Spoken Language Systems II
Richard Stern, Chairperson

The papers in this session were concerned with higher-level processing in speech recognition systems and, in some cases, the interface between the speech-recognition and natural-language components of a spoken language system. The session consisted of talks from four DARPA sites, Dragon Systems, SRI International, BBN Systems and Technologies, and MIT Lincoln Laboratory. These talks were followed by a period of free discussion.

Dean Sturtevant described recent work at Dragon Systems toward the implementation of more efficient *stack decoding algorithms* that develop and rank phrase and sentence hypotheses. For example, redundant computation of scores for partial hypotheses in the stack were eliminated by storing state information that characterized the running score on a word-by-word basis. Various ways of pruning the least likely phrase candidates from the stack, and limiting the number of hypotheses on the stack were also discussed. Speech recognition was performed using phoneme-in-contecxt word models which included characterizations of stress and the likelihood of pre-pausal lengthening, as well as dependencies on the preceding and succeeding phonemes.

Sturtevant also described several ways in which scores for partial hypotheses of varying length were tabulated. One of the more novel scoring strategies is a confidence-based measure, in which scores for a given sentence hypothesis are compared with an "expected score" obtained by averaging over many utterances that span the same path in the word/phrase network.

Some very preliminary results were described using small data sets. Operating in speaker-adapted mode, the stack decoder performed well in recognizing 7-digit strings. Recognition accuracy using a 100-word vocabulary from a radiology domain was not as good, however, as many correct hypotheses were pruned from the stack. Dragon apparently did not have time to evaluate the effectiveness of the specific innovations introduced into the decoding process.

In the second paper, Yen-Lu Chow (and Richard Schwartz) of BBN described a new *N-Best Algorithm*, which finds the best N sentence hypotheses in a computationally efficient fashion. This work was motivated by a desire to analyze incoming speech using a series of knowledge sources, performing a first pass of analysis with knowledge sources that are inexpensive in terms of computing and storage resources, and subsequently applying the more costly knowledge sources only to the more likely sentence hypotheses.

The N-Best algorithm is developed by modifying the time-synchronous Viterbi decoder. At each state, separate records are maintained for sentence hypotheses with different word sequence histories, and the probability of each partial hypothesis is tabulated. The algorithm keeps and maintains the records of the N hypotheses that have probabilities that are within a threshold of the probability of the most likely word sequence at a given state in the network. As new incoming words are considered, sentence hypotheses are generated by extending theories from the current state of the network, causing the number of hypotheses to temporarily extend beyond N. The list of theories is pruned back to N at the end of an analysis frame, or any time at which the number of hypotheses considered becomes too large. This strategy has been empirically found to incur a computational cost that is proportional to $\sqrt{N}$. While this algorithm requires at least N times the memory of each state of the hidden Markov model, the total number of states is typically much smaller than the amount of memory needed to represent all the different acoustic models.

In order to demonstrate that the correct hypothesis has a reasonable likelihood of falling within the best N generated, Chow presented some results of speech recognition experiments using sentences from the resource-management database, using a statistical class grammar with an approximate perplexity of 100, and no grammar (with perplexity 1000). Cumulative distribution functions of the rank of the correct hypothesis showed that using the fairly weak statistical class grammar, 99 percent of a test set of 215 sentences were within the 24 best hypotheses, and the average rank of the correct hypothesis was 1.8. These results indicate that the N-best strategy is a promising one. (With no grammar, the correct answer was on the final list of hypotheses only about 80 percent of the time, and the average rank of the correct hypothesis was 9.3).

In the third talk, Robert Moore of SRI International (with Hy Murveit) described their recent advances in integrating natural-language constraints into HMM-based speech recognition. The SRI system includes an HMM-based speech recognizer with a finite-state grammar, plus a unification parser that can develop a semantic interpretation of the recognized sentence.

Since the last DARPA speech and natural language meeting and this meeting eight months ago, the SRI group has achieved a speed-up by a factor of 11 in parsing time. This speed-up was accomplished by two changes to the system. First, the parser is used as a filter to the output of the recognizer, rather than as a mechanism for providing word predictions to the recognizer. (This change produced a speed-up by a factor of 4.9). Second, the parser was changed to be state-based rather than stack-based, which eliminated the need for a large number of redundant computations. (This latter change resulted in a speedup by a factor of 2.25.) When tested on 24 sentences from the resource-management domain, the parser achieved a mean parsing time of 12 seconds on a Sun 4/280 running in Prolog, on an 884-word subset of the 1000-word vocabulary. The word accuracy for these sentences was 88.4 percent.

Moore and Murveit were also able to increase the coverage of the resource-management corpus from 36 percent to 91 percent of a training set (providing 85 percent coverage of an independent test set), at the expense of an increase in parsing time by a factor of 3.5. They subsequently used an all-word first-pass matching algorithm to recover some of that time by reducing the number of words considered in each grammar state, although the speed-up factor produced by this innovation was not provided.

The final talk of the session was presented by Doug Paul, who proposed an interface specification that would be used for linking speech-recognition modules with natural-language modules of spoken language systems. Janet Baker, Charles Hemphill, and Lynette Hirschman also advised Paul in this work, although they do not necessarily each concur with all of the provisions of the specification. While the specification was developed primarily to link these two types of modules together, it can also be used in "stand-alone mode" to supply simulated output from a speech recognizer to a natural-language module, or simulated grammatical constraints to a speech recognizer. The specification includes both an integrated mode (in which both the speech recognizer and the natural-language processor contribute to the search control) and a decoupled mode (in which the speech recognizer operates independently of any language constraints and the flow of information is strictly bottom-up). The system can also output a list of the best *N* sentence hypotheses.

The basic specification assumes three components: (1) a stack decoder (such as that originally proposed by IBM and later modified by Dragon and others), (2) a module in the recognition system that can estimate the probabilities of acoustic data for partial sentence hypotheses, and (3) a natural-language processor that can estimate probabilities for the syntactic and semantic content of partial sentence hypotheses. Processing proceeds in left-to-right fashion, with partial word hypotheses generally extended in best-first order, considering both the acoustic-phonetic and semantic/syntactic likelihoods of the new words considered. The specification was deliberately written in a fashion that would minimize the number of restrictions placed on both components without sacrificing accuracy. The data-format specification also includes optional features such as different types of phrase-scoring functions, a provision for fast-match capability in either module, more detailed acoustic evaluation of best candidates in a second pass, and linguistic features (such as prosodics, markers for phrase/sentence boundaries, etc.), some error checking, and search aborts. It is assumed that the modules will communicate with each other using UNIX pipes. Paul's written report, which was widely distributed prior to the meeting, includes a detailed specification of the communication syntax. Some sample stack decodes were discussed in the presentation.

Many of the comments during the discussion period concerned Paul's proposed interface specification. While the proposal appears to have accomplished its goal of defining a workable interface standard to allow modular development of speech-recognition and natural language-processing modules across sites, its attractiveness to sites that are already developing both types of modules will clearly depend on how easily their system architectures can be adapted to accommodate the proposed standard. For example, such an adaptation might be more difficult or less advantageous for sites that have tighter coupling between the speech and natural language modules than the "parameter passing" model assumed in the specification. Nevertheless, there was a clear consensus that the effort to develop a standard interface is worthwhile (to the extent that the standard does not inhibit the development of effective but unconventional architectures by individual sites), and that adherence to a standard should be encouraged where feasible.