(2)

# Segment-based Acoustic Models for Continuous Speech Recognition

## Progress Report: 1 October 93 – 31 December 93



submitted to
Office of Naval Research
and
Advanced Research Projects Administration
11 February 1994

by
Boston University
Boston, Massachusetts 02215

**Principal Investigators**

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

Dr. J. Robin Rohlicek
Division Scientist, BBN Inc.
Telephone: (617) 873-3894



**Administrative Contact**

Maureen Rogers, 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 extending previous work in segment-based modeling and by introducing a new hierarchical approach to representing intra-utterance statistical dependencies. These techniques, which have high computational costs because of the large search space associated with higher order models, are made feasible through rescoring a set of HMM-generated N-best sentence hypotheses. We expect these different modeling techniques to result in improved recognition performance over that achieved by current systems, which handle only frame-based observations and assume that these observations are independent given an underlying state sequence.

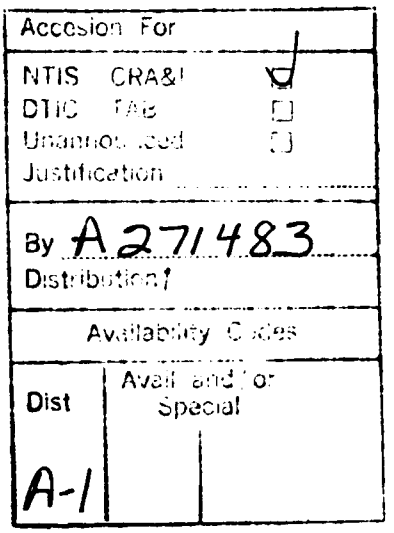
In the past quarter, the bulk of the effort on the project was centered around our participation in the ARPA speech recognition benchmark tests on the Wall Street Journal (WSJ) corpus. These efforts and other accomplishments of this project include:

- further investigation of different variations of mixture distributions, including segment-level mixtures and frame-level mixtures tied across clusters of segment regions;
- development of an algorithm for fast search of a word lattice for multi-pass recognition scoring;
- implementation of several major software changes in our recognition system to accommodate the amount of data associated with the recent release of Wall Street Journal training data;
- implementation of scoring and estimation algorithms for a sentence-level mixture language model, which was used in our recognition system for the benchmark tests;
- implementation of a duration model with parameters that are adapted according to speaking rate;
- participation in the ARPA Wall Street Journal 1993 Benchmark tests, where we achieved good performance on the hub test sets (14.3% word error on the 64k vocabulary test and 5.4% word error on the 5k vocabulary test for our best system).

On the November 1993 ARPA benchmark speech recognition tests, we demonstrated reductions of 10-20% in word error rate over baseline HMM performance using the SSM in the N-best rescoring formalism. In addition, only 3 of 8 sites achieved significantly better performance than our combined BBN-BU HMM-SSM system on the 64k vocabulary hub test set.

# 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:** Segment-Based Acoustic Models for Continuous Speech Recognition
**Grant or Contract Number:** ONR-N00014-92-J-1778
**Reporting Period:** 1 October 1993 - 31 December 1993

# 1 Productivity Measures

- Refereed papers submitted but not yet published: 0
- Refereed papers published: 1
- Unrefereed reports and articles: 0
- Books or parts thereof submitted but not yet published: 0
- Books or parts thereof published: 0
- Patents filed but not yet granted: 0
- Patents granted (include software copyrights): 0
- Invited presentations: 0
- Contributed presentations: 1
- Honors received:
  Dr. J. R. Rohlicek was elected to the Speech Processing Committee of the IEEE Signal Processing Society.
- Prizes or awards received: none
- Promotions obtained: none
- Graduate students supported $\geq$ 25% of full time: 4
- Post-docs supported $\geq$ 25% of full time: 0
- Minorities supported: 1 woman

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: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 October 1993 - 31 December 1993

# 2 Summary of Technical Progress

## Introduction and Background

In this work, we are interested in the problem of large vocabulary, speaker-independent continuous speech recognition, and primarily in the acoustic modeling component of this problem. In developing acoustic models for speech recognition, we have conflicting goals. On one hand, the models should be robust to inter- and intra-speaker variability, to the use of a different vocabulary in recognition than in training, and to the effects of moderately noisy environments. In order to accomplish this, we need to model gross features and global trends. On the other hand, the models must be sensitive and detailed enough to detect fine acoustic differences between similar words in a large vocabulary task. To answer these opposing demands requires improvements in acoustic modeling at several levels: the frame level (e.g. signal processing), the phoneme level (e.g. modeling feature dynamics), and the utterance level (e.g. defining a structural context for representing the intra-utterance dependence across phonemes). This project addresses the problem of acoustic modeling, specifically focusing on modeling at the segment level and above. The research strategy includes three main thrusts. First, phone-level acoustic modeling is based on the stochastic segment model (SSM) [1, 2], and in this area our main efforts involve developing new techniques for robust context modeling, mechanisms for effectively incorporating segmental features, and models of within-segment dependence of frame-based features. Second, high-level models are being explored in order to capture speaker-dependent and session-dependent effects within the context of a speaker-independent model. In particular, we are investigating hierarchical structures for representing the intra-utterance dependency of phonetic models, and more recently language models for representing topic dependency and language dynamics, recognizing that higher-order models of correlation can extend to the language domain as well as the acoustic domain. Lastly, speech recognition is implemented under the N-best rescoring paradigm [3], in which the BBN Byblos system is used to constrain the SSM search space by providing the top N sentence hypotheses. This paradigm facilitates research on high-order models through reducing development costs, and provides a modular framework for technology transfer that has enabled us to advance state-of-the-art recognition performance through collaboration with BBN.

## Summary of Technical Results

In the first year of this project, we have focused on improving the performance of the basic segment word recognition system and porting the system to the Wall Street Journal task domain. In brief, the accomplishments of that period included: improvements to the N-Best rescoring weight estimation algorithm; investigation of different mechanisms for improving the baseline acoustic model, including distribution clustering [4], mixture modeling at different time scales [5, 6], theoretically consistent models based on context-dependent posterior distributions, automatic distribution mapping estimation, and hierarchical models of intra-utterance phoneme dependence; implementation of baseline n-gram and sentence-level mixture language models; and improvements to the SSM rescoring mechanism such as allowing optional silence insertion.

The research efforts during this quarter, supported in part by an ONR AASERT award, have primarily involved software development on the BU recognition system to handle new issues raised by the recently released Wall Street Journal training and development test corpora, as well as substantial effort directed toward training and evaluating the system in the 1993 ARPA benchmark tests. These efforts and other research developments are summarized below.

**Mixture distributions in the SSM.** We continued investigation of mixture distribution modeling at both the segment and frame levels, shifting our focus primarily to "untied" mixtures at the segment level. (By "untied," we mean that the component distributions in the mixtures are not shared across all models, though they may be shared by clusters of models.) We have investigated performance of these systems on the male subset of the Resource Management (RM) task. Using context-independent (CI) models, we established baseline results for this system and explored some issues of parameter allocation. We achieved good results with segmental mixtures using up to 64 mixture components with diagonal covariance Gaussian distributions. In addition, we were able to combine frame-level Gaussian mixtures with segment-level mixtures to improve performance over the single-Gaussian case. Performance for the best of the above CI models is comparable to or better than the best frame-level tied mixture system; however, experiments with context-dependent (CD) segment-level mixture models currently yield worse performance than the best CD frame-level tied-mixture system. To improve CD segment-level mixture performance, we are investigating alternative initialization procedures, including the use of divisive clustering of segment-level distributions and allocating the number of components in a mixture in proportion to the amount of training for a model. We expect that these methods will improve both the segment-level and frame-level untied mixture models.

**Lattice search algorithms for multi-pass recognition scoring.** In tests we ran in the process of system development, we found that the SSM score was weighted more heavily than the HMM score, and that our overall performance improved when we rescored more hypotheses. These results

encouraged us to consider word lattice rescoring, which would enable us to more efficiently consider many sentence hypotheses. In addition, it will be an efficient rescoring framework for the more complex models that we hope to move to. Unlike other word lattice rescoring algorithms that use previous search passes to define a limited grammar [7], we plan to use time-stamped lattices to enable greater computation reduction, following a paradigm similar to that we have used in N-best rescoring. Within the lattice rescoring framework, we have developed a local search algorithm that leverages our earlier work in phoneme recognition search [8]. We plan to implement both optimal and local lattice search algorithms, and assess the performance/speed trade-offs experimentally.

**Adaptive segment duration models.** An advantage of the segment model is that it allows for explicit segment duration modeling, and an advantage of the N-best rescoring framework is that earlier recognition hypotheses can be used to estimate speaking rate or pre-pausal position, factors that can have a big effect on segment duration. We have made use of these two advantages of our recognition framework in an adaptive duration model. The model (developed under ARPA-NSF grant IRI-8905249) uses Gamma distributions with parameters that depend on the phone identity and pre-pausal position, and then adapts those parameters according to a maximum likelihood estimate of speaking rate from the HMM 1-best hypothesis [9]. Under the ARPA-ONR project, this model was ported to our WSJ recognition system. Unlike in our earlier experiments on the RM task, we did not see any improvement in performance over the static relative frequency duration distributions. We conjecture that this is due to the fact that the large WSJ training set made it possible to benefit from the greater number of free parameters in the relative frequency model. Our next step will be to port the more detailed clustered duration models also developed in [9], which we hope will use the WSJ training data to better advantage.

**Mixture language modeling.** One of the important questions in language modeling today is how to effectively represent the long-term structure of language, i.e. how to capture dependence over longer sequences of words than can be modeled with a simple n-gram. To address this problem, we developed a sentence-level mixture language model (LM) that represents the topic-dependent structure of language with separate n-gram language model mixture components. We investigated a few different numbers of LM mixture components, using automatic agglomerative clustering based on content word similarity to provide the initial LM clusters. The LM component models were then estimated by iteratively assigning paragraphs (or sentences) to the most likely cluster and re-estimating the cluster n-gram statistics, and finally smoothing the cluster n-grams with a general model using deleted interpolation. To reduce computation, the initial estimation algorithm did not use a full expectation-maximization (EM) training, but we plan to implement this in the future. We found a reduction in perplexity due to mixture modeling, and therefore used a 5-component mixture LM in the ARPA benchmark evaluations. We are currently assessing recognition performance gain for these models, which was not explicitly measured previously given computer resource limitations

before the benchmark tests.

**Recognition benchmark system development.** Several modifications were made to the recognition system to meet the challenge of the November 1993 WSJ benchmark tests. Various system improvements, including triphone score caching and other software enhancements to reduce paging, were introduced to speed up the recognition search so that we could increase the size of the N-best lists that we rescored from N=20 to N=100. Routines were also added to compute cepstral derivatives during recognition and training and to more efficiently store phoneme sequence label files, in order to reduce disk storage costs for the data. New silence models were introduced and optional silence insertion in rescoring was added, which gave a small improvement in performance. A more general resegmentation capability was added to our system to allow us to make use of training data from other sites and more easily change dictionaries. In the past, we have used phone boundaries to constrain the dynamic programming search and thereby reduce resegmentation costs, and in the new algorithm these boundaries are estimated from phone distributions and word boundaries. In this way, we were able to use training data segmented by SRI International, which assumed a different dictionary and phone set than that which we were using. Of course, in the process of making all these changes, several bugs were discovered and fixed.

On the November 1993 ARPA benchmark speech recognition tests, we demonstrated reductions of 10-20% in word error rate over baseline HMM performance using the SSM in the N-best rescoring formalism. In addition, only 3 of 8 sites achieved significantly better performance than our combined BBN-BU HMM-SSM system on the unconstrained vocabulary hub test set.

### Future Goals

Based on the results of the past year and our original goals for the project, we have set the following goals for the next six months: (1) implement the lattice search algorithm and assess performance/speed trade-offs; (2) further develop the hierarchical model formalism and assess the trade-offs between linear and non-linear models of dependence; (3) extend the language modeling work to handle new vocabulary items; and (4) investigate unsupervised adaptation in the WSJ task domain.

## References

[1] M. Ostendorf and S. Roukos, "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Trans. Acoustics Speech and Signal Processing, Dec. 1989.*

[2] S. Roucos, M. Ostendorf, H. Gish, and A. Derr, "Stochastic Segment Modeling Using the Estimate-Maximize Algorithm," *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 127–130, New York, New York, April 1988.

[3] M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz, J. R. Rohlicek, "Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses," *Proc. of the DARPA Workshop on Speech and Natural Language*, pp. 83-87, February 1991.

[4] A. Kannan, M. Ostendorf and J. R. Rohlicek, "Maximum Likelihood Clustering of Gaussians for Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, to appear.

[5] A. Kannan and M. Ostendorf, "A Comparison of Trajectory and Mixture Modeling in Segment-based Word Recognition," *Proc. of the Inter. Conf. on Acoust., Speech and Signal Processing*, Vol. II, pp. 327-330, April 1993.

[6] O. Kimball and M. Ostendorf, "On the Use of Tied Mixture Distributions," *Proceedings of the ARPA Workshop on Human Language Technology*, pp. 102-107, 1993.

[7] H. Murveit, J. Butzberger, V. Digalakis and M. Weintraub, " Large-Vocabulary Dictation Using SRI's Decipher Speech Recognition System: Progressive Search Techniques," *Proc. of the Inter. Conf. on Acoust., Speech and Signal Processing*, Vol. II, pp. 319-322, 1993.

[8] V. Digalakis, M. Ostendorf and J. R. Rohlicek, "Fast Search Algorithms for Phone Classification and Recognition Using Segment-Based Models," *IEEE Transactions on Signal Processing*, pp. 2885-2896, December 1992.

[9] C. Fong, *Duration Modeling for Speech Synthesis and Recognition*, Boston University M.S. Thesis, 1993.

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: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 October 1993 - 31 December 1993

# 3 Publications and Presentations

A conference presentation associated with this project, and a journal paper documenting our prior related work in recognition appeared during this period:

"ML Estimation of a Stochastic Linear System with the EM Algorithm and its Application to Speech Recognition," V. Digalakis, J. R. Rohlicek, and M. Ostendorf, *IEEE Transactions on Speech and Audio Processing*, October 1993, pp. 431-442.

"Beyond HMMS: A Unified View of Stochastic Modeling," M. Ostendorf, presented at the December 1993 IEEE Workshop on Speech Recognition.

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: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 October 1993 - 31 December 1993

# 4 Transitions and DoD Interactions

This grant includes a subcontract to BBN, and the research results and software is available to them. Thus far, we have collaborated with BBN by combining the Byblos system with the SSM in N-Best sentence rescoring to obtain improved recognition performance, and we have provided BBN with papers and technical reports to facilitate sharing of algorithmic improvements. On their part, BBN has been very helpful to us in our WSJ porting efforts, providing us with WSJ data and consulting on format changes.

The recognition system that has been developed under the support of this grant and of a joint NSF-ARPA grant (NSF # IRI-8902 24) is currently being used for automatically obtaining good quality phonetic alignments for a corpus of radio news speech under development at Boston University. The alignment effort is supported by the Linguistic Data Consortium, through a grant that allowed us to add cross-word phonological rules to the segmentation software.

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: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 October 1993 - 31 December 1993

# 5 Software and Hardware Prototypes

Our research has required the development and refinement of software systems for parameter estimation and recognition search, which are implemented in C or C++ and run on Sun Sparc workstations. No commercialization is planned at this time.