Session 10: Evaluation of Systems on the Resource Management Database

George Doddington, Chair
Texas Instruments
P. O. Box 655474, MS238
Dallas, TX 75265

This session presented papers on a number of different approaches to improving speech recognition performance, developed and evaluated on the DARPA resource management speech database. The session began with a summary presentation of results by Dave Pallett. Most systems perform comparably, with word-level total error rates in the range of about 6 percent on the June 1990 RM2 speaker-independent test set. The best performance, 4.3 percent error, was achieved at CMU.

BBN, in exploring training and speaker adaptation techniques, found that a large amount of speech from a few (12) speakers is effective for speaker-independent training, provided that the speakers' data are not pooled before training. This technique facilitates rapid speaker adaptation and incremental increases in the training database.

AT&T and CMU presented papers on exploration of enhanced acoustic features to improve recognition performance, with emphasis on higher-order differences in speech parameters. AT&T used multivariate Gaussian mixture models, while CMU expanded their discrete model to 4 codebooks. When a semi-continuous observation model was used in conjunction with these 4 codebooks, a combined total error rate reduction of 30 percent was achieved, yielding 4.3 percent word error.

MIT Lincoln Laboratory also demonstrated performance with a semi-continuous or tied mixture observation model, and introduced the "semiphone", a context-dependent phonetic model with the reduced time scope from the traditional triphone, which promises significant reduction in number of models.

SRI presented two papers, the first exploring training set issues and the second noise robustness. SRI found that speaker-independent recognition performance could be significantly improved (word error was reduced by 20 percent) by tripling the size of the training data set with a large amount of data from a few (12) speakers. This supports the finding of BBN. SRI also were able to achieve significant performance improvement from separated sex-specific recognition models and from corrective training.

SSI explored tree-structured MMI encoding and found slight improvement over traditional minimum distortion encoding. These results are preliminary, however, and error rates were in general significantly higher than those for other systems.