Segment-Based Acoustic Models for Continuous Speech Recognition

Mari Ostendorf *J. Robin Rohlicek*
Principal Investigators

Boston University
Boston, MA 02215

BBN Inc.
Cambridge, MA 02138

PROJECT GOALS

The goal of this project is to develop improved statistical models for speaker-independent recognition of continuous speech, together with efficient search algorithms appropriate for use with these models. The current work on acoustic modeling is focussed on: stochastic, segment-based models that capture the time correlation of a sequence of observations (feature vectors) that correspond to a phoneme; hierarchical stochastic models that capture higher level intra-utterance correlation; and multi-pass search algorithms for implementing these more complex models. In addition, we have extended the effort on models of high order statistical dependence to language modeling. This research has been jointly sponsored by ARPA and NSF under NSF grant IRI-8902124 and by ARPA and ONR under ONR grant N00014-92-J-1778.

RECENT RESULTS

Recent results on this project are summarized below with names of the students primarily responsible for the work indicated in parentheses.

- Ported the BU recognition system to the Wall Street Journal (WSJ) task and Switchboard task, obtaining results similar to HMM systems on those tasks. Also implemented several software changes to handle large vocabularies and allow for larger N-best lists by using more efficient score caching, as well as to accommodate the full amount of training data available. (F. Richardson, S. Tibrewal, A. Kannan)

- Continued investigation of mixture distribution modeling at both the segment and frame levels, shifting our focus primarily to "untied" segmental mixture systems. We have established baseline results and investigated various parameter allocation choices for these models in experiments on the Resource Management task. For context-independent models, performance is found to improve over uni-modal and tied-mixture systems, through combining segmental and frame-level mixtures. Further work on initialization is needed for estimating context-dependent models. (O. Kimball)

- Implemented a new duration model that uses speaking-rate adapted parameters.

- Developed a sentence-level mixture n-gram language model to handle topic-related language dynamics, and evaluated recognition performance with this model on the 5k WSJ task in the N-best rescoring framework, obtaining a slight improvement over standard trigrams. (R. Iyer)

- Developed the theoretical framework for an automatic mapping of distributions to arbitrary subsets of a variable-length segment feature matrix, as an alternative to the linear-time frame mapping currently used in the SSM.

- Developed the theoretical framework for a hierarchical model of intra-utterance observation correlation.

- Developed a new algorithm for fast search of a word lattice for multi-pass recognition scoring. (F. Richardson)

PLANS FOR THE COMING YEAR

- Implement and evaluate the hierarchical stochastic model of intra-utterance dependencies, first in TIMIT classification and later in the WSJ system if initial experiments are successful.

- Investigate unsupervised adaptation in the WSJ task domain.

- Investigate algorithms to improve recognition accuracy for telephone speech.

- Assess accuracy/speed trade-offs for different lattice search algorithms for the WSJ task.

- Extend work in mixture language modeling to capture more language dynamics and/or task domain change through adaptation.