WHITE PAPER
ON
SPOKEN LANGUAGE SYSTEMS

Dr. John Makhoul, Chairman
BBN Systems and Technologies Corporation

Dr. Fred Jelinek
IBM TJ Watson Research Center

Dr. Larry Rabiner
AT&T Bell Laboratories

Dr. Clifford Weinstein
MIT Lincoln Laboratory

Dr. Victor Zue
MIT

I. SCOPE

ULTIMATE GOAL

Spoken language is the most natural and common form of human-human communication, whether face to face, over the telephone, or through various communication media such as radio and television. In contrast, human-machine interaction is currently achieved largely through keyboard strokes, pointing, or other mechanical means, using highly stylized languages. Communication, whether human-human or human-machine, suffers greatly when the two communicating agents do not "speak" the same language. The ultimate goal of work on spoken language systems is to overcome this language barrier by building systems that provide the necessary interpretive function between various languages, thus establishing spoken language as a versatile and natural communication medium between humans and machines and among humans speaking different languages.

GRAND CHALLENGES

Spoken language systems differ widely in their capabilities and requirements. Three grand challenges for spoken language systems include:

• INTERACTIVE PROBLEM SOLVING -- interactive command, control, and information retrieval using voice input/output -- The system would require full integration of speech recognition and natural language understanding for input, and may require natural language generation and speech synthesis for output. Example applications include database query (e.g., airline reservations, library search, yellow pages with voice input), command and control, resource management (such as battle management workstation or logistics support), computer-assisted instruction, and aids for the handicapped.

• AUTOMATIC DICTATION (transcription) -- The challenge lies in the system's ability to transcribe arbitrary spoken input with virtually unlimited vocabulary and types of sentence construction.

• AUTOMATIC TRANSLATION -- multi-language voice input/output with automatic translation -- Example applications include automatic interpreter for multi-language speeches and meetings, translating telephone, and NATO field communications.

While these challenges constitute long-term goals for spoken language systems, there are challenging but achievable shorter-term goals that would have significant economic impact. One near-term challenge would be to develop robust, operational voice-operated data entry and query systems with limited language understanding capabilities in actual applications.

The grand challenges listed above require advances in speech processing (recognition and synthesis), natural language processing, and automatic translation. This report on spoken language systems has been written largely from a speech recognition point of view; the modeling of natural language, however, has such a great impact on the performance of a speech recognition system that language modeling becomes an important area for speech recognition research as well. The reader is referred to another document which covers the natural language processing areas in more detail. Also, natural speech synthesis is an important area that requires treatment beyond that allotted to it in this report.

MISSING SCIENCE

The areas of missing science fall in three general categories:

• Complete modeling of the speech signal and its variabilities to facilitate efficient information extraction for recognition and synthesis. These variabilities include phonetic and other linguistic effects, inter- and intra-speaker variabilities (including health condition and emotional state), and environmental acoustic variabilities.

• Automatic acquisition and modeling of linguistic phenomena, including domain-dependent and domain-independent knowledge (lexicon, syntax, semantics, discourse, pragmatics, task structure), especially the modeling of actual spoken language.

• Developing human factors methods for the design of user-friendly spoken language systems, including the use of clarification dialogues and the efficient training of users.

Statistical methods capable of modeling signal variability in parameter space as well as time, such as hidden Markov models, have put the speech recognition problem on a solid theoretical basis and have resulted in significant advances in continuous speech recognition in the last decade. The performance of such systems, however, is still far from adequate for the ultimate goals stated above and far inferior to human performance. One can improve the performance of current systems somewhat through better signal processing and feature extraction, and through extensions of the existing theoretical paradigms. However, significant improvement in performance will require a more comprehensive modeling of the speech signal and its variabilities, including possibly the development of new theoretical paradigms. A prerequisite to developing improved speech models is acquiring the knowledge of how to extract the needed information from the speech signal and how to build appropriate recognition structures that can take advantage of this information. To be useful, this knowledge must be developed in the context of building advanced speech recognition systems. An important aspect to operational speech

recognition systems will be their robustness to speaker and environmental variabilities. Of special interest to certain military applications, for example, is robustness to high levels of noise and stress. Methods that would adapt automatically and quickly to changes in speaker or in environment characteristics will need to be developed.

More comprehensive models of the speech signal will also benefit the automatic synthesis of speech from text. Current commercial synthesis devices may be adequate for some applications, but their synthetic speech quality limits their wide use. The ability of a machine to produce natural speech quality will be an important output modality for an advanced interactive human-machine interface. Speech output with natural quality will require significant research into improved speech signal models, including proper modeling of prosody, many aspects of which depend on the linguistic constructs of the text to be synthesized.

For humans, the speech understanding decisions depend on the acoustic content of the speech signal and the listener's expectation of what might be said. It is the purpose of the language model in the human to sharpen that expectation and to effect the understanding of the message. Similarly, in automatic speech understanding, it is the purpose of the language model to constrain the possible sequences of words, leading to improved recognition performance, and to interpret what was said if understanding of the message is desired. Even though much work remains to be done to solve the speech recognition problem as such, a major barrier to full realization of an advanced spoken language system arguably rests with the development of a mature natural language understanding technology. The speech recognition problem has benefited from the fact that the problem is well defined, where the input is the speech signal and the output is a set of words. Given the input and desired output, automatic methods have been developed for modeling various speech phenomena. These automatic methods have been crucial in advancing the state of the art in speech recognition. Furthermore, the performance of a speech recognition system can be evaluated by simply measuring the word error rate, for example, thus allowing for systematic and objective evaluations that can be used to improve system design. In contrast, the natural language understanding problem has not been as well defined. While the input here is taken to be a set of words, the output (i.e., the meaning of the utterance) is not well defined for all cases, nor is it well modeled computationally. One result has been the lack of rigorous evaluation of the performance of natural language systems. Partial theories for modeling meaning exist, such that limited language understanding systems have been built which may be useful in certain applications. The building of such systems has required the enormously labor-intensive process of developing grammars and semantic rules that map an input sentence into its meaning representation. The extent to which existing theories for modeling language are complete, however, has not been rigorously tested. While new linguistic theories may be needed to model a larger range of linguistic phenomena, there is a dire need to develop automatic or semiautomatic methods for the modeling of linguistic phenomena.

It is important to note that the language modeling problem is significant for speech recognition whether complete understanding of the input speech is required, as in the problem solving application, or merely transcription of what has been said, as in the dictation application. Statistical language models, for example, have been quite successful for the dictation application, without requiring understanding on the part of the machine. The translation problem, however, is directly affected by progress in modeling of syntax, semantics and discourse, especially in interactive applications over the telephone.

One of the major obstacles to the fielding of spoken language systems is often the lack of an ergonomically sound design. It is therefore important to develop a human factors technology that is appropriate for the design of user-friendly spoken language

systems. In support of possibly deficient language models, or genuine ambiguity on the part of the user, it would be important to develop graceful and effective methods for machine generation of cooperative clarification dialogues with the user to resolve possible errors or ambiguities. It would also be useful to develop methods for training users of spoken language systems to learn the limitations of such systems in a relatively short period of time. The learning by users of the capabilities and limitations of complex systems is, of course, a generic problem in other fields as well. For some applications, spoken language input and output will need to be integrated with other input/output modalities such as typing, graphics, and pointing (by mouse or touch screen).

BARRIERS TO PROGRESS

The most important barriers to progress are the areas of missing science mentioned above. Our fundamental lack of understanding of spoken language must be overcome through concentrated and substantial research efforts. In addition, there are barriers to progress in the form of computing, data, human resources, and support:

- Lack of very fast computing and large on-line storage for research.

- Unavailability of adequate and relevant speech and language databases and corpora.

- No comprehensive educational and training programs for scientists and engineers in speech and natural language.

- Lack of long-term research programs of sufficient size to build and test complete experimental systems.

Many of the advances in speech recognition in the last decade have benefited directly from the availability of faster computing. Significant additional advances can still be achieved simply by increasing computational power which would allow experimentation with more compute-intensive ideas. It is estimated that future spoken language systems may require computing speeds of 100 gigaflops or more. However, the ready availability of machines for research which compute at rates of 100-1000 megaflops would advance the state of the art considerably. These should be general-purpose machines that are easily programmable in a higher-level language for research purposes. In addition, on-line storage capabilities of at least 100 gigabytes would be needed to store data for regularly performed experiments. (See the Appendix for a detailed analysis of computing and storage needs.)

It is rather difficult to model phenomena that one is not able to observe adequately. Another deficiency in resources, in fact, has been the dearth of speech and natural language corpora that manifest the various speech and linguistic variabilities. Efforts have begun to define some of the needed corpora. It is estimated that hundreds of hours of speech and billions of words may be required to represent all the natural language phenomena of interest for the different applications. Additionally, labor-intensive special linguistic labeling of large portions of the collected data will be needed for training and test purposes, especially if semi-automatic modeling of spoken linguistic phenomena is to be accomplished. (See the Appendix for a detailed analysis of the need for large coropora of speech and text.)

The third barrier to progress is another resource problem: the lack of properly trained scientists and engineers in spoken language research. First and foremost, a solid background is needed in a number of abstract and applied mathematical disciplines,

including probability, statistics, linear systems, theory of computation and algorithms (including formal grammars, parsing, and search algorithms), logic, pattern recognition, and information theory. This background should be acquired by both speech scientists and computational linguists. In addition, speech scientists need to be trained in signal processing and speech communication, including phonetics, speech production and perception, speech analysis/synthesis, and speech recognition. Computational linguists need to be trained in phonology, morphology, syntax, semantics, and discourse, with special emphasis on data-driven, empirically-based linguistics. Of benefit to all also would be courses in psycholinguistics, cognitive science, and human factors (ergonomics). Currently, programs do not exist which offer the areas mentioned above in a coherent fashion. It is not typical for training in computational linguistics, for example, to include courses in probability and statistics. Nor are speech scientists usually properly trained in computer science. Many prominent schools do not even offer certain basic courses, such as pattern recognition.

Finally, it is obvious that insufficient funding would be a serious barrier to progress. The funding would have to be sufficient to support several research groups with critical mass on a long-term basis, i.e., groups with sufficient size to be able to build and test complete systems. Support for smaller groups that concentrate on research issues in specific areas is also necessary; however, such groups will need to integrate their work into the larger systems. Support will also be required for the purchase of adequate computing facilities and for the specification and collection of massive corpora for the purpose of system training and testing.

POTENTIAL BREAKTHROUGHS

Within the next decade, a major potential breakthrough is that large-vocabulary continuous speech recognition systems will have improved sufficiently to allow them to be integrated in some everyday applications. To make this happen, the following component technical breakthroughs are needed and are likely:

- A reduction in the word error rate for such systems by a factor of about five from the present, brought about possibly by improved signal representations, better modeling of various linguistic units, and enhanced methods for estimation of model parameters.

- The ability to handle out-of-vocabulary words in a graceful manner.

- Fast system training to new speakers and new environments.

- The integration of speech with limited natural language understanding capability for interactive applications.

In speech synthesis, a potential breakthrough is synthesis of speech from text with more natural quality than currently possible, including more natural-sounding intonation and prosody. However, work in this area would have to be increased significantly for advances to take place.

For breakthroughs in the natural language and machine translation areas, the reader is referred to another document on the topic.

II. BACKGROUND

ASSESSMENT OF THE FIELD

We will limit our discussion here to the assessment of speech recognition systems. A separate document will assess more thoroughly the natural language and translation areas. Speech synthesis systems from arbitrary text currently have a rather stylized synthetic quality, with rather unnatural intonation patterns.

The performance of continuous speech recognition systems is typically measured in terms of total word error rate, including insertions and deletions. For small vocabularies of less than 20 words, usually the digits plus some control words, speaker-independent performance (i.e., no special training needed for each speaker) has been measured at less than 1% word error rate. Systems with medium-size vocabularies of 100-200, a constrained grammar with perplexity (average branching factor) of on the order of 10, have achieved a word error rate of less than 1% in speaker-dependent mode. Both types of systems are available commercially.

Large-vocabulary systems of 1000 words, with grammars of perplexity 60, show continuous recognition performance of 5-10% word error rate. Systems with larger vocabularies typically are not operated in continuous mode, but rather the words are spoken in isolation, which tends to decrease the error rate. Very large-vocabulary systems of 20,000 words, spoken in isolation in speaker-dependent mode, perform at a 5% word error rate with a perplexity of 200. Large vocabulary systems are, for the most part, laboratory systems (some commercial isolated-word systems exist).

The results mentioned above have been obtained largely in relatively controlled environments. Performance typically degrades under hostile acoustic conditions, especially in noisy military platforms; however, good performance has been obtained for restricted tasks over dialed-up telephone lines and moderate amounts of background noise.

Most of the grammars employed with speech recognition systems are quite constrained in their ability to model natural language. For interactive applications, where understanding of the input is necessary, the grammars utilized are typically small and tree-like, with well-defined semantics. The integration of speech recognition with existing natural language understanding has started only recently.

RELATIONSHIP TO OTHER FIELDS

The design of spoken language systems depends on the integration of several of the following technologies, depending on the application: speech recognition and synthesis, natural language understanding and generation, automatic translation, and human factors engineering. It also depends on advances in computer architecture and accompanying software. Progress in all these areas is necessary for the reliable fielding of advanced systems.

An area that is intimately related to speech recognition is that of machine learning and self-organizing systems. In fact, the recent advances in speech recognition rest almost exclusively on the development of computational models (e.g., hidden Markov models and others) for which automatic training (i.e., learning) methods have existed for some time and are often taken for granted. These learning algorithms estimate the values of the model parameters directly from data, typically in a few iterations, and are able to generalize the models to unseen data. Because the performance of speech recognition systems can be

measured rigorously in terms of word error rate, the speech recognition problem, therefore, could serve as a convenient testbed for the comparative testing of other learning algorithms, such as those associated with artificial neural networks.

Another area that is closely related to speech recognition is that of speaker recognition, with its two branches: speaker verification (of a claimed identity) and speaker identification (of an unknown speaker) from a given speech utterance. The state of the art in these two areas already exceeds human performance doing the same task. It appears that humans are far better at recognizing what is being said, irrespective of who is talking, than at recognizing who is talking. As in speech recognition, the most successful approaches to speaker recognition have been those that characterize statistically the short-term spectral characteristics of a speaker. We expect that by working on recognizing speakers from their voices, we should be able to learn more about how to adapt a speech recognition system to the voice of a particular speaker.

Speaker verification is typically used for secure access to information media (such as the telephone) or to physical locations. The performance of a speaker verification system is often measured by the average of the rate of false rejection of correct talkers (customers) and the rate of false acceptance of incorrect talkers (imposters). The state of the art is less than 1% average error rate; this number is relatively independent of the number of talkers that the system can handle. Relative to speech recognition, speaker verification technology is considered quite mature and commercial products already exist. Partly because of human factors issues, these products do not appear to be in widespread use as yet.

In contradistinction with speaker verification where the system can prompt the user to say a particular utterance, in speaker identification the identity of the speaker must be determined independent of what the speaker utters, which is inherently a more difficult task. Two possible applications of this technology are the automatic identification of speakers when transcribing the proceedings of a meeting or conference (human transcribers are good at transcription but not at identifying the speakers), or for identifying speakers for intelligence purposes. The performance of speaker identification systems depends on a large number of factors which include the number of speakers in the set to be identified, the class of communication channels being used, the total amount of speech and the number of conversations for each speaker used in training the speaker models, and the amount of data used in the identification process. If, for example, we wish to identify among 20 speakers with speaker models developed from a total of 60 s of speech from different conversations and 20 s are used for identification, we expect to achieve at least 90% correct identification. With as little as 10 s for training and 2 s for identification and with communication taking place over highly variable radio channels, nearly 70% correct identification has been achieved.

CENTERS OF EXCELLENCE

The following US research organizations have strong capabilities in two or more of the needed technologies: AT&T Bell Laboratories, Bolt Beranek & Newman Inc., Carnegie-Mellon University, IBM T.J. Watson Research Center, Institute for Defense Analyses, ITT, Lincoln Laboratory, Massachusetts Institute of Technology, SRI International, and Texas Instruments.

III. RESEARCH OPPORTUNITIES

SCIENTIFIC OBJECTIVES

The scientific objectives flow from the areas of missing science above:

- Develop comprehensive models of the speech signal that take into account linguistic, as well as speaker, environmental, and channel variabilities. The ultimate goal here would be a robust continuous speech recognition system that can achieve on natural speech, using vocabularies of 50,000 words or more, word error rates of 1% or less.

- Perform rigorous testing of the quality of existing language modeling methods to natural language data, and develop more comprehensive models of natural language which include the phenomena not accounted for. As an aid in this process, develop methods for the automatic acquisition of linguistic knowledge. Incorporate the results in complete spoken language systems. The ultimate goal would be a system that is capable of understanding at least 95% of the sentences given to it by a user, and graceful handling of utterances not fully understood.

- Study the interaction of users with spoken language systems and develop criteria for the design of user-friendly systems (e.g., how to deal with out-of-vocabulary words and with new semantic concepts). Design procedures for training users on the capabilities of the system quickly, and for providing helpful clarification dialogues.

To develop the needed comprehensive models of the speech signal mentioned above, a significant program of research into various aspects of speech modeling will need to be undertaken. Areas of research include auditory modeling, acoustic phonetics, interaction of linguistic structure and speech at the phonetic and prosodic levels, adaptation to a speaker and to an environment, and new mathematical modeling techniques (e.g., maximum mutual information, stochastic segment models, and artificial neural networks).

MEASURES OF PROGRESS

It is important to monitor the progress of spoken language technology through periodic performance and evaluation tests. Speech recognition performance, for example, can be measured in terms of word error rate or sentence error rate under various test conditions. Performance of language understanding systems is more difficult to measure; it could be measured in terms of syntactic and semantic error rate, i.e., the number of sentences not understood correctly by the machine, provided that criteria for measuring correctness are well defined. However, what is really important in many cases is the accomplishment of certain tasks efficiently; therefore, number of tasks accomplished correctly within a certain time period would be another measure of performance. Many of these measures, especially those relating to natural language modeling, will have to be defined more rigorously. The performance measures, once defined, should be monitored on a regular basis as technology progresses.

Proper evaluation of progress will require the specification, collection, labelling, and distribution of sizeable databases and corpora on which system training and testing will be performed. Different types of coropora will need to be collected for the different applications. These corpora should be acquired and made available as a national resource. The National Bureau of Standards has been very active in the evaluation of speech recognition technology; they have helped in specifying testing standards (in collaboration with academia and industry) and in acting as a repository and distributor of national

databases in speech. It would be appropriate for NBS to play a similar role for work in spoken language systems.

Monitoring technology progress is important, but ultimately acceptance by the user will be the real measure of progress in spoken language technology.

IV. IMPACT

POTENTIAL IMPACT

Successful spoken language systems will redefine and revolutionize the way humans interact with machines. Such systems could be used anywhere humans come in contact with machines: at home or at school, in the car, over the telephone, in the factory, in the private and the public sectors, etc. The potential benefits are a simpler interface with certain machines and higher productivity. (Spoken language input, for example, could eliminate the need for delayed and typically tedious data entry.) The economic impact will be such that spoken language systems could result in a dominant share of the information processing industry into the next century.

Of special importance is the potential impact of spoken language understanding on military platforms which incorporate complex operational systems that interface with human operators. Such systems currently abound in DOD in the form of logistical support systems, command and control systems, tactical and battle mangagement systems, and systems where the eyes and hands are otherwise busy (such as in avionics). Spoken language technology will also facilitate the automatic analysis of massive amounts of voice communications for intelligence and security purposes. The latter applications would benefit signifcantly even with modest improvements in speech recognition accuracy.

TRANSITION TO THE REAL WORLD

We are rather fortunate that the area of spoken language systems has always had general appeal, not only for the general public, but also in the commercial and financial communities. This is attested to by the proliferation of companies in the last two decades in the areas of speech recognition and natural language access to database management systems. Therefore, the main thing that is needed in transferring spoken language technology to the real world would be simply to show feasibility by demonstrating the technology in real-world applications. Successful demonstrations would then spark commercial interest. The process of building such systems would also help identify where some of the crucial problems are, and hence impact the research towards more advanced systems.

An important aspect to such demonstrations is that they occur in real time, i.e., the user should not have to wait a long time for the machine to understand what is being said. This implies that the hardware, perhaps including special purpose accelerators, needs to be fast enough to accomplish the understanding in real time. Cost of the total system is, of course, another important aspect to transferring the technology to the real world. However, hardware costs would be expected to fall fast enough so that it cannot be viewed as a major limitation at this time.

A key ingredient to transitioning speech recognition technology to the real world is robust system operation under a variety of conditions, including the use of different microphones, background noise, reverberation, speaking rate, and various aspects of

spontaneous speech. Even more important is language-related robustness, e.g., handling out-of-vocabulary words and various linguistic constructs: a robust spoken language system would be able to respond properly to many different phrasings of requests and commands. There are also important robustness issues related to the user-machine interface, for which human factors technology can be employed profitably. All issues of robustness need to be dealt with in the context of systems working under operational conditions.

The US is generally ahead in the basic requisite technologies for spoken language systems, but other countries (e.g., Japan) are generally ahead in the commercial use of these technologies (especially in speech recognition). One reason for this imbalance in the development and the use of the technology in the US may be that the US consumer is very demanding in terms of requiring a higher degree of convenience and ease of use of the product. If the human factors issues can be dealt with satisfactorily, there is in the near term a potentially substantial market for speech recognition systems with small or medium-sized vocabularies. The development of more advanced systems with large vocabularies and natural language understanding capabilities will increase the economic impact significantly in the future.

TRANSITION TO DOD

Because the commercial and government sectors are often different in their needs and requirements, special issues typically need to be addressed in transferring the technology to the government sector, especially the military part of it. Here, it would be important to coordinate research efforts with the potential user organizations, such as the Services, and to couple demonstrations to specific application areas, so that the user organizations would adopt the technology effectively as it is being developed. Below, we present three specific areas in which spoken language technology could be used profitably.

1. Avionics

Limited-vocabulary (few hundred words) isolated or continuous speech recognition systems could be useful for a pilot to manipulate various functions in his fighter aircraft or helicopter. This application is characterized by a hands-busy, eyes-busy environment. Speech recognition would provide an additional crucial input modality for the pilot to communicate with his cockpit's computers without moving his hands from the controls or his eyes from scanning the space outside his aircraft. The payoff for allowing the pilot to keep his eyes on his target, rather than on his instruments, could be substantial. In this application, the speech recognition system must have very high recognition accuracy (>98% word accuracy) and be robust to varying noise and stress conditions.

2. Command and Control

Here, the need is for large-vocabulary (few thousand words) continuous speech recognition integrated with natural language technology for understanding purposes. However, the environment is typically not as severe as in the avionics application. The system must allow the speaker to speak in a natural, goal-directed manner, with graceful handling of out-of-vocabulary words and linguistic constructs. Typical applications include resource management and battle management, which are characterized by multi-modal interaction, including text, tables, menus, pointing, and graphics. The use of voice could obviate the need for complicated menu-based requests. The use of natural spoken language could also ameliorate the need for extensive user training of such complex systems.

3. Intelligence

Automatic spotting of key words and phrases in continuous speech can be used in surveillance and message classfication applications; it addresses the problem of reducing the workload and enhancing the efficiency of intelligence analysts. These applications are characterized by different types of noise and channel distortions and by the need to perform speaker-independent recognition of the key words. Because of the vast amounts of speech to be processed, however, even modest recognition accuracy could reap enormous benefits. There is a potential to integrate the technologies for speech recognition, speaker recognition, and language modeling, and apply them to message classification -- of which a simple but useful form is classification into messages of interest and messages of no interest. Of substantial utility also is unconstrained recognition of speech in a specific limited domain, such as monitoring of speech used in air traffic communication.

V. CONCLUSIONS AND RECOMMENDATIONS

CONCLUSIONS

Advanced spoken language systems will require the integration of the following technologies: speech recognition and natural language understanding for spoken input, and speech synthesis and natural language generation for spoken output. If the input and output languages are different, then automatic translation is also needed. In addition, speaker recognition could be an important capability for certain applications. Each of these component technologies has separate and relatively independent applications of its own, and work on each of them can be justified and should be continued independently. However, the synergism created by the integration of these technologies into spoken language systems promises to revolutionize the way humans interact with machines and to enhance communication among humans speaking different languages.

Speech recognition technology has matured sufficiently to allow for limited applications with vocabularies of hundreds of words, where the allowable grammatical constructs are limited and their semantics are well defined. Large-vocabulary applications will require reducing the word error rate further by a factor of about five. This reduction is achievable by developing more complete models of signal variabilities, more realistic language models, and utilizing larger speech and natural language databases. To increase the chances of practical utility, speech recognition technology needs to be integrated into real-world applications to guide the research. Special robustness issues need to be addressed by devoting specific effort to demonstrating speech interfaces to military and intelligence applications.

The state of the art in natural language understanding and automatic translation appears now to be sufficient for certain applications, such as limited database query systems or automatic translation of limited domains with the aid of a post-editor. However, advanced systems that allow a larger range of natural linguistic constructs will require the development of more comprehensive linguistic models, the development of methods for the automatic acquisition of linguistic and domain knowledge, and the evaluation of progress through rigorous evaluation procedures.

The integration of speech and natural language technologies, which has just barely begun, should be an ongoing endeavor that not only will incorporate the latest from the two technologies, but will also solve many problems unique to it, such as the proper utilization of the different sources of knowledge to limit the search space and result in the highest

understanding rate. The research for these integrated spoken language systems will require the availability of multi-disciplinary teams of highly trained individuals in the relevant disciplines, substantial computing power and storage, and adequate spoken language databases.

Research in speech synthesis will have to be actively supported at a significantly higher level if natural speech synthesis from text is to become a reality.

RECOMMENDATIONS

It may be the case that whoever controls the development of human-machine interfaces may own the key to controlling the information technology going into the next century. The importance of spoken language systems is in providing a simple, cost-effective interface to machines. Therefore, the US should make an effort at maintaining and enhancing its leadership position in the design of advanced spoken language systems.

First and foremost, the US must support and maintain a strong R&D program in spoken language systems in several academic and industrial centers of excellence. The program must foster basic research in the various disciplines but focus on the design, implementation, and testing of complete working systems. For the latter to take place, several research groups of sufficient size and longevity will need to be maintained. Since advanced spoken language systems do not exist as yet, many of the design issues cannot be foreseen in advance. Therefore, it will be important to promote a bootstrapping operation in which we would be able to learn from the fielding of complete working systems so that better ones can be designed, implemented, and fielded, and so on.

Second, the barriers to progress mentioned above must be overcome. Specifically, we must: make available fast computing and large on-line storage facilities for research; collect and distribute adequate speech and natural language corpora; and encourage multidisciplinary academic programs, including cooperative efforts with industry. The latter aspect may be especially important for universities that are not likely to build complete working systems. One avenue for cooperation would be for students to perform their thesis work in collaboration with industry, where the student incorporates the thesis work in an existing working spoken language system. Another avenue would be to sponsor efforts for developing certain standard software packages and complete modular systems, which implement many of the more mature aspects of the technology (e.g., standard recognition algorithms, parsers, etc.). Research at universities and other small research groups could then be performed within the context of complete working systems.

Third, it would be beneficial to nurture certain cooperative efforts with other countries, especially in the areas of automatic translation and development of common databases and corpora for the purposes of system development and evaluation.

Finally, in the application of spoken language technology to DOD, three areas should be targeted for the integration of this technology: avionics, command and control, and intelligence. The pilot's associate program should integrate the use of voice in the design of an intelligent pilot-machine interface in the cockpit. Resource or battle management offers a particularly rich command and control environment for integrating speech recognition and natural language understanding technologies with other modalities to aid the user in performing complex tasks. To assess their utility in these applications, spoken language systems must be built and tested, preferably under realistic field conditions. Lastly, research in the automatic spotting of key words and phrases and its integration with language modeling techniques for the purpose of message classification will need to be supported at a higher level if significant progress is to be made in this area.

APPENDIX. COMPUTATIONAL, DATA, AND STORAGE REQUIREMENTS

It is important to point out at the outset that the fast computing that is needed for research in spoken language systems does not derive so much from the need to achieve real-time with compute-intensive algorithms, for, given a particular algorithm, it has always been possible to build special-purpose hardware to perform the computations in real-time. Rather, it is the process of discovery of the algorithms and variations thereupon, and the testing of the various combinations of algorithms on real data, that require a fast, flexible, easily programmable computing environment. Below, we first give an example of real-time computing needs based on one of the most successful speech recognition algorithms to date. Then, we estimate future needs in terms of speech corpora that will be required to develop high-performance speech recognition systems. Based on these two data points, we estimate the computing needs to perform research in this area. We then estimate the amount of text corpora that will be needed for research in language modeling, and finally, we estimate the amount of on-line storage that will be needed for spoken language system research.

REAL-TIME COMPUTING

Assume that we are performing phonetic recognition of continuous speech using hidden Markov models for the different phonetic contexts. Because the acoustic realization of each phoneme depends not only on the identity of that phoneme but also on the identities of the left and right phonemes at least, for best recognition performance one needs a separate model for each of the possible triphone contexts, which we shall call "phones". If we assume that there are 50 phonemes in English, then the total number of triphones or phone models needed theoretically is 50**3=125,000. Not all of these will occur in actual speech because of the constraints on the allowable phoneme sequences in a language. However, because many of these constraints do not apply across word boundaries, a substantial fraction of the possible triphones can occur in continuous speech.

Assume that each phone is modeled by a finite-state hidden Markov model, and associated with each state is a probability density defined over the space of input vectors (typically representing the input spectrum every 10 ms). Further, assume that the multidimensional probability density is modeled as a mixture of Gaussian densities. For each frame of speech, i.e., every 10 ms, the recognition process consists in computing the likelihood of a sequence of phones using the given models and finding that sequence that maximizes the likelihood given the sequence of input spectra. Now, let:

P = Number of different phones
S = Number of states in each phone model
G = Number of Gaussians per state
D = Number of dimensions (components) in spectral vector
L = Number of multiply-adds per likelihood computation
F = Number of speech frames per second

The total number of computations per second of speech is then the product of all terms defined above:

C = P*S*G*D*L*F

Substituting the following realistic values for the different parameters: P=4000, S=3, G=3, D=25, L=10, and F=100, we obtain C=9*10**8 multiply-adds per second, i.e.,

approximately 10**9 or 1 gigaflops to perform the recognition in real time. (Here we have assumed that the covariance matrix of the Gaussian mixtures is diagonal.)

New phonetic models that are being considered currently will require an order of magnitude increase in computation. Further, if natural language understanding is incorporated in the whole process, one could expect another order of magnitude increase in computation. Therefore, future spoken language systems might require computing speeds of about 100 gigaflops for real-time operation.

SPEECH CORPORA

The estimation of the parameter values of phone models, a process known as training, requires the availability of sufficient speech data for training purposes. Typically, at least 50 samples of each phone are needed to estimate a robust model of that phone, i.e., one that does almost as well on test data as on training data. (Robustness here is with regard to the amount of training; there are other forms of robustness that are not intended in this discussion.) To get an idea of how much speech is needed to collect sufficient training data, we give below an accounting of the number of triphones found in the speaker-dependent training data of the standard DARPA 1000-word resource management task. The training data for each speaker totals 600 sentences, or about 30 minutes of speech. The total number of distinct triphones in the training set is 2524, out of a total of approximately 18,000 triphones in the 30 minutes. The table below shows a table where the first column gives ranges for the number of occurrences of triphones and the second column gives the number of triphones whose occurrences in the training set are in that range.

No. Occurrences	No. Triphones
1	518
2-5	1110
6-10	378
11-20	243
21-50	200
>50	75
	2524

For example, 518 distinct triphones occur only once in the whole training set, and 1110 triphones occur two to five times. Only 75 triphones have more than 50 instances; these triphones will be the ones that will have robust models. Therefore, most of the triphones will not be well modeled when only 30 minutes are used for training. The 518 triphones that occurred only once constitute less than 3% of the total number of triphones in the 30 minutes of data. This means that 3% is approximately the probability that we will see a new triphone by adding a single phoneme to the training data. In every additional second of speech (which contains approximately ten phonemes), the probability of seeing a new triphone is about 25% only. We conclude then that, to observe a much larger set of triphones and sufficiently to estimate robust models, will require the collection of many hours of speech.

To give an example of the importance of the amount of training on speech recognition performance, we give an example using hidden Markov models with the DARPA 1000-word database. Using the standard word-pair grammar supplied with this

database (this grammar has a perplexity of 60), the word error rate for a typical speaker under several conditions are summarized in the table below.

	WORD ERROR RATE	
Training Set (minutes)	Test Data	Training Data
15	8.9%	1.1%
30	5.5%	1.6%

Note that with 15 minutes of training, the word error rate on an independent data set is 8.9%, which drops to 5.5% when 30 minutes of training speech is used (this is considered a significant reduction in word error rate). To show that we would still expect a significant reduction in word error rate by increasing the training set size further, we look at the last column in the table above. This column gives the word error rate when the recognition is performed on the training data itself rather than on an independent test set. In each of the two rows in the table, the difference between the two word error rates is a measure of the robustness of the recognition system: the less the difference the more the robustness. With a smaller training set, one would expect a lower error rate on the training data but a larger error rate on test data, as is indeed the case in the table. By increasing the training to more than 30 minutes, the two columns in the table will trace two converging curves, with the difference between them decreasing with increasing training set size. Because of the large difference between the 5.5% and the 1.6% for 30-minute training, we will need a much larger amount of training data for the two word error rates to become similar. Increasing the training set size to several hours would be expected to reap benefits in reduced error rates and a more robust system.

The arguments made above, both in terms of the number of triphones that are likely to be observed with new data and in terms of improvement of performance with increased training, point to the need for substantial amounts of speech data for training purposes. To be able to observe a larger number of triphones and sufficiently to estimate robust models will likely require the collection of hundreds of hours of speech. While it may not be practical to collect this amount of speech from each speaker, it would not be unreasonable, for example, to collect a few hours from each of twenty speakers for speaker-dependent research, and one hour from each of 1000 speakers for speaker-independent research.

COMPUTING FOR RESEARCH

Much of the time in speech recognition research is spent in trying out different algorithms and variations thereof. For each experiment, one must estimate the models using the available training data and then test the models using independent test data. If it takes 1 gigaflops to do real-time recognition, then given a 1-gigaflops machine, one could do recognition on ten hours of speech data in ten hours of actual time, which means that one could do one experiment overnight. However, a 100-megaflops machine would require 100 hours for the same experiment, which means that one would not run many of those experiments.

Clearly, there may be ways to cut down the computation significantly. In fact, for many research groups today, a large fraction of the researcher's time is spent speeding up their algorithms; otherwise they would not be able to do any interesting experiments. However, as we have seen in the examples above, the computing needs could be justifiably increased substantially, for example to include a larger number of phonetic contexts or to use larger amounts of training data.

TEXT CORPORA

Text corpora or transcriptions of spoken corpora are needed in the development of language models for spoken language systems. The language model is that part of the recognizer that uses a priori knowledge of sentence generation to facilitate the decision about what was said. It must be capable of determining the likelihood that any possible sequence of words was (will be) uttered. The resulting knowledge is referred to as a grammar. We use the term "grammar" here whether it involves familiar concepts such as "noun", "verb", "subject", "noun phrase", "predicate", etc., or is defined in terms of relative frequency of occurrence of word sequences.

For applications where understanding of the speech is not necessary, such as automatic dictation, the currently most successful language model is based simply on the relative frequency of occurrence of word trigrams taken from several corpora of text totaling 250 million words with a vocabulary of 20,000 words. This trigram language model is unexpectedly powerful, but it has a number of limitations: (1) It can base its predictions only on a very short past (the last two words) of the utterance. (2) The amount of text on which it is based is miniscule (2.5 x 10**8) when compared to the number of possible trigrams from a 20,000 word vocabulary (8 x 10**12). (3) The language model reflects a narrow field of discourse: office correspondence in the data processing industry. (4) This method of model construction is inflexible as it requires huge data bases for each different recognizer application. It seems clear that future language models should in part be based on more conventional grammatical approaches. Unfortunately, the utility of currently existing grammars is quite limited for a variety of reasons, some of which are: (1) Limited coverage: Many naturally occurring sentences receive either a great many analyses or none at all. (2) Lack of statistical characterization: A sentence is deemed either legal or not; estimation of its likelihood is not attempted. (3) Linguistic phenomena often receive attention in accordance with their intrinsic intellectual interest, rather than with the frequency of their use in natural discourse. (4) Performance has been typically judged more via counterexample or by system demonstration than by rigorous testing on a large corpus of data.

For progress in realistic language modeling, it is clear that large data bases will have to be acquired, independent of the language modeling approach taken. To observe all or most of the naturally occurring linguistic phenomena in a 20,000-word vocabulary, for example, would require in principle billions of words of text. A practical long-term goal to shoot for, however, would be to collect 100 million words of text to be used as a national resource. The corpora must consist of a variety of sources so as to contain statistically significant samples of grammatical phenomena of English occurring in many applications, e.g., during text creation, dialog, interactive problem solving, etc. To enable automatic "rule" discovery and evaluation of grammar quality and development progress, the text should be linguistically annotated as to its linguistic structure (syntactic structure at a minimum, and semantic structure if possible). An initial effort to collect 10 million words over a period of two years would be a reasonable initial goal.

In support of the need for large corpora, we give the following concrete research fact. IBM has started annotating a corpus of debates from the Canadian parliament. After processing a text of 500,000 words, it has been observed that the annotation utilizes a (so far) unused context-free production rule at the rate of approximately one per additional sentence! Thus at least an order of magnitude larger sample will be required to exhibit a satisfactory fraction of linguistic phenomena occurring in even such a relatively homogeneous discourse domain.

STORAGE REQUIREMENTS

Computer on-line storage will need to support the ready availability of large amounts of speech and text data for training and test purposes. This is data that will be used on an on-going basis for the development and testing of algorithms and, therefore, should reside for the most part on-line. Currently, speech researchers utilize several gigabytes of on-line storage for their work. Advanced work in spoken language systems will deal with at least an order of magnitude increase in the amount of data to be analyzed. Therefore, 100 gigabytes of on-line storage is a reasonable estimate of what would be needed for future work.