Implicit Trajectory Modeling through Gaussian Transition Models for Speech Recognition

Hua Yu and Tanja Schultz
Interactive Systems Lab, Carnegie Mellon University, Pittsburgh, PA 15213
hyu@cs.cmu.edu

Abstract

It is well known that frame independence assumption is a fundamental limitation of current HMM based speech recognition systems. By treating each speech frame independently, HMMs fail to capture trajectory information in the acoustic signal. This paper introduces Gaussian Transition Models (GTM) to model trajectories implicitly. Comparing to alternative approaches, such as segment modeling and parallel path HMM, GTM has the advantage that it integrates seamlessly with the HMM framework; it can model a large number of trajectories and there is no need to define a topology a priori. Preliminary experiments on Switchboard, a large vocabulary conversational speech recognition task, have shown promising results.

1 Motivation

Hidden Markov model (HMM) has been the dominant approach in automatic speech recognition for years. Several assumptions are made in HMMs, one of which is frame independence: all speech frames are conditionally independent given the hidden state sequence. This makes HMMs ineffective in modeling trajectories.

Speech production differs from a random process in that articulators move along a low dimensional manifold. As a result, speech trajectory is relatively smooth in the feature space. But HMM, as a generative model, does not necessarily generate a smooth trajectory, due to the conditional independence assumption. This is best illustrated by considering a gender independent HMM, using Gaussian Mixture Model (GMM) as output density function (Figure 1). Let us assume certain mixture components are trained mostly on male speakers, while other components of the same mixture are trained mostly on females. Sampling the HMM produces a frame sequence randomly switching between male and female at any time.

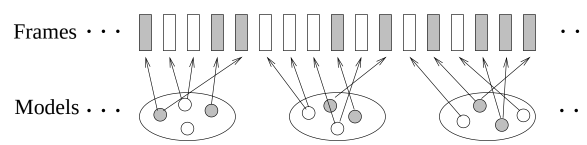


Figure 1: HMM-GMM as a generative model (shaded area stands for male speech, non-shaded area stands for female.)

Variations in speaker, context and speaking mode can all produce completely different trajectories for the same phone. If modeled by a single state sequence as in a regular HMM, trajectories will be all mixed up, resulting in a model with poor discrimination between trajectories.

Segment models attempt to exploit time-dependencies in the acoustic signal (Ostendorf et al., 1996), by modeling trajectories either parametrically or non-parametrically. Since these approaches typically fall outside the HMM framework, they can not take full advantage of the efficient HMM training and recognition algorithms.

Iyer et al. proposed the parallel path HMM to better model multiple trajectories (Iyer et al., 1998). This stays within the HMM framework. However, the number of parallel paths is normally quite limited (two or three). Choosing the right number of paths is also an unsolved problem.

In this paper, we propose a new model, namely Gaussian Transition Model (GTM), which attempts to capture dependency between adjacent frames by modeling Gaussian transitions. GTM is able to model a large number of trajectories. It also fits nicely within the HMM framework.

2 Gaussian Transition Model

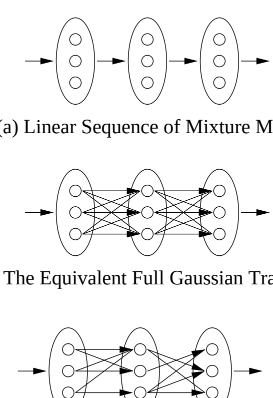
To introduce the idea of GTM, let us consider the probability of a sequence of frames $\boldsymbol{O} = (\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T)$ given a sequence of Gaussian mixture models $\boldsymbol{M} = (\boldsymbol{m}_1, \cdots, \boldsymbol{m}_T)$:

$$p(\boldsymbol{O}|\boldsymbol{M}) = \prod_{t=1}^{T} p(\boldsymbol{o}_t|M_t) = \prod_{t=1}^{T} \sum_{k_t} \pi_{tk_t} g_{tk_t}(o_t) \quad (1)$$

$$= \sum_{k_1} \sum_{k_2} \cdots \sum_{k_T} \prod_{t=1}^{T} \pi_{tk_t} g_{tk_t}(\boldsymbol{o}_t) \quad (2)$$

where $g_{tk}(\cdot)$ is the kth Gaussian in the tth model, π_{tk} is the mixture weight.

This is illustrated in Figure 2, where (a) shows the mixture model sequence and (b) shows the equivalent full Gaussian transition network. Think of each Gaussian g_{tk_t} as a modeling unit by itself, $\prod_t g_{tk_t}$ represents a unique trajectory, weighted by $\prod_t \pi_{tk_t}$. In the traditional HMM-GMM, all possible trajectories are allowed. Say some Gaussians model male speech while others model female speech, HMM-GMM allows trajectories that hop between the two genders in the middle of an utterance!



(a) Linear Sequence of Mixture Models

(b) The Equivalent Full Gaussian Transition Model

(c) Pruned Gaussian Transition Model

Figure 2: Gaussian Transition Network

GTM restricts the set of allowable trajectories by modeling transition probabilities between Gaussians in adjacent states:

$$a_{ij} = P(q_t = g_j | q_{t-1} = g_i)$$

where a_{ij} is the probability of transition from Gaussian g_i to Gaussian g_j, subject to the constraint $\sum_j a_{ij} = 1$. Figure 2(c) shows a GTM after pruning away unlikely transitions.

2.1 HMM-GMM as a Special Case

It can be shown that traditional HMM-GMM is a special case of GTM where $a_{ij} = \pi_j$ i.e. transition probability a_{ij} equals mixture weight of the destination Gaussian, independent of the identity of the source Gaussian. In other words, transition models are tied for all Gaussians in the same mixture.

2.2 GTM and Pronunciation Modeling for Sloppy Speech

As mentioned before, sloppy speech has a trajectory different from carefully articulated speech. Explicit pronunciation modeling (by adding alternative pronunciations to the lexicon) has so far been difficult, since many reductions are too subtle to be classified as either phoneme substitutions or deletions. Partial reduction or partial realization may actually be better modeled at a sub-phoneme level. Gaussian transition models can be thought of as pronunciation networks at the Gaussian level. In this sense, GTM provides a way to model alternative pronunciations implicitly. GTM provides finer model resolution, compared to pronunciation modeling at either phoneme level or state level (Saraclar et al., 2000).

3 Training GTM

When viewing each Gaussian as a state by itself, GTM can be readily trained using the existing Baum-Welch algorithm. Following notations of (Rabiner, 1989),

$$\gamma_t(i) = P(q_t = g_i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_j \alpha_t(j)\beta_t(j)}$$

$$\begin{aligned} \xi_t(i,j) &= P(q_t = g_i, q_{t+1} = g_j) \\ &= \frac{\alpha_t(i) a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_{t+1}(j)}{\sum_i \sum_j \alpha_t(i) a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_{t+1}(j)} \end{aligned}$$

where α is the forward probability and β is the backward probability, $b_j(\boldsymbol{o})$ is now a single Gaussian. The update formula is:

$$a_{ij} = \frac{\sum_t \xi_t(i,j)}{\sum_t \gamma_t(i)}$$

In practice, GTM training faces two major issues: insufficient data and pruning.

3.1 Trainability

GTM can take a large number of parameters. First, transition between two mixtures of n components each requires n^2 transition probabilities. Second, in an LVCSR system with thousands of mixture models, transition can happen between many of them.

Hence data sufficiency becomes a concern. In our experiments, we choose to model only frequent transitions. For everything else, we revert to the traditional HMM-GMM model: $a_{ij} = \pi_j$.

3.2 Pruning

Even in conventional HMM training, it is common to ignore transition probabilities. Their contribution to the overall score is quite small, in comparison to observation probabilities in a continuous HMM (which is several orders of magnitude larger). The same is true for Gaussian transition probabilities. While GTM offers better discrimination between trajectories, all trajectories are nonetheless still permitted. Pruning unlikely transitions leads to a model that is both more compact and more prudent. In reality, however, we need to exercise great care in pruning so as not to prune away unseen trajectories (due to a limited training set).

4 Experiments

Experiments are carried out on the Switchboard (SWB) task using the Janus system (Soltau et al., 2002). The test set is a 1 hour subset of the 2001 Hub5e evaluation set. Acoustic training uses a 66 hours subset of the SWB data. We use a 15k vocabulary and a trigram language model trained on SWB and CallHome. The front-end has vocal tract length normalization, cepstral mean normalization, an 11-frame window to derive delta and double-delta, linear discriminant analysis and semi-tied covariance with a single class. The acoustic model has roughly 6000 mixtures with a total of 86K Gaussians, on average 14 Gaussians per model.

We apply a two-tiered strategy to cope with the data sufficiency issue. Before training, we count the number of transitions for each model pair on the training data, using Viterbi alignment. Only transitions with counts above a certain threshold are modeled with GTM. Of about 6000 mixture models, a total of 40K model pairs (out of a potential 6K×6K=36M) has been observed. It turns out that most of the transitions (72%) are within the same model (corresponding to self-loop in HMM). We choose to model the most frequent 9400 model pairs with GTM. Not surprisingly, most of the 6000 same-model pairs are among those chosen. During training, we also apply a minimum count criterion: a transition model is updated only if the Gaussian receives enough training counts.

One iteration of Baum-Welch training gives significant improvement in term of likelihood. Log likelihood per frame improves from -50.67 to -49.18, while conventional HMM training can only improve less than 0.1. Considering the baseline acoustic model has already been highly optimized, this indicates improved acoustic modeling.

GTM transitions are pruned if their probabilities fall below a certain threshold (default is 1e-5). Table 1 shows word error rates for GTM models pruned against different thresholds. It is encouraging that a 0.5% gain is obtained after pruning away almost 2/3 of all transitions.

Pruning Threshold	Avg. # Transitions per Gaussian	WER (%)
baseline	14.4	34.1
1e-5	9.7	33.7
1e-3	6.6	33.7
0.01	4.6	33.6
0.05	2.7	33.9

Table 1: Word Error Rates on Switchboard

5 Future Work

In this paper, we have presented Gaussian Transition Model, a new approach to model trajectories within the HMM framework. Preliminary experiments have shown encouraging improvements.

There are several possibilities for further improvements. First, when modifying the decoder to use GTM, we used Viterbi approximation at word boundaries, which means trajectory information is lost upon word transition. Second, we plan to extend GTM to handle deletions in sloppy speech, a major challenge in LVCSR.

References

R. Iyer, H. Gish, M. Siu, G. Zavaliagkos, and S. Matsoukas. 1998. Hidden markov models for trajectory modeling. In *Proc. ICSLP*.

M. Ostendorf, V. Digilakis, and O. Kimball. 1996. From hmms to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing.*

L. Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. In *Proc. IEEE, 77 (2), 257-286.*

M. Saraclar, H. Nock, and S. Khudanpur. 2000. Pronunciation modeling by sharing gaussian densities across phonetic models. *Computer Speech and Language*, 14(2):137–160, April.

H. Soltau, H. Yu, F. Metze, C. Fügen, Y. Pan, and S. Jou. 2002. ISL meeting recognition. In *Rich Transcription Workshop*, Vienna, VA.