market--these include IBM, AT&T, Wang, ADP, Northern Telecom, and ECS Telecommunications. Two new product entries of 1982-83 include Votan's VX Series and Texas Instrument's Voice Management System (VMS) for the TI Professional Computer.

D. SPEECH CHIP TECHNOLOGY

The largest market for speech processing technology is now represented by the speech chip market, which isa currently dominated by speech synthesis chips, vocabulary Read Only memories and encoding/modulation chips. Speech recognition chips are also becoming an important aspect of this market segment as more advanced algorithms and increased processing power are becoming available and are being placed into VLSI-based systems. For example, the 32-bit TMS-320 signal processing microcomputer chip from Texas Instruments has an operating speed of five million instructions per second--faster than many mainframe computers of the 1970's era. Key technological contributors include the development of advanced algorithms for Linear Predictive Coding and the placement of those algorithms into a small, integrated circuit. Bell Labs has made progress towards a Digital Speech Processor (DSP), which will surely be placed in silicon and possibly be used in one form or another with their 32-bit microprocessor, the Bellmac 32A. Other speech chip manufacturers include National Semiconductor, American Microsystems, Inc. (AMI), General Instruments, Hitachi, Intel, Matsushita, Motorola, Interstate Electronics, Nippon Electric, and Votrax.

Semiconductor technology has become the industry of the future now--all future products will be designed around the latest VLSI (or VHSIC, Very High Speed Integrated Circuits, for military applications) techniques. Complex systems will be placed on semiconductor chips. Microcomputers will evolve into micro-mainframes. What once occupied a large amount of space on a circuit board can now be placed into one or more semiconductor chips. The key to industry advances that will make speech processing applications practical and cost effective will be the ability of "silicon designers" to effectively integrate complex systems on individual or multiple silicon chips (alternate technologies notwithstanding). Chip-set equivalents of the current board-level products are beginning to appear and will substantially reduce the OEM costs (from about $20 to $1 per word for voice recognition).

There are, however, some problems that must be overcome by VLSI designers. Designing VLSI is a complex, multilayered process. Apart from the difficulty of showing that the chips do what they are meant to do, the question arises whether these ever more complex designs can be implemented, or understood when implemented. Functional solution and technical implementation are intertwined, and consequently new design techniques must be used to overcome the "complexity barrier." Another consideration is that not all algorithms may be appropriate for implementation in silicon. Top-down, hierarchical designs are being used, along with many Computer Aided Design (CAD) software tools.

Japanese scientists at Toshiba Corp., for example, have developed a machine that uses a highly focused beam of ions to trace circuits on silicon chips. The ion-beam machine can draw lines so fine that Toshiba expects to be able to increase the number of electronic devices on a single 0.25 inch-square chip from the current 64,000 to about 4 million. Dense chips like these are needed in order to increase the vocabulary storage capacity of speech processing systems. The lead of Japanese (NEC and Matsushita), both in terms of technological capabilities and overall acceptance and total per capita sales of speech recognition products is seen as a driving force for technological innovations in speech recognition chips.

The current market for voice recognition chips is small, but in the next five years or so the market for speech recognition chips will grow--possibly equaling the market for speech synthesis chips by the early 1990's. But in order to be accepted, connected speech recognition chips will need a vocabulary of at least 2,500 to 5,000 words.

III. APPLICATIONS

There are five major applications areas for speech processing technology:

* Consumer Systems
* Commercial Systems
* Government/Military Systems
* Industrial Systems
* Rehabilitative Systems

At the present time, consumer systems have been the major speech synthesis market; it is also the most segmented with speech synthesizers used in learning aids, clocks, appliances, toys, personal computers, vending machines, language translators and automotive systems. There are some areas of the consumer market that are becoming applications are as for speech recognition systems. These include door locks and security systems, as well as personal computers. The recently introduced Voice Management System for the TI Professional Computer (which could also be placed under the commercial systems category) is one example of speech (speaker dependent) recognition for the personal computer.

Commercial systems include voice entry and recognition systems. Applications areas include data base access, catalog ordering, telephone network access and consumer bill paying (Electronic Funds Transfer Systems). The major portion of this market involves speech generation and Touch-Tone(R) access. Speech recognition, on the otherhand will be needed and used in many instance where the Touch-Tone(R) telephone dial is not satisfactory for inputting information. Voice is deemed more practical if such speech can be understood and the vocabulary that is used is fairly extensive--100 words or more. One of the reasons for the practicality of speech input is the complexity of the numerical combinations of dial tones. Other commercial applications for speech recognition include automatic telephone transaction systems (banking and credit authorization) and data entry for word processing.

In the future, voice recognition is expected to allow highly flexible direct voice interactions with processors, making it possible to convert spoken word directly into a written page. This technology can do more than reduce the ened for typing copy into a keyboard--it will potentially allow a system to carry out the commands spoken to it. Although the great dream of automatic dictation will probably not come until until the early 1990's, restricted capabilities in speech recognition systems may find application in the office in the very near future (limited voice recognition word processing is available from Interstate Electronics Corporation). The most successful systems development in thearea of automatic dictation (connected speech recognition) or the Voice Activated Typewriter (VAT) is expected to come from the Japanese, since their language is better suited to machine recognition than English. IBM, however, has been conducting research into the area of connected speech recognition for almost 20 year and is expected to introduce the first commercial VAT in the United States.

The commercial market will eventually be dominated by the Voice Activated Typewriter (VAT). This technology is being developed by IBM and Matsushita and will possibly be introduced between 1985 and 1987. The VAT will include a keyboard for editing and correcting inaccurately recognized words; a controller accepting voice input or input from keyboards or other peripherals (such as digitizers); memory; and an editing video display terminal. The VAT will allow input of continous spoken speech as opposed to isolated words that are now possible or the limited vocabularies (50 to 150 words) of recently introduced speech recognition/voice response systems.

IBM's current research is carried out relative to task domains which greatly restrict the sentences that can be uttered. Task domains are of two kinds: those where the allowed sentences are prescribed a priori by a grammar and those related to natural tasks (such as text of business letters and patent applications). The experimental environment has also been restricted to a very quiet room. The system is tuned to a specific talker, a script is read, and false starts are eliminated. The basic CSR system used at IBM's T.J. Watson Research Center consists of an acoustic processor, which transcribes speech into a string of phonetic symbols, followed by a linguistic decoder that translates the potentially garbled phonetic string into a string of words. IBM has demonstrated 93 percent recognition accuracy with a 1,000-word vocabulary.

A commercially feasible system, however, would require about 5,000 words, although a core language of 2,500 words would be useful in the "executive suite." Japanese and German languages will probably be the first and easiest to develop because of the language structure. The first VAT's should appear between 1985 and 1990.

The government and military markets are the largest areas for speech recognition equipment today and are expected to grow rapidly throughout the 1980's as the demand to unburden pilots, especially in military aircraft, and to produce training systems that do not require the constant attention and response of personnel other then the trainee. A large number of potential systems for military applications are under study or development at the present time. These include: Automatic Speaker Verification, secure access control applications, word recognition for militarized tactical data systems and on-line cartographics systems, and speech recognition/voice response systems for the "cockpit of the future."

The industrial market, consisting of factory floor data entry, package sorting, machine-tool control and programming, inspection, quality control and serial marking systems, is, and will remain, one of the most viable areas for spech recognition rather than generation. This is because of the pervasive need for hands-free inputting of data and increased productivity requirements that have brought abought the concept of factory automation. As systems become more sophisticated, speech synthesis equipment will be installed to warn of production difficulties, as well as verification of input data. Japanese manufacturers may not be able to claim dominance int he field of speech recognition, but they are the most aggressive in putting it to practical use.

The embryonic market of speech recognition products has been generated by the need of the handicapped--utilization of voice entry for raising and lowering beds, dialing telephone, turning on lights or controlling wheel chairs are potential applications of speech technology for the rehabilitative market. Products include IBM's audio typewriter, the Kurzweil Reading Machine, Maryland Computer Services' Total Talk Terminal and Votrax Speech Synthesizers.

IV. INTEGRATION--THE KEY TO THE FUTURE OF SPEECH PROCESSING TECHNOLOGY

Speech communication has become a focus for the convergence of many diverse scientific inquiries. Emergence of "natural languages," speech recognition/synthesis subsystems, and Logically Integrated Software Architecture (LISA) for microcomputers is indicative of efforts to humanize the man-machine interface. Artificial Intelligence (AI) will be the foundation for bridging the man-machine communications barrier. Purely syntactic approaches have been abandoned in favor of semantics. Language understanding, visual-pattern recognition, and chess playing appear to tractable problems. Structuring knowledge, learning, and problem solving are tasks that must be solved for an adaptive system to access and manipulate information.

Future developments will see speech processing in every part of the home and business: a spoken word will release a door lock, turn on the hall light, the radio, the microwave oven, the stereo. Voice input will be commonplace in the automobile of the future. Voice maps will tell us what landmarks to expect and where to turn. Voice input will replace familiar driving controls. The house will given a voice to scare intruders--keys will be obsolete as the "house" will be able to recognize the owner's voice. Offices of the future will use voice activated word processors, copiers and computer terminals. Voice mail will be commonplace. For example, reports will be routinely accompanied with voice annotations.

And, as the computer speaks and understands speech, confusion will arise about just how intelligent the computer is. From a purely engineering point of view, the design of the human brain represents a much more complex engineering problem than computer design. We certainly will see systems that will be able to pass for human in brief telephone conversations. Although, no current computers have integrated pointing and speech with typed input, important improvements in user interfaces could stem from combinations of input modalities (such as voice input/output, multiple windows, interactive graphics, natural language, inference, image recognition, etc.). Future interfaces must take into account the user's current work practices and expertise, as opposed to obtaining maximum power from scarce and expensive hardware and adapting the user to machine

oriented-software. Software is "first, last and forever." Psychological studies must simulate man-computer dialogues and adaptive input devices, such as speech recognition, will be utilized to meet those communication needs.

The Intelligent Interface Machine (IIM), which has been proposed by the Japanese as the Fifth Generation Computer, will be based on the next generation VLSI, that is, ULSI (Ultra Large Scale Integration). New parallel architectures, such as data-flow machines, and languages, like LISP and PROLOG, will be used in forms that vary from personal computers to supercomputers. The IIM will offer natural language access, the ability to learn and infer, and an understanding of the data it stores. The IIM should be able to program itself, listen to and obey spoken commands, and treat images and graphics the same as words, that is, "see" and "understand" an image. The computer would be able to carry on an intelligent conversation with a person (question-and-answer session) and would make judgements that will enhance the thinking capacity of its human masters. Another goal of the IIM is to translate foreign languages.

The IIM-type machine or ultimate man-machine interface (UMMI) must be able to correct linguistic errors; recognize and adjust to the particular user; correct spelling or syntactic error; let the user know when it does or does not understand; ask the user to select between interpretations of what was said; follow the focus of the user as it changes during a dialogue; identify objects from its database from user descriptions; offer answers to user questions; and be able to produce appropriate output, such as text, graphics or voice.