Hidden Markov Models for Trajectory Modeling

Rukmini Iyer Herbert Gish Man-Hung Siu George Zavaliagkos Spyros Matsoukas

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

ABSTRACT

Current state-of-the-art statistical speech recognition systems use hidden Markov models (HMM) for modeling the speech signal. However, it is well known that HMM's do not exploit the time-dependence in the speech process, since they are limited by the assumption of conditional independence of observations given the state sequence. Alternative techniques, such as segment modeling approaches, can effectively exploit time-dependencies in the acoustic signal by discarding the observation independence assumption. However, losing the basic HMM structure is often a high computational price to pay for improved acoustic models. In this paper, we introduce the **parallel path HMM** that exploits the time-dependence in speech via parametric trajectory models while maintaining the HMM framework. We present preliminary results on Switchboard, a large vocabulary conversational speech recognition task, demonstrating both improved modeling and potential for improved recognition performance.

1. Introduction

Hidden Markov models (HMM) are the most popular approach to statistical speech recognition [1]. It is well known that HMM's can only exploit the time-dependence in the speech process in a limited way, 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 [2], can effectively exploit time-dependencies in the acoustic signal by relaxing the HMM independence assumption. In particular, parametric trajectory models [3, 4] explicitly represent the temporal evolution of the speech features as a Gaussian process with time-varying parameters. However, segment modeling approaches typically fall outside the framework of the HMM's, hence, are unable to take advantage of the efficient HMM training and recognition algorithms.

In this paper, we describe a new approach that exploits time-dependence in speech using parametric trajectory models, but does so while maintaining the basic HMM framework. The paper is organized as follows. In Section 2, we first investigate trajectories in the speech signal as modeled by the HMM, and introduce the notion of a **parallel path HMM**. Section 3 describes in detail the training and recognition procedures with the new HMM structure, while raising the important modeling and parameter sharing questions that need to be tackled. In Section 4, we present preliminary results on the Switchboard task [5], and finally, Section 5 concludes with a discussion of the future research issues for the parallel path HMM framework.

2. Trajectories in HMM

As preliminary work, we first analyzed the time-sequence produced by the BBN Byblos speech recognition system [6], a state-of-the-art speaker-independent HMM system. Figure 1 presents, for phone "W", the means of the most likely terms of the HMM state mixture distributions and the corresponding input cepstra that generated the particular outputs as a function of time. Only the first cesptral feature and the corresponding mean are represented in this figure for clarity, a reasonable comparison given the HMM mixture distributions used diagonal covariances.

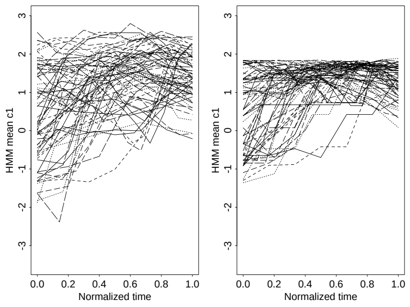

Figure 1: *Plots of the most likely terms of the HMM state mixture distribution and the input cepstral feature in normalized time-sequence.*

While the plots clearly demonstrate the existence of trajectories in the speech signal, they also indicate that the HMM is, by and large, producing trajectories that are representative of the input data. Thus, while the HMM is *creating* trajectories because the data follows trajectories, it is *modeling* these trajectories without the knowledge of the inherent trajectory structure in the input features.

If we assume that a phone can be represented as a single trajectory, then the HMM essentially needs to model the inherent variability around this trajectory; we refer to this variability as the **intra-trajectory** variability. However, we know from our segment modeling experience [9] that a single trajectory representation is often not sufficient to explain all the observed variability. Variations in speaking rate, context, speaker and even gender can result in completely different trajectories for the same phone. We refer to this variability across all trajectories representing the same phonetic unit as the **inter-trajectory** variability. In a regular HMM, the state mixture distributions model both the intra- and the inter-trajectory variability. This poor modeling strategy manifests itself by the HMM output hopping between trajectories with a limited likelihood penalty.

Recently, there have been two new approaches that aim at overcoming this HMM limitation. In [7], each phonetic unit is represented as a regular left-to-right 9-state HMM modeled by a single Gaussian per state. The mean values of the HMM state sequence then represent the mean trajectory, which is now subject to some parameterized transformation (for example, a random shift) that is global to that segment. However, the estimation of this random shift requires the speech segment boundaries (both duration and HMM state alignments); therefore, this model is used only in re-scoring N-best lists generated by an HMM. A more general approach has been proposed in [8], where a mixture of Gaussians per state is used to model the intra-trajectory variability, while the inter-trajectory variability is represented as a randomized shift modeled by a second mixture of Gaussians. In this paper, we propose an alternative approach, that creates a parallel collection of HMM's for each phonetic unit, each parallel path representing a smaller set of trajectories. Our approach, besides modifying the HMM topology, also differs from both [7, 8] in that trajectory variations are no longer restricted to random shifts.

3. Parallel Path HMM

We first describe the **parallel path HMM** in Section 3.1, before discussing the initialization and parameter sharing issues raised by adopting such a modified HMM structure in Sections 3.2 and 3.3.

3.1. HMM Topology

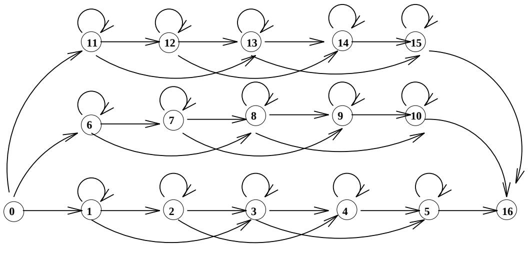

Figure 2: *Left-to-right HMM topology with $S = 5$ states in time sequence and $M = 3$ parallel paths. Transitions are not allowed across parallel paths.*

In the parallel path HMM, each phonetic unit is represented by a collection of M HMM's that model the inter-trajectory variability. Each of these M HMM's have S states in time, each HMM representing a smaller set of trajectories that model the intra-trajectory variability via a mixture of K Gaussians per state. Henceforth, in this paper, we will use the term *parallel states* to refer to a 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 $s = ((M-1)*S+1)$, 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 one.

In the topology represented in Figure 2, the segmental structure is imposed by disallowing transitions across parallel paths. Each parallel path can be trained using the regular HMM training algorithms. The multiple paths easily fit into this framework simply as additional states with transition probability constraints. In fact, no change to the training or decoding procedure is required.

3.2. Initializing HMM Training

The key to creating the parallel path HMM's lies in establishing the sets of trajectories that will form the basis of the parallel collection of HMM models. A random initialization can prove to be very inefficient, besides potentially nullifying the effects of modeling the inter-trajectory variability. In this work, we use the parametric trajectory segment models [3, 4, 9] for initializing the parallel path HMM training.

Parametric Trajectory Models: As mentioned earlier, parametric trajectory models [3, 4] explicitly represent the temporal evolution of the speech features as a Gaussian process with time-varying parameters. Given a speech segment with a duration of N frames, where each frame is represented by a D dimensional feature vector, the segment can be expressed in matrix notation as:

$$\mathbf{C} = \begin{bmatrix} c_{1,1} & \dots & c_{1,D} \\ c_{2,1} & \dots & c_{2,D} \\ \vdots & & \vdots \\ c_{N,1} & \dots & c_{N,D} \end{bmatrix} = \begin{bmatrix} \underline{\mathbf{C}}_1 & \dots & \underline{\mathbf{C}}_D \end{bmatrix} \quad (1)$$

and modeled as a Gaussian process with time varying parameters as in:

$$\mathbf{C} = \mathbf{Z}\mathbf{B} + \mathbf{E} \quad (2)$$

where $\mathbf{Z}$ is a $N \times R$ design matrix that specifies the type of model to use, $\mathbf{B}$ is a $R \times D$ trajectory parameter matrix that requires estimation, $\mathbf{E}$ is a residual error matrix that also provides the trajectory covariance Σ, and R is the number of parameters in the trajectory model. In this paper, we use quadratic trajectories, $R = 3$, for all our experiments.

The likelihood of an observed segment k, $L(\hat{\mathbf{B}}_k, \hat{\Sigma}_k)$ with estimated trajectory mean $\hat{B}_k$ and covariance $\hat{\Sigma}_k$ given the model mean B and model covariance Σ can be expressed as:

$$\begin{aligned} L(\hat{\mathbf{B}}_k, \hat{\Sigma}_k | \mathbf{B}, \Sigma)) \; = & \quad (3) \\ (2\pi)^{-\frac{DN_k}{2}} \; |\Sigma|^{-\frac{N_k}{2}} \cdot \exp\left(-\frac{N_k}{2} \mathrm{tr}\left[\Sigma^{-1}\hat{\Sigma}_k\right]\right) \cdot & \\ \exp\left(-\frac{1}{2} \mathrm{tr}\left[\mathbf{Z}_k(\hat{\mathbf{B}}_k - \mathbf{B})\Sigma^{-1}(\hat{\mathbf{B}}_k - \mathbf{B})'\mathbf{Z}_k'\right]\right). & \end{aligned}$$

The above formulation can be further extended to estimating parameters of a mixture of M parametric trajectory models [4, 9]. The parameters to be estimated now include the means and covariance of the trajectory mixture components, $\hat{B}_m$, $\hat{\Sigma}_m$, as well as the mixture weights, $\mathrm{p}(m)$ for $0 \leq m < M$. We initialize the trajectory mixtures using a non-parametric K-means approach described in detail in [9]. The mixture parameters are further iteratively re-estimated using the Expectation-Maximization (EM) algorithm.

Parallel Path HMM Training: The parametric mixtures of trajectories are introduced into the HMM training to initialize the parallel path HMM. Training the parallel path HMM now comprises the following steps.

1. Phonetically label the training data via Viterbi decoding with a regular S state HMM trained on the same data. In addition to providing segment boundaries, the Viterbi alignment also produces the state sequence alignment $s_k(t) : 0 \leq t \leq N, 1 \leq s \leq S$ for each N-frame segment observation k.
2. Given the training segmentations, for each phone j, estimate a mixture of M parametric trajectory models, $\{B^j_m, \Sigma^j_m, p^j(m) : 0 \leq m < M\}$.
3. For each segment observation k, estimate the most likely trajectory mixture component $\hat{m}_k$
$$\hat{m}_k = \underset{m}{\operatorname{argmax}}\, p^j(m) L(B_k, \Sigma_k | B^j_m, \Sigma^j_m), \qquad (4)$$
where j refers to the phone identity, $\{B^j_m, \Sigma^j_m, p^j(m)\}$ are the means, covariance and prior probability of the m^{th} component of phone j, and $L()$ is estimated using Equation 3.
4. For each segment observation k, re-label the state sequence s_k to reflect the mixture component selected, i.e.
$$\hat{s}_k(t) = S \times \hat{m}_k + s_k(t). \qquad (5)$$
Note that the state alignments for the parallel paths do not change.
5. The new state sequence labels are used to bootstrap the parallel path HMM with $M \times S$ number of states, followed by a regular HMM training using the EM algorithm.
6. Finally, we re-label the training data using the newly trained parallel path HMM's to get more consistent state alignments, and re-train.

3.3. Parameter Sharing

Sharing of mixture distributions and weights is important for robust training of a context-dependent large vocabulary HMM recognition system. Sharing is typically decided by a tree clustering algorithm, where a tree is built by first creating branches for each phone and then 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 until two sets of clusters are determined, one that specifies which triphones of a state share mixture components (referred to as the shared **codebook**), and another that specifies which triphones of a state share the entire distribution: both the codebook and the mixture weights (referred to as the shared **pdf**). Note that the clustering algorithm is designed to disallow sharing across states.

Clustering is an even more critical issue for the parallel path HMM, where the data is additionally partitioned across parallel paths. There are several clustering alternatives available for the parallel path HMM's. We list two choices here, one where no parameter sharing is allowed across parallel states, and another where parameters can be shared across parallel states based on a tree-based clustering algorithm:

- We can adhere to the current clustering paradigm, where a clustering tree is grown per state in the HMM topology. This explicitly imposes the trajectory information to have precedence over contextual variations. In this scheme, no parameters are shared across parallel states.
- The more flexible clustering choice includes the parallel path information as simply an additional question during the tree growing procedure. More specifically, trees are grown per 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 $M = 2$, this is equivalent to first 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 then comparing it to the context-related splits. Note that sharing distributions across parallel states as a result of the clustering procedure is not equivalent to having a single merged state. For one, we still maintain separate state transition probabilities for the different paths, and secondly, there can be different parallel states that can precede and follow the current state.

4. Experiments

Recognition results are reported on the Switchboard and Callhome corpora [5] using the BBN Byblos System [6]. The test set comprises 7 Switchboard and 7 Callhome conversations drawn from the NIST 1997 Large Vocabulary Speech Recognition evaluation data set. Acoustic training for all the experiments use an in-house 18 hour subset of the Switchboard data. The baseline training and recognition dictionary comprises 25,000 words.

Initial phonetic segments (or labels) are obtained using a 5-state regular HMM trained on the same 18 hour subset. Input features include 14 cepstral coefficients, the normalized energy, and their first and second order differences. Parametric trajectory models with quadratic trajectories, diagonal covariances and 2 mixture components are trained on these initial labels. Due to time limitations, we only report proof-of-concept results for a 2 parallel path case, initialized with the parametric trajectory mixtures. We did not re-label the data to improve the state alignments for the parallel HMM paths in the experiments reported in this paper.

Table 1 compares a regular 5-state HMM topology with a 10-state parallel path HMM with 5 states in a sequence and 2 parallel paths. Two clustering configurations are investigated for the parallel path framework: (i) a total set of 1000 codebooks with 64 diagonal Gaussian mixtures per codebook, and (ii) a total set of 1600 codebooks with 40 Gaussian mixture distributions per codebook. Clustering trees for both cases are grown using questions based on contextual cues; no parameter sharing across parallel states is allowed. Recognition performance with 2 parallel paths degrades compared to a regular HMM. Closer analysis shows small improvements in training likelihoods. We hypothesize that the performance degradation is a result of our sub-optimal clustering strategy, where we impose the trajectory/path information to take precedence over context information. Thus, by maintaining the same number of parameters, we effectively reduced context-dependence in favor of the parallel states. However, our intuition that fewer Gaussians will be required to model the intra-trajectory variability with the parallel path HMM's appears to be true, although more experiments are required to confirm this, given the small difference in results.

Table 1: *Word error rate on BBN development test set. No parameter sharing across parallel states.*

Acoustic Model	# of parameters	WER (%)
5-state regular HMM	1000x64	53.7
10-state parallel HMM	1000x64	54.0
10-state parallel HMM	1600x40	53.9

In the next series of experiments (refer to Table 2), we modified the clustering algorithm for the parallel path HMM's to allow clustering across parallel states. The tree growing questions now include a path-dependent question. With the new clustering, we observe a small improvement in both training and recognition likelihoods. We also see a small improvement in recognition performance with the parallel state HMM using an identical configuration (same set of decoding weights and pruning thresholds, parameters include 1000 codebooks with 64 Gaussians per codebook) as the baseline HMM.

Table 2: *Word error rate on a BBN development test set. Parameter sharing across parallel states allowed.*

Acoustic Model	WER (%)
5-state regular HMM	53.7
10-state parallel HMM	53.2

5. 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. The new approach provides a more structured framework for distributing as well as increasing the number of HMM parameters. Preliminary results with 2 parallel paths have shown encouraging improvements.

There are several straightforward advances that will further improve recognition performance with the new models. These include increasing the number of parallel paths, increasing acoustic training to include the full 160 hour Switchboard data set, and re-labeling the training data to correct the state alignments for the parallel paths. At the same time, there are several interesting research questions that need to be investigated. The parametric trajectory models is just one of several alternative models available for initializing the parallel path HMM. Non-parametric models may also provide an interesting alternative, especially since they are more consistent with the HMM. It would also be worthwhile evaluating the impact of improved segment models on the parallel path training.

6. REFERENCES

1. L. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” in *Proc. IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

2. M. Ostendorf, V. V. Digilakis and O. A. Kimball, “From HMM's to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition,” in *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, 1996.

3. H. Gish and K. Ng, “A Segmental Speech Model with Applications to Word Spotting,” in *Proc. ICASSP*, vol. 2, pp. 447-450, 1993.

4. H. Gish and K. Ng, “Parametric Trajectory Models for Speech Recognition”, in *Proc. ICSLP*, pp. 466-469, 1996.

5. J. J. Godfrey, E. C. Holliman and J. McDaniel, “Switchboard: Telephone Speech Corpus for Research and Development,” in *Proc. ICASSP* vol. 1, pages 517-520, 1992.

6. J. Billa etal, “Multilingual Speech Recognition: The 1996 Byblos Callhome System,” *Proc. Eurospeech* vol. 1, pp. 363-366, 1997.

7. J. Goldberger, D. Burshtein and H. Franco, “Segmental Modeling Using a Continuous Mixture of Non-Parametric Models,”, in *Proc. EUROSPEECH*, pp. 1195-1198, 1997.

8. G. Zavaliagkos and S. Matsoukas, “Convolutional Density Estimation,” Manuscript in preparation.

9. M. Siu, R. Iyer, H. Gish and C. Quillen, ”Parametric Trajectory Mixtures for LVCSR,” in this *Proc. ICSLP*, 1998.