MODELING TRAJECTORIES IN THE HMM FRAMEWORK

Rakmini Iyer  Owen Kimball  Herbert Gish

GTE/BBN Technologies
70 Fawcett Street, Cambridge, MA 02138
riyer@bbn.com

ABSTRACT

Most state-of-the-art statistical speech recognition systems use hidden Markov models (HMM) for modeling the speech signal. However, limited by the assumption of conditional independence of observations given the state sequence, current HMM’s poorly model the trajectory constraints in speech. In [1], we introduced the parallel path HMM, where each phonetic unit is represented by a parallel collection of HMM’s that model the phone trajectory variability. The trajectory constraint is imposed by disallowing transitions across parallel paths. In this paper, we investigate improvements to two critical components of this new framework: (i) initializing the sets of trajectories per phone that will form the basis of the parallel collection of HMM’s, and (ii) evaluating alternative parameter sharing strategies related to distributing the number of model parameters. Recognition results on Switchboard, a large vocabulary conversational speech recognition task, demonstrate 0.7-1.0% absolute performance improvements with the parallel path HMM in the N-best rescoring paradigm.

1. INTRODUCTION

Hidden Markov models (HMM) are the most popular approach to statistical speech recognition [2]. Although quite powerful given their simplicity, HMM’s are unable to fully capture the time-dependence in the speech process due to the assumption of conditional independence of observations given the hidden state sequence. Alternative techniques, such as parametric and non-parametric constrained-mean trajectory segment modeling approaches [3], can effectively exploit time-dependencies in the acoustic signal by relaxing the HMM independence assumption at the cost of losing the efficient HMM training and recognition algorithms.

In [1], we introduced a new model, the parallel path HMM or the PP-HMM, that exploits time-dependence in speech via segmental trajectory constraints, but does so while maintaining the basic HMM framework. In the PP-HMM topology, shown in Figure 1, each phonetic unit is represented by a collection of M HMM’s that model the trajectory variability in the phone. The trajectory constraint is imposed by disallowing transitions across parallel paths. Each path has S states in time, and models the per-path trajectory variability via a mixture of K Gaussians per state. To clarify terminology, henceforth, in this paper, we will use the term parallel states to refer to the set of states in the same position in different parallel paths of an HMM, i.e. the collection of states \( \{ s + m \times S, 0 \leq m < M \} \) where \( 1 \leq s \leq S \). There are two pseudo-states, \( s = 0 \) and \( s = ((M - 1) \times S) + 1 \), that have only have exit and entry transitions respectively. Also, we will use the term regular HMM to distinguish the normal S-state HMM from the \( S \times M \)-state parallel path model.

![Figure 1: Left-to-right HMM topology for a phone, with \( S = 5 \) states in time sequence and \( M = 3 \) parallel paths. Transitions are not allowed across parallel paths.](image)

In this paper, we investigate improvements to two critical components of this new HMM framework, initializing the parallel paths in training and parameter sharing (or clustering) of parameters given parallel states. In Section 2, we discuss the impact of different trajectory initializations on the PP-HMM training. Next, Section 3 evaluates alternative parameter sharing strategies related to distributing the number of model parameters in the new framework. Section 4 describes the experimental paradigm and results, and finally, Section 5, concludes with a discussion of the future research issues for the PP-HMM.

2. HMM TRAJECTORY INITIALIZATION

A crucial step in creating the PP-HMM’s is initializing the sets of trajectories that will form the basis of the parallel collection of HMM models. In the initialization process discussed in [1], a regular HMM phonetically “labels” the acoustic training data via Viterbi decoding. Given these HMM phone segmentations, a context-independent segmental trajectory mixture model [4] labels each phone instance in training with one of the parallel path labels, while retaining the original state alignments. The new path labels and old state alignments bootstrap the initial PP-HMM training. These newly trained PP-HMM’s are then used to re-label the training data to improve the initial state...
alignments for better modeling of the state mixture distributions. Early experiments with the model demonstrated that the re-labeling step improves the parallel path performance; so we use it experiments throughout the rest of this paper.

In this section, we discuss three issues that impact HMM trajectory initialization: (i) using parametric vs. non-parametric segment models to label the training data, (ii) effect of using phone duration as a factor in trajectory initialization, and (iii) using context-independent vs. context-dependent trajectory initial estimates.

2.1. Parametric vs. Non-Parametric Trajectories

In [4], we presented a non-parametric segmental k-means algorithm, where exemplar trajectories for a phone are first formed by linearly interpolating the observed phone segments to a fixed length. Then, these interpolated segments are assigned to a centroid trajectory using a weighted Euclidean distance. If we denote \( L(x_i|D) \) as the interpolated version of the segment \( x_i \) of length \( n_i \) to a new length \( D \), and \( c_j \) as the \( j^{th} \) centroid, then the distance metric \( d(x_i, c_j) \) is defined as

\[
d(x_i, c_j) = \frac{||L(x_i|D) - c_j||^2}{D} + \frac{||L(c_j|n_i) - x_i||^2}{n_i}
\]

where \( D \) is selected as a fixed percentile of the segment durations. The centroids are iteratively re-estimated until there is no further decrease in the total Euclidean distortion. Given these non-parametric k-means centroids, the HMM path label for each phone exemplar can be selected using Equation 1.

Alternatively, the k-means centroids can be used to seed parametric trajectory mixture models [5] that explicitly represent the temporal evolution of the speech features as a Gaussian process with time-varying parameters. The mixture parameters of this model are iteratively re-estimated using the Expectation-Maximization (EM) algorithm. Given these parametric trajectory mixtures, the HMM path label for each phone exemplar can be selected based on the most likely mixture component.

Using non-parametric trajectories is computationally cheaper, although the parametric trajectories can result in smoother trajectory representations. On the other hand, parametric trajectories are sensitive to segment boundaries; bad initial segmentations can result in poor parametric trajectory modeling as shown in [4].

2.2. Impact of Phone Duration

The distance metric \( d(x_i, c_j) \) can be further modified to include the duration of the segments as an additional factor. Namely,

\[
d(x_i, c_j) = \frac{||L(x_i|D) - c_j||^2}{D} + \frac{||L(c_j|n_i) - x_i||^2}{n_i} + w ||n_i - d_j||^2,
\]

where \( d_j \) is the average duration of the centroid \( c_j \) and \( w \) is a pre-defined weight which is selected empirically and is the same for all phones. Using duration in the distance metric significantly impacts the final non-parametric trajectory estimates as shown in Figure 2.

While the use of duration information has had a positive impact on phone classification accuracy [5], the impact of duration on the PP-HMM performance is more equivocal; we discuss this in detail later in Section 4.

2.3. Context-Dependent Trajectories

One possible improvement to the method discussed above is to use a context-dependent initialization of the trajectory models. The above context-independent initial labeling runs the risk of having the context-dependent HMM training, initialized from these labels, getting stuck in a local optimum unable to capture context-specific trajectory behavior. On the other hand, if different context models have independent initial distributions for their paths, clustering path models becomes more challenging since we no longer have a common definition of the question of membership in a “path” for use in growing trees. Context-dependent initialization is an active area of our parallel-path research, although we do not yet have results to report in this paper.

3. PARAMETER SHARING

Sharing of mixture distributions and weights is important for robust training of a context-dependent large vocabulary HMM recognition system. In most speech recognition systems, sharing is typically decided by a tree clustering algorithm, where a separate tree is built for each state of each phone. Using a combination of contextual cues (for e. g. “is the right phone a fricative?”, “is the left phone a vowel?”) and acoustic data, further branches are added automatically to the tree using some distance metric until training counts at the leaves fall below a specified threshold. Two sets of clusters are determined from the trees using two different count thresholds, one that specifies which triphones of a state share mixture components, and another that spec-
ifies which triphones of a state share the entire distribution: both the codebook and the mixture weights.

In the PP-HMM, the path information is asked as simply an additional question during the tree growing procedure [1]. More specifically, trees are grown for each regular state \( s : 1 \leq s \leq S \), and mixture distributions and/or weights can now be shared across parallel states. For the simplest case of \( M = 2 \) paths, this is equivalent to evaluating the goodness of the split achieved by the question “should the observations at this node be split based on their parallel path identity?”, and comparing it to the usual context-related splits.

There are two reasons why the clustering procedure with the 2-path model may select the path question before linguistic questions. As described in Section 2, the parallel states are initialized using a data-driven k-means algorithm that attempts to minimize a Euclidean distortion metric. Our tree growing procedure selects splits based on a metric that minimizes the variability of the data in the split groups. Minimizing distortion in k-means and minimizing variability in clustering can produce very similar decisions in splitting data. Furthermore, the clustering metric decision is weighted by the amount of data in the resulting split groups which tends to favor balanced data splits. In the parallel path model with two paths, path splits in clustering lead to exactly half the data going to each child node.

A partial solution to this problem is to select the path question only if the total gain in the clustering metric from the current split and the next best split (one step lookahead) is greater than the total gain from the best context split and its next best split. Additionally, we hypothesize that bias towards path questions due to their balanced splits will decrease as we increase the number of paths since the clustering questions will continue to split off just one of the \( M \) paths at a time.

### 4. EXPERIMENTS

Recognition results are reported on the Switchboard and Callhome corpora [6] using the BBN Byblos System [7]. Acoustic training for all the experiments use 60 hours of the older Switchboard data. The baseline dictionary contains 25,000 words. We report results on two different test sets: (i) a Switchboard (SW) test set that comprises 7 conversations held-out from the training, and (ii) a Switchboard II and Callhome (SW2+CH) test set that comprises 7 conversations drawn from each corpora from the NIST 1997 Large Vocabulary Speech Recognition evaluation data set. Differences in time periods, speaker collection, and in the case of Callhome, differences in style result in a mismatch between the SW2+CH test set and the training.

We present word error rate results using two testing methodologies. In one case, we decode a lattice, and in another, we rescore the N-best obtained from a regular 1-path HMM system. In the rescoring formalism, the top \( N \) sentence hypotheses (in our case \( N = 100 \)) are rescored by the \( M \)-path models, and a weighted combination of scores from different knowledge sources is used to re-rank the hypotheses. Two sets of N-best rescoring experiments are reported: (i) we replace the regular HMM score with the 1-path score to observe a direct comparison between the two models (N-best-1), and (ii) we retain the 1-path scores and add the 2-path score as an additional knowledge source to test if the model contributes different information relative to the 1-path HMM (N-best-II).

We initialized from 1-path HMM labels and trained \( M \)-path models as described above. Input features include 14 cepstral coefficients, the normalized energy, and their first and second order differences. The regular HMM used to phonetically label the training was a 5-state model. Parametric trajectory models with quadratic trajectories, diagonal covariances and 2 or 4 mixture components are seeded with the k-means centroids.

#### 4.1. Initialization

Table 1 compares a regular 5-state HMM topology with a 10-state PP-HMM with 5 states in a sequence and 2 parallel paths. Two initialization methods are investigated for the parallel path framework: one where duration is used as a factor in the segmental k-means distortion metric, and one where it is not. Parametric trajectory models are used to generate the HMM path labels.

<table>
<thead>
<tr>
<th>Method</th>
<th>Test Set</th>
<th>Models (WER %)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1-path w/ dur</td>
</tr>
<tr>
<td>Decoding</td>
<td>SW</td>
<td>34.1</td>
</tr>
<tr>
<td></td>
<td>SW2+CH</td>
<td>45.9</td>
</tr>
<tr>
<td>N-best-I</td>
<td>SW</td>
<td>34.1</td>
</tr>
<tr>
<td></td>
<td>SW2+CH</td>
<td>45.9</td>
</tr>
<tr>
<td>N-best-II</td>
<td>SW</td>
<td>34.1</td>
</tr>
<tr>
<td></td>
<td>SW2+CH</td>
<td>45.9</td>
</tr>
</tbody>
</table>

There are several inferences that can be drawn from Table 1.

- The 2-path model almost always outperforms the baseline 1-path model on the SW test set that is matched with the training. However, the gain fails to carry over on the mismatched test set, except under the N-best-II formalism. It may be that using parallel paths for specifying parameter sharing while keeping the clustering thresholds the same, effectively reduces context-dependence. Context information is data-independent while the path is derived from the training data. This could result in the path model performing better in the matched case relative to the mismatched case.

- The gain in performance for the 2-path model in N-best-I relative to decoding seems to indicate increased confusability during decoding, potentially due to the same reason outlined earlier. However, it is encouraging to note that the 2-path model does contribute new information resulting in a total absolute gain of 1.0% on SW and 0.6% on SW2+CH.

- The use of duration results in small gains in the N-best-II formalism relative to the model without duration. However, given the mixed results in the other
two testing paradigms, we feel that more experiments are needed to confirm this gain. We disregard duration in the rest of the paper.

In the next series of experiments (refer to Table 2), we investigate the effect of using a non-parametric trajectory initialization instead of the parametric model used in Table 1. The non-parametric trajectory initialization appears

Table 2: Word error rate on BBN development test sets. Testing impact of using parametric vs. non-parametric 2-path trajectories.

<table>
<thead>
<tr>
<th>Method</th>
<th>Test Set</th>
<th>Models (WER %)</th>
<th>1-path</th>
<th>2-path</th>
<th>2-path</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>param</td>
<td>non-param</td>
<td></td>
</tr>
<tr>
<td>Decoding</td>
<td>SW</td>
<td>34.1</td>
<td>33.9</td>
<td>34.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SW2+CH</td>
<td>45.9</td>
<td>46.5</td>
<td>46.0</td>
<td></td>
</tr>
<tr>
<td>N-best-I</td>
<td>SW</td>
<td>34.1</td>
<td>33.6</td>
<td>34.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SW2+CH</td>
<td>45.9</td>
<td>46.2</td>
<td>46.1</td>
<td></td>
</tr>
<tr>
<td>N-best-II</td>
<td>SW</td>
<td>34.1</td>
<td>33.4</td>
<td>33.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SW2+CH</td>
<td>45.9</td>
<td>45.6</td>
<td>45.4</td>
<td></td>
</tr>
</tbody>
</table>

more stable across test sets relative to the parametric models, although the differences in WER are small. Since the k-means initialization is computationally cheaper, we use it in the clustering experiments described next.

4.2. Clustering

Table 3 reports recognition performance using non-parametric initialized 2-path models with and without the lookahead strategy in clustering as described in Section 3. Clustering

Table 3: Word error rate on BBN development test sets. Testing impact of using lookahead strategy in clustering.

<table>
<thead>
<tr>
<th>Method</th>
<th>Test Set</th>
<th>Models (WER %)</th>
<th>2-path w/o lookahead</th>
<th>2-path w lookahead</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decoding</td>
<td>SW</td>
<td>34.0</td>
<td>34.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SW2+CH</td>
<td>46.2</td>
<td>46.1</td>
<td></td>
</tr>
<tr>
<td>N-best-I</td>
<td>SW</td>
<td>34.1</td>
<td>34.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SW2+CH</td>
<td>46.1</td>
<td>46.1</td>
<td></td>
</tr>
<tr>
<td>N-best-II</td>
<td>SW</td>
<td>33.2</td>
<td>33.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SW2+CH</td>
<td>45.4</td>
<td>45.2</td>
<td></td>
</tr>
</tbody>
</table>

with lookahead does not appear to have a significant impact on performance, although the path questions were asked lower in the tree relative to using clustering without lookahead.

5. DISCUSSIONS AND FUTURE WORK

In this paper, we have presented the parallel path HMM, a new approach that combines the advantages of segmental models with HMM's while maintaining the basic HMM structure. The segmental information is directly used for bootstrapping the training of the HMM model. Recognition results on Switchboard demonstrate 0.7-1.0% absolute performance improvements with the new model.

Some straightforward advances that may further improve recognition performance with the M-path models include using context-dependent trajectory initialization and increasing the number of parallel paths. We also need to investigate more closely the differences in decoding vs. resoring that may be a result of the current parameter clustering and smoothing strategies. The PP-HMM framework provides several alternatives for adding structure to the parameter architecture. For example, increasing the number of parallel paths would imply modeling the phone trajectory variability in greater detail, thus requiring fewer mixtures at the state-level.

6. REFERENCES