HEURISTIC STRUCTURAL MODIFICATIONS TO THE HMM FOR EFFICIENT RESOURCE UTILIZATION

Yong-Beom Lee[*]

Samsung Advanced Inst. Tech. / HIC Lab
Kiheung-Eup Yongin City
Kyunggi-Do 449-712 KOREA

J.R. Deller, Jr.[†]

Michigan State Univ. / 2120 EB
Dept. Elec. & Computer Engr.
East Lansing MI 48824 USA

ABSTRACT

Embedded speech processing systems require stringent memory allocation and computing resources. To minimize such resources, a simple, flexible HMM evaluation technique is presented which employs a state-space formulation in conjunction with a simplified likelihood measure. The method offers several advantages including the ability to reduce redundant computation and memory allocation across models, and a flexible structure that can exploit known results concerning state-space systems. Although performance is insignificantly effected in preliminary experiments, these benefits are achieved at the cost of a weaker coupling between the two stochastic processes that define the HMM. We augment the method with a Markov chain model of the observations to compensate for the weaker state coupling. Preliminary experiments are used to analyze recognition performance and as a basis for discussion.

1. INTRODUCTION

This work is motivated by the increasing need to conserve computational resources in embedded speech recognition systems, a problem of increasing commercial and research interest (e.g., [1, 2]).

The *hidden Markov model* (HMM) is used pervasively in speech-recognition technologies to automatically model the extraordinary variability of speech acoustics, and complex language structures (e.g. [3]). Nevertheless, it is widely accepted that some of the statistical assumptions about speech that underlie the HMM are tenuous. Among these are exponentially distributed state durations, and the assumption that elements of the observation sequence are locally (within-state) independent.[1] A current summary of some of the attempts to compensate for HMM deficiencies is found in the paper by Ogawa and Kobayashi [4].

The state structure of the HMM is intuitively appealing in its apparent relationship to the time-varying properties of speech, which, in turn, are reflective of the dynamic "states" of the speech-production system. The Bakis state constraints on the HMM [3] further reconcile the operation of the mathematical model with the physical system. In spite of its intuitive appeal, there is much empirical evidence that the state structure plays a relatively insignificant role in overall performance of the HMM. This is not surprising since the observation sequences decoded by an HMM are ordinarily much longer than the number of states in the model. This means that most of the self-transition loops in the model must have probabilities close to unity since there can only be a small number of between-state transitions among the large number of total transitions. In fact, the self-transition probability of the final (Bakis) state is exactly unity so that a path search may remain in the last state indefinitely without state transition costs. This leads to the frequently-observed phenomenon in which the final state has high probabilities of generating a vast number of the symbols, with earlier states able to generate relatively few. In such a case, there is apparently enough information in the observation string itself to maximize likelihood without much recourse to the state structure. Clearly, an HMM with such a skewed allocation of symbols would not accurately represent the sequence of underlying physical states of the speech system. This phenomenon lends support to both the idea that the state structure is relatively insignificant, and to the notion that maximum likelihood (ML) optimization (in HMM training and decoding) might not be the best approach if the goal is to create an HMM whose states accu-

[*] Work performed while YBL was at Michigan State U.

[†] JD's effort was supported in part by the National Science Foundation of the United States under Cooperative Agreement No. IIS-9817485. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

[1] The present discussion is centered on the Moore form of the HMM [3], but similar findings would pertain to the Mealy form.

rately model physical states. Several researchers have attempted to address the problem of "unrealistic" distributions of observations in order to maximize whatever performance benefits may accrue from the state structure (e.g., [5, 6, 7]).

This work reported in this paper draws upon both of the ideas raised in the previous paragraph in the development of a resource-efficient HMM evaluation algorithm. However, the issue of meaningful state distributions issue is only briefly mentioned in the final discussion due to space limitations (see [8]). In this paper, a simple relaxation of the inherent HMM coupling between the state and observation processes yields a decoding method which is both computationally cost-effective, and allows sharing of memory resources to an adjustable degree (Section 2). This method has been shown to have negligible impact on performance in spite of the potential for state path violations. To help compensate for the weaker state coupling, we augment the method with a Markov chain model of the observations (Section 3). Sections 4 and 5 present preliminary experiments and analysis.

2. A TIME-INVARIANT STATE-SPACE FORMULATION OF THE HMM

Formulation. The *forward-backward* (F-B) *algorithm* of Baum *et al.* [9] is one of two popular techniques for training, and subsequent evaluation of the HMM. Lee [8] has shown that the recursive, scalar operations comprising the F-B algorithm can be reformulated into linear, but *time-varying*, state-space model (SSM) in which the forward or backward partial sequence probability sequences – typically denoted $\alpha_t(i)$ and $\beta_t(j)$ at observation t – comprise the states of the SSM. Each of these SSM states, in turn, is uniquely related to one of the states of the HMM. A simplified linear, *time-invariant* SSM governing the HMM dynamics is proposed by Deller and Snider [9] in which each of the SSM states, at time t, represents a nonstationary probability of generating O_t from one of the HMM states. We shall refer to this model as the *time-invariant approximation to the HMM* (TIA-HMM). The TIA-HMM, is governed by a time-invariant state-space system whose defining equations are as follows: Let $\mathcal{M}$ denote an HMM with N states $q_i, i \in [1, N]$ and M discrete observation symbols $s_k, k \in [1, M]$. At each observation time t, we define the state probability vector $\boldsymbol{x}(t)$, and the observation probability vector, $\boldsymbol{y}(t)$, as follows (primed vectors indicate the transpose):

$$\boldsymbol{x}'(t) \overset{\text{def}}{=} \begin{bmatrix} x_1(t) & x_2(t) & \cdots & x_N(t) \end{bmatrix} \qquad (1)$$

$$\boldsymbol{y}'(t) \overset{\text{def}}{=} \begin{bmatrix} y_1(t) & y_2(t) & \cdots & y_M(t) \end{bmatrix} \qquad (2)$$

where, $x_i(t)$ is the probability of being in q_i at discrete time t given the model $\mathcal{M}$, $P(q_i \text{ at } t \mid \mathcal{M})$, and $y_k(t)$ is the probability of generating symbol s_k at discrete time t given the model $\mathcal{M}$, $P(s_k \text{ at } t \mid \mathcal{M})$. In these terms, the dynamics of the TIA-HMM are as follows:

$$\begin{aligned} \boldsymbol{x}(t+1) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{u}(t)\delta(t) &\qquad (3)\\ \boldsymbol{y}(t) &= \boldsymbol{B}\boldsymbol{x}(t) &\qquad (4) \end{aligned}$$

where, $\boldsymbol{A}$ is the $N \times N$ state-transition matrix associated with the HMM whose (i, j) element, $a_{ji} = P(q_j \text{ at } t+1 \mid q_i \text{ at } t)$ for any t; $\boldsymbol{B}$ is the $M \times N$ observation probability matrix whose (k, j) element, $b_{kj} = P(s_k \mid q_j)$; and $\boldsymbol{u}(0)$ is some vector such that when $\boldsymbol{x}(0)$ is defined as zero, $\boldsymbol{x}(1)$ takes the proper initial values, with $\boldsymbol{u}(t)$ arbitrary but finite for all $t \neq 0$, and $\delta(t)$ is the Kronecker sequence. $y_{O_t=s_k}(t) = y_k(t)$ corresponds to the k^{th} element of vector $\boldsymbol{y}(t)$. Here s_k is the symbol realized by O_t. The evaluation score for the TIA-HMM with respect to observations $\mathcal{O} = \{O_t\}_{t=1}^{T}$ is given by a simplified likelihood measure

$$\mathcal{L}(\mathcal{O} \mid \mathcal{M}) \overset{\text{def}}{=} \prod_{t=1}^{T} P(O_t \mid \mathcal{M}) = \prod_{t=1}^{T} y_{O_t}(t). \qquad (5)$$

The likelihood $\mathcal{L}(\cdot)$ inherently assumes independence of the elements O_t from the history of the state path, but retains pointwise dependency of O_t upon state residence at time t. As a result, the observation string $\mathcal{O}$ is more weakly linked to the state structure than is prescribed by the customary HMM assumptions.

As noted by Mitchell *et al.* [10], the TIA- HMM likelihood computation (5) can even include "illegal" paths (paths that violate the Bakis constraints). However, Lee [8] has analyzed the "illegal path" problem, posing analytical arguments for the relative insignificance and unlikely occurrence of this phenomenon. Notably, the strong diagonal dominance of the HMM state transition matrix, characteristic of the Bakis model, mitigates the potential problem.

Resource Benefits. The TIA-HMM comprises a compact and flexible structure whose dynamics are readily analyzed using decades of results on linear state-space systems. In particular, the system can (almost always) be transformed to have a diagonal state-transition matrix, thus reducing the number of operations to decode an observation string of length T from $\mathcal{O}\{N^{3/2}T\}$ (Bakis HMM) [or $\mathcal{O}\{N^2T\}$ (ergodic HMM)] to $\mathcal{O}\{NT\}$. The process effectively decouples the computations associated with the states, rendering each state computation is equivalent to a simple single pole filter. Further, since the poles for the entire population of states across all HMMs in a task are real, positive and bounded by unity, with most ≈ 1, many redundant "filter" structures exist both within and across HMMs. It is possible, therefore, to achieve significant computational savings through a process of merging statistically similar states across models. Depending on the circumstances, it is possible to control the trade-off between speech recognition performance and computing resource requirements by manipulating the number of merged states. In spite of its departure from the conventional HMM statistical assumptions, it has been shown, in limited experiments [9], to achieve very good recognition performance, even when the ability to "compress" models is exploited.

3. ENHANCING THE TIA-HMM

In the F-B algorithm, one of the main effects of the "coupling" of the state and observation probability computations is that, in forward decoding for example, observation O_t may not be attributed to state q_i, unless q_i or one of its predecessor states can produce each of the observations $\{O_\tau\}_{\tau=1}^{t-1}$. Computation of the joint probability of states and observation strings provides a "veto" mechanism to prevent state sequences that cannot legitimately produce the observed string. The assumptions of state-conditional dependence only, allows the F-B algorithm to, in principle, compute the joint probability of the entire set of observations up to any time t. Effectively, the F-B includes a check on appropriate time ordering of the observations.

Illegal paths can occur in TIA-HMM decoding because the state probabilities at time t are computed independently of the *past* observation string $\{O_\tau\}_{\tau=1}^{t-1}$. However, state probabilities are not without effect in the TIA-HMM. In fact, the state probabilities tend to modulate the conditional observation probabilities in a manner that makes it very unlikely (but not impossible) for a state with a low probability at time t to contribute to the likelihood of O_t. Thus, to the extent that the HMM training produces a viable state model, the illegal path problem is alleviated. To the extent that illegal paths remain problematic, however, the problem can be ascribed, as above, to a lack of attention to appropriate ordering of observation sequences. Thus, to include additional time-ordering information about the observations in a model, we supplement the TIA-HMM state equations with a stationary Markov chain model of successive observations. This enhancement is motivated a similar procedure due to Dai *et al.* [11].

In Dai's approach, the conventional F-B computations are complemented by a Markov chain in which the states are elements of a vector-quantized codebook of observation symbols.[2] The revised optimization criterion is to seek a HMM which produces a ML score in the conventional F-B HMM sense, in conjunction with the likelihood based on the Markovian relation between symbols. Similarly to Dai's approach, we pose a new criterion for TIA-HMM evaluation. In the revised method, the evaluation of a given TIA-HMM is based on the likelihood

$$\mathcal{L}'(\mathcal{O} \mid \mathcal{M}, \mathcal{M}') = \prod_{t=1}^{T} P(O_t \mid \mathcal{M}) P(\mathcal{O} \mid \mathcal{M}')$$

$$= \prod_{t=1}^{T} \left[P(O_t \mid \mathcal{M}) P(O_t \mid O_{t-1}, \mathcal{M}') \right] \qquad (6)$$

in which $\mathcal{M}'$ denotes the Markov chain. The negative log-likelihood is therefore

$$\mathcal{L}(\mathcal{O} \mid \mathcal{M}, \mathcal{M}') = \qquad (7)$$

$$-\sum_{t=1}^{T} \log \left[P(O_t \mid \mathcal{M}) P(O_t \mid O_{t-1}, \mathcal{M}') \right].$$

4. EXPERIMENTAL RESULTS

To analyze the performance of the augmented TIA-HMM evaluation, an isolated-word English alphabet recognition test was performed so that the results could be compared meaningfully with Dai's. A qualification is that the speech corpus in Dai's work is a British English corpus to which we did not have access, so that the standard TI-46 corpus [12] is employed. The selected corpus is composed of 26 utterances of each alphabetic character from each of 16 speakers, eight of each sex, for a total of 10,816 spoken utterances.

A speaker-independent recognition task used the entire corpus for four males and four females as training data, and the other eight persons' data for testing. Consequently, each HMM represents an alphabetic character trained by 208 utterances from eight speakers. An equal set of utterances from the other eight speakers was used for testing.

All data were sampled at 12.5kHz using 16-bit quantization. Tenth order mel-cepstral vectors were generated as spectral features over 256-point Hamming windows of the data. The LBG algorithm (e.g. [3]) with "center splitting" was used to create a static vector-quantized codebook of size 128. This codebook was used to quantize the speech utterances for training and testing data. In addition, ϵ-smoothing was applied to the Markov chain $\mathcal{M}'$. Each of the 26 discrete utterances was modeled using a five-state Bakis HMM with one forward skip allowed. The conventional F-B

[2] Dai's technique and the present work is focused on the discrete observation model [3], but each is generalizable to accommodate continuous observations.

algorithm was used to obtain HMM parameter matrices $\mathcal{M}$. For recognition, the Viterbi algorithm was applied for $P(\mathcal{O} \mid \mathcal{M})$ evaluation. The TIA-HMM state model was diagonalized for efficient computation of $\prod_{t=1}^{T} P(O_t)$ [9].

Recognition results for four different methods are shown in Table 1. In this table, "TMM" refers to the *temporal Markov model*, Dai's name for the supplementary model $\mathcal{M}'$ [11].

Model Form	Recognition Rate
HMM (F-B decoding)	82.4 %
HMM-TMM	85.7 %
TIA-HMM	80.7 %
TIA-HMM-TMM	84.4 %

Table 1. Recognition results with four different HMM evaluation methods for the TI-46 spoken-English alphabet corpus.

5. DISCUSSION AND CONCLUSIONS

To the extent that this relatively simple experiment may be used to draw general conclusions, several points are in evidence. First, the small degradation in performance between the conventional F-B HMM and the TIA-HMM, suggests that if observations are to be treated as nominally independent, then the state conditioning *might* produce a marginal improvement in certain tasks. Given the large number of acoustically confusable alphabet characters and the relatively short duration of most utterances in the present task,[3] this result is not entirely unexpected. However, in spite of its appeal as a "physically meaningful" configuration, whether the stronger state conditioning in the F-B case is enough to appropriately model dependencies in the observation string is called into question by considerable improvements in both F-B and TIA-HMM cases caused by the use of TMM. Apparently local correlations in the observation strings are more significant than the more global nonstationarities in the observation probabilities effected by state conditioning. This is especially remarkable in light of the fact that the the ϵ-smoothed Markov constraint between successive observation symbols could conceivably degrade recognition performance by effectively allowing "illegal paths," even in the F-B case.

All of these remarks are contingent upon the F-B training having produced models whose states are "meaningful" in terms of truly representing nontrivial clusters of statistically stationary observations, rather than simply representative of a structure that satisfies the ML optimization criterion. In the extreme case, for example, in which all but a few observations s_k are generated exclusively by the final state, then the Markov chain model $\mathcal{M}'$ would be of much greater significance than the "hidden" structure of the model. In this sense, the training paradigm suggested by Lee [8] could have a significant impact on these results. This issue will be pursued in future work.

Finally, the TMM augmentation renders the TIA-HMM a very competitive alternative to the conventional HMM. The benefits of the TIA-HMM, which have been discussed in [9] and further researched in [8], could be significant in many applications.

[3] This implies that the probabilities of inappropriate states might not have the opportunity to sufficiently damp out to prevent illegal paths in the TIA-HMM evaluation.

6. REFERENCES

[1] K. Reinhard *et al.*, "Optimization of HMMs for embedded systems," *Proc. Intl. Conf. Spoken Lang. Proc., Denver*, 2002, Pub'd on CD-ROM.

[2] S. Astrov, "Memory space reduction for HMMs . . . ," *Proc. Intl. Conf. Spoken Lang. Proc., Denver*, 2002, Pub'd on CD-ROM.

[3] J.R. Deller, Jr. *et al.*, *Discrete Time Processing of Speech Signals* (2d ed.), IEEE Press, 2000.

[4] T. Ogawa and T. Kobayashi, " . . . [S]tate observation dependency in partly HMMs," *Proc. Intl. Conf. Spoken Lang. Proc., Denver*, 2002, Pub'd on CD-ROM.

[5] S.J. Young and P.C. Woodland, "State clustering in HMM-based CSR," *Computer Speech & Lang.*, vol. 8, pp. 369–383, 1994.

[6] R. Cordóba *et al.*, "State clustering improvements . . . ," *Proc. Intl. Conf. Spoken Lang. Proc., Denver*, 2002, Pub'd on CD-ROM.

[7] S. Watanabe *et al.*, "Constructing shared state HMMs . . . ," *Proc. Intl. Conf. Spoken Lang. Proc., Denver*, 2002, Pub'd on CD-ROM.

[8] Y.B. Lee, *Evaluation and Improvement of the HMM by State-Space Modeling*, Ph.D. thesis, Michigan State Univ., 2000.

[9] J.R. Deller, Jr. and R.K. Snider, "Reducing . . . computation in HMM evaluation," *IEEE Trans. Speech & Audio*, vol. 1, pp. 465–471, 1993.

[10] C. Mitchell *et al.*, "Comments on [4]," *IEEE Trans. Speech & Audio*, vol. 2, pp. 542–543, 1994.

[11] J. Dai *et al.*, "Stochastic modeling of temporal information in speech for HMMs," *IEEE Trans. Speech & Audio*, vol. 2, pp. 285–348, 1994.

[12] G.R. Doddington and T.B. Schalk, "Speech recognition," *IEEE Spectrum*, vol. 18, pp. 26–32, 1981.