

NIST-DARPA Interagency Agreement: Spoken Language Program

David S. Pallett, Principal Investigator

National Institute of Standards and Technology
Room A216, Building 225 (Technology)
Gaithersburg, MD 20899

PROJECT GOALS

1. To coordinate the design, development and distribution of speech and natural language corpora for the DARPA Spoken Language research community.
2. To design, coordinate implementation, and analyze the results of performance assessment benchmark tests for DARPA's speech recognition and spoken language understanding systems.

RECENT RESULTS

1. Acquired hardware and installed the MIT/LCS-developed "TINA" and SRI-developed "DECIPHER"(tm) ATIS systems.
2. Revised the NIST speech recognition software to incorporate phonologically-motivated string alignment procedures, and prepared an ICASSP'93 paper to document the advantages of this approach.
3. Developed a speech data quality assurance software package to measure S/N and other properties. Shared this software package with other sites, including SRI, for use in monitoring quality for the WSJ-CSR corpora.
4. Acquired and made use of recordable CD-ROM technology for preliminary, limited, distribution of speech corpora.
5. Participated, with SRI, in annotation and "bug fixes" for the ATIS MADCOW-collected corpora.
6. Prepared for, and implemented benchmark tests for: (1) the Resource Management corpus (final test set, September), (2) the WSJ-CSR corpus (November), (3) the "dry run stress test" (December), (4) the ATIS MADCOW corpus (November), and (5) the "dry run end-to-end" evaluation (January).

PLANS

1. Produce and distribute the next phase of the WSJ-CSR corpora, "WSJ-CSR Phase II, Part 1", and some portion of Part 2, on pressed CD-ROM in collaboration with the Linguistic Data Consortium.
2. Coordinate collection, screening and processing of the next portion of ATIS MADCOW data, to be collected with the 46-city OAG- derived relational database.
3. Implement benchmark tests in the WSJ-CSR and ATIS domains, as required by the DARPA Program Manager and Coordinating Committee.