Prosody, Performance Evaluation, Databases, and ISAT

Victor Zue
MIT

This session began with a one-hour invited tutorial on prosody given by Dr. Janet Pierrehumbert of AT&T Bell Laboratories, which was followed by several short reports on database collection and performance evaluation activities dealing with various aspects of the spoken speech system development effort. Issues and discussions on methods of formal evaluation were particularly relevant for natural language processing, where formal evaluation is in many respects an elusive goal.

Speakers and topics included (in the order of presentation):

Martha Palmer (UNISYS): Report on the Natural Language Evaluation Workshop

Beth Sundheim (NOSC): Report on the upcoming MUCK-II (Message Understanding Conference-2)

Victor Zue (MIT): Status of the proposed National Speech Database

Mitch Marcus (U. Penn): A proposal for the collection of a Parse-Tree Bank

Dave Pallett (NIST): Report on the existing TONE database of elicited spontaneous speech

George Doddington (TI): The next generation database for the DARPA Spoken Language System task

Ralph Weischedel (BBN): Status of a white paper to ISAT Study Group on Natural Language Processing, and

Roy Byrd (IBM): A proposal for a large annotated on-line dictionary.

A brief description of the invited talk and other presentations in this sessions is included below.

1. Janet Pierrehumbert's Invited Talk on Prosody

When we recognize speech, we don't just recognize segments. We recognize the segments together with their syllable structure, stress pattern, division into words and phrases, and tune. "Prosody" designates these additional aspects of the representation of the sound.

Prosody can be represented as a tree structure to which the segments and the elements of the tune are associated. This structure has pervasive effects, of which two were discussed in the talk. First, each level in the tree imposes well-formedness constraints; the example discussed was the nasal homorganic rule, which is imposed by the syllable structure and therefore is inapplicable when the nasal and the following obstruent are not in the same syllable. Second, segments are pronounced in different ways depending on their prosodic position. One result is that distributions of phonetic measures for different segments taken in isolation can overlap substantially even when the segments are well distinguished in context. For example, /s/ is longer than /z/ in any given context, but the total distribution of /s/-lengths substantially overlaps that of /z/-lengths. A similar result obtains for the degree of voicing of /s/ and /z/.

The aim of automatic speech recognition should be to recognize the entire phonological description, that is, the segments together with their prosody. Word internal prosody is likely to be handled by storing its phonetic consequences in the lexical representation. Effects of sentence-level prosody, especially gradient ones, will need to be handled by parsing the sound structure since they do not lend themselves to precompilation. These include effects on the voicing and degree of constriction of obstruents, effects on the formant patterns of vowels, and effects due to intonational control of laryngeal configuration. Building a phonological parser will be relatively straightforward, because there is no recursion and some parts of the grammar are well-understood. However, segmentation is implicit, phonetic properties vary statistically, and there is a pervasive lack of statistical independence (just as in syntactic parsing).

One aspect of the prosodic system is intonation (also "tune" or "melody"). A brief review of the English intonation system along the lines of Pierrehumbert (1980) and Beckman and Pierrehumbert (1984) was supplied. It was shown that intonation is a quasi-independent channel of speech, with different tunes used to convey different messages about how the utterance fits into the discourse. Because of this, F0 (the primary phonetic correlate of intonation) cannot be expected to resolve segmental ambiguities. However, parsing of the entire intonation pattern may help to establish the entire phonological structure. Intonation is also important because of the light it sheds on aspects of meaning which are important in natural language processing.

The last part of the talk reviewed the kinds of meaning conveyed by prosody. Subtopics treated were grouping, focus, and the intentional structure of discourse. Although prosodic phrasing is not isomorphic to syntactic constituent structure (as usually conceived of), the syntactic structure does constrain the prosodic phrasing. Further research on the relationship is needed in order to exploit prosodic phrasing information in parsing. Pitch range also indicates groupings in the discourse structure. Experiments by

Silverman show that pitch range and pause duration can disambiguate otherwise ambiguous paragraph structure. Examples were presented pointing out the relationship of focus to truth value, scope of adverbs, pronominal interpretation, and the contrast between given and new information.

Lastly, a theory of the meaning of tunes along the lines of Pierrehumbert and Hirschberg (1989) was sketched. It was argued that speakers use different tunes to convey to the hearer how each utterance relates to their mutual beliefs and to previous and subsequent utterances. Conclusions about the speaker's emotions and attitudes may be derived from the usage of tunes in context, and are therefore highly context dependent. In view of the modest state of technical formulation of these aspects of meaning, it would be most feasible and productive to begin a serious research effort on systems involving only very limited semantic domains.

2. Martha Palmer: Report on the Natural Language Evaluation Workshop

Martha Palmer, of UNISYS, described a two-day workshop that she recently organized on the topic of Evaluation of Natural Language Processing Systems. Fifty people participated in the meeting, which was held near Philadelphia last December. The workshop was structured around two main topics, black box evaluation and glass box evaluation. The first day working groups concentrated on defining criteria for black box evaluation - to measure system performance on a given task in terms of well-defined I/O pairs, such as database creation or database retrieval applications from text input. On the second day, working groups focused on defining criteria for glass box evaluation -- intended to examine the internal workings of the system. For example, glass box performance evaluation for a system that is supposed to perform semantic and pragmatic analysis should include the examination of predicate-argument relations, referents, and temporal and causal relations. Glass box evaluation is seen as subsuming black box evaluation of individual components, and is especially important for providing direction for research efforts.

There was general agreement that parsers would benefit from a standard testing procedure, and that this could best be accomplished by gathering a large corpus of marked text, some of which could be held as a secret test set. The parsers would have to be able to map from their own output to the corpus markings so that their parses could be matched against the data automatically. It is much more difficult to determine how semantics and discourse components should be evaluated. The discourse group at least agreed on a list of discourse phenomena and on the importance of getting together to continue discussing methods of evaluation. The semantics group could not even agree on a list of phenomena, much less methods of evaluation.

3. Beth Sundheim: Report on the Status of the upcoming MUCK-II Conference

Beth Sundheim, of NOSC, presented plans for an evaluation of NLP systems that have focused on the issues of text understanding as exemplified in short texts from military messages. The evaluation will conclude with the second Message Understanding Conference (MUCK-II), to be held at the Naval Ocean Systems Center in June 1989. The plan includes definition of bodies of text to use as development and test data, namely the narrative lines from a type of naval message known as OPREP-3 Pinnacle Front Burner, and definition of a simulated database update task that requires NLP systems to fill a template with information found in the texts. Documentation related to the naval messages and examples of filled templates have been prepared to assist NLP system developers. Systems would be judged on both detection (filling the correct slots) and identification (providing the correct contents). It is anticipated that developers of a number of different NLP systems will participate in the evaluation. MUCK-II will provide a forum for presenting and interpreting the results and for critiquing the test design. Two critical questions are whether the task can sufficiently be accomplished in the allotted time such that significant results can be obtained, and whether the task is designed well enough such that results will truly measure systems with respect to the desired evaluation criteria.

4. Victor Zue: A National Speech Database

Victor Zue from MIT talked about a proposed National Speech Database that would serve to greatly enhance the availability of recorded speech data suitable for training and testing recognizers. A working group of 17 individuals representing both academic and industrial sites has met twice to discuss details of the plan, and has produced a draft proposal. Approximately 1000 hours of speech would be collected representing both read and spontaneous speech, and both task-specific and task-independent subject material. Recordings would be made using up to four distinct microphone environments, but there are no plans to collect degraded speech. Orthographic transcriptions would be provided for all of the recorded material, and a subset would also be phonetically transcribed. A 5-year multi-site project would be coordinated through the Spoken Language Systems group at MIT-LCS.

5. Mitch Marcus: A National Parse-Tree Bank and Database

Mitch Marcus from U. Penn described a proposal for a database to aid in developing and evaluating existing natural language systems. The database would include annotation of some higher level linguistic structure. Parts of speech and a skeletal parse would be provided for all the data. Ambiguous attachments of modifiers would presumably be left unmarked. General Electric has volunteered some seed money to initiate the project.

There is hope that the grammar annotations can be acquired semi-automatically at a rate of about 2 million words per year. Automatic identification of parts of speech with a 1.6% error rate has been demonstrated in a system developed by AT&T Bell Labs, using the Brown corpus and a probabilistic Markov model. The availability of such automatic methods would accelerate considerably the labelling process. Several large bodies of text are available in the public domain, from sources such as the Library of America, GPO, and on-line computer documentation, and these would be exploited.

6. Dave Pallett: Report on the status of the TONE Database

David Pallett, from the National Institute of Standards and Technology (NIST) described the TONE (Task Oriented Naturally Elicited) database. This database was collected at NOSC in an experiment to elicit spontaneous speech from naval personnel performing a problem-solving task requiring database query. TI digitized the recorded speech data and developed two sets of transcriptions for the 338 spontaneous utterances in this database: one that is "raw" or "as read", and a second set that is "cooked" or "cleaned-up." The "cleaned up" version accounts for such phenomena as marked and unmarked false starts and back-ups, filled pauses, mispronunciations and semi-grammatical sentences. Patti Price, from SRI, described proposals that have been circulated within the DARPA community for the use of this database and the transcriptions for research in spontaneous goal-directed speech.

7. George Doddington: Next Generation Spoken Language Systems Database

George Doddington discussed initial draft guidelines for the development of the next-generation Spoken Language Systems speech research database. The data would promote progress in phonetic modelling, higher level modelling of the speech mechanism, and language modelling. Of primary concern was the requirement that the task be natural, i.e., unconstrained by fixed vocabulary or syntax. The speech should involve or simulate interactive human/machine problem solving, so that the domain would be naturally constrained. The speaker's focus of attention should not be on the speech act itself but rather on the task which the speech serves. The database should be sufficiently representative and general so that it can support SLS technology development that is useful in general, beyond the specific task domain of the database. Approximately 100 hours of speech would be collected from around 1000 speakers.

8. Ralph Weischedel: Status of the Natural Language ISAT Report

Ralph Weischedel from BBN reported on the status of a white paper to the ISAT Study Group in order to help DARPA identify future directions and opportunities for Natural Language. Potential breakthroughs will occur in the areas of interactive dialogue,

automatic translation, architectures for coordinating syntax, semantics, and pragmatics, algorithms that capitalize on parallel hardware, and the acquisition of a substantial grammar that could accept most sentences in common English. Since the writing of the report is still in progress, he urged members of the research community to contact the writers of the white paper and make their contributions.

9. Roy Byrd: A Consortium for Lexicon Research

Roy Byrd from IBM talked about a proposal for a Consortium for Lexical Research. The goal is to establish a large repository of lexical data which would include over 100,000 dictionary entries. The entries would include standard pronunciations, subcategorization frames for verbs, and selectional restrictions. Some proper nouns such as personal names and place names would also be included. Some tools for conveniently expanding the data set would be provided as well. He suggested that the database should be administered at an academic rather than an industrial site, although others were not in agreement on that issue. The database would permit both industrial and academic sites to pursue broad-based natural language processing problems within a common framework, and would prevent a duplication of effort on a very hard problem.

Demos and Tapes

Patti Price
SRI International

This evening session was well-attended and included both demonstrations (the DragonDictate speech recognition system and the NIST DARPA TIMIT CD-ROM and CD-ROM reader) and videotapes. The videotapes demonstrated speech recognition systems (from Lincoln Laboratories and from Carnegie Mellon University), natural language understanding (from New York University), and, in keeping with the theme of the meeting, initial studies illustrating suggestive samples of how natural language and speech recognition could work together (from MIT and from Stanford). Demonstrations and videotapes can provide an important complement to the technical sessions. There is nothing so convincing as trying or observing a system yourself in a demonstration. Videotapes can simulate this experience and serve as a promotional tool for new technologies or new concepts. As speech and natural language technologies are moved into systems and applications, we will likely see an increasing number of demonstrations and videotapes. Below is a summary of each of the presentations.

Janet Baker of Dragon Systems demonstrated the DragonDictate system. This system is a real-time, high accuracy, large open vocabulary (8000 words) product that features a seamless boundary of vocabulary size. That is, no explicit training is required: the user interface allows the user to enter new words as needed and, if the vocabulary size is exceeded, a word infrequently used is removed. A 70,000 word phonemic dictionary (based on the Random House Unabridged Dictionary) is integrated into the user interface. On-line adaptation continuously refines a speaker's acoustic-phonetic models and statistical natural language models to improve performance and throughput over time. Conference participants were allowed to talk to the system.

Dave Pallett of the National Institute of Standards and Technology and colleagues presented a demonstration of the DARPA TIMIT CD-ROM and CD-ROM player. The DARPA TIMIT speech database was designed to provide acoustic phonetic speech data for the development and evaluation of automatic speech recognition systems. The TIMIT corpus consists of a combination of dialect calibration sentences, phonetically compact sentences, and natural phonetic sentences. The database includes 630 adult talkers (male and female) of various ages representing the major dialects of American English. A demographically proportional training subset of the database including 420 talkers (over 438 megabytes) is contained on the CD-ROM. A CD-ROM containing the test material is planned for future release. Each sentence on the CD-ROM includes (1) a binary version

of the speech waveform, (2) an ASCII file containing a time-aligned broad acoustic-phonetic transcription, and (3) an ASCII file containing an orthographic transcription. Documentation on the CD-ROM and its use, and on the design of the database was distributed.

Bonnie Webber of the University of Pennsylvania showed a videotape based on work she participated in with Larry Fagan and others from the Medical Computer Science Group at Stanford University. The videotape demonstrated the use of the Speech Systems Incorporated speech recognition system in an application where voice is commonly used for data input: physicians dictating summaries of observations and procedures. The prototypes demonstrated focus not on the speech recognition technology, but on the specification of habitable constraints using graphics to indicate the legal sentences at different points in the system. A paper describing the goals of the project was distributed.

Ralph Grishman of New York University showed a videotape of the Proteus Natural Language Understanding System, a system aimed at the understanding of narrative messages in limited domains. Improving the current state-of-the-art for such systems will require a better understanding of how to capture and utilize domain information, and how to effectively combine the various sources of information (syntactic, semantic, and discourse) to create a robust language analyzer. The videotape focussed on CASREP messages describing the failure, diagnosis, and attempted repair of shipboard equipment. Much of the information to be acquired, particularly the relation between the individual events in the narrative, is implicit. Hence, a thorough understanding of these messages requires substantial knowledge of the equipment involved. The equipment knowledge base is captured through a model which incorporates structural and functional information about the equipment. The natural language analyzer uses the simulation capabilities of this model to determine the implicit causal relations between the events in a message. For example, if a message states "Compressor won't start. Shaft was sheared." the language analyzer could use the model to verify that the shearing was a plausible cause of the failure of the compressor.

Doug Paul of MIT Lincoln Laboratories presented a videotape demonstrating robust isolated word recognition, connected speech recognition in the resource management domain, and voice control of a flight simulator. The isolated word recognition system was trained on the 105-word TI-style database, which includes the words spoken in a variety of styles (fast, slow, loud, etc.). The system is capable of making fine distinctions (e.g., "no" vs. "go") quite reliably. The videotape showed the word "sensor" spoken in many different styles, always correctly recognized. The connected speech recognition system features continuous observation HMM models and uses function-word-dependent triphone models. It provides state-of-the-art performance for both speaker-dependent

and speaker-independent training modes. The simulated fighter aircraft taking off, flying a landing pattern, and landing.

Jim Glass of MIT showed a videotape describing the "Knowledgeable Navigator" (KN) performance task. The KN is envisioned to be a system with a geographical area knowledge base concerning roads and landmarks (e.g., hotels, restaurants, historic sites, stores) with the ability to provide assistance on getting from one location to another. The current implementation of the KN uses a modified version of a system for the Boston area developed by Jim Davis at the Media Laboratory of MIT. The videotape, using typed input that will eventually be replaced with speech, showed two samples of proposed interactions: asking for the location of an object, and asking for directions from one object to another. In both cases, the system responded with graphical, and textual information, as well as with synthetic speech.

Kai-Fu Lee of Carnegie-Mellon University showed a short videotape of the Sphinx connected speech, speaker-independent speech recognition system. The videotape showed the microphone being passed around the room to several talkers in an auditorium reading sentences from the Resource Management set. This system provides state-of-the-art speaker-independent recognition in near real-time.

Integration of Speech and Natural Language

Clifford J. Weinstein
MIT Lincoln Laboratory

This session dealt directly with the primary theme of the Spoken Language Systems Workshop, which was integration of speech and natural language systems.

The seven papers showed promising results and ideas both in system integration, and in speech and natural language techniques specifically appropriate for speech understanding systems.

Some of the key themes and issues were:

* What is the speech/natural language interface? Word lattices with probabilities took a central role in all papers which dealt specifically with integrating speech recognition with natural language parsing.

* How should constraints (e.g., agreement constraints) in the grammar be handled? Unification grammars were a central theme, although other techniques were introduced.

* How should the syntax-constrained search be efficiently managed? Strategies described included chart-parsing, use of probabilities for pruning, and incremental expansion of the grammar.

* Can higher-level dialog information be utilized to improve speech understanding? Positive results in dialog-based prediction were presented.

* What is the role of intonation? A theory of the relationship between syntactic and intonation structure was described.

* What are the pertinent performance measures? Measures were described and results presented for: syntactic and semantic coverage, perplexity reduction, search computation, word error rate, and semantic accuracy.

The first paper, "Integration of Speech and Natural Language," presented by Salim Roukos and David Stallard of BBN Systems and Technologies Corporation, described the structure of the BBN Spoken Language System and presented results on the DARPA

Resource Management task. The system uses HMM word models to produce a lattice of possible words, a unification grammar for syntax, and an extension of a chart parser (originally developed for text input) to find all parses for all the syntactically possible word sequences in the lattice. The semantic component then selects the semantically-meaningful response with the highest acoustic score. Performance was assessed on the DARPA RM database by defining a training corpus of 791 sentences and a test corpus of 200 sentences. Performance results on syntactic coverage, semantic coverage, and overall word error rate are given in the conference proceedings paper.

"Progress in Spoken Language Systems at UNISYS" was then described by Lynette Hirschman of UNISYS Paoli Research Center. This paper gave an overview of spoken language systems research at UNISYS. The work builds on existing text processing research at UNISYS, including a broad coverage natural language processing system called PUNDIT. As a step toward speech/natural language integration, a preliminary integration of PUNDIT with the MIT SUMMIT speech recognizer was developed, using a word lattice as the speech/natural language interface. A partitioning technique for increasing the efficiency of the PUNDIT parser's search through the word network was described, and preliminary performance results were presented. A plan was outlined for moving from the current word lattice speech/natural language interface to a more fully interleaved system. Experience with PUNDIT was cited which indicated that perplexity is not a good measure of the search pruning available from a natural language system. Finally, experience in porting PUNDIT from a message processing task to the RM task was described, which indicated that PUNDIT was portable with modest effort.

"TINA: A Probabilistic Syntactic Parser for Speech Understanding Systems" was described by Stephanie Seneff of the MIT Laboratory for Computer Science. This paper described a new natural language system designed specifically for speech understanding applications, and presented initial results on coverage, overgeneration, portability, and trainability. In TINA, a probabilistic grammar network is constructed based on example sentences. The parser uses best-first search strategies which take advantage of the probabilities in the grammar network. New techniques were described for dealing with agreement constraints, long distance movement, and gaps. The TINA system was originally brought up using a subset of the TIMIT database, and then ported to the RM database. This development was expedited by the aspect of TINA that allows a grammar to be acquired automatically from a set of parsed sentences. TINA performance on TIMIT and RM sentences are given in the proceedings paper. Plans for integrating TINA with SUMMIT to form a complete speech understanding system were described.

The paper presented by Robert Moore of SRI International on "Integrating Speech and Natural Language Processing" focussed on a dynamic grammar network approach for

increasing the efficiency of speech/natural language integration in a speech understanding system. The approach is equivalent to word-lattice parsing, but reduces computation by using a natural language parser (a stack-based unification grammar parser) to incrementally generate the grammar-state-transition table used in the standard HMM system. In this system, the recognizer does all the tracking of input positions, and the parser need not deal with word-boundary uncertainty; also, the HMM recognizer can potentially reduce its search by taking advantage of grammatical constraints. Results were presented on a subset of the DARPA RM database, which indicated that the dynamic grammar technique could produce substantial reductions in parsing time and recognition search space, relative to other word-lattice parsing techniques running on similar tasks.

The next paper, "Intonation and Syntax in Spoken Language Systems" by Mark Steedman of University of Pennsylvania, dealt with the relationship of intonational structure to syntactic surface structure. It argued that prosody can be derived directly from surface structure, using a combinatory extension of categorical grammar. According to this theory, the syntax and intonation structures of English are identical, and have the same grammar. The potential implication for spoken language understanding systems is that phonological information could be used in concert with parsing, rather than as a separate information source.

A paper on "Chart Parsing of Stochastic Spoken Language Systems" was presented by Charles Hemphill of Texas Instruments. This paper described a chart parser applied to layered stochastic regular grammars (RG), and presented the results of recognition experiments which calibrate the system with respect to an existing system using finite state automata (FSA). It was shown that training algorithms, pruning, and garbage collection could be successfully incorporated into a chart parser for stochastic RGs. The RG chart parser showed promising results (as compared to FSA) in computation time, effect of pruning on accuracy, and memory utilization. It was indicated that the layers of RGs used in these experiments are compatible with a unification grammar (UG) framework, and that further development of efficient parsers for stochastic UGs is an essential next step in the research.

The final paper of the session described the MINDS system, in which tracking of all the information in a dialog is utilized to improve speech understanding performance by predicting the probable content of the next utterance. The paper, entitled "The MINDS System: Using Context and Dialog to Enhance Speech Recognition," was presented by Sheryl Young of Carnegie-Mellon University. In the MINDS system, the predictions generated by the dialog model can range from very specific to very general, depending on what constraints are available at a given point in the dialog. A layered approach is used which allows the system to adjust the constraints based on scores produced by the

recognizer. The content predictions are translated into dynamically-generated grammars and lexicons which are used by the recognizer (an adapted version of the SPHINX system). Results were shown on a limited naval resourcement management task (dealing with ships which have problems), where predictions based on the dialog are successful in substantially reducing perplexity, and in increasing both recognition word accuracy and semantic accuracy.

Speech Results on Resource Management Task

David S. Pallett
National Institute of Standards and Technology

The session on "Resource Management Task Speech Recognition Results" included presentations discussing progress in speech recognition using the DARPA Resource Management Speech Database. Benchmark Test results were presented by DARPA contractors at BBN, CMU, MIT Lincoln Laboratory, MIT Laboratory for Computer Science, and SRI. Summary tabulations of a portion of these test results are presented in this report. Additional preliminary results using this database were presented by two other organizations: AT&T Bell Laboratories and IBM Watson Research Laboratories.

OVERVIEW

Two sets of Benchmark Test material were selected from the Resource Management Database prior to this meeting: one set of test material for speaker dependent systems included 25 sentence utterances from each of the 12 speakers in the Speaker Dependent Evaluation ("tdde") subset, and the other set for speaker independent systems included 30 utterances from each of 10 speakers in the Speaker Independent ("tdie") subset. The speaker dependent test material was used by BBN and MIT Lincoln Laboratory for their speaker dependent systems. The speaker independent test material was used for tests of speaker independent systems at AT&T Bell Laboratories, CMU, MIT Laboratory for Computer Science, MIT Lincoln Laboratory, and SRI.

Schwartz, from BBN, noted that the speaker-dependent hidden Markov model (HMM) system used at BBN for their Benchmark Tests was based on studies that included a comparison of methods for smoothing discrete probability functions. Reference (3) describes experiments conducted prior to the DARPA Benchmark Tests that indicate the best method (in those experiments) was triphone co-occurrence smoothing, a method based on deriving a probabilistic co-occurrence matrix between different vector-quantized spectra.

Paul reported Benchmark Test results at MIT Lincoln Laboratory (2) for the "Lincoln stress resistant HMM" continuous speech recognition system for both speaker-dependent and speaker-independent tasks, using both Benchmark Test sets. The system used for results reported earlier, the "June 88" system, was a "continuous observation density HMM with triphone (left and right context sensitive phone) models. For this system, word- context-free [WCF] triphones (i.e., the triphone contexts included word boundaries, but excluded the phone on the other side of the [word] boundary) were used.

More recently, word-context-dependent models were also trained, providing the "recognizer with a set of models for [both] the observed word boundaries and a set of WCF models to be used for word boundaries allowed by the grammar but not observed in the training data. This reduces the number of phones extrapolated by the recognizer." Paul's results show that the speaker-dependent system was improved significantly by the addition of the word boundary triphone model. However, the speaker-independent system results were worse than for the WCF system, and "the word-context-dependent system appears to be too detailed a model for the available speaker independent training data".

Paul also reported results using both test sets for his speaker-independent system, trained on 109 speakers (3990 sentence utterances), to provide a comparison between the two test sets. For the case of the word-pair grammar, the word error was 11.4% for the speaker-dependent test set, and 9.8% for the speaker-independent test set, possibly suggesting that the speaker-independent test set may be somewhat easier to recognize than the speaker-dependent test set. Finally, Paul also provided comparisons of performance with speaker-independent systems trained on both 72 speakers (2880 sentence utterances) [the LL (72) system of Table 2] and 109 speakers [LL (109) in Tables 1 and 2], with better results occurring for the better-trained systems.

Recent improvements in the CMU SPHINX speech recognition system were described by Lee (1). These enhancements include function-phrase modeling, between-word coarticulation modeling and corrective training. "Function word/phrase dependent models" have been incorporated to more explicitly model "the most difficult vocabulary", involving a set of 42 function words and each of the 105 phones contained in these function words. Because "function words are hardest to recognize when they occur in clusters, such as is the, that are, and of the", Lee et al. identified a set of 12 such phrases, modified the pronunciations of these phrases according to phonological rules, and modeled the phones in them separately. The new system also incorporates "generalized triphone models", created from triphone models using a clustering procedure. One benefit of this procedure is that it provides an "ideal means for finding the equilibrium between trainability and sensitivity", which is important in view of the limitations on the amount of available training material in the Resource Management Database. CMU found that "generalized triphones outperformed triphones, while saving 60% memory". Like others reporting results at this meeting (e.g., Paul at MIT Lincoln Laboratory (2) and Weintraub et al. at SRI (4)) CMU's recent work included procedures to account for between-word coarticulation. Because the number of triphone models grows sharply when between-word triphones are considered, CMU clustered a set of 7057 triphones (that was generated from an original set of 2381 within-word triphones by considering the between-word triphones) into 1000 generalized triphone models. More

complex connections are then needed to link adjacent words together in the sentence model, and the recognition algorithm must be modified. "Corrective training" [introduced by Bahl et al.] was included in the CMU speaker-independent system by using cross validation procedures and a combination of a dynamic programming algorithm to align reference sentences with misrecognized sentences in the cross-recognized training set to produce an ordered list of likely phrase substitutions. This list of phrase substitutions was then used to randomly hypothesize near-miss sentences for reinforcement learning, improving correct words and suppressing near-misses.

The SRI speaker-independent continuous speech, large vocabulary speech recognition system, DECIPHER, integrates "speech and linguistic knowledge into the HMM framework". Phonological modeling is explicitly accounted for by developing phonological rule sets based on measures of coverage and overcoverage of a database of pronunciations in order to maximize the coverage of pronunciations observed in a corpus, while minimizing the size of the pronunciation networks. The DECIPHER system incorporates probabilities into the network of word pronunciations. A number of different lexicons were studied, including those used at BBN and CMU. For the SRI lexicons studied, the mean number of pronunciations per word ranged from 1.0 to 4.2. The studies showed that careful design of the dictionary of pronunciations can yield performance improvements (i.e., automatically deriving a dictionary of most common pronunciations proved superior to the case for a dictionary carefully designed by hand by an expert linguist). Modelling a small number of multiple probabilistic pronunciations (e.g., a mean number of 1.3 pronunciations per word) showed greater performance improvements than for the case of a larger number (e.g., 4.3 pronunciations per word), perhaps because in the latter case the pronunciation networks become "too bushy". SRI attributes the success of their approach to modelling multiple pronunciations to the incorporation of constraints on the pronunciation networks. SRI's system also incorporated consideration of coarticulatory effects across word boundaries. In this implementation, "modeling acoustic variations across words was limited to initial and final phones in words with sufficient training data" (i.e., "provided that 15 occurrences of a (previous/next) phone occurred in the training database").

In contrast to the other systems described in this Session, all of which are HMM systems, the MIT SUMMIT System (5) is a phonetically based system that "attempts to express the speech knowledge within a formal framework using well-defined mathematical tools. Features and decision strategies are discovered and trained automatically, using a large body of speech data." (5) At this time, the SUMMIT system does not explicitly make use of context-dependent models. Zue et al. (5) note that the SUMMIT system, making use of 75 phoneme models, might be compared to an early version of the CMU SPHINX

system that "achieved a word recognition rate of 84% and 93% using 48 and 1,000 models" [on an earlier test set]. Zue et al. go on to note that their "result of 87% on 75 models is quite competitive using a very different approach to speech recognition than hidden Markov modelling."

Preliminary results presented by Pieraccini of CSELT on behalf of the group at AT&T Bell Laboratories described a series of recognition experiments to determine the extent to which standard modelling techniques for continuous density hidden Markov models could be applied. AT&T's studies included the effects of sampling rate (downsampling to rates of 6.67 kHz and 8 kHz), and frame shifts of 15 and 10 msec. In general, superior results occurred for the higher sampling rate due to the larger bandwidth of the signal, while the shorter frame shifts produce better temporal resolution at the acoustic level.

Picheny from the IBM Watson Laboratories presented some informal results on two large-vocabulary continuous speech tasks: the 5,000 word office correspondence task and the DARPA [991 word] Resource Management task.

SUMMARY DARPA-SITE BENCHMARK TEST RESULTS

In Tables 1. and 2., the quantity "Corr" is the percentage of words in the reference string that are correctly recognized in the system's output hypothesis strings. "Sub" is the percentage of words resulting in substitution errors, and "Del" and "Ins" are the percentages of words resulting in deletion or insertion errors, respectively. The total word error percentage, "Err" = "Sub" + "Del" + "Ins". (The term "Word Accuracy" [%] is sometimes used to refer to the quantity 100 - "Err".) "Sent Err" refers to the percentage of sentences in the test material that are recognized without errors of any kind.

The data presented in Tables 1 and 2 are derived from implementation of the DARPA NIST (NBS) standard scoring software package (6) on results reported to NIST. Although the scoring software provides a great deal of data, it was thought desirable to provide a concise summary of results for the February 1989 Benchmark Tests using a consistent format, as in these tables. Test results are shown for two conditions: (1) the "word-pair grammar", a non-probabilistic list of allowable word pairs, and (2) "no grammar", in which all words are treated as equally probable.

Because differences between the results for different systems are small, the statistical significance to be attributed to these differences is not known at present. Future modifications to the scoring software may incorporate provisions for statistical tests such as McNemar's and the Cochran Q-test (7) so that significances may be assigned to these differences.

For the speaker-independent systems, results are grouped according to the amount of system training material used in preparing for these tests.

ACKNOWLEDGEMENTS

At NIST, John Garofolo was responsible for preparation and distribution of the test material and implementation of the NIST standardized scoring software for the Benchmark Tests. It is to his credit that this task was completed in a very timely manner, despite a number of hardware and software problems.

The discussion of system attributes presented in the "Overview" portion of this paper was prepared by the author using material in references 1-5. Any inaccuracy is solely my responsibility: the reader is referred to these references for more details.

REFERENCES

1. Lee, K-F., Hon, H-W., and Hwang, M-Y., "Recent Progress in the Sphinx Speech Recognition System", in Proceedings of the February 1989 DARPA Speech and Natural Language Workshop Philadelphia, February 21-23, 1989.

2. Paul, D.B., "The Lincoln Continuous Speech Recognition System: Recent Developments and Results", in Proceedings of the February 1989 DARPA Speech and Natural Language Workshop, Philadelphia, February 21-23, 1989.

3. Schwartz, R. et al., "Robust Smoothing Methods for Discrete Hidden Markov Models", to appear in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Glasgow, Scotland, May 23-26, 1989.

4. Weintraub, M. et al., "Linguistic Constraints in Hidden Markov Model Based Speech Recognition", in Proceedings of the February 1989 DARPA Speech and Natural Language Workshop, Philadelphia, February 21-23, 1989.

5. Zue, V. et al., "The MIT SUMMIT Speech Recognition System: A Progress Report", in Proceedings of the February 1989 DARPA Speech and Natural Language Workshop, Philadelphia, February 21-23, 1989.

6. Pallett, D.S., "February 1989 DARPA Speech Recognition Resource Management Benchmark Tests, Amended January 23, 1989", notes outlining February 1989 DARPA Benchmark test procedures [distributed

at the February 1989 DARPA Speech and Natural Language Workshop, Philadelphia, February 21-23, 1989].

7. Gillick, L. and Cox, S.J., "Some Statistical Issues in the Comparison of Speech Recognition Algorithms", to appear in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Glasgow, Scotland, May 23-26, 1989.

Table 1. SPEAKER-DEPENDENT TEST SET

Average results for 300 sentence utterances (12 speakers)

a. Word-Pair Grammar

	Corr	Sub	Del	Ins	Err	Sent Err
BBN Triphone	97.5	2.0	0.5	0.6	3.1	21.0
LL Dependent	96.8	2.7	0.5	1.0	4.2	28.0
LL Independent (109)	90.7	6.8	2.5	2.1	11.4	47.7

b. No Grammar

	Corr	Sub	Del	Ins	Err	Sent Err
BBN Triphone	87.0	10.0	3.0	0.8	13.8	66.0
LL Dependent	89.6	8.4	1.9	2.8	13.2	60.3
LL Independent (109)	72.6	20.5	6.9	3.8	31.2	87.0

Table 2. SPEAKER-INDEPENDENT TEST SET

Average Results over 300 sentence utterances (10 speakers)

a. Word-Pair Grammar

Systems trained on 109 speakers	Corr	Sub	Del	Ins	Err	Sent Err
CMU (109)	94.5	4.4	1.1	0.6	6.1	34.3
SRI (109)	91.6	5.9	2.5	0.4	8.8	42.7
LL (109)	91.3	6.4	2.2	1.2	9.8	44.3
Systems trained on 72 speakers						
SRI (72)	91.1	6.4	2.5	0.3	9.2	43.7
LL (72)	90.3	6.8	2.9	1.3	11.0	50.0
MIT (72)	87.6	10.3	2.1	1.2	13.6	54.7

b. No Grammar

Systems trained on 109 speakers	Corr	Sub	Del	Ins	Err	Sent Err
CMU (109)	80.2	17.4	2.4	4.8	24.5	76.7
SRI (109)	73.8	21.0	5.2	1.5	27.7	87.0
LL (109)	75.3	18.7	6.1	3.1	27.9	82.0
Systems trained on 72 speakers						
SRI (72)	70.6	23.4	6.1	1.8	31.3	87.7
LL (72)	72.3	21.0	6.7	3.0	30.7	86. 3
MIT (72)	49.8	40.0	10.3	3.5	53.8	97.4

Natural Language

Robert C. Moore
SRI International

In this session there were presentations of natural-language processing research being carried out at New York University, University of Pennsylvania, Carnegie Mellon University, Unisys Corporation, SRI International, BBN Laboratories, and IBM.

Ralph Grishman of NYU began the session by describing work on analyzing telegraphic narratives in the Proteus project. In this project, the sentence structure of military messages (RAINFORMs) is analyzed and implicit elements are recovered. This is done by applying a general grammar of English, with relaxation of grammatical constraints. The system handles omitted subjects, omitted forms of "be," and omitted case and complement markers. Local semantic constraints are used to recover the missing elements. A penalty is imposed for missing elements, and a best-first parsing strategy is used to minimize the number of elements assumed to be missing. Omitted essential arguments are treated like anaphors. Cause/effect and enablement/action patterns are used to guide the resolution strategy.

Work going on at Penn was presented by Aravind Joshi and Bonnie Webber. Joshi described his theoretical work on equivalence of grammatical formalisms. Recent results include the fact that Mark Steedman's combinatory grammars and Joshi's lexicalized tree-adjoining grammars (TAGs) both fall in the class that Joshi had previously named "mildly context sensitive." Joshi also described work on parsing lexicalized grammars and parsing idioms using TAGs. Webber described work on cooperative response generation and discourse processing. She argued that cooperative response generation is partly reactive and partly reflective, proceeding from plan inference and evaluation and employing the declarative specifications of response content decisions. In describing the work at Penn on discourse processing, Webber concentrated on clausal reference. She stated that clausal reference comes from discourse segments on the "right frontier" of the discourse structure, and is constrained to be consistent with the interpretation of the matrix clause. This theory of clausal reference is being applied to Italian. Webber concluded by noting that both the work on cooperative response generation and the work on discourse processing share a concern with "larger" entities: events and situations.

Wayne Ward of CMU talked about the natural-language aspects of understanding spontaneous speech in the context of a speech interface to a spreadsheet system. Problems include elliptical and telegraphic utterances, false starts and corrections, "um"s,

"ah"s and other nonword utterances. None of these are handled well by current recognition techniques, but it is necessary to accommodate them to have a habitable system. The approach being taken in this work is to explicitly model spontaneous nonword speech events, to use a grammar only to provide local constraints, and to use semantics to provide global constraints. An outstanding remaining problem is to know how much to delete when a false start is hypothesized.

Following Ward's talk, there was a brief open discussion period. Much of the discussion focused on the question of whether sentence level grammar would be needed to analyze spontaneous speech.

Martha Palmer presented work at Unisys on the Pundit system, integrating syntax, semantics, and pragmatics. She first talked about the application of Pundit to message routing and analysis with the RAINFORM messages. Next she discussed the limitations of Pundit's linguistic coverage. The main part of the talk was devoted to recent work on portability: extending the linguistic coverage, partially porting Pundit to the resource management domain, and developing tools to enhance portability including a lexical entry tool and a semantic rule editor. Future directions of this work include developing a cross-domain lexicon, incorporating more general word meanings that can be specialized for a particular domain; closer integration of the lexical entry tool and the semantic rule editor; and the addition of a capability to access a knowledge base, to include more complex domain information in deriving interpretations. Finally, Palmer described work on evaluation of Pundit, including defining a performance task for Pundit, and plans for the upcoming MUCK-II evaluation meeting.

Jerry Hobbs of SRI talked about recent progress in the Tacitus project. He began by presenting the basic idea behind Tacitus, abductive inference to the best explanation, and outlining the major current research issues: making abduction more efficient, choosing among almost equally plausible competing explanations, and encoding more knowledge in the Tacitus knowledge base. He presented several tools that have been implemented recently, including a lexical acquisition component, a knowledge acquisition component, and a type hierarchy editor. He described work on creating a knowledge base for interpreting terrorist reports, and he concluded by mentioning some current problems: the difficulty of parsing and interpreting very long sentences in the terrorist reports and fragmentary and run-on sentences in the RAINFORM messages.

Robert Bobrow of BBN gave a talk on enhancing the portability of natural-language interfaces. He began by making the conjecture that the major limiting factor in the application of current natural-language systems to real-world interface problems was in fact their lack of portability. He then listed three ways that portability could be improved: by reducing the amount and complexity of application specific knowledge required, by

matching the interface to the capabilities and needs of the users, and by reducing the marginal cost of porting through the use of knowledge embedded in the application system. The bulk of the talk focused on the KNACQ knowledge acquisition system for domain model concepts. The talk concluded with discussion of future plans, and an evaluation of KNACQ on the CASES order-of-battle task, where a five to ten fold increase in productivity in porting a natural-language system to this domain was reported.

The last speaker in the session was Roy Byrd of IBM, who talked about the construction of lexical knowledge bases for word sense disambiguation. The goal of this work is to build a system that can choose the correct sense for a word, as listed in a conventional dictionary, based on the linguistic context the word occurs in. To accomplish this goal, Byrd and his co-workers are creating a hierarchically-linked lexical knowledge base derived from conventional dictionaries, large corpora, linguistic knowledge, and "preserved lexical knowledge" (i.e., manual corrections). Particular techniques discussed included finding synonymous senses in a thesaurus from the largest synonym intersection, extracting genus terms for verbs and nouns from dictionaries, and extracting "sense property vectors" (selectional restrictions) from definitions of words.

Speech

Kai-Fu Lee
Carnegie Mellon University

This session contained five speech talks. Four of these talks described recent progress and results in speech recognition, and were given by DARPA contractors (Signition, Xerox, SRI/Berkeley/BBN, and NOSC). The remaining talk was an invited guest talk from AT&T Bell Laboratories.

The first talk was given by George Zweig of Signition. In this talk, Dr. Zweig presented recent results of his cochlear mechanics research. Improvements in the speech preprocessing algorithms (e.g., moving Fourier transforms) provide a means for ambiguity reduction, one that is especially important in noisy environments. For example, speech sounds of differing intensities (such as unvoiced and voiced speech) should be preprocessed differently. His work makes the remarkable prediction that quiet sounds (e.g., unvoiced plosives) are amplified by active (energy producing) mechanisms that function somewhat like a "hydromechanical laser." Other nonlinear signal processing mechanisms, expected to be useful to the speech recognition process, are currently under investigation.

The second talk was given by Meg Withgott and Francine Chen of Xerox. Several results from basic research were described. Dr. Withgott described an experiment using Gary Kopec's new LPC spectral similarity measure, which showed substantial improvement in co-channel speech recognition, especially when the target-interference ratio worsens. Next, Dr. Withgott described the work by Julian Kupiec on probabilistic language models. The first model is based on local context, or probabilities of parts-of-speech sequences. The second is based on word recurrence modeling using a word cache. These techniques have two important advantages: it has no need for annotated data, and it is vocabulary-independent. Finally, Dr. Chen described her recent work on automatic discovery of contextual factors in phonological variation. In this work, each phoneme is partitioned using an automatically generated decision tree that asked questions about lexical contexts of the phonemes. At each node, the question that maximally increased mutual information is used to subdivide the partition represented by the node. The resulting trees show that phonetic realization is caused by many contextual factors, which suggests that proper modeling of these contexts should improve the current speech recognizers.

The third talk was given by Hy Murveit of SRI and Tony Stoltzle of UC Berkeley. Dr. Murveit provided an overview of the HMM Viterbi Beam Search hardware search

machine jointly designed by SRI, Berkeley, and BBN. The first prototype of this machine is expected to be completed by this summer. It will support up to a 3000-word task in real-time, and can accommodate various HMM-based recognition algorithms. Dr. Stoltzle gave a detailed description of the hardware, which is partitioned into a word processing subsystem, and a language processing subsystem. The language processor hypothesizes words according to grammar, while the word processor searches these word hypotheses time-synchronously. Techniques such as sequential memory access and pipelining make possible the execution of the search algorithm in real time.

Next, Steve Levinson from AT&T Bell Laboratories described the recent work on speaker-independent phonetic transcription, and the application of this work to large-vocabulary recognition. The basic philosophy behind this work is that there is a set of phonetic symbols which are mental constructs but not directly accessible. Dr. Levinson's work attempts to implement this idea in a framework where these phonetic symbols are associated with hidden Markov states, while the variability of the measurable acoustic signal is captured by the observable state-dependent random process. This phonetic transcription system was implemented using a CVDHMM to model each phone, and between-phone transitions to model phonotactics. A preliminary result of 64% phonetic accuracy was reported. Combined with a lexical access module and a parser, this system achieved a 87% accuracy on the resource management task using the finite state grammar.

Finally, Steve Nunn from NOSC described the Battle Group Tactical Trainer (BGTT) as a potential candidate for speech recognition application. BGTT is a war game simulation at the Naval Post-Graduate School. In BGTT, a commander issues commands and queries in natural spontaneous speech. A human transcriber then translates these commands and queries into a form understandable by the BGTT program. Since this is a task-oriented spontaneous speech real application, it may be a good source of natural speech data, as well as a basis on which a testbed could be developed.

Natural Language

Martha Palmer
Unisys Defense Systems

This session consisted of four talks from BBN (William Crowther), ISI (Robert Kasper), New Mexico State University (Roger Hartley), and the University of California at Berkeley (Robert Wilensky).

BBN

This talk described a large relational data base of common facts. It consists of half a million relations, most of which have been derived automatically from a machine-readable dictionary. There is no domain model, and there is no attempt at representing the meaning of any of the vocabulary items. The aim is to encode the information found in the dictionary as being relevant to a particular item in a form that allows it to be retrieved. The system is not designed to allow a user to ask questions such as "Which employees are over 35?". It cannot draw inferences about the data it stores, it can simply feed it back. An example was shown of the system being given the dictionary definition of "rattlesnake", and of the relationships it derived. The system was able to represent most of the definition in a coherent fashion, except for the information about the tail. The user goes over the relationships that are generated to weed out half truths, contradictions and circular definitions. The dictionary that was used is 600 pages long, and is estimated to contain 1.5 million relationships. The system uses a bootstrapping method where it starts with a small hand generated data base, and uses that as the semantics to parse as many dictionary entries as it can. This will result in a larger data base, so another pass will be made that can parse more entries, and so on, until the system has as many relationships as it can. In the second phase of the project, the system was able to create .5 million relationships (approximately one third of the contents), estimated to be 80% correct. It is predicted that the current version of the system would capture 1 million relationships from the same dictionary, approximately two thirds of the estimated 1.5 million relationships. There have also been some attempts to try this method with text from the encyclopedia, but that is so difficult that not much progress has been made. This is a goal for future research, as is finding suitable applications, e.g., pruning parses in a natural language processing system.

ISI

This talk presented the recent developments in the Penman Interface. This interface is aimed at helping users port the Penman generation system to new application domains,

based on the experience ISI has had porting Penman to four application areas within ISI in 1988. In generating a description of a goal in an application, the Penman system relies on the complex interaction of literally hundreds of features used to represent the information. These features can be linguistic (syntactic and semantic) or propositional. In porting to a new application, it is necessary to map the application domain model to the UPPER MODELS, which are abstract taxonomies that sit on top of the domain model, and to provide a bridge between the application domain model and the collections of Penman features. The domain model might be explicit in an expert system application but implicit in a relational data base application. It was obviously too difficult to expect someone not familiar with the system to be able to make the correct decisions about the importance of every possible feature, so an intelligent interface had to be developed that could simplify this process. A new interface notation, Sentence Plan Language (SPL), was developed that allows control of features on multiple levels, including domain facts. There are many defaults and macros built into SPL that allow the user to quickly specify clusters of features that normally appear together (e.g., how to refer to an object by a proper name), and to specify invariants (e.g., default to present tense). With SPL, the user can impose constraints on how something should be expressed from a linguistic point of view, as well as specifying what can be expressed from a propositional point of view. In addition to SPL, ISI also developed a tool for traversing the UPPER MODEL, the upper model construction tool, that provides descriptions of the upper model hierarchy. In combination, these two components, the upper model construction tool and SPL, make Penman relatively easy to use for simple applications, without limiting the power of a large purpose grammar.

New Mexico State University
Computing Research Laboratory

This presentation gave a quick overview of several different research areas at NMSU, and a short description of MGR, Model Generative Reasoning. Some of the general areas of research that are being focused on at NMSU include:

ViewGen, a computational model of belief ascription;

Problem solving by abductive assembly;

Management of alternative hypotheses;

Temporal and spatial reasoning;

Graphics based knowledge acquisition.

In MGR, knowledge is represented as partially ordered lattices and includes facts, definitions, and models. There are a small number of well-defined operations

implemented as graph joining operations which are intended to capture the semantic relations of justification and interpretation. These are:

specialize, Sp
merge, Mr
fragment, Fr
generalize, Gr

Combinations of these four operators, applied to a given set of facts, allow MGR to generate models which constitute explanatory hypotheses as well as models containing relative temporal and spatial information. By making the hypothesis generation and evaluation mechanisms explicit, MGR is claimed to implement a theory of abduction as well as support a general theory of commonsense reasoning.

Berkeley

This presentation summarized briefly several research areas at Berkeley, and went into two of them in more detail. The overall aims of the Berkeley work are 1) producing better interfaces, 2) natural language processing, and 3) building autonomous planning agents. There are two main systems that serve to focus the research efforts; a UNIX Consultant (UC) and a text understanding system. UC is an on-line system that answers questions about UNIX commands. As such, it must be cooperative and sensitive to the user's goals. It makes plans to answer the questions, and gives advice on how to get around any problems in the plan. Recent development in UC has focused on knowledge acquisition, and an important addition has been a knowledge acquisition component, UC Teacher, that reads on-line UNIX man pages and augments the knowledge base. This required extensions to the KODIAK representation language so that it could support the required learning techniques. Berkeley has also been experimenting with knowledge acquisition for UC through reading the dictionary, through being told, and through understanding metaphor. Learning through metaphor touches on another important area, the incremental acquisition of new metaphoric word senses. Recent progress has been made in getting the system to recognize a new metaphorical use by having it find a similar metaphorical use it already knows about. For example, the system can understand "kill a process" by relating it to what it already knows about "terminate a conversation." Other recent accomplishments include addressing theoretical issues involved in advancing their theory of inference, implementing a portion of a new grammar, exploring operationality with respect to Explanation Based Learning algorithms, and a speed up in the EBL algorithm.

Performance Task Progress

John Makhoul
BBN Systems and Technologies

As part of the DARPA Spoken Language Systems Program, an effort has been underway at several sites to demonstrate spoken language input to a machine in real time for a variety of tasks. This part of the program has three immediate aims:

1. Demonstrate the feasibility and utility of spoken language systems.
2. Collect realistic data for natural language research.
3. Collect speech from a real human-machine interaction.

The first item is important if the program is to have real-world utility. The second item is aimed at collecting sample English interactions between users and machines. This data will allow the natural language community to focus on research areas that are especially relevant to human-machine interaction using language. The third item of collecting speech samples from user-machine voice interactions will give speech researchers data that they can use in designing realistic speech recognition systems. The overall aim, of course, is to advance the technology of spoken language systems and render them practical for real-world applications.

Six sites (CMU, Dragon, Unisys, MIT, BBN, SRI) each gave a 10- minute presentation about their plans and progress in developing a real-time task. The tasks chosen by the different sites are:

CMU -- Spreadsheet task, including data entry, planning, and programming using voice.

Dragon -- Dictation task with continuous speech as input and text as output.

Unisys -- Maintenance of complex electronic/mechanical equipment, where the maintenance person will be able to get information using voice without having to stop what he is doing or looking at.

MIT -- Knowledgeable Navigator: voice interaction with a system that will have knowledge of the physical environment of a geographical area and will be able to provide assistance on how to get from one location to another within that area.

BBN -- Personnel database query system: using spoken input, the user will be able to ask questions and retrieve information from an existing personnel database.

SRI -- Airline travel planning: using voice, the user will be able to interact with an electronic Official Airlines Guide to access information and make flight, hotel, and car reservations.

The six sites are at various stages of system development; most had just started their work. Of crucial importance to this part of the program is the existence of hardware that can perform the recognition of continuous speech in real time. It is expected that several of the sites will have such hardware by the end of 1989. Initial demonstrations will likely take place during the fall of 1989.

Technology Transfer

Aravind K. Joshi
University of Pennsylvania

There are two aspects of technology transfer: (1) Transfer of technology form the research laboratory environment to the industrial/commercial environment. Here we are concerned with the transfer from feasibility studies and pilot implementations to practical systems with an aim towards building a product, with specific application domains and users in mind. (2) Transfer from the industrial/commercial environment to the actual user environment. Here we are concerned with how to get the clients to accept the product, which embodies, the new technology. Although the session was entitled 'Technology Transfer', the organizers of the Workshop clearly had the second aspect of transfer in mind. This is clearly reflected in the choice of the speakers and their topics for this session. Success in transfer of the first kind is necessary but it does not automatically guarantee success in transfer of the second kind. Success of transfer of the second kind is ultimately, of course, the real measure of transfer of technology.

Two of the speakers (Bates and Danis) have described specific tools and techniques that help reduce the time and cost of installing and customizing a natural language or speech system. One of the speakers (Baker) has described in detail the experience of moving a range of speech systems to the customers and the successes and failures in this effort.

Bates has described a tool called Learner, which is a software tool that creates domain dependent knowledge bases the Parlance system (a natural language system developed by BBN) needs. Thus Learner learns the knowledge Parlance needs, and thus helps create the appropriate Parlance configurations. The speed-up in creating these configurations ranges from 5 to 15 (compared to the time taken when this process was done "by hand". These results are quite impressive, given that Parlance is a fairly sophisticated natural language system.

Danis has described the development of behavioral strategies to deal with problems encountered by novice users in their failure to be recognized by TANGORA (a speech system developed by IBM). The goal here is, in some sense, to 'customize' the customer. The experiments for determining the aspects which require modification by the novice speaker and the behavioral strategies needed to accomplish this are essential because as Danis notes the large vocabulary, discrete word speech systems are not at present "walk-up and use" systems and require the users to modify their speech habits. The results of these experiments, I think, will also inform us about the development of more

robust systems, thus this effort has relevance not only for transfer of technology but also for further developments of the technology.

Baker has given a historical account of the success or failures of a range of products from Dragon Systems, e.g., Voice Scribe-64, Demos, Voice Scribe-1000, Dragon Write 5000, and Dragon Dictate. Baker considers a transfer successful only if the product is accepted by the user group for which it was designed (i.e., the user group which the designer had in mind when the product development was undertaken). Sometimes a product may get accepted by a user group for which the product was not really targeted, and it may not get accepted by the user group for which it was designed. In this case, Baker counts the transfer unsuccessful. Voice Scribe-64 was such a product, according to Baker. With this definition of success, Baker claims 3 successes and 3 failures. One important conclusion that follows from Baker's presentation is as follows: Given that we are interested in the transfer of the second kind there is not much correlation between improvements in technology and technology transfer. Baker has discussed in detail some of the factors that determine successful transfer and has illustrated his ideas with reference to specific products of Dragon Systems. I believe that Baker's analysis of the success and failure of the specific products will hold up for other speech systems products of the recent past and immediate future.

In summary, the session dealt with (1) some of the tools and techniques necessary to make language and speech systems acceptable to the users by reducing the installation and customizing period, and also to make the users modify their behavior to some extent to increase their success with the system, and (2) the history of one company's success or failure of a range of its products, with respect to technology transfer, a transfer is successful only if the product is accepted by the user group for which it was intended and not by some other user group. I believe the session was very informative to the entire language and speech community represented at the workshop, perhaps especially to those (the session chair included) who think of technology transfer only in the first sense.