# SESSION 4: SPEECH I

*Richard F. Lyon*

Apple Computer, Inc.,
20450 Stevens Creek Boulevard, MS 76-2H
Cupertino, CA 95014

## INTRODUCTION

This session focussed on robustness in speech recognition. The first two papers considered the effects of working with a population of real users in telephone-services environments, while the final two papers looked at front-end technologies, namely microphone arrays and representations of acoustic information.

## SUMMARY OF PRESENTATIONS AND DISCUSSION

The first paper, "Field Test Evaluations and Optimizations of Speaker Independent Speech Recognition for Telephone Applications," by Gagnoulet and Sorin of CNET, was presented by Christel Sorin. This paper discussed various ways of improving system usability and performance by optimizing both the dialog ergonomy and the recognition technology within the constraints of low-cost real-time implementation. Techniques discussed included use of field data in training, increasing the number of parameters, automatic adjustments of the HMM structure, and better rejection procedures. A brief discussion of the rejection rate versus error rate tradeoff ensued; nobody had any good data or ideas on how to make this tradeoff, so when one person suggested that the rejection rate should be adjusted to keep the error rate under 5 percent, we said OK and moved on.

The second paper, "Collection and Analysis of Data From Real Users: Implications for Speech Recognition/ Understanding Systems," by Judith Spitz and the AI Speech group at NYNEX, concentrated on analyzing user response characteristics as a function of the prompts used, and on comparing user versus laboratory speech characteristics with respect to their effects on recognition performance. Since NYNEX has gone to the trouble of collecting lots of good data, including TIMIT data run through the telephone network, there was some discussion of the possibility of distributing some of their data, such as the Network-TIMIT data and telephone services data, through NIST. Legal issues are the most serious problem at this point for the telephone services data, since it is not possible to get explicit consent from the talkers.

The third paper, "Autodirective Microphone Systems for Natural Communication with Speech Recognizers," by Flanagan, Mammone, and Elko of Rutgers University, was presented by Jim Flanagan. He surveyed recent advances and opportunities in steerable-beam microphone arrays with automatic source tracking. An audio tape demonstrated excellent-quality recording from a single speaker in a 300-seat auditorium using a 2D array on the ceiling. A video tape showed the 1D array used in the HuManNet system. The relative merits of noise cancellation filters and steerable beams were discussed, and it was suggested that noise cancellation may actually be a much more useful technique when combined with a steerable microphone array.

The final paper, "Signal Representation, Attribute Extraction, and the Use of Distinctive Features for Phonetic Classification," by Meng, Zue, and Leung of MIT's Laboratory for Computer Science, was presented by Helen Meng. This presentation covered results of careful experimental comparisons of different front-end representations (e.g. auditory, Mel-cepstrum, DFT, etc.) and various ways of incorporating acoustic attributes and distinctive features, in the context of multilayer perceptron based vowel classification. The dual auditory model (mean rate plus synchrony representations) worked best as the front end (especially in noise, but by an insignificant margin in some other cases). Significant computational savings was possible by reducing the front-end output to a few simple acoustic attributes, and the loss in accuracy was small and probably insignificant. Using distinctive features was said to provide for the possibility of better phonological-level generalization; the loss in accuracy of incorporating features (followed by a second MLP to do the phoneme classification) was not significant. Discussion followed on possible explanations for why the auditory model works as well as it does; nobody had a suitable explanation, but the conjecture that it was primarily due to the synchrony information was shown to be not supported by the data, since the rate-only model worked almost as well.