Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses

M. Ostendorf† *A. Kannan*† *S. Austin*‡ *O. Kimball*†
R. Schwartz‡ *J.R. Rohlicek*‡

† Boston University
44 Cummington St.
Boston, MA 02215

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

ABSTRACT

This paper describes a general formalism for integrating two or more speech recognition technologies, which could be developed at different research sites using different recognition strategies. In this formalism, one system uses the N-best search strategy to generate a list of candidate sentences; the list is rescored by other systems; and the different scores are combined to optimize performance. Specifically, we report on combining the BU system based on stochastic segment models and the BBN system based on hidden Markov models. In addition to facilitating integration of different systems, the N-best approach results in a large reduction in computation for word recognition using the stochastic segment model.

INTRODUCTION

While most successful systems to date have been based on hidden Markov models (HMMs), there may be utility in combining the HMM approach with some other very different approach. For example, the research group at Boston University is exploring the use of the Stochastic Segment Model (SSM) [9,11] as an alternative to the HMM. In contrast to the HMM, the SSM scores a phoneme as a whole entity, allowing a more detailed acoustic representation. In previous work [6], it was demonstrated that the SSM is effective in the task of phoneme recognition, with results on the TIMIT database using context-independent phoneme models that are comparable to context-dependent HMMs. Thus, there is a good possibility that, with the proper use of context, the performance may surpass that of the HMM system. Unfortunately, the computation required for the SSM is considerably greater than that for HMMs, making it impractical to implement the standard optimal dynamic programming search algorithms.

In this paper, we introduce a general formalism for integrating different speech recognition technologies, which also enables evaluation of word recognition performance with the SSM. In this approach, one recognition system uses the N-best search strategy to provide a list of sentence hypotheses. A second system (presumably more complex) is used to rescore these hypotheses, and the scores of the different systems are combined, giving a new ranking of the sentence hypotheses. If the errors made by the two systems differ, then combining the two sets of scores would yield an improvement in overall performance (either in terms of the percent of correct sentences or the average rank of the correct sentence). The N-best formalism offers a means of reducing the computation associated with combining the results of two systems by restricting the search space of the second system. It therefore also 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.

In the following section, we describe the integration methodology in more detail. Next, we present experimental results combining the stochastic segment model with the BBN Byblos system, including a result that incorporates statistical grammar scores as well as a benchmark result using the word-pair grammar. Finally, we conclude with a discussion of possible implications and extensions of this work.

INTEGRATION STRATEGY

The basic approach involves

1. computing the N best sentence hypotheses with one system;
2. rescoring this list of hypotheses with a second system; and
3. combining the scores to improve overall performance.

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. Details of each step, based on the use of these two systems, are given below.

N-Best Scoring

The idea of scoring the N best sentence hypotheses was introduced by BBN as a strategy for integration of speech and natural language [3]. Given a list of N candidate sentences, a natural language system could process the different hypotheses until reaching one that satisfied the syntactic and semantic constraints of the task. An exact, but somewhat expensive algorithm for finding the N best sentence hypotheses was also described in [3]. Since then, several sites have adopted the N-Best strategy for combining speech recognition with natural language. In addition, more efficient approximate scoring algorithms for finding the N Best sentences have been developed (e.g., [12,13]). These algorithms introduce only a short delay after finding the 1-Best hypothesis for finding the N-Best hypotheses.

This same N-best scoring paradigm can be used for the integration of different recognition techniques. The main difference is that, for the rescoring application, it is useful to have the word and/or phoneme boundaries associated with this hypothesis. Since the recognition algorithm involves maximizing the joint probability of the HMM state sequence and the observed data, the boundaries can be obtained from the traceback array typically used in decoding.

Rescoring

Rescoring the list of hypotheses is a constrained recognition task, where the phoneme and/or word sequence is given and the phonetic segmentation is optionally given. Here we use a stochastic segment model in rescoring, but any acoustic model would be useful in this formalism. (For example, a neural network model of phoneme segments is used in [1].) The constrained recognition search is particularly useful for segmental acoustic models, which have a significantly larger recognition search space than frame-based hidden Markov models.

If the phoneme segmentations are given and assumed fixed, the computation required for rescoring is extremely small. If the phoneme segmentations are not given for the N hypotheses, then rescoring is essentially automatic segmentation. The maximum likelihood segmentation is given by a dynamic programming algorithm, typically with minimum and maximum phoneme duration constraints, as in [9]. Scoring a sentence with the optimal segmentation for a model will yield better results than scoring according to the segmentation determined by a different model, but the cost in computation is significant (roughly a factor of 300 more than using fixed segmentations). Since we have found the stochastic segment model performance to be fairly sensitive to boundary location, we anticipate that optimal segmentation may be very important. A compromise strategy is to find the optimal segmentation subject to the constraint of being within a fixed number of frames of the HMM segmentation. The constrained dynamic programming solution appears to suffer no loss in performance and saves a factor of 30 in computation relative to the unconstrained algorithm.

A slight variation of the segmentation algorithm involves searching for the optimal phone sequence and its segmentation, given a word sequence. In other words, we allow alternative pronunciations in rescoring a sentence hypothesis. We hypothesize that the use of alternative pronunciations will significantly improve SSM word recognition performance, mainly because SSM phoneme recognition performance is much higher on the carefully hand-labeled TIMIT database than it is on the Resource Management Task (in which case we assume that the phone sequence assigned by the BBN single pronunciation recognizer is "correct"). However, we have not investigated this question on a dictionary with a sufficiently rich set of pronunciations. The additional cost of modeling multiple pronunciations should be relatively small.

Combining Scores

An important issue is how to combine the scores from the systems so as to optimize the performance of the overall system. In this initial work, we chose to use a linear combination of HMM log acoustic score, log grammar score, number of words in the sentence (insertion penalty), number of phonemes in the sentence, and SSM log acoustic score. This is a simple extension of the current HMM system ranking, which uses the first three of these five measures.

We estimate the set of weights that optimizes a generalized mean of the rank of the correct answer:

$$m(\mathcal{S}) = |\frac{1}{S}\sum_{i=1}^{S} r(i)^p|^{1/p} \qquad (1)$$

where $r(i)$ is the rank of the correct answer in sentence i of a set $\mathcal{S}$ of S sentences, and p determines the type of mean. For example, $p = 1$ specifies the average, $p = 2$ specifies the root-mean-square, $p = -1$ specifies the harmonic mean, and $p = -\infty$ only counts the percent correct. For speech recognition applications $p = -\infty$ would be appropriate, but for speech understanding applications, $p = 1$ might be more useful. In practice we find that the different values of p did not have a significant impact on the results.

Estimation of the weights is an unconstrained multidimensional minimization problem. The algorithm used here is Powell's method [10], which iteratively minimizes the generalized mean (Equation 1) by optimizing the weights in successive conjugate directions. Because the algorithm seemed to be sensitive to local optima, we determine the weights by trying several different initial points. This strategy gave an increase in performance.

EXPERIMENTAL RESULTS

The recognition experiments were based on the Resource Management (RM) corpus. Both the BBN Byblos system and the BU stochastic segment models were trained on the speaker-independent SI109 corpus. Both systems used feature vectors comprised of 14 mel-warped cepstral coefficients and the respective derivatives; the BBN system also used power and second derivatives of the cepstra.

The basic BBN Byblos system is essentially the same as originally described in [2]. These experiments used context-dependent but not cross-word triphone models. The models are gender-dependent; the system scores a sentence with both male and female models and then chooses the answer that gives the highest score. With few exceptions, the correct speaker sex is chosen. The Byblos system was used to generate the top 20 sentence hypotheses for each utterance. Experiments with larger numbers of hypotheses suggested that the additional rescoring computation was not warranted. This was due to the fact that, using the HMM models, the correct sentence was almost always included within the top 20 hypotheses.

Two different SSM systems were used to rescore these hypotheses: one context-independent and one using left-context phone models. In both cases gender-dependent models are used, where the speaker sex was that chosen by the BBN system. The model structure from the best case system found in previous studies [5] was used. This system is based on independent samples, frame-dependent feature transformations, and five distributions per model. Infrequently observed classes are modeled with a frame-dependent, model-independent tied covariance matrix, otherwise a model- and frame-dependent covariance matrix is used. Using more sophisticated estimation techniques, as well as generalized triphones [8], would likely yield significant improvements for context-dependent models. In addition, recent work in time correlation modeling [7] could be used to improve performance, and this will be integrated into a later version of the system.

Results for two different test sets are described below. First, we investigated different score combinations on the February 1989 RM test set. Second, we report results on the February 1991 RM benchmark test set, where the previous test set is used to estimate weights for combining the scores.

Different Score Combinations

In the first set of experiments, the N-best hypotheses were generated using the Byblos system with a fully-connected statistical bi-class grammar [4]. In this experiment, we used a grammar with 548 classes that had a perplexity of 23 on the test set. This system finds the correct sentence in the top 20 hypotheses 98% of the time. These sentences were rescored using the two different stochastic segment models. For each sentence hypothesis, the total score included the log HMM acoustic score and/or the log SSM acoustic score (either context-independent or context-dependent). In addition, all score combinations included log grammar scores, word and phoneme count. The weights for different combinations of scores were designed as described in the previous section, using the generalized mean optimization criterion with $p = -1$. Table 1 summarizes the performance of several different system combinations.

System	% sent corr	avg sent rank
CI SSM, fixed seg	56.3	2.84
CI SSM, opt seg	64.3	2.37
CD SSM, opt seg	68.0	1.86
CD HMM, N-best	71.3	1.73
CD HMM, optimized	75.7	1.75
CD HMM + CI SSM	78.8	1.68
CD HMM + CD SSM	79.3	1.56

Table 1: Percent sentence correct and average rank of correct sentence when it is in the top 20. Results are based on the Feb. 1989 test set using a statistical class grammar.

The table shows improved performance for more complex versions of the stochastic segment model. Using the fixed segmentations yields significantly lower performance for the segment model, so all further experiments use the constrained optimal segmentation. The simple left-context model results in improved performance over the context-independent model, both alone and in combination with the HMM. The HMM which uses triphone models outperforms the SSM which uses left-context models, but the performance of the two systems is close in comparing percent sentence correct in the top N for $N > 4$ (see Figure 1).

Table 1 also shows the improvement associated with the rescoring formalism. First, since the N-best search algorithm is sub-optimal, simply rescoring the hypotheses with the original HMM (referred to in the table and figure as an "optimized HMM") yields some improvement in performance. More importantly, the results show that even at the lower level of performance of the SSM, combining the HMM and SSM scores yields improvement in performance, particularly through raising the rank of the correct sentence. This is shown more clearly in Figure 1, which illustrates the cumulative distribution function of percent of sentences correct in the top N hypotheses.

As mentioned previously, this is a preliminary result, so we expect additional improvement – both for the SSM alone and the combined systems – from further research in SSM context modeling.

Benchmark Results

A second experiment involved testing performance of the scoring combinations on the February 91 benchmark test set. In this case, the 20 best sentence hypotheses were generated using the word-pair grammar. These sentences were

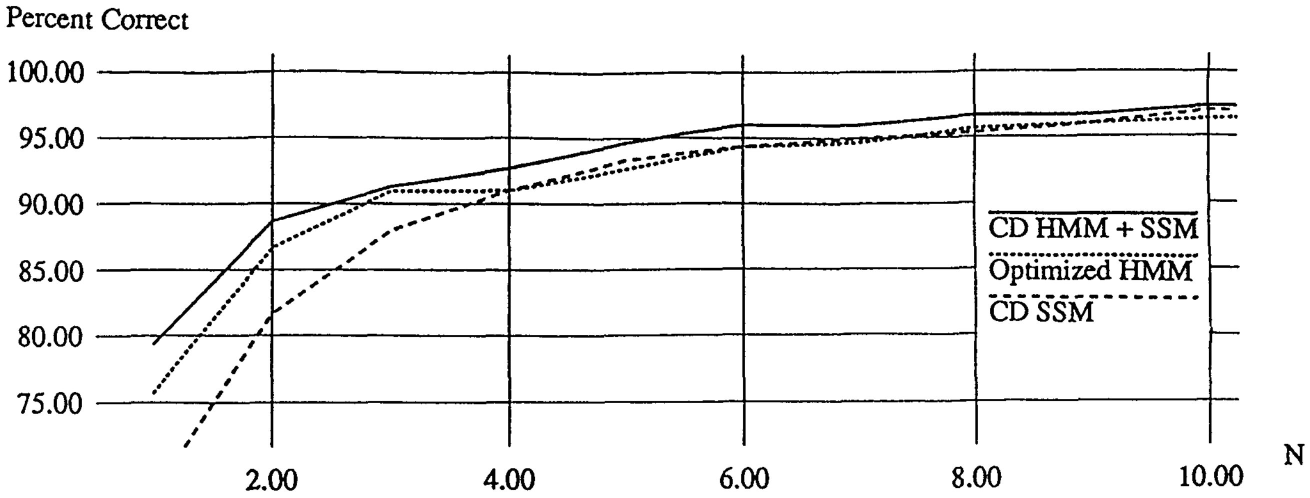

Figure 1: Cumulative distribution function of percent sentences correct in the top N hypotheses for: (a) optimized HMM, (b) context-dependent SSM, and (c) combined HMM and context-dependent SSM.

rescored using the context-independent SSM with the constrained optimal segmentation algorithm. The scores used were log HMM and SSM scores and word and phoneme counts; no grammar scores were used in this experiment. Weights were trained using the February 1989 test set. Although $p = -\infty$ would be appropriate for this task, we used $p = -1$ because of the sensitivity of the search to local optima. In Table 2, we show benchmark test results for different combinations of HMM and SSM, with performance on the February 1989 test set given for comparison. For each case, we give the percent of the sentences recognized correctly as the top choice and the average rank of the correct answer when it is in the top 20. The HMM results reported here may be lower than other results reported in this proceedings, since we are using a simpler version of the Byblos system (specifically without cross-word phonetic models). As before, we find that the context-dependent HMM is outperforming the context-independent SSM, and that rescoring yields a small improvement in performance, mainly in average sentence rank.

System	N-Best HMM	Optimal HMM	CI SSM	HMM +SSM
Avg sent rank				
Feb 89	2.13	2.15	3.07	2.11
% sent corr				
Feb 89	67.7	69.7	50.0	70.0
Feb 91	72.3	73.0	52.7	73.0
% word err				
Feb 91	5.4	5.3	9.7	5.6

Table 2: Percent sentence correct and average rank of correct sentence when it is in the top 20. Results are reported for development (Feb. 1989 test set) and benchmark (Feb. 1991 test set), using a word-pair grammar, but no grammar scores.

DISCUSSION

In summary, we have introduced a new formalism for integrating different speech recognition technologies based on generating the N best sentence hypotheses with one system, rescoring these hypotheses, and combining the scores of the different systems. This N-best rescoring formalism can be useful in several ways.

Specifically, it makes practical the implementation of a computationally expensive system such as the Stochastic Segment Model, and has allowed us to investigate the utility of the SSM for word recognition. The results reported here are the first reported on the Resource Management task for the SSM. Our initial results were much lower than would be predicted from phoneme recognition results on the TIMIT database, underscoring the need for additional system development. The rescoring formalism will facilitate further research in SSM word recognition, particularly in the utilization of recent techniques developed for time correlation modeling and context modeling. Research in context-modeling is particularly facilitated by the rescoring formalism, since the computation time is the same order of magnitude as context-independent models.

More generally, the rescoring formalism enables cross-site collaboration and fast evaluation of potential improvements in speech understanding associated with integration of different knowledge sources. It provides a simple mechanism for integrating even radically different recognition technologies, enabling higher performance than either technique alone. The results reported heıe yield some improvement in performance, but we anticipate a greater effect with future improvements to the SSM. Improvements can also be gained from further research on score combination, since the weight estimation technique was found to be very sensitive to initial starting points. In addition, scores from very different types of knowledge sources could be combined to improve the performance of a speech understanding system. For example, if scores are combined after natural language processing, it would be possible to include a score which represents the prosodic consistency of a parse [14]. This is one of many possible areas for future research.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge John Makhoul for many valuable comments. This research was jointly funded by NSF and DARPA under NSF grant number IRI-8902124.

REFERENCES

1. S. Austin, J. Makhoul, R. Schwartz and G. Zavaliagkos, "Continuous Speech Recognition Using Segmental Neural Nets," this proceedings.

2. Y. Chow *et al.*, "BYBLOS: The BBN Continuous Speech Recognition System," *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 89–92, 1987.

3. Y.-L. Chow and R. Schwartz, "The N-Best Algorithm: An Efficient Procedure for Finding Top N Sentence Hypotheses," *Proceedings of the Second DARPA Workshop on Speech and Natural Language*, pp. 199–202, October 1989.

4. A. Derr and R. Schwartz, "A Simple Statistical Class Grammar for Measuring Speech Recognition Performance," *Proceedings of the Second DARPA Workshop on Speech and Natural Language*, pp. 147–149, October 1989.

5. V. Digalakis, M. Ostendorf and J. R. Rohlicek, "Improvements in the Stochastic Segment Model for Phoneme Recognition," *Proceedings of the Second DARPA Workshop on Speech and Natural Language*, pp. 332–338, October 1989.

6. V. Digalakis, M. Ostendorf and J. R. Rohlicek, "Fast Search Algorithms for Connected Phone Recognition Using the Stochastic Segment Model," manuscript submitted to *IEEE Trans. Acoustic Speech and Signal Processing* (a shorter version appeared *Proceedings of the Third DARPA Workshop on Speech and Natural Language*, June 1990).

7. V. Digalakis, J. R. Rohlicek and M. Ostendorf, "A Dynamical System Approach to Continuous Speech Recognition," this proceedings, also to appear in the *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, May 1991.

8. K.-F. Lee, "Context-dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition," *IEEE Trans. Acoustic Speech and Signal Processing*, Vol. ASSP-38(4), pp. 599–609 , April 1990.

9. M. Ostendorf and S. Roukos, "A Stochastic Segment Model for Phoneme-based Continuous Speech Recognition," *IEEE Trans. Acoustic Speech and Signal Processing*, Vol. ASSP-37(12), pp. 1857–1869, December 1989.

10. W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, *Numerical Recipes*, Cambridge University Press, Cambridge 1986.

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

12. R. Schwartz and S. Austin, "Efficient, High Performance Algorithms for N-Best Search," *Proceedings of the Third DARPA Workshop on Speech and Natural Language*, pp. 6–11, June 1990.

13. F. K. Soong and E.-F. Huang, "A Tree-Trellis Based Fast Search for Finding the N-Best Sentence Hypotheses in Continuous Speech Recognition," *Proceedings of the Third DARPA Workshop on Speech and Natural Language*, pp. 12–19, June 1990.

14. C. W. Wightman, N. M. Veilleux and M. Ostendorf "Using Prosodic Phrasing in Syntactic Disambiguation: An Analysis-by-Synthesis Approach," this proceedings, 1991.