# Real-Time Speech Recognition Systems

Hy Murveit and Mitch Weintraub
SRI International

## Objective:

SRI and U.C. Berkeley have begun a cooperative effort to develop a new architecture for real-time implementation of spoken language systems (SLS). Our goal is to develop fast speech recognition algorithms, and supporting hardware capable of recognizing continuous speech from a bigram or trigram based 20,000 word vocabulary or a 1,000 to 5,000 word SLS systems.

## Recent Accomplishments

- We have designed eight special purpose VLSI chips for the HMM board, six chips at U.C. Berkeley for HMM beam search and viterbi processing, and two chips at SRI for interfacing to the grammar board.
- SRI and U.C. Berkeley have completed simulations of each of the special purpose VLSI chips.
- U.C. Berkeley has completed the fabrication of one printed circuit board, for the HMM output probability computation.
- By reimplementing SRI's Natural Language Parser from PROLOG to C, SRI was able to decrease the parse time by a factor of three.
- SRI modified SRI's language processing algorithms to implement a continuum between SRI's Dynamic Grammar Network (DGN) and BBN's N-best algorithm, to balance the computational load between the HMM speech recognizer and the SLS parser.
- SRI completed the implementation of SRI's standard VQ front-end on the Berkeley dual TMS320C25 board, and in C.
- Implemented corrective training to improve recognition performance; on the standard training set this improves speaker-independent perplexity 60 performance from 6.7% error to 5.1% error, and for a larger training set (about 11,000 sentences), improves speaker-independent recognition from 5.3% error to 4.1% error.

## Plans

- Complete the construction of the current hardware design, and develop software tools to support this architecture.
- Develop a large vocabulary recognizer to fully use the capabilities of this design.
- Design, implement, and evaluate algorithms for real-time grammar processing computation.
- Evaluate the current architecture to determine the computational and algorithmic bottlenecks.
- Replicate the system and port to a DARPA and NASA site.