



STATE-OF-THE-ART SPEECH RECOGNITION

U.S. RESEARCH AND BUSINESS UPDATE

Janet M. Baker

INTRODUCTION

We are standing today on the edge of a new era. Not an era which opens as a stage curtain with orchestral fanfare on opening night, but more like a majestic mountain landscape emerging from the early morning mist. Practical speech recognition, available from multiple vendors in multiple forms, has been successfully pioneered in numerous applications. The scale and scope of these applications has recently begun to grow appreciatively, in large part, due to the confluence of several key factors. The technology itself has improved sufficiently to provide for satisfactory operational performance, both in terms of accuracy and clear user benefits. Significant capabilities can now be more easily integrated into valuable widely used applications. With the recent availability of powerful inexpensive memory and processors (often packaged in personal computers), high performance recognition has become more affordable.

The research tasks, while ever formidable and challenging, are visibly chiseling away old problems (though always revealing new ones!). Through a succession of contributions, by researchers worldwide, the technology base has become progressively stronger. Despite significant unknowns, the prospects of success are now undeniable. This presentation proposes to illustrate, by example, a sample of notable applications and research programs through which we catch glimpses of the enticing vistas ahead.

APPLICATION SAMPLER

The adage that the "proof of the pudding is in the eating" applies as well to speech recognition capabilities as to culinary skills. Over the past decade, productive speech recognition applications have been emerging to show slow but steady growth. Most of these have centered around "hands busy, eyes busy" industrial workers engaged in inspection, sortation, and simple data entry. Within this past year or two, new applications have started burgeoning into a much broader array of areas serving the professional and business sectors, as well as casual consumers. Although not yet reflected in sizable industry revenues, this growth is laying the necessary foundation for the future. The firmness of this foundation will determine the size and time-scale for significant market development.

Described here is a representative sampling of successful applications, utilizing a variety of leading vendor products. Spanning the past several years, these applications are notable because recognition provides high value under operational conditions, where the alternatives are less effective, more costly, or more time-consuming. A rapid proliferation of applications is becoming progressively evident, a tribute to better products at lower prices, easier integration into existing systems, and improved customer awareness and understanding of the technology.

Factory applications for speech recognition are responsible for having furnished the first exemplars of technology success. Initial systems a decade ago, many times more costly

Dragon Systems, Inc., Chapel Bridge Park, 90 Bridge Street, Newton, MA 02158, U.S.A.;
Tel.(617) 965-5200, Fax. (617) 527-0372 Copyright © Dragon Systems, Inc. 1987

and difficult than those today, had to be cost-justified primarily on the basis of significant labor-savings. An individual worker sorting packages moving by on conveyor belts can manipulate these objects while simultaneously entering destination or other data, by voice with a headset microphone. The same operations using manual data entry often take twice as long, or necessitate a two-person operation. In inspection operations, workers while engaged in physically handling objects, or examining them through a microscope for example, also need to record measurements, inspection verifications, and status information.

In November, 1983, General Electric Co. introduced speech recognition into its operations for inspecting printed circuit boards manufactured by its Aerospace Control Systems Department (ref 1). Operators conducting careful visual inspections of these assemblies, use speech recognition to report specific defects (e.g. "solder bridge") as well as to perform application commands (e.g. "print", "pass", "complete", etc.). Equipped with small vocabulary, discrete speaker-dependent speech recognition capabilities supplied by Votan, 14 workstations, serving 3 shifts of inspectors, were installed. Within a 6-week period of continual use, operators were averaging an improvement in throughput of 30% over previous manual methods. Most users transitioned easily to the new technology and quickly became, and continue to be, strong advocates.

At Burlington Industries, inspectors must visually examine yards of textiles steadily streaming by on frames slanted at 45 degrees in front of them (ref 2). Using gooseneck mounted microphones, inspectors can enter defect information without stopping operations to perform manual data entry. Using 2 voice-switchable subvocabularies (38 words and 24 words, respectively), operators can enter defects (e.g. "broken pick", "cut filling", etc.), as well as control the equipment itself ("reverse the frame", "speed frame up", etc.). In this installation with equipment supplied by Texas Instruments, small vocabulary, discrete speech recognition amply serves the time-critical requirements for reporting. Since 1984, about 35 units, divided between two factory installations, have been in production use daily by 3 shifts of operators. Throughput improvements with voice input are at least 25% over previous keyed input methodology. By actually naming the defects, rather than first converting to numeric codes, operators achieve greater accuracy and consistency in the fabric grading process itself.

Remote machine command/control as well as voice data entry/retrieval capabilities are frequently cited as major areas of opportunity for speech recognition. Furthermore, it has long been appreciated that aids to the handicapped have often served as precursors of goods and services broadly applicable to the general public. The following application is an interesting example of this kind of combination.

Competing in the 1986 Pacific Cup Race from San Francisco, CA to Hawaii was a sailboat named "The Awesome" with a crew of 8 people including a blind skipper. Using a system developed locally by Aerodnetics (ref 3), the skipper Lynn Olsen, communicated verbally with both sets of navigation instrumentation on board ship. The Long Range Navigation (LORAN) System and Brooks and Gatehouse Instrumentation System furnish position, compass, wind, speed, depth readings, as well as dead reckoning positional information. Wearing a wireless headset microphone, the skipper, from anywhere on board ship, could interrogate all the instrumentation functions by using speech recognition for input and text-to-speech for output. Both speech capabilities were furnished by an IBM multifunction personal computer (PC) product attached to a modified PC (e.g. no keyboard or display monitor) plugging into the navigation equipment. Reliable "standby" and "wake-up" command features allowed "hands-free" operation. The freedom of movement and self-sufficiency afforded by this system was quickly appreciated by the sighted crew members as well. Additional installations are presently underway for speech I/O hook-ups to other equipment, including satellite navigational systems.

In the Fall of 1986, Xerox Corp. completed a 100% inventory parts audit, using speech recognition to catalogue over 2.2 million parts throughout the U.S (ref 4). This entire process took about 2 months, significantly less time than required for previous sampled audits using manual keypunch methods. Using Dragon System's VoiceScribe (TM)-1000 speech recognition (up to 1000 word active vocabulary, discrete speaker-dependent),

installed in Xerox's PC, naive end-users were fully trained and productively auditing parts within 1\2 hr. Full "turn-key" application development and speech integration was accomplished in the month prior to field shipments to 105 district regions.

Inventory audits were conducted at over 15,000 sites, in diverse environments including offices, warehouses, even service representative vehicles parked at loading docks. Easily transported from one site to another, single units cost-effectively inventoried many sites within a region. Although systems were shipped with headset noise-cancelling microphones, some regions substituted lapel or other microphones to accommodate local user preferences. Given the choice of voice vs. keyboard data entry, users strongly preferred voice. Files generated on the PCs were regularly transferred to a centralized mainframe computer for consolidation of the national audit. Individual districts kept back-up copies of these files, including their automatically generated summary statistics, and found the immediate availability of this information quite useful in better managing their local resources.

In light of rapidly rising medical costs coupled with a chronic shortage of skilled transcriptionists, and the need for fast turn-around, medical facilities have been expressing strong interest in speech recognition to help address these problems. Although commercial speech recognition technology today does not yet satisfy the requirements of very large vocabulary, speaker-dependent, unconstrained free-text entry, it can provide significant benefit for more modest tasks. The practicality of dictating many types of routine medical reports is presently gaining acceptance at a growing number of medical facilities. Routine reports are produced for laboratory tests, X-rays, electrocardiograms and stress tests, patient status and progress records, as well as for surgical procedures.

An illustrative example of this type of application is a system presently being beta-tested at several Boston, MA hospitals (ref 5). Alan H. Robbins, M.D., a professor of radiology at Tufts University School of Medicine, has developed a radiology report generator allowing voice data entry using the Kurzweil Voice System (up to 1000 word active vocabulary, discrete speaker-dependent) attached through an RS232 port to a personal computer. Included are five predefined vocabulary sets suitable for reporting on general radiology, contrast studies, imaging, neuroradiology, and mammography. Since according to Robbins, up to 80% of such reports are routine in nature, a single word or phrase can be invoked to insert several associated words or even whole paragraphs. Although such macro command facilities are easily implemented through standard keyboard entries, voice data entry in many cases is far easier and more convenient for non-typists, especially those working in the dark! Robbins reports that routine reports generated by voice average 67 seconds vs. 56 seconds with standard dictation alone. Benefits of real-time recognition include immediate review capabilities and report availability.

Timely reports are essential for rapid communication between medical personnel specialized in different disciplines. Especially in time-critical situations, there is no time for standard transcription services. Furthermore, it is desirable that important information be computer-readable so that it can be rapidly disseminated to medical staff wherever they are located.

Speech recognition plays a central role in a recent medical application for multiple networked users in a hospital setting. The Arizona Heart Institute (ref 6,7), which specializes in open-heart surgery, has focussed on the development of a "paperless" intensive care unit (ICU). Up to 16 nurses can simultaneously monitor and report on patient conditions using speech recognition. Nurses wear wireless headset microphones which transmit to speech recognizers hosted on personal computers. These PCs in turn are attached to a local area network providing immediate communication and information access to staff throughout the hospital. Small vocabularies (up to 50 words simultaneously active) have been used with a number of different recognizers, including Votan, Texas Instruments, Kurzweil Applied Intelligence, and Dragon Systems, and larger vocabularies (up to 660 words) with Kurzweil and Dragon recognizers.

Audio prompts and feedback using PC-based text-to-speech, in addition to local screen displays, provide users on-line formatted reporting, recognition verification, and medical database information. The report generation software itself provides many significant functions by verifying that medications and the dosages administered match physician

orders, for example, as well as automating error-prone fluid input/output computations to maintain proper electrolyte balance, etc. The preliminary results, including more timely reporting, on-line information accessibility, and high user acceptance of speech I/O, appear quite promising for this integrated application.

PRICING

In any of the applications cited above, speech recognition is incorporated as a subsystem, often as a PC plug-in board. Generally speaking, prices today are somewhat lower than for comparable capabilities 2 years ago. In other cases, prices have remained relatively steady but the functionality of vendor products, at a given price, has increased substantially. A recent review article (ref 8) including vendor-supplied pricing, showed that for the recognition subsystems (speaker-dependent, isolated word, with 50 to 1000 word/phrase active vocabularies) used in the above applications, retail pricing now ranges between about \$1000. and \$1200. Exceptions include the Kurzweil KVS3000 (1000 isolated word active vocabulary) with base pricing of about \$6000 (\$12,000+ including radiology vocabulary software), and for high performance continuous speech recognition (up to 100 word active vocabulary), the Interstate Vocaline CSRB and Verbex 4000 for about \$4000.

CONSUMER PRODUCTS

The consumer area in recognition has recently started attracting significant attention as well, especially for cellular telephones and toys. Cellular telephones can pose significant hazards by distracting drivers. Hands-free dialing coupled with speaker-phone communication are clearly advantageous in this environment. With some products, a microphone is mounted on the sun-visor or steering column; in others the speaker uses one hand to hold a handset to talk while driving. In the latter case, recognition is easier, and performance improved, because the signal-to-noise ratio (SNR) is so much higher with a close-talking microphone. Vehicle noises, including engine sounds, the radio, background speech, traffic, etc., are typically highly variable, and can pose significant interference with speech utterances, especially for those detected 6+ inches from the talker. The first of these products have come to market, with a flood of additional products expected, riding 1) on the new popularity of cellular telephones, and 2) government safety regulations.

Very low recognition cost in a compact stand-alone form also imposes severe constraints on the speech processing capabilities in these units. Typical consumer prices for products containing recognition, range from about \$100+ for toys to \$200+ for voice dialer telephones up to \$400+ for cellular telephone recognizer add-ons. To achieve adequate performance, the recognition task is usually simplified to a very limited set of active commands. U.S. vendors supplying these capabilities today include AT&T, Audec Corp., Innovative Devices, Interstate Voice Products, and Voice Control Systems. A number of other vendors as well are developing these and other consumer product capabilities for products ranging from automotive comfort control (for windows, radio tuning, etc.) to remote appliance control.

Even tighter cost constraints, but much looser performance requirements, apply to the consumer toy market. In light of the \$25 million U.S. market for talking toys in 1986, speculation is running high on market prospects of up to \$100 million for the listening bears, dolls, robots, etc. soon to be promoted for the 1987 holiday season. Although the infamous volatility of this marketplace imposes a high variance on any projections, short or long-term, the acceptance for recognition capabilities, in general, may well be accelerated as a consequence of better consumer market conditioning.

VOICEWRITING AND OTHER RESEARCH PROGRAMS

A prime objective (sometimes referred to as the "holy grail"), for many recognition researchers and developers is to achieve genuine "voicewriting" capabilities for the automatic transcription of free-text spoken language. The initial large vocabulary speech recognition systems more than a decade ago, primarily government-funded (ref 9), were housed in roomfuls of mainframe computers, array processors, etc. Starting in the same time frame, and also using massive resources, IBM Research Division has successfully demonstrated a sequence of large-vocabulary free-text entry recognition systems [ref 10,11).

Despite the host of challenges remaining ahead, we have, finally, entered a new era which will revolutionize the way in which communications and business are conducted. Today the first large vocabulary free-text desktop workstations have been shown publicly, performing with good accuracy in near real-time.

In November, 1985, Dragon Systems publicly demonstrated its DragonWriter VoiceScribe (TM) large-vocabulary (2000 words, discrete, speaker-dependent) free-text entry recognition prototype system running in near real-time, on a standard 512K IBM PC/AT (6MHz clock) containing one audio input peripheral board (ref 12,13). Similar computationally efficient system configurations were announced for full 5000 to 20,000 word, discrete, speaker-dependent, free-text recognition products, now nearing completion.

Shortly thereafter in April, 1986, IBM publicly demonstrated its Tangora large vocabulary (5000 words, discrete, speaker-dependent) free-text entry recognition running on an IBM PC/AT containing multiple boards with proprietary processing hardware (ref 14). Subsequent research enhancements, including additional hardware housed in an expansion box, have boosted free-text recognition capabilities up to a full 20,000 words.

Both systems successfully demonstrated, for the first time, the technical feasibility of 1) integrating very large vocabulary recognition, running concurrently with 2) full natural language processing capabilities, 3) in a desktop computer. Both systems incorporate stochastic processing methodologies for assessing and combining the contributions of multiple knowledge sources in a consistent manner. Neither system was intended to represent a finished product. In fact, none of the commercially available systems incorporate natural language processing capabilities yet; one or more are likely to be available in the foreseeable future however.

In addition to IBM and Dragon Systems, other companies maintaining visibility in very large vocabulary recognition, include Kurzweil AI, and Speech Systems, Inc. Kurzweil AI also advocates using multiple knowledge sources referred to as "experts" for their future Kurzweil Voiceworks (TM) 5000 to 20,000 isolated word products. At Speech Tech87, a development prototype for a desktop system was shown running on a combined hardware/software configuration with a PC/AT connected via RS232 to external hardware.

In contrast to the discrete, large active vocabulary speech recognition capabilities now shown by IBM, Dragon, and Kurzweil, Speech Systems has focussed on providing phonetic representations of continuously spoken utterances where each is typically drawn from small active vocabularies, contained within a 20,000 word on-line dictionary. In a 2-step process, incoming speech is first reduced to a string of phonetic codes by a hardware peripheral, the Phonetic Engine (TM), and then transmitted over a terminal line for subsequent decoding on a general purpose computer.

Voice Processing Corp. is also concentrating on research in detailed acoustic-phonetic representations, but primarily for continuous small vocabularies (e.g. digit strings), recognized speaker-independently, for high bandwidth and telephone communications. Recent research at Bell Labs (Murray Hill) has centered on connected digit recognition over telephone lines, using both Hidden Markov Models (HMM) and template-based

approaches. Tests have confirmed better performance with HMMs for speaker-dependent and multi-speaker recognition (ref 15).

The Defense Advanced Research Projects Agency (DARPA) Strategic Computing Program is aimed at creating a new generation of machine intelligence capabilities. Advanced speech recognition research is a key technology in this program. State-of-the-art component technologies are being developed and integrated by a number of leading speech research laboratories throughout the U.S. A series of phased demonstrations are scheduled to assess progress toward ambitious goals. Specific application task domains provide an operationally relevant focus for these demonstrations. Tools and speech databases for training and test form an integral part of this program, facilitating algorithm assessments and comparisons.

Two specific task directions have been identified. Each encompasses a different complex of characteristics and trade-offs. For both areas, significant effort must be expended to support the most advanced algorithms with real-time hardware architectures.

Task 1: - "Robust" small-vocabulary, speaker-dependent, isolated word recognition suited to challenging environments where the speaker may be exposed to significant physical and/or psychological stresses. In fighter cockpits, for example, such systems must deal not only with effects of high noise, but also with highly variable voice distortions due to facemask characteristics, accelerations up to several G's, speaker stress, fatigue, etc.

Task 2: - Very large vocabulary (up to 10,000 words), continuous natural language, speaker independent/adaptive recognition operable in quiet to moderate noise environments. Among the tasks envisioned for this capability is natural language database query, providing timely access to complex scenario information including battle management.

The contractors presently engaged in these programs include: Bolt, Beranek, and Newman (HMM-based continuous speech recognition), Carnegie-Mellon University (New Generation System integrator, acoustic-phonetics, parsers, lexical access, parallel architecture), Dragon Systems (Computationally efficient multi-processor architecture, natural language application interface), Lincoln Laboratories (Robust HMM-based speech recognition), Massachusetts Institute of Technology (acoustic-phonetics, cochlear modeling), National Bureau of Standards (acoustic features, performance test procedures), Naval Ocean Systems Center (application integration), Stanford Research Institute (phonological rules), Schlumberger Palo Alto Research (cochlear modeling), and Texas Instruments (robust speech, multi-processor architecture, and databases).

Significant progress has already been achieved in each of the research areas cited above (ref 16-18). Preliminary full system tests have also been demonstrated on several recorded test databases, as well as with "live testing, for up to 1000 word continuous speech recognition tasks (electronic mail and natural language database interface), both speaker-independently and speaker-dependently, with and without language model/syntax constraints. Subsequent tests and reports of these will be made available in the future.

FUTURE DIRECTIONS

Speech recognition research and commercial developments are evolving far more rapidly now than at any time in the past 2 decades. The speech community worldwide has been favored by the honor of participating in this birthing process. The attendant pleasures and pains are intense.

With appropriate application integration and user interfacing, the benefits of state-of-the-art speech recognition are many, even for small to medium vocabulary, isolated word/phrase input. With speech as an additional input modality, most people will be

better able, both physically and psychologically, to access and work with information as well as tools, in their daily lives.

With the advent of much larger vocabulary and other sophisticated recognition capabilities, entirely new sets of meaningful opportunities materialize, many of them just in time to help resolve existing serious problems in the workplace. Despite the ever-increasing demands for documentation, reports, and communications, secretarial staffs continue to shrink. Nursing schools are closing their doors due to falling enrollments (with a 4:1 salary differential between doctors and nurses, serious students apply to medical school preferentially). Skilled legal and medical transcriptionists are also in short supply. Consequently costly turn-around times stretch out in producing necessary communications in these domains, as well as for the financial, business, etc.

Furthermore, the complexity of many work settings has raised the standard of skills required to work effectively at all levels, from office and factory workers, to senior

business executives and professionals. Like it or not, with diminished support, the work force as a whole, has had to become progressively more self-sufficient. The introduction of PCs has in many ways both helped to address some of these issues while exacerbating the workplace complexity problems. Speech recognition can help off-load the general lack of typing and computer skills. Two prime areas of application will be natural language database interfacing, followed by automatic dictation transcription or voicewriting. Interfacing to natural language databases using speech recognition will provide very broad capabilities, ranging from the general public in applications providing instructional and informational services as well as to senior executives and professionals in industry and government, especially for timely data access and retrieval.

FURTHER READING AND OTHER RESOURCES

This report is intended only as a snapshot. The picture is necessarily incomplete. It is guaranteed to change. Interested observers and participants in this area are encouraged to consult primary resources; namely, people and publications. In the U.S., the two major technical publications are the IEEE Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) and the Journal of the Acoustical Society of America (JASA). The largest technical conference for speech recognition is the IEEE ICASSP itself. The Fall and Spring Meetings of the Acoustical Society of America also include a number of technical speech sessions. "Speech Technology" magazine (Media Dimensions, NYC) is a quarterly directed toward applications; the two largest speech industry newsletters are "Voice Processing" (Probe Research, Morristown, N.J.) and "Voice News" (Stoneridge Technical Services, Rockville, M.D.). The largest applications shows are Speech Tech (by Media Dimensions), in the Spring, and the American Voice I/O Society (AVIOS) conference (Palo Alto, CA), in the Fall. A number of smaller specialized speech shows are also being held separately for medical, military, factory, and office automation applications.

PERFORMANCE

In operational environments, performance must be judged in terms of a combination of 1) recognition accuracy, 2) applications design/integration, and 3) user interface characteristics. Ideally all three should be optimized. A growing body of useful information and "know-how", regarding design and human factors considerations, is available through speech vendors, application designers, conferences, and publications.

Since speech recognizers operate on data influenced by diverse variables (voice quality, speaking rate, background noise, microphone filtering, etc), accuracy determinations can be difficult and time-consuming. Notoriously poor predictors include 1) subjective demo perceptions, 2) unsupported manufacturer PR claims, 3) prices, 4) fond hopes or wishful thinking, and 5) "tests" run in violation of manufacturer recommended procedures (especially for training), or accepted experimental design rules.

For rigorous comparisons of both research algorithms and commercial systems, a number of well-defined pre-recorded speech databases, available through the National Bureau of Standards (NBS), have become de facto industry standards. Although neither these databases nor any others, can be considered appropriate for a broad range of situations, rank orderings of algorithms and commercial products on these have been consistent with operational experience. General testing guidelines (ref 19) have been developed through the joint cooperative efforts of the speech recognition community. On-going performance evaluation and database activities are conducted through NBS as well as the IEEE Working Group on Speech I/O Performance Evaluation. Collaboration of these groups, at an international level, with organizations having similar concerns, such as the NATO RSG-10 Study Group, is strongly endorsed (ref 20).

POST SCRIPT

This paper is believed to be the first published paper to be completely transcribed using speech recognition. Every word of this paper, including references, punctuation, and word processing/formatting commands, has been dictated by the author and automatically transcribed using the Dragon VoiceScribe(TM)-1000 speech recognition system, running interactively, near real-time, on a 16 MHz 80386-based personal computer with standard word processing software. This text contains a total of 1538 different lexical items, of which 593 occurred twice or more.

BIBLIOGRAPHY

1. D. Nelson, Proc. Sp. Tech-86 (Media Dimensions, NYC), p62.
2. J. Pierce, personal comm., 1987.
3. J. Hall, personal comm., 1986-87
4. E. Olson, Sp. Tech-87(Media Dimensions, NYC), p95.
5. B. Dooley, Mgmt. Info. Sys. Week, 2 Feb., 1987, p10.
6. E. Diethrich, Sp. Tech in Health Care Conf.(Inst. for Med. Rec. Econ., Inc., Boston) San Francisco, 26-27 Aug., 1987.
7. D. Frazier, personal comm., 1987.
8. P. Wallich, IEEE Spectrum, April, 1987, p55.
9. D. Klatt, JASA Vol.62, No.6, Dec., 1977, p1345.
10. L. Bahl, J.K. Baker, et al, Proc. IEEE ICASSP-78, Tulsa, p422.
11. A. Averbuch, L. Bahl, et al, Proc. IEEE ICASSP-87, Dallas, p701
12. B. Davis, Wall St. Jour., 18 Nov., 1985, p1.
13. J. M. Baker, IEEE ICASSP-86, Tokyo, oral pres.
14. A. Averbuch, R. Bakis, et al, Proc. IEEE ICASSP-86, Tokyo, p53.
15. L. Rabiner, J. Wilpon, B. Huang, Proc. IEEE ICASSP-87, Dallas, p101.
16. E. Corcoran, IEEE Spectrum, April, 1987, p50.
17. L. Baumann, ed., Proc. DARPA Sp. Recog. Wkshop, Palo Alto, Feb., 1986.
18. L. Baumann, ed., Proc. DARPA Sp. Recog. Wkshop, San Diego, Mar., 1987.
19. D. Pallett, J. Res. NBS, Vol. 90, No. 5, Sept-Oct, 1985.
20. J. M. Baker, D. Pallett, and J. Bridle, Proc. IEEE ICASSP-83, Boston, p527.