

**AD-A259 780**



2

**Segment-based Acoustic Models  
for Continuous Speech Recognition**

**Progress Report: July - December 1992**

submitted to  
Office of Naval Research  
and  
Defense Advanced Research Projects Administration  
22 December 1992

by  
Boston University  
Boston, Massachusetts 02215

**S** **DTIC** **D**  
**ELECTE**  
DEC 29 1992  
**A**

**Principal Investigators**

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

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

**Administrative Contact**

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

This document has been approved  
for public release and sale; its  
distribution is unlimited.

**92-32841**



2188

92 12 28 010

## Executive Summary

This research aims to develop new and more accurate acoustic 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 are more costly than traditional approaches because of the large search space associated with higher order models, are made feasible through rescoreing a set of HMM-generated N-best sentence hypotheses. We expect these different acoustic modeling methods 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 first six months of the project, in coordination with a related DARPA-NSF grant (NSF no. IRI-8902124), we have:

- Improved the N-best rescoreing paradigm by introducing score normalization and more robust weight estimation techniques.
- Investigated techniques for improving the baseline stochastic segment model (SSM) system, including context clustering for robust parameter estimation, tied mixture distribution, a two level segment/microsegment formalism, and multiple pronunciation word models.
- Extended the classification and segmentation scoring formalism to context-dependent modeling without assuming independence of observations in different segments, which opens the possibility for a broader class of features for recognition.

Our current best results represent an 18% reduction in error over the last six months; we currently report 3.95% word error on the October 1989 Resource Management test for the SSM alone, and 3.1% word error for the combined SSM-HMM system. On the recently released September 1992 test set, our performance figures are 7.3% and 6.1% word error, respectively. In addition, we see much room for further improvement, as these models still rely on an assumption of conditional independence assumption and do not take full advantage of the segment formalism.

Accession For	
NTIS CRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and/or Special
A-1	

## **Contents**

<b>1 Productivity Measures</b>	<b>4</b>
<b>2 Summary of Technical Progress</b>	<b>5</b>
<b>3 Publications and Presentations</b>	<b>9</b>
<b>4 Transitions and DoD Interactions</b>	<b>10</b>
<b>5 Software and Hardware Prototypes</b>	<b>11</b>

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 July 1992 – 31 December 1992

## 1 Productivity Measures

- Refereed papers submitted but not yet published: 0
- Refereed papers published: 0
- Unrefereed reports and articles: 2
- 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:  
Served on the IEEE Signal Processing Society Speech Technical Committee
- Prizes or awards received: 0
- Promotions obtained: 0
- Graduate students supported  $\geq 25\%$  of full time: 0
- Post-docs supported  $\geq 25\%$  of full time: 0
- Minorities supported: 0

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 July 1992 - 31 December 1992

## 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 specifically in the acoustic modeling component of this problem (as opposed to language modeling). 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. New signal processing or feature extraction techniques can provide more robust features as well as capture more acoustic detail. Advances in segment-based modeling can be used to take advantage of spectral dynamics and segment-based features in classification. Finally, a new structural context is needed to model the intra-utterance dependence across phonemes.

This project will address some of these modeling problems, specifically advances in segment-based modeling and development of a new formalism for representing inter-model dependencies. The research strategy includes three thrusts. First, speech recognition is implemented under the N-best rescoring paradigm [4], in which the BBN Byblos system is used to constrain the segment model search space by providing the top N sentence hypotheses. This paradigm facilitates research on the segment model through reducing development costs, and provides a modular framework for technology transfer that has already enabled us to advance state-of-the-art recognition performance through collaboration with BBN. Second, we are working on improved segment modeling at the phoneme level by developing new techniques for robust context modeling with Gaussian distributions, and a new stochastic formalism - classification and explicit segmentation scoring - that more effectively uses segmental features. Lastly, we plan to investigate hierarchical structures for representing the intra-utterance dependency of phonetic models in order to capture speaker-dependent and session-dependent effects within the context of a speaker-independent model.

Of the different approaches to acoustic modeling for speech recognition, statistical models have the advantage that they can be automatically trained and have yielded the best performing systems to date. We have chosen to base our work on a statistical approach, but with the goal of developing new models rather than following the traditional hidden Markov model (HMM) [1] approach. HMMs have two disadvantages that our work attempts to address: they require frame-based features and they assume that observations are conditionally independent given the Markov state sequence. (Of course, HMMs also have many advantages associated with efficient automatic training and recognition algorithms, which our work can benefit from to some extent.) The Stochastic Segment Model (SSM) [5, 6] is an alternative to the HMM for representing variable-duration phonemes. The SSM provides a joint Gaussian model for a sequence of observations. Assuming each segment generates an observation sequence  $Y = [y_1, \dots, y_L]$  of random length  $L$ , the model for a phone  $\alpha$  consists of 1) a family of joint density functions (one for every observation length), and 2) a collection of mappings that specify the particular density function for a given observation length. Typically, the model assumes that segments are described by a fixed-length sequence of locally time-invariant regions (or regions of tied distribution parameters). A deterministic mapping specifies which region corresponds to each observation vector. In research supported by NSF and DARPA, under NSF grant number IRI-8902124, we achieved improved SSM recognition performance through advances in context modeling, time-correlation modeling and speaker adaptation. In addition, we developed search algorithms that greatly reduce the complexity of recognition. Our results demonstrate the potential of segment-based models, though much remains to be taken advantage of in this formalism.

## Summary of Recent Technical Results

In the first half of Year 1, we have focused on improving the performance of the basic segment word recognition system. Through this grant and work sponsored by a related DARPA-NSF grant (NSF no. IRI-8902124), we have already accomplished many of the goals for Year 1, including:

*Improved N-Best rescoring techniques:* Early this year, we developed a grid-based search to avoid local optima in the weight optimization criterion, together with methods for choosing among different local optima to obtain more robust results [3]. More recently, we have found that normalization of scores by sentence length prior to the linear combination allows us to obtain more robust weights and has reduced our error rates by roughly 10% on the October 1989 test set.

*Developed a method for clustering contexts to provide robust context-dependent model parameter estimates:* We investigated both agglomerative and divisive clustering methods for grouping triphone labels into classes for tying covariance parameters, finding both methods work well and provide a factor of two reduction in storage and run-time memory costs. In this work, we introduced a new divisive clustering criterion based on a likelihood ratio test, which is a variant of the agglomerative measure suggested in [2].

*Extended the classification and segmentation scoring formalism:* An important step forward in building a formalism for using posterior distributions in classification is our recent development of a mechanism to handle context-dependent models without requiring the assumption of independence of features spanning different phone segments. The context-dependent model was derived using a maximum entropy criterion in estimating a combined function of posterior probability terms. This formalism will allow the use of acoustic measurements over a longer time span and facilitate hierarchical modeling. We evaluated the context-dependent model and determined that the current approach for computing segmentation scores, which is not context-dependent, needs to be extended to a more detailed model as well.

In addition to the original research plan, we have also investigated other areas for improving recognition performance, including:

*Evaluated a new time warping (distribution mapping):* In previous phone recognition research sponsored by NSF, we found that a slightly modified distribution mapping led to recognition performance improvements. Recently, we have confirmed that this warping leads to improved performance on the Resource Management word recognition task, reducing error rate on our development test set by 8%.

*Investigated the use of different phone sets and multiple-pronunciation networks:* A facility for generating multiple pronunciations, developed under NSF grant number IRI-8805680 for obtaining high quality phonetic alignments of speech, was extended to the Resource Management recognition applications. No improvements have been obtained as yet, but the algorithm for estimating robust probabilities in pronunciation networks is still under development.

*Investigated the use of tied mixture distributions:* Though many HMM recognition systems now use tied mixture distributions, the trade-offs between tied mixture and full covariance modeling had not been fully investigated. In our SSM implementation of tied mixtures at the frame level, we evaluated different covariance assumptions and training conditions and found that detailed, full-covariance models were in fact useful for this task, contrary to the results others have reported. We achieved a 10-15% reduction in word error over our previous best results on the Resource Management task.

*Extended the two level segment/microsegment formalism:* The use of two level segment models, which can be thought of as mixture distributions below the segment level but above the frame level, was previously introduced and evaluated for context-independent phone recognition. Here it has been extended for use in word recognition with context-dependent models. In evaluating the trade-offs associated with modeling trajectories vs. mixtures, we found that mixtures are more useful for context-independent modeling but representation of a trajectory is more useful for context-dependent modeling. However, these microsegment mixtures were not tied, and results from our tied mixture studies at the frame level suggest further experiments.

Our current best result is based on the tied-mixture system, which achieves 3.95% word error on the October 1989 test set (compared to 3.8% for BBN's Byblos system and 2.2% for LIMSI's HMM system, the best reported HMM results) and 7.3% word error on September 1992 test set (a respectable result for this difficult test set). Our best combined HMM-SSM result on the October 1989 test set is 3.1% word error, based on the microsegment SSM. This system has not yet been evaluated on the September 1992 test set, but with improved score normalization and the tied-mixture SSM, our combined HMM-SSM result on this data is 6.1% word error, a 13% reduction in our previous error rate.

## 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 remainder of Year 1: (1) continue system developments in multiple pronunciation networks and segmentation scoring; (2) move to a new recognition task, either the DARPA ATIS or 5000-word Wall Street Journal tasks; and (3) focus on development of the hierarchical model formalism, and implementation of robust training algorithms.

## References

- [1] L.R. Bahl, F. Jelinek, and R.L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, PAMI-5(2):179-190, March 1983.
- [2] H. Gish, M. Siu, R. Rohlicek, "Segregation of Speakers for Speech Recognition and Speaker Identification", *Proceedings IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 873-876, May 1991.
- [3] A. Kannan, M. Ostendorf, J. R. Rohlicek, "Weight Estimation for N-Best Rescoring," *Proc. DARPA Speech and Natural Language Workshop*, pp. 455-456, February 1992.
- [4] 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.
- [5] M. Ostendorf and S. Roukos, "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Trans. Acoustics Speech and Signal Processing*, Dec. 1989.
- [6] S. Roukos, 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.

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 July 1992 - 31 December 1992

### 3 Publications and Presentations

Two conference papers were written during the first half of Year 1, as listed below. Copies of these papers are included with the report.

- "Continuous Word Recognition Based on the Stochastic Segment Model," M. Ostendorf, A. Kannan, O. Kimball and J. R. Rohlicek, *Proceedings of the 1992 DARPA Workshop on Continuous Speech Recognition*, to appear. (This work was presented at the conference by John Makhoul from BBN, since the Principal Investigators of this grant were unable to attend the meeting.)
- "A Comparison of Trajectory and Mixture Modeling in Segment-based Word Recognition," A. Kannan and M. Ostendorf, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, to appear April 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 July 1992 - 31 December 1992

## 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 made our improvements in weight estimation for score combination available to BBN, which will be useful for their work in segmental neural network rescoring.

The recognition system that has been developed under the support of this grant and of a joint NSF-DARPA grant (NSF # IRI-8902124) has been used for automatically obtaining good quality phonetic alignments for a corpus of radio news speech under development at Boston University in collaboration with researchers at SRI International and MIT. The subset of the corpus that has been phonetically aligned has been given to Colin Wightman at the New Mexico Institute of Mining and Technology, and others have expressed interest in obtaining the data. We also have plans to request support from the Linguistic Data Consortium to use this software to phonetically align the remainder of the corpus.

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 July 1992 - 31 December 1992

## 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.

# Continuous Word Recognition Based on the Stochastic Segment Model\*

Mari Ostendorf, Ashvin Kannan, Owen Kimball, J. Robin Rohlicek †

Boston University  
44 Cummington St.  
Boston, MA 02215

† BBN Inc.  
10 Moulton St.  
Cambridge, MA 02138

## ABSTRACT

This paper presents an overview of the Boston University continuous word recognition system, which is based on the Stochastic Segment Model (SSM). The key components of the system described here include: a segment-based acoustic model that uses a family of Gaussian distributions to characterize variable length segments; a divisive clustering technique for estimating robust context-dependent models; and recognition using the N-best rescoring formalism, which also provides a mechanism for combining different knowledge sources (e.g. SSM and HMM scores). Results are reported for the speaker-independent portion of the Resource Management Corpus, for both the SSM system and a combined BU-SSM/BBN-HMM system.

## 1. INTRODUCTION

In the last decade, most of the research on continuous speech recognition has focused on different variations of hidden Markov models (HMMs), and the various efforts have led to significant improvements in recognition performance. However, some researchers have begun to suggest that new recognition technology is needed to dramatically improve the state-of-the-art beyond the current level, either as an alternative to HMMs or as an additional post-processing step. One such alternative that has shown promise is the stochastic segment model (SSM). The SSM has some of the advantages of the HMM, including the existence of well understood training and recognition algorithms based on statistical methods, and the SSM can borrow from many of the gains achieved by HMMs. However, the SSM has the additional advantage that it can accommodate more general features sets and less restrictive probabilistic assumptions.

In this paper, we will overview a continuous word recognition system based on the SSM, which serves as a test-bed for further development of this acoustic modeling formalism. We begin by introducing the general formalism for modeling variable-length segments with a stochastic model, and describing the specific assump-

tions currently implemented and used in the September 1992 evaluation. Next, we describe our current approach to modeling context-dependent variation, a recent advance in the system based on divisive clustering. We then review the N-best rescoring formalism for recognition, together with our current approach for estimating the weights for score combinations. Finally, we present experimental results in speaker-independent word recognition on the Resource Management task, and conclude with a summary of the key features of the system and a discussion of possible future developments.

## 2. GENERAL SSM DESCRIPTION

The Stochastic Segment Model (SSM) [1, 2] is an alternative to the Hidden Markov Model (HMM) for representing variable-duration phonemes. The SSM provides a joint Gaussian model for a sequence of observations. Assuming each segment generates an observation sequence  $Y = [y_1, \dots, y_L]$  of random length  $L$ , the model for a phone  $\alpha$  consists of 1) a family of joint density functions (one for every observation length), and 2) a collection of mappings that specify the particular density function for a given observation length. Typically, the model assumes that segments are described by a fixed-length sequence of locally time-invariant regions (or regions of tied distribution parameters). A deterministic mapping specifies which region corresponds to each observation vector.

The specific version used here assumes that frames within a segment are conditionally independent given the segment length. In this case, the probability of a segment given phone  $\alpha$  is the product of the probability of each observation  $y_i$  and the probability of its (known) duration  $L$ :

$$p(Y|\alpha) = p(Y, L|\alpha) = p(L|\alpha) \prod_{i=1}^L p(y_i|\alpha, T_L(i)),$$

where the distribution used corresponds to region  $T_L(i)$ . The distributions associated with a region  $j$ ,  $p(y|\alpha, j)$ , are multivariate Gaussians. The phone length distribution  $p(L|\alpha)$  can be either parametric (e.g., a Gamma distribution) or non-parametric; the results reported here

\*This research was jointly funded by NSF and DARPA under NSF grant number IRI-8902124, and by DARPA and ONR under ONR grant number N00014-92-J-1778.

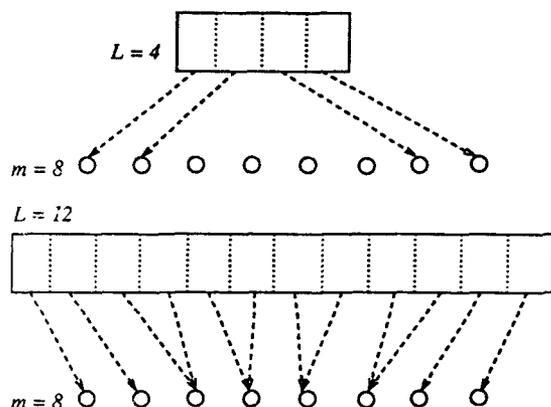


Figure 1: Illustration of mapping from observations to distribution regions for  $m = 8$  regions and  $L = 4$  and 12 frames.

are based on a non-parametric smoothed relative frequency estimate.  $T_L(i)$  determines the mapping of the  $L$ -long observation to the  $m$  regions in the model. The function  $T_L(i)$  in this work is linear in time, excluding the initial and final frames which map to the initial and final regions, as illustrated in Figure 1. This function represents a slight modification from previous work, where the warping was linear in time for the entire segment. The endpoint-constrained warping yields an 8% reduction in error over the strictly linear warping.

The segment model that uses the assumption of conditional independence (given segment length) of observations can be thought of as a hidden Markov model with a particular complex topology, or a hidden Markov model with a constrained state sequence. The conditional independence assumption has the consequence that the model does not take full advantage of the segment formalism; it captures segmental effects only in the duration distribution and the length-dependent distribution mapping. However, it has been useful for exploring issues associated with robust context modeling and word recognition system implementation, which will facilitate incorporation of acoustic models with less restrictive assumptions (e.g. [3]).

The parameter estimation algorithm for the SSM is an iterative procedure analogous to "Viterbi training" for HMMs, which involves iteratively finding the most likely segmentation and the maximum likelihood (ML) parameter estimates given that segmentation. Given a set of parameters, new phone segmentations for the training data are found using a dynamic programming algorithm to maximize the probability of the known word sequence.

Given phone segmentations, maximum likelihood (ML) parameter estimates are computed for the mean and covariance associated with each region, using all the observation frames that mapped to that region according to  $T_L$ . In this work, where initial segmentations were provided by the BBN HMM, only a few training iterations were needed.

### 3. CONTEXT CLUSTERING

Robust context modeling is an important problem in speech recognition in general, but particularly for the segment model in that it typically requires more parameters and therefore suffers from poorly estimated models for underrepresented contexts. To obtain robust estimates for context-dependent models in the SSM, covariance parameters are tied across similar classes [4]. Simple examples of classes for tying include left-context, right-context and hand-specified linguistically motivated subsets. Recently, we have investigated the use of automatic clustering techniques to determine the classes for tying. This approach is motivated by previous work in context clustering [5, 6], but differs from other approaches in that we cluster continuous rather than discrete distributions, in the specific clustering criterion used, and in that the goal of clustering is to determine classes for covariance parameter tying.

Divisive clustering is performed independently on the observations that correspond to each region of a phone, with the goal of finding classes of triphones that can share a common covariance. More specifically, the clustering algorithm is a binary tree growing procedure that successively partitions the observations (splits a node in the tree), at each step minimizing a splitting criterion over a pre-determined set of allowable questions. The questions used here are linguistically motivated, related to features such as the place and manner of articulation of the immediate left and right neighboring phones of the triphone. To reduce computation and simplify the initial implementation, we use only questions relating to individual features; that is, neither compound questions nor linear combinations of features are used.

An important aspect of divisive clustering is the node splitting criterion. As we wish to cluster together data which can be described with a common Gaussian distribution, we evaluate a two-way partition of data in a node according to a likelihood ratio test along the lines of [7] to choose between one of two hypotheses:

- $H_0$ : the observations were generated from two different distributions (that represent the distributions of the child nodes), and
- $H_1$ : the observations were generated from one dis-

tribution (that represents the distribution of the parent node).

Define a generalized likelihood ratio,  $\lambda$ , as the ratio of the likelihood of the observations being generated from one distribution ( $H_1$ ) to the likelihood of the observations in the partition being generated from two different distributions ( $H_0$ ). For Gaussians [7],  $\lambda$  can be expressed as a product of the quantities  $\lambda_{COV}$  and  $\lambda_{MEAN}$ , where both these terms can be expressed in terms of the sufficient statistics of the Gaussians.  $\lambda_{MEAN}$  depends on the means of the distributions while  $\lambda_{COV}$  depends on their covariances. Since the purpose of clustering is only to obtain better covariance estimates (the triphone means are used directly in recognition), we use only the  $\lambda_{COV}$  factor in the splitting criterion. We define the reduction in distortion due to the partition as  $-\log \lambda_{COV}$ :

$$-\log \lambda_{COV} = \frac{n}{2} \cdot [\log |W| - \alpha \log |\hat{\Sigma}_l| - (1 - \alpha) \log |\hat{\Sigma}_r|],$$

where  $n_l$  and  $n_r$  are the number of observations in the left and right child nodes with  $n = n_l + n_r$ ,  $\hat{\Sigma}_l$  and  $\hat{\Sigma}_r$  are the maximum-likelihood estimates for the covariances given the observations associated with the left and right nodes,  $\alpha = \frac{n_l}{n}$ , and  $W$  is the frequency weighted tied covariance, viz.,  $W = \frac{n_l}{n} \hat{\Sigma}_l + \frac{n_r}{n} \hat{\Sigma}_r$ . We evaluate this quantity for all binary partitions allowed by the question set and over all terminal nodes, and then split the terminal node with the question that results in the largest reduction in distortion [8].

For the context clustering tree, it is assumed that all valid terminal nodes must have more than  $T_c$  observations, where  $T_c$  is an empirically determined threshold to indicate that a reliable covariance can be estimated for that node (we use  $T_c = 250$ , for vector dimension 29). The tree is grown in a greedy manner until no more splits are possible that result in valid child nodes. When the tree is grown, each terminal node has a set of observations associated with it that map to a set of triphone distributions. The partition of observations directly implies a partition of triphones, since the allowable questions refer to the left and right neighboring phone labels. Each node is associated with a covariance, which is an unbiased estimate of the tied covariance for the constituent distributions computed by taking a weighted average of the separate triphone-dependent covariances. During recognition, all distributions associated with a terminal node share this covariance.

Experimental results indicate that context clustering results in a slight improvement in performance over covariance tying classes given simply by the left and right phone labels, while at the same time reducing the num-

ber of covariance parameters (and storage costs) by a factor of two.

## 4. N-BEST RESCORING FORMALISM

In [9], we introduced a general formalism for integrating different speech recognition methodologies using the N-best rescoring formalism. The rescoring formalism is reviewed below, followed by a description of the estimation procedure for the score combination parameters.

### 4.1. N-best Rescoring in Recognition

Under the N-best rescoring paradigm, a recognition system produces the N-best hypotheses for an utterance which are subsequently rescored by other (often more complex) knowledge sources. The different scores are combined to rerank the hypotheses. This paradigm offers a simple mechanism to integrate very different types of knowledge sources and has the potential of achieving better performance than that of any individual knowledge source [9]. In addition, the rescoring formalism provides a lower cost mechanism for evaluating word recognition performance of the SSM by itself, through simply ignoring the scores of the HMM in reranking the sentences.

Although the scores from more than two systems can be combined using this methodology, we consider only two systems here. The BBN Byblos system was used to generate the N-best hypotheses, and the Boston University SSM system was used to rescore the N hypotheses. The BBN Byblos system [10, 11] is a high performance HMM system that uses context-dependent models including cross-word-boundary contexts. The HMM observation densities are modeled by tied Gaussian mixtures.

Word recognition by the SSM is performed by rescoring the candidate word sequences for each sentence hypothesis, given a phone/word segmentation from the HMM. A phone network for the constrained SSM search is created by concatenating word pronunciation networks and then expanding the entire network to accommodate triphone models, so triphone context is modeled across word boundaries without distinguishing between cross-word and non-cross-word contexts. A dynamic programming search through this network provides the optimum SSM phone sequence and segmentation, and the desired new score. The segmentation is constrained to be within  $\pm 10$  frames (100 ms) of the original HMM segmentation, allowing for insertion and deletion of phones associated with alternate pronunciations. The 10 frame constraint was chosen to significantly reduce computation,

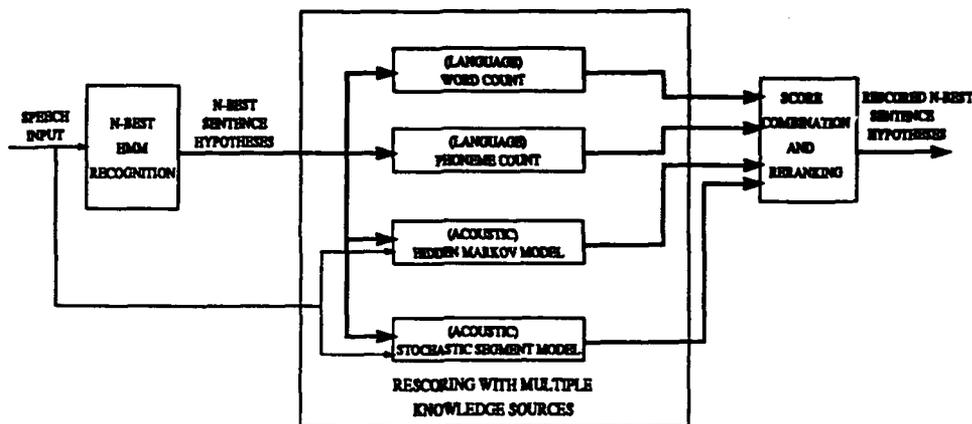


Figure 2: The N-best rescoring formalism, illustrated with the knowledge sources used in this work.

without affecting recognition performance. In addition, phoneme-dependent minimum and maximum segment lengths constrain the possible segmentations.

Once the N-best list is rescored by the different knowledge sources (such as the SSM), it is reordered according to a combination of the scores from the different knowledge sources. In this work, we use a linear combination of "scores", specifically the SSM log acoustic probability, the number of words in the sentence (insertion penalty), the number of phones in the sentence, and optionally, the HMM log acoustic probability.

## 4.2. Score Combination

N-best rescoring requires estimation of the weights used in the score combination. Different optimization criteria may be useful for finding the weights, depending upon the application. For recognition, where the goal is to minimize word error, the optimization criterion for score combination minimizes average word error in the top ranking hypothesis.<sup>1</sup> Estimation of the weights is an unconstrained multi-dimensional minimization problem, that we initially [9] approached using Powell's method. However, we noticed that optimization was sensitive to the large number of local minima in the error function, and therefore introduced an alternative procedure [12], reviewed below.

We begin by evaluating the error function at a large number of points in the weight-space, specifically, on a multi-dimensional lattice spanning the range of probable weights to determine the set of weights that results in the best performance for the test set used for weight

<sup>1</sup>For speech understanding applications where natural language processing may take the top N sentences in order of their rank, the generalized mean of the rank of the correct sentence (proposed in [9]) is a more appropriate optimization criterion.

training. Note that the error function is piece-wise constant over the weight space; a particular ranking of the hypotheses corresponds to a region (cell) in weight space defined by a set of inequalities that describe a polytope. In the hope of obtaining a more robust estimate, we find an approximate center for each of the lowest error cells and choose the cell with the largest "volume". The "center" of a cell is found by: 1) measuring the amount of slack for the different coefficients along the coordinate axes such that the weight remains within the cell, 2) computing a new "center" that is the midpoint defined by the slacks, 3) moving to the new "center" and iterating this procedure a few times. The product of the slacks in the different coordinate directions at the final "center" is an approximate indicator of the "volume" of the cell. Weights which correspond to the final "center" of the chosen cell are used for combining scores in the test set.

We use the February 89 and October 89 speaker-independent (SI) test sets to estimate weights that were used to combine scores for the evaluation test set (September 92). As the error function for male speakers differs significantly from female speakers, we estimate gender-dependent weights. In [12], where we studied error function mismatch for different test sets, we recommended weight estimation on a large number of speakers for robust estimates. Therefore, we trained weights on two test sets (February 89 and October 89) for this evaluation. As we shall see from the experimental results, test set mismatch is still somewhat of a limitation.

## 5. RM EXPERIMENTS

Results are reported on the speaker-independent Resource Management task, which has a vocabulary of 991 words. The SSM models are trained on the SI-109, 3990

Wt. Training	Test Set	SSM	HMM	SSM+HMM
Feb 89	Oct 89	4.4	3.8	3.3
Feb, Oct 89	Sep 92	8.5	6.7*	7.0
Sep 92	Sep 92	7.7	—	5.9

Table 1: Performance for word-pair grammar case (in average word error percentage). \* indicates that weights were trained only on the Feb 89 set.

utterance SI training set. The training was partitioned to obtain gender-dependent models; the specific gender used by the SSM in recognition was determined by the BBN system for detecting gender. The recognition dictionary is the standard lexicon, with a small number of words having multiple pronunciations.

The BU SSM system uses frame-based observations of spectral features, including 14 mel-warped cepstra and their first differences, plus the first difference of log energy. The segment model uses a sequence of  $m = 8$  multivariate (full) Gaussian distributions, assuming frames are conditionally independent given the segment length.

In our experiments, we use  $N = 20$  for the N-best list. The correct sentence is included in this list about 98% of the time by the Byblos system, under the word-pair grammar condition. The SSM uses no grammatical information other than the constraints imposed by the BBN N-best hypotheses. The Byblos system uses either the no-grammar condition or the standard RM word-pair grammar for the N-best list generation.

Performance of our system on the October 89 development test set and the September 92 evaluation test set for the word-pair grammar and no grammar case is shown in Table 1 and Table 2 respectively. The results represent the average word error rate in the top ranking hypothesis. The "SSM" system is the BU-SSM system while the "SSM+HMM" system uses the HMM scores of the Byblos system in the score combination also. The "HMM" system alone includes HMM rescoring to address approximations made in the N-best search and to simplify the use of cross-word models in the HMM.

The results for the October 89 test set (Table 1) clearly show performance gains associated with combining the HMM and the SSM, and this result is among the best reported. However, there was actually a degradation in performance in combining the two systems for the September 92 test set using the word-pair grammar, in contrast to our results on other test sets. To see if this was due to weight mismatch, we optimized weights on the September 92 test sets to see the best possible perfor-

Wt. Training	Test Set	SSM	HMM	SSM+HMM
Feb 89	Oct 89	19.2	—	17.5
Feb, Oct 89	Sep 92	24.5	23.3*	22.3
Sep 92	Sep 92	23.5	—	21.7

Table 2: Performance for no grammar case (in average word error percentage). \* indicates that weights were trained only on the Feb 89 set.

mance of our system. These numbers (last row in table) show that degradation in performance is due in part to weight mismatch. However, our results, like those of others, suggest that this evaluation test set is indeed very different from the two test sets that we have used to develop our system.

## 6. CONCLUSIONS

In summary, we have described the Boston University continuous speech recognition system and presented experimental results on the Resource Management task. The main features of the system include the use of segment-based acoustic models, specifically the SSM and the N-best rescoring formalism for recognition. The recent developments incorporated in this version of the system, include a new distribution mapping (time warping function), the use of divisive clustering for robust and efficient context modeling, and a more robust weight estimation technique.

Our previous experimental results on the speaker-independent Resource Management corpus yielded much lower error rates than we observed for the September 92 test set, both for the SSM system and the combined HMM-SSM system. In assessing the results of the different participating systems and listening to the speech in the September 92 test set, we feel that the system result could be improved by robust modeling of pronunciation variation. Other system improvements that we hope to pursue include extension of the clustering algorithm to accommodate more complex questions and a bigger window of context, assessment of the benefits of shared mixture distributions, and more effective use of the segmental framework either through time correlation modeling [3] and/or segmental features in a classification/segmentation framework [13], and possibly unsupervised adaptation.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge BBN, especially George Zavaliagos, for their help in providing the N best sentence hypotheses. We also thank John Makhoul

for presenting this work at the September 1992 DARPA Continuous Speech Recognition Workshop. Finally, we thank Fred Richardson of Boston University for his help in system development.

## References

1. M. Ostendorf and S. Roukos, "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, pp. 1857-1869, December 1989.
2. S. Roukos, M. Ostendorf, H. Gish and A. Derr, "Stochastic Segment Modeling Using the Estimate-Maximize Algorithm," *Proc. of the Inter. Conf. on Acoust., Speech and Signal Proc.*, pp 127-130, April 1988.
3. V. Digalakis, J. R. Rohlicek, M. Ostendorf, "A Dynamical System Approach to Continuous Speech Recognition," *Proc. of the Inter. Conf. on Acoust., Speech and Signal Proc.*, pp. 289-292, May 1991.
4. O. Kimball, M. Ostendorf and I. Bechwati, "Context Modeling with the Stochastic Segment Model," *IEEE Trans. Signal Processing*, Vol. ASSP-40(6), pp. 1584-1587, June 1992.
5. K.-F. Lee, S. Hayamizu, H.-W. Hon, C. Huang, J. Swartz and R. Weide, "Allophone Clustering for Continuous Speech Recognition," *Proc. of the Inter. Conf. on Acoust., Speech and Signal Proc.*, pp. 749-752, April 1990.
6. L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo and M. A. Picheny, "Context Dependent Modeling of Phones in Continuous Speech Using Decision Trees," *Proc. DARPA Speech and Natural Language Workshop*, pp. 264-269, February 1991.
7. H. Gish, M. Siu, R. Rohlicek, "Segregation of Speakers for Speech Recognition and Speaker Identification", *Proc. of the Inter. Conf. on Acoust., Speech and Signal Proc.*, pp. 873-876, May 1991.
8. A. Kannan, "Robust Estimation of Stochastic Segment Models for Word Recognition", Boston University MS Thesis, 1992.
9. 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. DARPA Speech and Natural Language Workshop*, pp. 83-87, February 1991.
10. F. Kubala, S. Austin, C. Barry, J. Makhoul, P. Placeway, R. Schwartz, "BYBLOS Speech Recognition Benchmark Results," *Proc. DARPA Speech and Natural Language Workshop*, pp. 77-82, February 1991.
11. R. Schwartz, and S. Austin, "Efficient, High Performance Algorithms for N-Best Search", *Proc. DARPA Speech and Natural Language Workshop*, pp. 6-11, June 1990.
12. A. Kannan, M. Ostendorf, J. R. Rohlicek, "Weight Estimation for N-Best Rescoring," *Proc. DARPA Speech and Natural Language Workshop*, pp. 455-456, February 1992.
13. O. Kimball, M. Ostendorf and J. R. Rohlicek, "Recognition Using Classification and Segmentation Scoring," *Proc. DARPA Speech and Natural Language Workshop*, pp. 197-201, February 1992.

# A COMPARISON OF TRAJECTORY AND MIXTURE MODELING IN SEGMENT-BASED WORD RECOGNITION

Ashvin Kannan

Mari Ostendorf

Electrical, Computer and Systems Engineering  
Boston University  
Boston, MA 02215, USA

## ABSTRACT

This paper presents a mechanism for implementing mixtures at a phone-subsegment (microsegment) level for continuous word recognition based on the Stochastic Segment Model (SSM). We investigate the issues that are involved in trade-offs between trajectory and mixture modeling in segment-based word recognition. Experimental results are reported on DARPA's speaker-independent Resource Management corpus.

## 1. INTRODUCTION

In earlier work, the Stochastic Segment Model (SSM) [1, 2] has been shown to be a viable alternative to the Hidden Markov Model (HMM) for representing variable-duration phones. The SSM provides a joint Gaussian model for a sequence of observations. Assuming each segment generates an observation sequence of random length, the model for a phone consists of 1) a family of joint density functions (one for every observation length), and 2) a collection of mappings that specify the particular density function for a given observation length. Typically, the model assumes that segments are described by a fixed-length sequence of locally time-invariant regions (or regions of tied distribution parameters). A deterministic mapping specifies which region corresponds to each observation vector.

A framework has recently been proposed for modeling speech at the microsegment level (a unit smaller than a phone segment) [3], in addition to the segment and frame level. Initial experiments with context-independent (CI) phone classification suggested that microsegment models provided a significant gain over the standard SSM when both models assumed conditional independence of frames given the phone segmentation. In this paper, we modify the microsegment framework for word recognition, extend it to context-dependent (CD) modeling using mixture distributions, and investigate the trade-offs of using more distributions per microsegment (model length) versus more mixture components. We present experimental results on the Resource Management task, and conclude with

a discussion of our results and possible future work.

## 2. MICROSEGMENT FRAMEWORK

The framework consists of two levels: the upper level represented by phones and the lower level represented by microsegments (MS). Each phone-length segment is divided into a fixed number of MS-sized regions. A region is characterized by a set of MS models, each an independent-frame SSM with a fixed number of distributions (multivariate Gaussians) representing a variable-length sequence of frame-level observations. The number of distributions (or MS model length) may vary across regions but is constant for different MS models representing the same region. We use a deterministic linear warping to obtain the MS-level segmentation within a phone segment, since dynamic segmentation did not lead to improved performance [3] and is much more expensive.

The sequence of MS labels can be modeled using a variety of techniques. In [3], the sequence is modeled as a first-order Markov chain, an assumption that was also used in this work for CI models. For CD models, however, the computation was too costly given the minimal benefit over independent MS regions. Consequently, for the CD MS system, we represent only marginal probabilities of the microsegment regions, which is equivalent to a mixture distribution at the microsegment level. Thus the probability of an observed segment  $Y$  given phone  $\alpha$  is defined as:

$$p(Y|\alpha) = \prod_i \sum_{a_i} p(Y_i|a_i, \alpha) p(a_i|\alpha) \quad (1)$$

where  $Y_i$  and  $a_i$  represent observations and MS labels respectively for MS region  $i$ . The components of the MS mixture are MS models  $p(Y_i|a_i)$  and the probabilities  $p(a_i|\alpha)$  which serve as mixture weights. In earlier work [3], it was found that tied-mixtures (sharing the mixture components across all phones) produced poor results, so tied mixtures were not explored here.

We implemented three MS systems and compared their performance with the 8-distribution long SSM. The (3,2,3) system used three MS regions in a segment with 3 distributions in the first and last MS region and

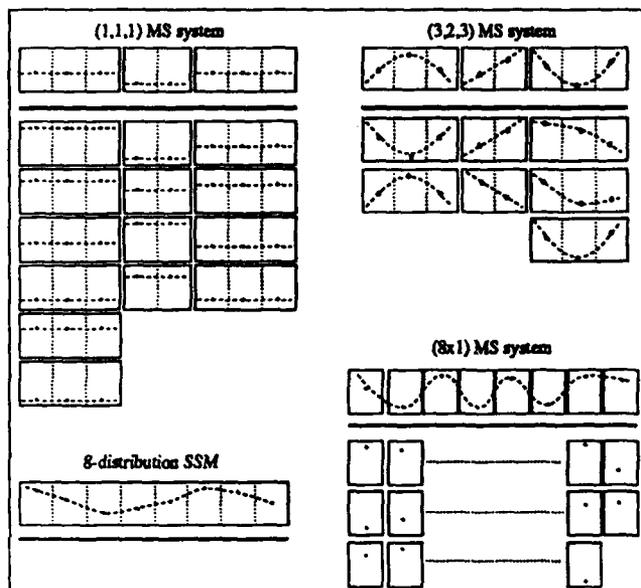


Figure 1: Trajectory assumptions (illustrated for one feature) for the SSM and MS systems. Clockwise from top-left, (1,1,1), (3,2,3), (8x1) MS systems and 8-distribution SSM. Mixture components (when present) are shown below the solid line.

2 distributions in the middle MS region. The (1,1,1) system used three regions with one distribution length each, and the (8x1) system used 8 regions each one distribution long. These systems make different assumptions about the modeling of trajectories of features of speech. The (3,2,3) system assumes that trajectories move within a region, while the (1,1,1) system assumes trajectories are fixed within a region but has more mixture components. The (8x1) system assumes no restriction on the trajectories, and has the same form as 8-distribution SSM except that the distributions are mixtures. These trajectory assumptions are schematically illustrated for one feature in Figure 1.

### 3. RECOGNITION

Implementation of the recognition search involves a dynamic programming or Viterbi search at the segment level, as for other SSM systems. For the microsegment framework, the difference from the standard SSM is the computation of the probability of a segment for a hypothesized phone label, which can be implemented either as a mixture distribution (as in Equation 1) or approximated by finding the most likely MS sequence. Both methods were investigated here.

The segment probability computation based on the dominant mixture components was investigated to reduce recognition search costs. Under this mode, the search jointly finds the most probable phone and MS sequence, replacing the probability  $p(Y|\alpha)$  by the ap-

proximation

$$p(Y|\alpha) \approx \max_A p(Y, A|\alpha) = \prod_i \max_{a_i} p(a_i|\alpha) p(Y_i|a_i, \alpha),$$

where  $A$  represents an MS label sequence for the phone  $\alpha$ . (Note that, for the Markov MS label sequence assumption,  $p(a_i|\alpha)$  is replaced by  $p(a_i|a_{i-1}, \alpha)$  and a MS-level dynamic programming search is needed.) As we allow for a variable number of microsegment components per region, choosing the dominant component of the mixture results in the grammar introducing differing penalties on phones with different numbers of mixture components. Therefore, the grammar is used in determining the best MS sequence but left out from the segment acoustic probability, i.e.,

$$p(Y|\alpha) \approx p(Y|\hat{A}, \alpha) \approx \prod_i p(Y_i|\hat{a}_i, \alpha), \quad (2)$$

and this algorithm is what is referred to here as "Viterbi" recognition. In experiments, it was observed that the grammar probabilities had no effect on recognition performance.

### 4. ESTIMATION OF MS PARAMETERS

Estimation of MS parameters involves estimating means and covariances of their associated Gaussians and the grammar probabilities for the MS units. We first describe the basic procedure and then describe extensions to context modeling.

#### 4.1. Basic procedure

Since the microsegments do not correspond to any linguistic unit, we need to automatically determine and label them in the training database. Training of MS parameters involves the following steps:

1. With the phone segmentation fixed, find initial estimates of MS models -
  - (a) Use binary divisive clustering on data to get initial means and partitions.
  - (b) Use K-means to improve partitions and define microsegments labels.
  - (c) Find maximum-likelihood estimates of mixture components with the partitions found in 1 (b).
2. Use segmental K-means to iteratively improve mixture component parameter estimates -
  - (a) Segment speech with current MS parameters.
  - (b) Find maximum-likelihood estimates of the MS parameters with the new segmentation.

These steps are described in more detail below.

## Initialization

Each MS region is initialized independently of other regions. For each  $m$ -distribution long MS region, an  $n$ -ary tree with one node for each phone is specified. Each node consists of all the observations from the training set that map to this particular phone and MS region according to the deterministic linear warping. To split a node in step 1 (a), K-means clustering with  $K=2$  is performed at the microsegment level (the mean of a cell is of dimension  $m \times k$ , where  $k$  is the dimension of the feature vector), using a Mahalanobis distance and a linear time warping to map observed frames to regions in the microsegment. A greedy-growing algorithm is used to split the node with the maximum reduction of node distortion. The reduction of node distortion is the difference between the total distortion of the parent node and the sum of the total distortions of the two child nodes, where the distortion of a node is defined as the sum of length-normalized microsegment distances from the mean.

The number of terminal nodes is constrained so that the number of free parameters are comparable across experiments. Specifically, for the CI experiments the number of terminal nodes is equal to three times the number of initial nodes, resulting in three times as many parameters as that used in the CI 8-distribution SSM. After the tree has been fully grown, K-means clustering is performed within each phone sub-tree, to obtain better estimates (Step 1(b)). The resulting clusters define the phone-dependent MS alphabet, referred to here as the CI MS alphabet. The means and covariances of the observations in the terminal nodes are the initial estimates for the CI MS models.

## Iterative segmentation/re-estimation

Once initial estimates for the MS models are available, a segmental K-means procedure is used to obtain better estimates. This involves iterating between segmenting speech into microsegments using the current MS parameters and finding new maximum-likelihood estimates for the MS models from the segmented speech.

Bigram and marginal probabilities of the MS labels ( $p(a_i|a_{i-1}, \alpha)$  and  $p(a_i|\alpha)$ , respectively) are given by the relative frequencies observed after each segmentation pass. The bigram probabilities, which are used only for experiments with the 3-region CI MS alphabet, are smoothed with the *a priori* probabilities. During recognition it was observed that the grammar score is two orders of magnitude smaller than the acoustic score of the microsegments and its exclusion does not affect recognition performance with the Viterbi search.

## 4.2. Context Modeling

Context modeling with microsegments is not practical with equivalents of "diphones" or "triphones", since the alphabet size is much larger than that for phones.

Instead we define context classes by the collection of triphones at the terminal nodes of the context tree grown using binary divisive clustering as in [4], but with the generalized likelihood ratio distance measure [5, 6].

Once we define context classes to use, we can model context using microsegments in different ways and two schemes were evaluated. First, we can retain the CI MS alphabet<sup>1</sup> and estimate models for these labels conditioned on the context classes. In this case, we estimate CD models from the MS observations that are assigned a CI label according to the training segmentation and also correspond to the specific context class. Alternatively, we can incorporate information of the context classes in the MS initialization process and obtain a CD MS alphabet. In this case, the MS tree growing procedure is modified to start with a node for each context class for each phone, with observations arising from that specific context class and that MS region. The tree is grown until we have the desired number of terminal nodes. The rest of the procedure is analogous to the estimation of CI MS acoustic models.

The current approach to estimating the CD MS alphabet results in many fewer free parameters than the context-dependent system based on the CI MS alphabet. In order to compare systems with similar numbers of free parameters, the MS tree growing algorithm was modified such that the tree is grown beyond the first-level "terminal" nodes (called "covariance nodes" and having at least 250 observations to estimate a full covariance) to a second-level set of terminal nodes ("mean nodes") based on a lower threshold, i.e. 50 observations. The mean nodes now constitute an "extended" alphabet and share the covariance of their parent covariance node.

## 5. EXPERIMENTAL CONDITIONS

Word recognition with the MS-based SSM is performed using the N-best rescoring formalism [2] on DARPA's Resource Management speaker-independent corpus with the word-pair grammar. Gender-dependent MS models are trained on the SI-109, 3990 utterance set. The systems use frame-based observations that include 14 mel-warped cepstra and their first differences, plus the first difference of log energy.

Development was performed on the February 1989 test set and results are also reported on the October 1989 test set. The experimental results for the different systems using Viterbi recognition are shown in Table 1. For the CI MS systems, we see that it is better to have more mixture components than mixtures

<sup>1</sup>For context-modeling experiments, "CI MS alphabet" refers to using the MS labels that were produced from the CI MS tree. In the strict sense, this is not really CI as during re-estimation of the models we use context-dependent variants of these labels. However, we use this nomenclature to differentiate this from the "CD MS alphabet" that is introduced later.

MS System	Average Word Error (%)		
	(8x1)	(3,2,3)	(1,1,1)
Context-independent	7.8	7.6	7.3
CD with CD MS alph.	-	6.3	6.5
CD with CI MS alph.	-	5.8	6.1

Table 1: Performance of the MS systems using Viterbi recognition on the February 89 test set. The 8-distribution SSM achieves 8.9% and 4.8% word error for CI and CD models respectively on this test set.

of sequences since the (1,1,1) system has the best CI performance. On the other hand, for CD systems, it is more important to model the trajectory, since the (3,2,3) system outperforms the (1,1,1) system. In addition, the 8-distribution CD SSM, which does not use mixtures and models the trajectories at the segment rather than the MS level has the best performance.

The initial experiments showed that the CI MS alphabet gave better performance than the CD MS alphabet. However, these systems were not comparable because of differences in the number of free parameters, so further experiments were conducted with the extended CD MS alphabet and the (3,2,3) case using a comparable number of means in both cases. The best CD alphabet system in this case had a maximum of five mean nodes per covariance node. Viterbi recognition for this system resulted in 6.1% word error for the February 89 task while mixture recognition resulted in 5.8%, which was also achieved with the CI alphabet. However, on an independent test set (October 89), the CD alphabet system performed poorly with both Viterbi and mixture recognition. Thus, we conclude that the CI alphabet gives more robust CD models.

We evaluated the best case MS systems, CI (1,1,1) system and the CD (3,2,3) system based on the CI alphabet, on the October 89 test set. The recognition performances were 7.0% and 6.0% respectively. The performance of a comparable 8-distribution SSM on this test set were 8.7% and 4.7% for CI and triphone systems respectively. (Lower error rates have been obtained with more recent system modifications.) Although the microsegment formalism does not yield performance improvements for the CD SSM, it does seem to be preferable in combination with the HMM scores from BBN's Byblos using the N-best rescoring formalism: the word error rate drops to 3.1% on the Oct89 test set from 3.4% for the 8-distribution triphone SSM. For comparison, the Byblos HMM error rate is 3.8%.

## 6. CONCLUSIONS

In summary, we have described a mechanism for implementing mixtures at a microsegment level and investigated trajectory assumptions for the acoustic modeling for continuous word recognition. Our results suggest

that there is a trade-off in using mixture models and trajectory models, associated with the level of detail of the modeling unit (e.g., CI vs. CD), although some level of trajectory constraints is useful even for CI models. The results support the use of whole segment models in the context-dependent case, and microsegment-level (and possibly segment-level) mixtures rather than frame-level mixtures.

In the "mixture" implementation of recognition, we used MS models which were not trained using a "true" mixture procedure, but with the segmentation produced by the dominant component of the best scoring mixture, i.e., with a Viterbi-style training. Performing mixture training may improve performance further. Another possible extension is to further investigate the use of tied microsegment mixtures. Although previous work suggested that tied MS mixtures were not useful, these results were based on region-dependent mixtures, which we have since found are not robust in recent experiments with frame-based mixtures in the SSM.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge BBN Inc. for their help in providing the N-best sentence hypotheses. We thank J. Robin Rohlicek of BBN and Vassilios Digalakis of SRI for useful discussions. This research was jointly funded by NSF and DARPA under NSF grant number IRI-8902124, and by DARPA and ONR under ONR grant number N00014-92-J-1778.

## REFERENCES

- [1] M. Ostendorf and S. Roukos, "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Trans. on Acoust., Speech and Signal Processing*, pp. 1857-1869, December 1989.
- [2] M. Ostendorf, A. Kannan, O. Kimball and J. R. Rohlicek, "Continuous Word Recognition Based on the Stochastic Segment Model," *Proceedings of the DARPA Workshop on Continuous Speech Recognition*, September 1992.
- [3] V. Digalakis, *Segment-Based Stochastic Models of Spectral Dynamics for Continuous Speech Recognition*, Boston University Ph.D. Dissertation, 1992.
- [4] K.-F. Lee, S. Hayamizu, H.-W. Hon, C. Huang, J. Swartz, R. Weide, "Allophone Clustering for Continuous Speech Recognition," *Proceedings IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 749-752, April 1990.
- [5] H. Gish, M. Siu, R. Rohlicek, "Segregation of Speakers for Speech Recognition and Speaker Identification", *Proceedings IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 873-876, May 1991.
- [6] A. Kannan, *Robust Estimation of Stochastic Segment Models for Word Recognition*, Boston University MS Thesis, 1992.