HIGH PERFORMANCE SPEECH RECOGNITION USING CONSISTENCY MODELING

Vassilios Digalakis
Hy Murveit
Mitch Weintraub

SRI International
Speech Research and Technology Program
Menlo Park, CA, 94025

PROJECT GOALS

The primary goal of this project is to develop acoustic modeling techniques that advance the state-of-the-art in speech recognition, focusing on those techniques that relax the hidden Markov model's improper independence assumptions. Such techniques should both improve robustness to systematic variations such as microphone, channel, and speaker, by conditioning state's acoustic output distributions on long-term measurements, as well as improve general acoustic calibration by removing improper short-term (e.g. frame to frame) independence assumptions.

In order to perform this work certain infrastructure needs to be developed. This includes the development of a state-of-the-art baseline recognition system for the development task (ARPA's Wall-Street Journal Task); the development of search techniques that allow experiments with computationally expensive techniques to have reasonable turn-around times; and the development of modular software that enables rapid prototyping of new algorithms.

RECENT RESULTS

- We have built a software library that implements the components of an HMM recognition system dealing with the observation distributions. The functional interface is designed to enable fast integration of new acoustic modeling techniques

- We introduced a new search strategy, called Progressive Search, that constrains the search space of computationally expensive systems using simpler and faster systems in an iterative fashion. Using the word graphs created during the initial recognition pass as grammars in subsequent recognition passes, we have been able to reduce recognition time of systems that use more complex acoustic models and higher order language models by more than an order of magnitude.

- We developed a less-traditional, continuous output distribution system where different allophones of the same phone share the same sets of Gaussians, but different Gaussians are used for different phones. Our phonetically-tied mixture system achieved a 16% reduction in error rate over a typical tied mixture system.

- We found that the different pronunciation dictionaries and the corresponding phone sets that the various sites used in the last CSR evaluations can account for differences in performance in the order of 10 - 15%.

- We developed new algorithms for local consistency by modeling the correlation between spectral features at neighboring time frames. This acoustic correlation is used to improve the accuracy of the acoustic model by conditioning the state output probabilities on the previous frame's observations.

- We have achieved a 31% reduction in error rate over our November evaluation system on the 5K, non verbalized punctuation development set. The improvement is the combined effect of the phonetically-tied mixtures, the improved pronunciation dictionaries and replacement of RASTA filtering with cepstral-mean removal on a sentence basis.

PLANS FOR THE COMING YEAR

- Continue exploring trade-offs in parameter tying for continuous distribution acoustic models. We will sample other points beyond tied-mixture, phonetically-tied mixture, and untied Gaussian-mixture systems.

- Explore techniques for modeling the global consistencies of speaker and channel effects across the speech acoustic models.

- Continue to develop search techniques that both allow us to perform experiments using computationally burdensome techniques, as well as those that allow us to implement these systems as real-time demonstrations.