Capacity and Complexity of HMM Duration Modeling Techniques

Michael T. Johnson

***Abstract*—The ability of a standard hidden Markov model (HMM) or expanded state HMM (ESHMM) to accurately model duration distributions of phonemes is compared with specific duration-focused approaches such as semi-Markov models or variable transition probabilities. It is demonstrated that either a three-state ESHMM or a standard HMM with an increased number of states is capable of closely matching both Gamma distributions and duration distributions of phonemes from the TIMIT corpus, as measured by Bhattacharyya distance to the true distributions. Standard HMMs are easily implemented with off-the-shelf tools, whereas duration models require substantial algorithmic development and have higher computational costs when implemented, suggesting that a simple adjustment to HMM topologies is perhaps a more efficient solution to the problem of duration than more complex approaches.**

***Index Terms*—Duration models, hidden Markov models, speech recognition.**

I. Introduction

A WELL-KNOWN limitation of the hidden Markov model (HMM) used for tasks such as speech recognition is that the underlying Markov assumption constrains the state occupancy duration to be exponentially distributed according to $P(d) = (1 - a_{ii})a_{ii}^{d-1}$, where d is the duration, and a_{ii} is the self-transition probability. Since this is often inconsistent with the known duration distributions of the observation sequences being modeled, there has been substantial research in improving the HMM's duration modeling capability, originating with the work of Ferguson [1] and Levinson [2]. Duration modeling has been shown to yield small but consistent improvement in speech recognition accuracies [3]–[5].

The approaches to HMM duration modeling can be broken into three categories:

- Hidden semi-Markov models (HSMMs), a form of segment model [6], sometimes called semi-HMMs. Here, the occupancy of each state is chosen directly from a specified duration distribution. This group includes both Ferguson's explicit duration HMM (EDHMM) [1], which learns a discrete duration distribution, and Levinson's continuously variable duration HMM (CVDHMM) [2], which learns a parametric duration distribution. There have also been algorithms developed to implement upper and lower bounds on duration without specific probabilistic modeling [7].
- Variable transition HMMs (VTHMMs). In these models, the transition probabilities of each state are a function of the state's current occupancy, allowing for arbitrary duration distribution. This approach has been introduced by multiple authors, including Ramesh and Wilpon's inhomogenous HMM (IHMM) [8], Sin and Kim's nonstationary HMM (NHMM) [9], and models by Vaseghi [10]–[12], Yoma *et al.* [13], [14], and Park *et al.* [15].
- Standard HMMs with more states and/or more complex state topologies, often coupled with state distribution tying, e.g., the expanded state HMM (ESHMM) [4], [16]–[18].

Regarding VTHMMs, it is straightforward to show that there is a one-to-one transformation between a set of variable transition probabilities $a_{ij}(d)$ and a corresponding discrete duration distribution $p_i(d)$ [19]. Provided that the exit transition probabilities of the two models are in the same ratios, the net probability $P(\mathbf{S}|\mathbf{O})$ of any given state sequence $\mathbf{O}$ under an arbitrary observation sequence $\mathbf{S}$ is equivalent under the EDHMM and the VTHMM approaches, so all VTHMM methods outlined above are essentially variations on Ferguson's original EDHMM with explicit discrete distributions. The number of duration parameters needed under these approaches varies depending on whether the representation is discrete or parametric but is typically small relative to the number of distribution parameters. A parametric HSMM approach could perhaps be viewed as yielding the most direct insight into a model's duration properties.

The duration modeling capacity of a standard HMM is controlled by the duration of the underlying Markov chain [20], [21], with an overall duration distribution that can be represented as a series-parallel network of exponential random processes [4], [22]. A number of specific topologies have been investigated in the context of ESHMM work, including the Type A, Type B, and Fergusson topologies [4], which are shown in Fig. 1, as well as other unique configurations suited to specific tasks, such as including independent paths within the topology to achieve multimodal duration distributions [18]. Under the Type A configuration, the resulting duration is a modified negative binomial distribution, and under the Fergusson topology with $D = d_{\max}$ substates, the result is equivalent to a VTHMM and, thus, also Ferguson's EDHMM. Russell and Cook [4] found similar accuracies on digit and word-recognition tasks when comparing Type B topology ESHMMs with explicit duration models.

Manuscript received October 27, 2004; revised December 5, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Israel Cohen.

The author is with the Electrical and Computer Engineering Department, Marquette University, Milwaukee, WI 53233 USA (e-mail: mike.johnson@mu.edu).

Digital Object Identifier 10.1109/LSP.2005.845598

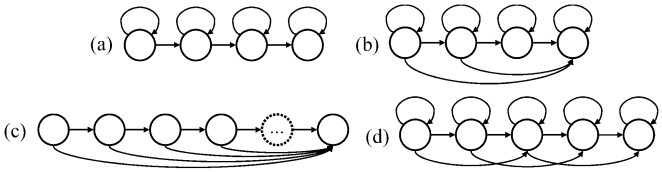

Fig. 1. HMM topologies. (a) Type A (no skip HMM). (b) Type B. (c) Fergusson. (d) One-skip HMM.

The impact on computational complexity due to any of the HSMM or VTHMM approaches is roughly an increase linearly proportional to the maximum number of states D[1]. There have been several excellent papers giving improvements to the forward–backward and Viterbi algorithms for EDHMMs [23], [24]. A detailed comparison of complexities for different models will be given in Section IV.

II. Gamma Distribution Experiments

HSMM re-estimation equations have been derived for a number of different parametric duration distributions, including, in particular, the Gamma distribution proposed in Levinson's original work [2]. Gamma distributions have been shown to match those typically seen in speech phonemes [25].

Since both explicit and parametric HSMMs can accurately model the Gamma distribution, we examine the comparative ability of standard HMMs and ESHMMs. Analytic distribution computation and parameter fitting is complex and topology dependent, so simulations were conducted using Markov chains of the desired topology with observation sequence lengths chosen from the specified Gamma distribution and parameters learned via the Baum–Welch algorithm. The distribution associated with the trained HMM was then determined by generating observation sequences and creating an empirical duration distribution. The topologies were designed to guarantee a minimum one-state duration capability, by allowing the entry state to transition to all other states.

Bhattacharyya distance, which is a simple symmetric metric between two distributions given by $\rho = -\log \int \sqrt{p_1(x)p_2(x)}$, is used to measure the distance between the original Gamma distribution and the distribution of the trained HMM. Other metrics could also be used, yielding similar results. To help visualize this metric, Fig. 2 shows a Gamma distribution with a mean of 5, with curve-fitted simulated distributions having distances of 0.11, 0.03, 0.01, and 0.002 superimposed. A "close fit" can be thought of as a distance in roughly the 0.001–0.01 range.

For the Gamma distribution experiments, the number of HMM states and ESHMM substates is varied from 1 to 6, while the mean of the Gamma distribution $(\eta = 1)$ is varied from 1 to 25. This range is similar to that of average phoneme durations given a typical 10 ms observation frame rate.

The experiments were run with Type B topology ESHMMs as well as with standard left-to-right HMMs with both no-skip (i.e., Type A) and one-skip topologies, illustrated previously in Fig. 1. Eight iterations of Baum–Welch were used for estimation, 20 000 observation sequences were used for constructing output histograms, and results were averaged over 100 runs. Results for all cases were identical to within visual discrimination. Illustrative results for the one-skip HMM are shown in Fig. 3. With one state, the distribution is purely exponential and matches the desired distribution very poorly; however, the ability of the model to track the target Gamma distribution improves rapidly as the number of total states is increased, then begins to converge.

[1]Vaseghi's work ([10, Eq. 15]) gives an expression for the Viterbi algorithm for a VTHMM without this increase. This algorithm does not maximize over the possible durations of the preceding state, which, while functional, does not yield the optimal state sequence under the model.

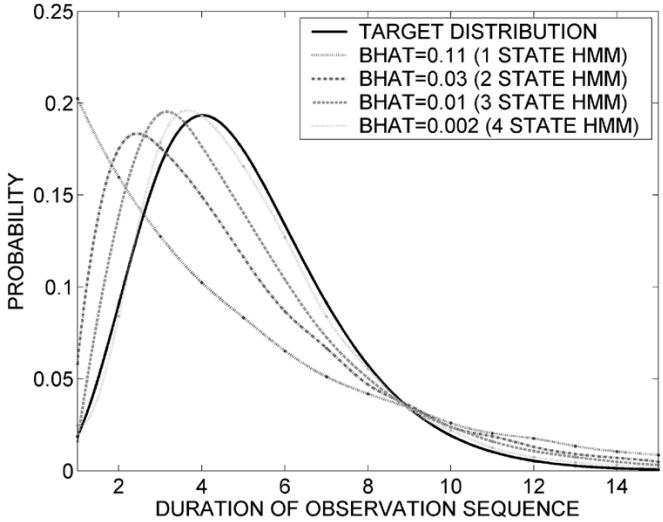

Fig. 2. Examples of distributions $(\mu = 5)$ with varying Bhattacharyya distances.

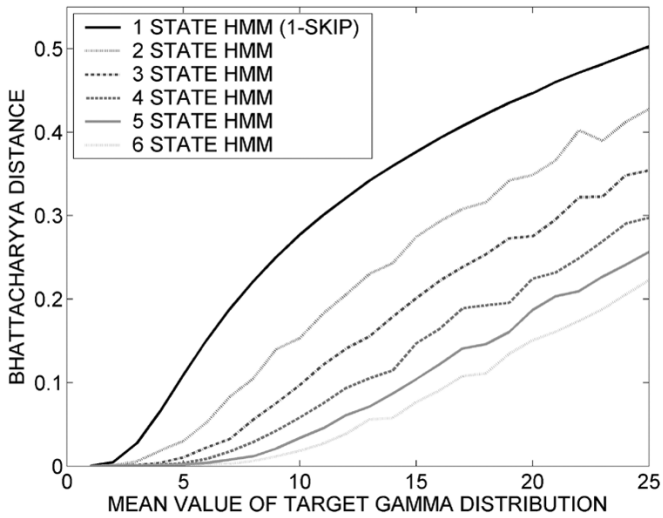

Fig. 3. Bhattacharyya distance between simulated HMMs and Gamma distribution.

III. TIMIT Phoneme Experiments

The same simulation mechanism from the previous section is used to see how well standard HMMs are able to model duration distributions of phonemes taken from the TIMIT corpus [26], a corpus which includes expertly labeled phoneme boundaries, giving good distribution approximations for read speech. The average duration of observation frames for the

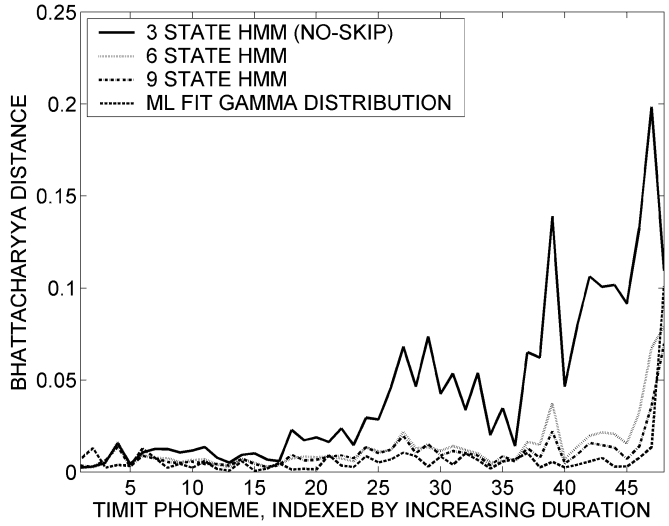

Fig. 4. Bhattacharyya distance between simulated HMMs and TIMIT distributions.

TIMIT phonemes, assuming a 10 ms observation frame rate, varies between two and 17, with over 70% of the phonemes having mean durations of between five and ten observations.

Identical experimental setup and simulation settings were used here, with no-skip HMMs, one-skip HMMs, and three-state Type B topology ESHMMs. The results were similar between the three approaches; with the no-skip HMM and ESHMM results nearly identical as before but the 1-skip HMM showing a slightly larger Bhattacharyya distance to the target distributions. One possible hypothesis is that this increase is due to the existence of multiple possible paths for sequences of the same net duration, leading to poorer transition estimation. Fig. 4 illustrates the results for the no-skip HMM case.

For ease of visualizing the results, the phonemes are ordered along the x axis in order of increasing average duration. For additional reference, the Bhattacharyya distance between the target phoneme distribution and a maximum likelihood (ML) fit two-parameter Gamma distribution is also displayed.

It can be easily seen that although a small number of states does in fact do a relatively poor job of fitting the phoneme distributions, the distance to the target distribution drops quickly as the number of states is increased. The distances converge to an asymptote close to those of the directly fitted Gamma distribution, and beyond about nine total states, there is little additional improvement. The average Bhattacharyya distance across all phonemes for nine or more states (all topologies) is approximately 0.0075.

IV. Complexity Analysis

The forward, backward, and Viterbi algorithms are the central elements of HMM training and testing. All have the same time complexity, for both standard HMMs and for any of the duration HMMs shown above. The notation used is as follows:

N total number of unique states in HMM set;
K average number of predecessor states;
T number of observations;
D maximum duration in HSMM models;
E number of expansion substates in ESHMM models;
B number of operations to compute an observation likelihood.

Note that the operations needed to compute an observation likelihood is an important component of the overall complexity. Using Gaussian mixture models, typically $B = O(MF)$ for diagonal covariances or $O(MF^2)$ for full covariances, where M is the number of mixtures, and F is the number of features in each observation.

The equation for the standard forward algorithm is

$$\alpha_j(t) = \left(\sum_i \alpha_i(t-1) a_{ij} \right) b_j(o_t), \quad \forall\, j, t.$$

The complexity of the standard algorithm is normally given as $O(N^2T)$; however, this is an approximation that is more accurately given as $O((B+K)NT)$ since observation likelihoods can be precomputed over all states and times and since the summation need only include predecessor states.

For the HSMM, the algorithm becomes

$$\alpha_j(t) = \sum_{\forall\, j, t.} \left(\sum_d \alpha_i(t-d) p_j(d) a_{ij} \prod_{s=t-d+1}^{t} b_j(o_s) \right),$$

Using a recursion to save some of the accumulated terms, as outlined in [24], the complexity of this algorithm can be given as $O((B+D)KNT)$. By precomputing all observation likelihoods and implementing the summation recursively so that there is only one new multiplier in the product term for each term in the summation, the total complexity can be reduced to $O((B+KD)NT)$. More recent work in improving the complexity can be seen in Yu and Kobayashi [23], who developed a new recursion using a duration-dependent forward variable $\alpha_j(t,d)$, with net complexity $O((B+K+D)NT)$. A similar recursion for the VTHMM approach has been given by Ramesh and Wilpon for their IHMM in [8], with complexity $O((B+KD)NT)$, and it is likely that this could also be reformulated after the Yu and Kobayashi approach.

For the ESHMM approach, the standard algorithm is used, and the impact on computational complexity is an increase in the number of states. Since the substate observation distributions are tied, a linear expansion topology, such as any of those shown in Fig. 1, gives negligible impact on the size of K and a linear increase on the value of N. This results in a net complexity of $O((B+KE)NT)$.

A summary of these complexities is given in Table I. All of the algorithms have an $O(BNT)$ term that involves computing observation likelihoods. Focusing on the post-observation computations and ignoring the common NT terms, the difference among the remaining terms is essentially $O(K)$ versus $O(KE)$ versus $O(K+D)$ versus $O(DK)$, as highlighted in the table. K is two to three for left-to-right HMMs, while D is typically 50 or more. Looking at the results in the previous sections, a reasonable value for E in a three-state configuration would perhaps be two or three. There is also a slight difference in the number of multipliers needed in the innermost loop of

TABLE I
COMPARISON OF ALGORITHM TIME COMPLEXITIES

Algorithm	Complexity
Standard HMM	$O((B+\boldsymbol{K})NT)$
ESHMM	$O((B+\boldsymbol{KE})NT)$
HSMM (fastest version [23])	$O((B+\boldsymbol{K}+\boldsymbol{D})NT)$
VTHMM (and typical HSMM)	$O((B+\boldsymbol{DK})NT)$

each iteration of the recursion, with only one required for standard HMM/ESHMM and about three to four per iteration for the more complex HSMM and VTHMM algorithms. Overall, this indicates that the algorithm speed (after precomputing all likelihoods) for the ESHMM is roughly an order of magnitude faster than that of the most efficient HSMM algorithms to date and one to two orders of magnitude faster than most VTHMM or HSMM algorithms currently in use. It should be noted that with any of these algorithms, the time to compute observation likelihoods is a large part of the overall complexity, substantially diminishing the differences between the different approaches.

Recent work [27] has compared empirical speech recognition accuracies for HMMs, HSMMs, and ESHMMs as a function of the computation speed. The results demonstrated very similar accuracies across all methods, with HMMs giving better results at low real-time factors and ESHMMs and HSMMs yielding small improvements at high real-time factors.

V. CONCLUSIONS

It has been demonstrated that either a standard HMM or an expanded state HMM, with a fairly small increase in total number of states, is able to closely model the distributions of actual phoneme durations, performing comparably to the parameterized Gamma distribution families typically used in HSMMs. This suggests that standard models, coupled with a moderate increase in overall topological complexity and state distribution parameter tying, are already well suited to handling nonexponential duration distributions. This is almost certainly a much better practical choice for duration modeling than development and implementation of more complex and computationally expensive models with explicit modifications to handle duration probabilities, for which off-the-shelf tools are not currently available.

REFERENCES

[1] J. D. Ferguson, "Variable duration models for speech," in *Proc. Symp. App. Hidden Markov Models Text Speech*, 1980.

[2] S. E. Levinson, "Continuously variable duration hidden Markov models for speech analysis," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1986, pp. 1241–1244.

[3] A. Anastasakos, R. Schwartz, and H. Shu, "Duration modeling in large vocabulary speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 628–631.

[4] M. J. Russell and A. E. Cook, "Experimental evaluation of duration modeling techniques for automatic speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1987, pp. 2376–2379.

[5] A. Bonafonte, J. Vidal, and A. Nogueiras, "Duration modeling with expanded HMM applied to speech recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, 1996, pp. 1097–1100.

[6] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, "From HMM's to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 360–378, Sep. 1996.

[7] H. Gu, C. Tseng, and L. Lee, "Isolated-Utterance speech recognition using hidden Markov models with bounded state durations," *IEEE Trans. Signal Process.*, vol. 39, no. 8, pp. 1743–1752, Aug. 1991.

[8] P. Ramesh and J. G. Wilpon, "Modeling state durations in hidden Markov models for automatic speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1992, pp. 381–384.

[9] B. Sin and J. H. Kim, "Nonstationary hidden Markov model," *Signal Process.*, vol. 46, pp. 31–46, 1995.

[10] S. V. Vaseghi, "State duration modeling in hidden Markov models," *Signal Process.*, vol. 41, pp. 31–41, 1995.

[11] ——, "Hidden Markov models with duration-dependent state transition probabilities," *Electron. Lett.*, vol. 27, pp. 625–626, 1991.

[12] S. V. Vaseghi and P. Conner, "On increasing structural complexity of finite state speech models," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1992, pp. 537–540.

[13] N. B. Yoma, F. R. McInnes, M. A. Jack, S. D. Stump, and L. L. Ling, "On including temporal constraints in viterbi alignment for speech recognition in noise," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 2, pp. 179–182, Feb. 2001.

[14] N. B. Yoma and J. S. Sanchez, "MAP speaker adaptation of state duration distributions for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 7, pp. 443–450, Oct. 2002.

[15] Y. K. Park, C. K. Un, and O. W. Kwon, "Modeling acoustic transitions in speech by modified hidden Markov models with state duration and state duration-dependent observation probabilities," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 389–392, Sep. 1996.

[16] M. J. Russell and R. K. Moore, "Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1987, pp. 2376–2379.

[17] A. E. Cook and M. J. Russell, "Improved duration modeling in hidden Markov models using series-parallel configurations of states," in *Proc. Inst. Acoust.*, vol. 8, 1986, pp. 299–306.

[18] X. Wang, L. F. M. T. Bosch, and L. C. W. Pols, "Integration of context-dependent durational knowledge into HMM-based speech recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, 1996, pp. 1073–1076.

[19] P. M. Djuric and J.-H. Chun, "An MCMC sampling approach to estimation of nonstationary hidden Markov models," *IEEE Trans. Signal Process.*, vol. 50, no. 5, pp. 1113–1123, May 2002.

[20] J. R. Norris, *Markov Chains*. Cambridge, U.K.: Cambridge Univ. Press, 1997.

[21] B. D. Hughes, *Random Walks and Random Environments*. Oxford, U.K.: Oxford Univ. Press, 1995, vol. 1.

[22] D. R. Cox, "A use of complex probabilities in the theory of stochastic processes," in *Proc. Cambridge Philosoph. Soc., Mathemat., Phys. Sci.*, vol. 51, 1955, pp. 313–319.

[23] S.-Z. Yu and H. Kobayashi, "An efficient forward-backward algorithm for an explicit-duration hidden Markov model," *IEEE Signal Process. Lett.*, vol. 10, no. 1, pp. 11–14, Jan. 2003.

[24] C. Mitchell, M. Harper, and L. Jamieson, "On the complexity of explicit duration HMM's," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 3, pp. 213–217, May 1995.

[25] D. Burshtein, "Robust parametric modeling of durations in hidden Markov models," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 548–551.

[26] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," in *Proc. Ling. Data Consort.*, 1993.

[27] J. Pylkkonen and M. Kurimo, "Duration modeling techniques for continuous speech recognition," in *Proc. ICASSP*, Jeju Island, Korea, 2004, pp. 385–388.