# REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188



| 1. AGENCY USE ONLY (Leave blank) | 2. REPORT DATE | 3. REPORT TYPE AND DATES COVERED |
|---|---|---|
| | December 1995 | Quarterly Report - 10/1/95 - 12/31/95 |

**4. TITLE AND SUBTITLE**
High-Order Modeling Techniques for Continuous Speech Recognition

**5. FUNDING NUMBERS**
8547-5 - BU Source #
ONR Grant #: N00014-92-J-1778

**6. AUTHOR(S)**
Mari Ostendorf

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
Trustees of Boston University
881 Commonwealth Ave.
Boston, MA 02215

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**

**11. SUPPLEMENTARY NOTES**

**12a. DISTRIBUTION/AVAILABILITY STATEMENT**
Approved for public release; distribution is unlimited

**12b. DISTRIBUTION CODE**

**13. ABSTRACT** *(Maximum 200 words)*

This research aims to develop new and more accurate stochastic models for speaker-independent continuous speech recognition by developing acoustic and language models aimed at representing high-order statistical dependencies within and across utterances, including speaker, channel, and topic characteristics. These techniques, which have high computational costs because of the large search space associated with higher order models, are made feasible through a multi-pass search strategy that involves rescoring a constrained space given by an HMM decoding. With these overall project goals, the primary research efforts and results over the last quarter have included: 1) developed much of the theory for two new models for adaptation, 2) further explored methods for robust dependence tree topology design and implemented training algorithms for hidden dependence tree models; 3) repeated sentence-level mixture language modeling experiments with new versions of NAB training set, showing improvements in both perplexity and word error rates; 4) developed software tools for using HTK in experiments on HMM topology design; and 5) furthered efforts on establishing a baseline HTK recognition system for a task of recognizing the Macrophone natural numbers data, on which we currently achieve 84% word accuracy

| 14. SUBJECT TERMS | 15. NUMBER OF PAGES 13 |
|---|---|
| | 16. PRICE CODE |

| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |
|---|---|---|---|
| unclassified | unclassified | unclassified | |

Boston University

College of Engineering
44 Cummington Street
Boston, Massachusetts 02215
617/353-2811

Electrical, Computer and Systems Engineering

February 17, 1996

Defense Technical Information Center
Building 5, Cameron Station
Alexandria, Virginia 22304-6145

Dear Sir or Madam,

Enclosed is a copy of the quarterly progress report for ONR research grant No. N00014-92-J-1778, "High-Order Modeling Techniques for Continuous Speech Recognition," for the period from October 1995 to December 1995. Please let me know if I can provide any additional information. I would also be happy to hear any feedback you have about the research.

Sincerely,

Mari Ostendorf
Associate Professor
617-353-5430



ENCL (8)

# High-Order Modeling Techniques for Continuous Speech Recognition

## Progress Report: 1 October 1995 – 31 December 1995

submitted to
Office of Naval Research
and
Advanced Research Projects Administration

by
Boston University
Boston, Massachusetts 02215

**Principal Investigator**

Dr. Mari Ostendorf
Associate Professor of ECS Engineering, Boston University
Telephone: (617) 353-5430

**Administrative Contact**

Maureen Rodgers, Awards Manager
Office of Sponsored Programs
Telephone: (617) 353-4365

# Executive Summary

This research aims to develop new and more accurate stochastic models for speaker-independent continuous speech recognition by developing acoustic and language models aimed at representing high-order statistical dependencies within and across utterances, including speaker, channel and topic characteristics. These techniques, which have high computational costs because of the large search space associated with higher order models, are made feasible through a multi-pass search strategy that involves rescoring a constrained space given by an HMM decoding. With these overall project goals, the primary research efforts and results over the last quarter have included:

- developed much of the theory for two new models for adaptation,
- further explored methods for robust dependence tree topology design and implemented training algorithms for hidden dependence tree models;
- repeated sentence-level mixture language modeling experiments with new version of NAB training set, showing improvements in both perplexity and word error rates;
- developed software tools for using HTK in experiments on HMM topology design; and
- furthered efforts on establishing a baseline HTK recognition system for a task of recognizing the Macrophone natural numbers data, on which we currently achieve 84% word accuracy.

As usual, substantial software maintenance and development efforts were also required during this period.

# Contents

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: High-Order Modeling Techniques for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 October 1995 – 31 December 1995

# 1 Productivity Measures

- Refereed papers submitted but not yet published: 1 revised and re-submitted
- Refereed papers published: 0
- Unrefereed reports and articles: 1 thesis proposal, 3 conference papers submitted
- Books or parts thereof submitted but not yet published: 1
- Books or parts thereof published: 0
- Patents filed but not yet granted: 0
- Patents granted (include software copyrights): 0
- Invited presentations: 1
- Contributed presentations: 2
- Honors received: none
- Prizes or awards received: none
- Promotions obtained: none
- Graduate students supported $\geq$ 25% of full time: 3
- Post-docs supported $\geq$ 25% of full time: 0
- Minorities supported: 2 women

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: High-Order Modeling Techniques for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 October 1995 – 31 December 1995

# 2 Project Summary

## 2.1 Introduction and Background

The goal of this work is to develop and explore novel stochastic modeling techniques for acoustic and language modeling in large vocabulary continuous speech recognition, particularly recognition of spontaneous speech. Although significant advances have been made in recognition technology in recent years, spontaneous speech recognition accuracy is still hardly better than 50%. More casual speaking modes introduce additional sources of variability that require improvements at all levels of the recognition process: signal processing, acoustic modeling, lexical representation and language modeling – both in terms of the baseline stochastic models and the techniques for adapting these models. In addressing these challenges, the general theme of the research in this project is high-level correlation modeling, i.e. representing correlation of observations beyond the level of the frame or the word to dependencies within and across utterances associated with speaker, channel, topic and/or speaking style. In particular, we will concentrate on three problems: development of hierarchical models of intra-utterance correlation of phones and model states, e.g. by extending the theory of Markov dependence trees; unsupervised adaptation of acoustic models within and across utterances based on these models; and language modeling triggered by acoustic and dialog-level cues. The research approach is to develop formal models of statistical dependence that overcome limitations of existing models, in combination with exploring fast search and robust parameter estimation techniques to address the added complexity of these models. By considering radically new but formal models, rather than minor variations of existing models or heuristic patches, this work offers the potential to address many of the most difficult problems in speech recognition, including recognition of spontaneous speech. By also building on the existing strengths of speech recognition technology, both in the theoretical foundation and in the use of multi-pass search, this work has the added advantage that advances will be more apparent and more easily transitioned to existing systems.

In FY95, the focus of the efforts on this project were in three main areas. First, standard n-gram training and dynamic cache language modeling techniques were extended for use in sentence-level mixture modeling, yielding significant reduction in perplexity though only small gains in recogni-

tion performance as yet. Second, an algorithm was developed and implemented for training discrete dependence trees with missing observations, and initial experiments have explored topology design issues and obtained improved prediction error using dependence trees in a simple adaptation experiment. Third, lattice search algorithms were implemented to reduce computation in segment model rescoring, including a local search algorithm suitable for the higher-order language and acoustic models explored in this work. Other results included development of a parametric segment model clustering algorithm, development of an algorithm for HMM topology design and exploration of auditory-based signal processing algorithms. In addition to these research advances, significant effort was devoted to software system improvements and participation in the ARPA speech recognition benchmarks, where BU achieved 11.6% error in the officially reported result and 10% error using the BBN benchmark system for first pass scoring. In the most recent quarter, we have continued to build on these results, as described in the next section.

## 2.2 Summary of Recent Work

The research efforts during this period covered a variety of problems as summarized below, including work supported in part by an ARPA AASERT award that is coordinated with this effort.

**Theoretical Development of Adaptation Models.** With the goal of developing algorithms for adaptation with small amounts of data, Ashvin Kannan developed two complementary approaches to representing correlation of observations in a Bayesian MAP estimation approach to adaptation. First, we developed a framework for adapting the parameters of a polynomial trajectory model [7] using a prior on a shift of the regression parameters that is tied over subclasses of models. This approach allows for consistent adaptation of parameters across a segment. To capture correlation of observations across broader classes of phones, we have begun development of a parsimonious model that represents shifts using a stochastic process evolving on a tree [6], where the tree structure in our case is simply that tree designed using distribution clustering. The new approach has a few remaining theoretical issues to be resolved, but it offers the possibility of adapting all parameters given data from only a subset of models.

**Intra-utterance phoneme dependence modeling.** We have extended our previous work in dependence tree models in the development and experimentation of the continuous case extension of the dependence tree model. Earlier, we introduced a hidden dependence tree model in which the discrete phone-state vector of an utterance represents the hidden modes of Gaussian mixture distributions. This approach allows us to make use of the entire continuous observation sequence as opposed to using only the observable phone-state in the discrete case. In the past quarter, Orith Ronen implemented the training algorithm for the continuous case. Experiments showed that the distributions based on explicit training of the continuous model gives significantly higher likelihood

of independent test data than using a discrete model in a continuous framework. (This difference is analogous to the difference between HMM tied mixture retraining vs. use of VQ codewords in a semi-continuous HMM without distribution mean retraining.) Since the discrete model is still used for tree topology design, we also continued to run experiments with the discrete dependence tree model design using the WSJ corpus, focusing on robust tree topology design. The experiments showed that the WSJ corpus gives better tree topologies than TIMIT, probably because the training set is much larger. Dependence trees that were automatically designed by the tree topology design algorithm generally capture correlation between phones due to manner of articulation (Z-S-SH, M-N-NX), but also other correlations that may by due to co-articulation effects (AA-R).

**Language Modeling.** In the area of language modeling, efforts involved reassessing the sentence-level mixture language models. The aim of the sentence-level mixture language model is to capture topic-dependent effects within sentences. In addition, we extended the model to use a dynamic cache, which provides a mechanism for capturing word dependence across sentences. Although the language modeling advances gave significant gains in perplexity, there were no gains in recognition in earlier experiments on the NAB task, specifically the November 1994. This was mainly due to a mismatch in the acoustic and LM dictionary, as well as training based on poorly conditioned data. Since a better version of the North American Business (NAB) news data was recently made available by LDC, Rukmini Iyer attempted to analyze and possibly improve upon our previous results on the November 1994 ARPA benchmark evaluations. She retrained the components of the mixture model using the new data and a dictionary matched with the acoustic dictionary used during the ARPA evaluations. We observed a 2.5% improvement in recognition accuracy over our previous baseline results in the H1-P0 segment model system. The static (not adapted) mixture model also performed better than a standard trigram model adapted in a supervised mode. The dynamic modeling techniques implemented in the framework of the mixture model, e.g. using weighted n-gram and content-word unigram caches, gave a small but significant gain and gave us our best performance of 10.8%, an overall improvement of 4.5% in recognition accuracy over the previous best performance for the segment model system. Further experiments are being conducted on the best case system which combines both segment model and HMM scores.

**HMM Topology Design.** In order to represent reduction phenomena that are so problematic in spontaneous speech, we are exploring automatic design of context-dependent HMM topologies, building on a maximum-likelihood variation of successive state splitting (ML-SSS) [5]. As a first step in this effort, Song Xing implemented software for translating ML-SSS models (provided by ATR) into HTK format and verified that HTK can be used with these models in recognition experiments on Japanese speech. He is currently running Switchboard training and testing with a 5k vocabulary, to establish a baseline for comparison of HTK tree clustering to ML-SSS topology design.

**Recognition of telephone speech.** In an ARPA-sponsored AASERT project coordinated with this effort, we have been developing a baseline system for recognizing word strings with natural numbers, based on a subset of the Macrophone corpus [8], for research on telephone channel compensation. In this past quarter, along with beta testing HTK V2.0, Rebecca Bates developed a bigram grammar for the baseline system for recognizing word strings with natural numbers. This addition to our system, coupled with improved acoustic models which are trained on the entire Macrophone training corpus, has reduced word error rates from 24% to 16%. Progressing towards the goal of testing different channel models, a unimodal Gaussian was found to be an appropriate initial model for the channel prior distribution to be used in MAP channel estimates (i.e. a Gaussian mixture model is unnecessary). MAP and ML channel estimation routines were incorporated into the HTK V2.0 source code and are currently being debugged and tested.

**Software Maintenance.** Again, some effort was devoted to software maintenance. In particular, effort was devoted to merging different versions of recognition decoder software by Ashvin Kannan. Rebecca Bates spend substantial effort in beta-testing and porting the HTK effort to V2.0, and Rukmini Iyer worked on porting all language modeling software to a new compiler, as well as on improving the speed of the LM training software to facilitate future research.

## 2.3 Future Goals

The goals of this project in the next quarter are:

- compare performance of non-parametric vs. polynomial trajectory models on the WSJ and Switchboard tasks;
- assess the use of the dependence tree model in recognition as an extra knowledge source in N-best rescoring;
- measure the entropy of Switchboard text using human subjects in an experiment modeled after Shannon's guessing game;
- obtain 5000 word HTK baseline on Switchboard task for comparison to ML-SSS topology design; and
- evaluate the current system and anticipated advances on the Switchboard spontaneous speech recognition task.

## References


[1] R. Iyer, M. Ostendorf and J. R. Rohlicek, "Language Modeling with Sentence-Level Mixtures," *Proc. ARPA Workshop on Human Language Technology,* March 1994, pp. 82-87.

[2] O. Ronen, J. R. Rohlicek and M. Ostendorf, "Parameter Estimation of Dependence Tree Models Using the EM Algorithm," *IEEE Signal Processing Letters,* Vol. 2, No. 8, August 1995, pp. 157-159.

[3] F. Richardson, M. Ostendorf and J. R. Rohlicek, "Lattice-based Search Strategies for Large Vocabulary Speech Recognition," *Proc. Int'l. Conf. on Acoust., Speech and Signal Proc.,* pp. 576-579, 1995.

[4] M. Ostendorf, F. Richardson, R. Iyer, A. Kannan, O. Ronen and R. Bates, "The 1994 BU NAB News Benchmark System," *Proceedings of the ARPA Workshop on Spoken Language Technology,* January 1995, pp. 139-142.

[5] H. Singer and M. Ostendorf, "Maximum Likelihood Successive State Splitting," *Proc. Int'l. Conf. on Acoust., Speech and Signal Proc.,* 1996, to appear.

[6] K.C. Chou, "A Stochastic Modeling Approach to Multiscale Signal Processing", PhD Thesis, MIT, May 1991.

[7] H. Gish and K. Ng, "A Segmental Speech Model with Applications to Word Spotting," in *Proc. Int'l. Conf. on Acoust., Speech and Signal Proc.,* 1993, pp. II-447-450.

[8] K. Taussig and J. Bernstein, "Macrophone: An American English Telephone Speech Corpus," *Proc. ARPA Spoken Language Technology Workshop,* 1994.

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: High-Order Modeling Techniques for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 April 1995 – 30 June 1995

# 3 Technical Transitions

Some technical transitions have occured because of our collaboration with BBN, where BBN has provided BU with N-best recognition hypotheses to help reduce our search and system building costs and BU has provided BBN with software as well as papers and technical reports to facilitate sharing of algorithmic improvements. In addition, BU student Rukmini Iyer worked at BBN as part of a graduate student co-op program, and she also participated in the 1995 Workshop on language modeling at Johns Hopkins University.

More generally, the results of this work are of interest to the speech research community and have been made available through timely dissemination in papers and presentations. The students trained on this grant also serve to transfer technology when they graduate.

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: High-Order Modeling Techniques for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 October 1995 – 31 December 1995

# 4 Publications and Presentations

During this reporting period, we wrote or revised several publications, as listed below:

**Refereed papers submitted but not yet published: (revised and resubmitted)**

"From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition," M. Ostendorf, V. Digalakis and O. Kimball, *IEEE Transactions on Speech and Audio Processing,* revised version resubmitted, to appear.

**Unrefereed reports and articles:**

"Design of a Speech Recognition System based on Acoustically Derived Segmental Units," M. Bacchiani, M. Ostendorf, Y. Sagisaka and K. Paliwal, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* to appear.

"A Dependence Tree Model of Phone Correlation," O. Ronen and M. Ostendorf. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* to appear.

"Maximum Likelihood Successive State Splitting," H. Singer and M. Ostendorf. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* to appear.

"Incremental Adaptation of Spectral Trajectory Models for Continuous Speech Recognition," A. Kannan, Ph.D. Thesis Proposal, December 1995.

**Books or parts thereof submitted but not yet published:**

"From HMMs to Segment Models: Stochastic Modeling for CSR," M. Ostendorf, in *Automatic Speech and Speaker Recognition - Advanced Topics,* ed. C. H. Lee and F. K. Soong, Kluwer Academic Publishers, revised version submitted, to appear 1996.

**Invited presentations:**

"New Approaches to Stochastic Modeling of Speech," J. Picone and M. Ostendorf, session overview, *Proc. of the IEEE Workshop on Speech Recognition,* Dec. 1995.

**Contributed presentations:**

"Unsupervised Learning of Non-Uniform Segmental Units for Acoustic Modeling in Speech Recognition, " M. Bacchiani, M. Ostendorf, Y. Sagisaka and K. K. Paliwal, presented at the *IEEE Workshop on Speech Recognition,* Dec. 1995.

"Speaker-Independent Successive State Splitting," H. Singer and M. Ostendorf, presented at the *IEEE Workshop on Speech Recognition,* Dec. 1995.

**On-line information:**

General information about the research in the Signal Processing and Interpretation Laboratory (SPI Lab), headed by Prof. Ostendorf, is available by browsing the SPI Lab WWW home page (http://raven.bu.edu/), which includes a description of this and related projects and a publication list. Technical reports and recent theses can be obtained by anonymous ftp to raven.bu.edu (in the pub/reports directory).

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: High-Order Modeling Techniques for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 June 1995 – 31 December 1995

# 5 Team Members

- Principal Investigator: Mari Ostendorf
- Graduate students:
  - Orith Ronen, Ph.D. candidate
  - Ashvin Kannan, Ph.D. candidate
  - Rukmini Iyer, M.S. 1994, Ph.D. candidate
- Undergraduate students
  - Greg Grozdits, B.S. candidate
- Visiting researcher: Song Xing (partial support)

This project is coordinated with work funded by an ARPA AASERT award on channel modeling for speech recognition which supported graduate student Rebecca Bates.