Recognition Using Classification and Segmentation Scoring*

Owen Kimball †, *Mari Ostendorf* †, *Robin Rohlicek* ‡

† Boston University
44 Cummington St.
Boston, MA 02215

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

ABSTRACT

Traditional statistical speech recognition systems typically make strong assumptions about the independence of observation frames and generally do not make use of segmental information. In contrast, when the segmentation is known, existing classifiers can readily accommodate segmental information in the decision process. We describe an approach to connected word recognition that allows the use of segmental information through an explicit decomposition of the recognition criterion into classification and segmentation scoring. Preliminary experiments are presented, demonstrating that the proposed framework, using fixed length sequences of cepstral feature vectors for classification of individual phonemes, performs comparably to more traditional recognition approaches that use the entire observation sequence. We expect that performance gain can be obtained using this structure with additional, more general features.

1. INTRODUCTION

Although hidden-Markov-model (HMM) based speech recognition systems have achieved very high performance, it may be possible to improve on their performance by addressing the known deficits of the HMM. Perhaps the most obvious weaknesses of the model are the reliance on frame-based feature extraction and the assumption of conditional independence of these features given an underlying state sequence. The assumption of independence disagrees with what is known of the actual speech signal, and when this framework is accepted, it is difficult to incorporate potentially useful measurements made across an entire segment of speech. Much of the linguistic knowledge of acoustic-phonetic properties of speech is most naturally expressed in such segmental measurements, and the inability to use such measurements may represent a significant loss in potential performance.

In an attempt to address this issue, a number of models have been proposed that use segmental features as the basis of recognition. Although these models allow the use of segmental measurements, they have not yet achieved significant performance gains over HMMs because of difficulties associated with modeling a variable length observation with segmental features. Many of these models represent the segmental characteristics as a fixed-dimensional vector of features derived from the variable-length observation sequence. Although such features may work quite well for classification of individual units, such as phonemes or syllables, it is less obvious how to use fixed-length features to score a sequence of these units where the number and location of the units is not known. For example, simply taking the product of independent phoneme classification probabilities using fixed length measurements is inadequate. If this is done, the total number of observations used for an utterance is $F \times N$, where F is the fixed number of features per segment and N is the number of phonemes in the hypothesized sentence. As a result, the scores for hypotheses with different numbers of phonemes will effectively be computed over different dimensional probability spaces, and as such, will not be comparable. In particular, long segments will have lower costs per frame than short segments.

In this paper, we address the segment modeling problem using an approach that decomposes the recognition process into a segment classification problem and a segmentation scoring problem. The explicit use of a classification component allows the direct use of segmental measures as well as a variety of classification techniques that are not readily accommodated with other formulations. The segmentation score component effectively normalizes the scores of different length sequences, making them comparable.

2. CLASSIFICATION AND SEGMENTATION SCORING

2.1. General Model

The goal of speech recognition systems is to find the most likely label sequence, $\mathbf{A} = a_1, ..., a_N$ given a sequence of acoustic observations, $\mathbf{X}$. For simplicity, we can restrict the problem to finding the label sequence, $\mathbf{A}$, and segmentation, $\mathbf{S} = s_1, ..., s_N$, that have the highest joint likelihood given the observations. (There is typically no

*This research was jointly funded by NSF and DARPA under NSF grant number IRI-8902124.

explicit segmentation component in the formulation for HMMs; in this case, the underlying state sequence is analogous to the segmentation-label sequence.) The required optimization is then to find labels $\mathbf{A}^*$ such that

$$\begin{aligned} \mathbf{A}^* &= \underset{\mathbf{A},\mathbf{S}}{\operatorname{argmax}}\, \mathrm{p}(\mathbf{A},\mathbf{S} \mid \mathbf{X}) \\ &= \underset{\mathbf{A},\mathbf{S}}{\operatorname{argmax}}\, \mathrm{p}(\mathbf{A},\mathbf{S},\mathbf{X}). \qquad (1) \end{aligned}$$

The usual decomposition of this probability is

$$\mathrm{p}(\mathbf{A},\mathbf{S},\mathbf{X}) = \mathrm{p}(\mathbf{X} \mid \mathbf{A},\mathbf{S})\mathrm{p}(\mathbf{S} \mid \mathbf{A})\mathrm{p}(\mathbf{A}) \qquad (2)$$

as is commonly used in HMMs and has been used in our previous segment modeling. However, we can consider an alternative decomposition:

$$\mathrm{p}(\mathbf{A},\mathbf{S},\mathbf{X}) = \mathrm{p}(\mathbf{A} \mid \mathbf{S},\mathbf{X})\mathrm{p}(\mathbf{S},\mathbf{X}).$$

In this case, the optimization problem has two components a "classification probability," $\mathrm{p}(\mathbf{A} \mid \mathbf{S},\mathbf{X})$, and a "probability of segmentation", $\mathrm{p}(\mathbf{S},\mathbf{X})$. We refer to this approach as *classification-in-recognition* (CIR).

The CIR approach has a number of potential advantages related to the use of a classification component. First, segmental features can be accommodated in this approach by constraining $\mathrm{p}(\mathbf{A} \mid \mathbf{X},\mathbf{S})$ to have the form $\mathrm{p}(\mathbf{A} \mid f(\mathbf{X}),\mathbf{S})$, where $f(\mathbf{X})$ is some function of the original observations. The possibilities for this function include the complete observation sequence itself, as well as fixed dimensional segmental feature vectors computed from it. A second advantage is that a number of different classifiers can be used to compute the posterior probability, including neural networks and classification trees, as well as other approaches.

To simplify initial experiments, we have made the assumption that phoneme segments are generated independently. In this case (1) is rewritten as

$$\mathbf{A}^* = \underset{\mathbf{A},\mathbf{S}}{\operatorname{argmax}} \prod_i \mathrm{p}(a_i \mid \mathbf{X}(s_i), s_i)\mathrm{p}(s_i, \mathbf{X}(s_i))$$

where a_i is one label of the sequence, s_i is a single segment of the segmentation[1], and $\mathbf{X}(s_i)$ is the portion of the observation sequence corresponding to s_i. Segmental features are incorporated by constraining $\mathrm{p}(a_i \mid \mathbf{X}(s_i), s_i)$ to be of the form $\mathrm{p}(a_i \mid f(\mathbf{X}(s_i)), s_i)$, as mentioned above.

There are a number of segment-based systems that take a classification approach to recognition [1, 2, 3]. With the exception of [2], however, these do not include an explicit computation of the segmentation probability. Our approach differs from [2] in the types of models used and in the method of obtaining the segmentation score. In [2], the classification and segmentation probabilities are estimated with separate multi-layer perceptrons.

2.2. Classification Component

The formulation described above is quite general, allowing the use of a number of different classification and segmentation components. The particular classifier used in the experiments described below is based on the Stochastic Segment Model (SSM) [4], an approach that uses segmental measurements in a statistical framework. This model represents the probability of a phoneme based on the joint statistics of an entire segment of speech. Several variants of the SSM have been developed since its introduction [5, 6], and recent work has shown this model to be comparable in performance to hidden-Markov model systems for the task of word recognition [7]. The use of the SSM for classification in the CIR formalism is described next.

Using the formalism of [4], $\mathrm{p}(\mathbf{X}(s_i) \mid s_i, a_i)$ is characterized as $\mathrm{p}(f(\mathbf{X}(s_i)) \mid s_i, a_i)$, where $f(\cdot)$ is a linear time warping transformation that maps variable length $\mathbf{X}(s_i)$ to a fixed length sequence of vectors $\mathbf{Y} = f(\mathbf{X}(s_i))$. The specific model for $\mathbf{Y}$ is multi-variate Gaussian, generally subject to some assumptions about the covariance structure to reduce the number of free parameters in the model. The posterior probability used in the classification work here is obtained from this distribution according to

$$\mathrm{p}(a_i \mid f(\mathbf{X}(s_i)), s_i) = \frac{\mathrm{p}(f(\mathbf{X}(s_i)) \mid a_i, s_i)\,\mathrm{p}(a_i, s_i)}{\sum_{a_i} \mathrm{p}(f(\mathbf{X}(s_i)) \mid a_i, s_i)\,\mathrm{p}(a_i, s_i)}.$$

There are more efficient methods for direct computation of the posterior distribution $\mathrm{p}(a_i \mid f(\mathbf{X}(s_i)), s_i)$, such as with tree-based classifiers or neural networks. However, the above formulation, which uses class-conditional densities of the observations, $\mathrm{p}(f(\mathbf{X}(s_i)) \mid a_i, s_i)$, has the advantage that we can directly compare the CIR approach to the traditional approach and therefore better understand the issues associated with using fixed-length measurements and the effect of the segmentation score. In addition, this approach allows us to take advantage of recent improvements to the SSM, such as the dynamical system model [6], at a potentially lower cost due to subsampling of observations.

2.3. Segmentation Component

There are several possibilities for estimating the segmentation probability, and two fundamentally different approaches are explored here. First we note that we can

[1] If s_i is defined as the start and end times of the segment, clearly consecutive s_i are not independent. To avoid this problem, we think of s_i as corresponding to the length of the segment.

estimate either $p(S \mid X)$ or $p(S, X)$ for the segmentation probability, leading to the two equivalent expressions in (1).

One method is to simply compute a mixture distribution of segment probabilities to find $p(s_i, X(s_i))$:

$$\begin{aligned} p(s_i, X(s_i)) &= \sum_j p(s_i, X(s_i), c_j) \\ &= \sum_j p(X(s_i)|s_i, c_j) p(s_i, c_j) \qquad (3) \end{aligned}$$

where $\{c_j\}$ is a set of classes, such as linguistic classes or context-independent phones. In order to find the score for the *complete* sequence of observations, the terms in the summation in (3) are instances of the more traditional formulation of (2). This method uses the complete observation sequence, as in [4], to determine the segmentation probabilities, as opposed to the features used for classification, which may be substantially reduced from the original observations and may lack some cues to segment boundaries, such as transitional acoustic events.

Another method for computing the segmentation probability, similar to that presented in [2], is to find the posterior probability $p(S \mid X)$. In this approach, we use distributions that model presence versus absence of a segment boundary at each frame, based on local features. The segmentation probability is written as

$$p(S \mid X) = \prod_i p(s_i \mid X(s_i)) \qquad (4)$$

and the probability of an individual segment of length L is

$$p(s_i \mid X(s_i)) = p(b_L \mid X(s_i)) \prod_{j=1}^{L-1} p(\overline{b_j} \mid X(s_i)), \qquad (5)$$

where b_L is the event that there is a boundary after frame L and $\overline{b_j}$ is the event that there is not a boundary after the jth frame of the segment. We estimate the frame boundary probabilities as

$$p(b_j \mid X(s_i)) = \frac{LK}{1 + LK}$$

where $K = p(b)/p(\bar{b})$ and

$$L = \frac{p(x_j, x_{j+1} \mid \overline{b_j})}{p(x_j, x_{j+1} \mid b_j)}.$$

The component conditional probabilities are computed as

$$p(x_j, x_{j+1} \mid \overline{b_j}) = \sum_\beta p(x_j, x_{j+1} | \beta) p(\beta) \qquad (6)$$

and

$$p(x_j, x_{j+1} \mid b_j) = \sum_{\beta_1} \sum_{\beta_2} p(x_j \mid \beta_1) p(x_{j+1} \mid \beta_2) p(\beta_1, \beta_2), \qquad (7)$$

where β ranges over the manner-of-articulation phoneme classes: stops, nasals, fricatives, liquids, vowels, and additionally, silence.

The two segmentation models presented have different advantages. The first method makes use of the complete set of SSM phone models in determining likely boundaries for each segment and hence may have a more complete model of the speech process. On the other hand, the second approach uses models explicitly trained to differentiate between boundary and non-boundary acoustic events. The best choice of segmentation score is an empirical question that we have begun to address in this work.

3. EXPERIMENTS

Experiments have been conducted to determine the feasibility of the recognition approach described here. First, we wished to determine whether fixed-length measurements could be as effective in recognition as using the complete observation sequence, as is normally done in other SSM work and in HMMs. This test would tell whether the segmentation score can compensate for the use of fixed-length measurements. Second, we investigated the comparative performance of the two segmentation scoring mechanisms outlined in the previous section.

3.1. CIR Feasibility

The feasibility of fixed-length measurements was investigated first in a phoneme classification framework. Since we planned to eventually test our algorithms in word recognition on the Resource Management (RM) database, our phone classification experiments were also run on this database. Since the RM database is not phonetically labeled, we used an automatic labeling scheme to determine the reference phoneme sequence and segmentation for each sentence in the database. The labeler, a context-dependent SSM, took the correct orthographic transcription, a pronunciation dictionary, and the speech for a sentence and used a dynamic programming algorithm to find the best phonetic alignment. The procedure used an initial labeling produced by the BBN BYBLOS system [8] as a guide, but allowed some variation in pronunciations, according to the dictionary, as well as in segmentation. The resulting alignment is flawed in comparison with carefully hand transcribed speech, as in the TIMIT database. However, our experience has shown that using comparable models and analysis, there is only about a 4-6% loss in classification performance (e.g., from 72% to 68% correct for context-independent models) between the two databases, and the RM labeling is adequate for making preliminary comparisons of classification algorithms. The final test of any classification algorithm is made under the CIR formalism in word recognition experiments, for which the RM database is well suited.

In classification, the observation vectors in each segment were linearly sampled to obtain a fixed number of vectors per segment, $m = 5$ frames. For observed segments of length less than five frames, the transformation repeated some vectors more than once. The feature vector for each frame consisted of 14 Mel-warped cepstral coefficients and their first differences as well as differenced energy. Each of the m distributions of each segment were modeled as independent full covariance Gaussian distributions. Separate models were trained for males and females by iteratively segmenting and estimating the models using the algorithm described in [4]. The testing material came from the standard "Feb89" and "Oct89" test sets. In classification experiments using the Feb89 test set, the percent correct is reported over the complete set of phoneme instances, 11752 for our transcription. Several simplifying assumptions were made to facilitate implementation. Only context-independent models were estimated, and the labels and segments of the observation sequence were considered independent.

On the Feb89 test set the classification results were 65.8% correct when the entire observation sequence was used and 66.4% correct when a fixed number of observations was used for each segment. This result indicates that, in classification, using fixed length measurements can work as well as using the entire observation.

Having verified that fixed-length features are useful in classification, the next step was to evaluate their use in recognition with the CIR formalism. In recognition, we make use of the N-best formalism. Although originally developed as an interface between the speech and natural language components of a spoken language system [9], this mechanism can also be used to rescore hypotheses with a variety of knowledge sources [10]. Each knowledge source produces its own score for every hypothesis, and the decision as to the most likely hypothesis is determined according to a weighted combination of scores from all knowledge sources. The algorithm reduces the search of more computationally expensive models, like the SSM, by eliminating very unlikely sentences in the first pass, performed with a less expensive model, such as the HMM. In this work, the BBN BYBLOS system [8] is used to generate 20 hypotheses per sentence.

Using the N-best formalism, an experiment was run comparing the CIR recognizer to an SSM recognizer that uses all observations. The classifier for the CIR system was the same as that used in the previous experiment. The joint probability of segmentation and observations, $p(\mathbf{X}, \mathbf{S})$, was computed as in Equation (3), using a version of the SSM that considered the complete observation sequence for a segment. That is, not just m, but *all* observation vectors in a segment were mapped to the distributions and used in finding the score. The weights for combining scores in the N-best formalism were trained on the Feb89 test set. In this case the scores to be combined were simply the SSM score, the number of words and the number of phonemes in a sentence.

In evaluating performance using the N-best formalism, the percent word error is computed from the highest-ranked of the rescored hypotheses. On the Feb89 test set the word error for both the classification-in-recognition method and the original recognition approach was 9.1%. To determine if these results were biased due to training the weights for combining scores on the same test data, this experiment was repeated on the Oct89 test set using the weights developed on the Feb89 test set. The performance for the CIR recognizer was 9.4% word error (252 errors in a set of 2684 reference words) and the performance for the original approach using the complete observation sequence was 9.1% word error (244 errors). The performance of the new recognition formalism is thus very close to that of the original scheme, and in fact the difference between them could be attributed to differences associated with suboptimal N-best weight estimation techniques [11].

3.2. Segmentation Score

As mentioned previously, some current systems use a classification scheme with no explicit probability of segmentation. We attempted to simulate this effect with the classification recognizer by simply suppressing the score for the joint probability of segmentation and observations. This is equivalent to assuming that the segmentation probabilities are equally likely for all hypotheses considered. Scores were computed for the utterance with and without the $p(\mathbf{X}, \mathbf{S})$ term on the Feb89 test set. When just the classification scores were used, word error went from from 9.1% to 10.8%, an 18% degradation in performance. Apparently, the joint probability of segmentation and observations has a significant effect in normalizing the posterior probability for better recognition.

Experiments were also run to compare the two methods of segmentation scoring described above. In the first method, based on equation (3), the same analysis described earlier was used at each frame (cepstra plus differenced cepstra and differenced energy) and the summation was over the set of context independent phones. In the second method, which computes p($\mathbf{S} \mid \mathbf{X}$) using equations (4) - (7), we modeled each of the conditional densities in (6) and (7) as the joint, full covariance, Gaussian distribution of the cepstral parameters of the two frames adjoining the hypothesized boundary. In order to reduce the number of free parameters to estimate in the Gaussian model, we used only the cepstral coefficients as features for each frame. On the Feb89 test set the first method had 9.1% combined word error for male and female speakers, while the second method had 11.0% word error. Using the best weights for the N-best combination from this test set, the segmentation algorithms were also run on the Oct89 test set. In this case, the word error rates for the two methods were 9.4% and 11.9%, respectively.

This result suggests that the boundary-based segmentation score yields performance that is worse than no segmentation score. However, the "no segmentation" case actually uses an implicit segmentation score in that the N hypotheses are assumed to have equally likely segmentations (while all other segmentations have probability zero) and in that phoneme and word counts are used in the combined score. Although we suspect that the marginal distribution model for segmentation scores may still be preferable, clearly more experiments are needed with a larger number of sentence hypotheses to better understand the characteristics of the different approaches.

4. DISCUSSION

In summary, we have described an alternative approach to speech recognition that combines classification and segmentation scoring to more effectively use segmental features. Our pilot experiments demonstrate that the classification-in-recognition approach can achieve performance comparable to the traditional formalism when frame-based features and equivalent Gaussian distributions are used, and that the segmentation score can be an important component of a classification approach. We anticipate performance gains with the additional use of segmental features in the classification component of the CIR model. We also plan to extend the model to incorporate context-dependent units.

Our initial experiments with the segmentation probability indicate that finding this component via marginal probabilities computed with a detailed model may be more accurate than estimating boundary likelihood based on local observations, although this conclusion should be verified with experiments using a larger number of hypotheses per sentence than the 20 used so far. A number of improvements can be made to both models, including using different choices for mixture components and eliminating some of the independence assumptions. Additionally, in the second method we plan to increase both the number of features per frame and the number of boundary-adjacent frames considered in computing the boundary probabilities. Eventually a hybrid method that combines elements of both approaches may prove to be the most effective.

References

1. S. Austin, J. Makhoul, R. Schwartz and G. Zavaliagkos, "Continuous Speech Recognition using Segmental Neural Nets," *Proceedings of the DARPA Workshop on Speech and Natural Language*, pp. 249-252, Feb. 1991.
2. H. C. Leung, I. L. Hetherington and V. Zue, "Speech Recognition Using Stochastic Explicit-Segment Modeling," *Second European Conference on Speech Communication and Technology*, Genova, Italy, September, 1991.
3. P. Ramesh, S. Katagiri and C. H. Lee, "A New Connected Word Recognition Algorithm based on HMM/LVQ Segmentation and LVQ Classification," *Proceedings IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 113-116, Toronto, May 1991.
4. M. Ostendorf and S. Roukos, "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Trans. on Acoust., Speech and Signal Processing*, Dec. 1989, pp. 1857–1869.
5. S. Roukos, M. Ostendorf, H. Gish and A. Derr, "Stochastic Segment Modeling Using the Estimate-Maximize Algorithm," *Proceedings IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp 127–130, New York, New York, April 1988.
6. V. Digalakis, J. R. Rohlicek, M. Ostendorf, "A Dynamical System Approach to Continuous Speech Recognition," *Proceedings IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 289–292, Toronto, May 1991.
7. O. Kimball, M. Ostendorf and I. Bechwati, "Context Modeling with the Stochastic Segment Model," to appear in *IEEE Trans. Signal Processing.*
8. F. Kubala, S. Austin, C. Barry, J. Makhoul, P. Placeway, R. Schwartz, "BYBLOS Speech Recognition Benchmark Results," *Proceedings of the DARPA Workshop on Speech and Natural Language*, pp. 77-82, February 1991.
9. R. Schwartz and Y.-L. Chow, "The N-Best Algorithm: An Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses," *Proceedings IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1857–1869, April 1990.
10. 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," *Proceedings of the DARPA Workshop on Speech and Natural Language*, pp. 83–87, Asilomar, CA, Feb. 1991.
11. A. Kannan, M. Ostendorf, J. R. Rohlicek, "Weight Estimation for N-Best Rescoring," this proceedings.