WHAT HMMS CAN'T DO

Jeff A. Bilmes

⟨bilmes@ee.washington.edu⟩
Dept. of Electrical Engineering, University of Washington
EE1-418, Box 352500, Seattle, WA 98195-2500, USA

ABSTRACT

Hidden Markov models (HMMs) are the predominant methodology for automatic speech recognition (ASR) systems. Ever since their inception, it has been said that HMMs are an inadequate statistical model for such purposes. Results over the years have shown, however, that HMM-based ASR performance continually improves given enough training data and engineering effort. In this paper, we argue that there are, in theory at least, no theoretical limitations to the class of probability distributions representable by HMMs. In search of a model to supersede the HMM for ASR, therefore, we should search for models with better parsimony, computational properties, noise insensitivity, and that better utilize high-level knowledge sources.

1. INTRODUCTION

More than any other statistical technique, the Hidden Markov model (HMM) has been most successfully applied to the ASR problem. Recent results have shown that they are remarkably good even for conversational text-to-speech [25] — the latest Switchboard word error rates are at around 13%. There are many HMM overviews [21, 8, 15, 10, 24]. In the classic paper [24], for example, an HMM is introduced in a generative way. For statistical speech recognition, one is not entirely concerned about how HMMs generate data, but instead, in how they discriminate between competitive utterances.

This paper argues that, within the paradigm offered by statistical pattern classification [9, 14], there is no theoretical limit to HMMs given enough hidden states, rich enough observation distributions, sufficient training data, adequate computation, and appropriate training algorithms. Instead, only a particular individual HMM used in a particular speech recognition system might be inadequate. This perhaps provides a reason for the continual speech-recognition accuracy improvements we have seen with HMM-based systems, and for the difficulty there has been in producing a model to supersede HMMs in all cases.

This paper does not argue, however, that HMMs should be the final technology for speech recognition. On the contrary, a main hope is that by examining a list of what HMMs can do, a better understanding of their limitations may be found so they ultimately will be abandoned in favor of a superior model. A main thrust should be searching for inherently more parsimonious models, ones that utilize knowledge, and that incorporate only the distinct properties of speech utterances relative to competing speech utterances with respect to this knowledge. The rest of this paper is thus devoted to what HMMs can do.

2. HIDDEN MARKOV MODELS

We begin immediately with the definition of an HMM.

Definition 2.1. Hidden Markov Model *A hidden Markov model (HMM) is collection of T discrete scalar random variables $Q_{1:T}$ and T other variables $X_{1:T}$ (either discrete or continuous, and either scalar- or vector-valued). These variables have a joint probability distribution $p(Q_{1:T}, X_{1:T})$ which factorizes with respect to the Bayesian network shown in Figure 1*

According to the rules of Bayesian network factorization [19], the definition implies the following conditional independence properties:

$$\{Q_{t:T}, X_{t:T}\} \perp\!\!\!\perp \{Q_{1:t-2}, X_{1:t-1}\} | Q_{t-1} \tag{1}$$

$$X_t \perp\!\!\!\perp \{Q_{\neg t}, X_{\neg t}\} | Q_t \tag{2}$$

for each $t \in 1 : T$, where $X_{\neg t} \triangleq X_{1:T} \setminus X_t$. The variable Q_t may take values in a finite set $\mathcal{Q}$ called the state space of the HMM, with has cardinality $|\mathcal{Q}|$.

Note that definition 2.1 does not limit the number of states $|\mathcal{Q}|$ in the Markov chain, does not require the observations $X_{1:T}$ to be either discrete, continuous, scalar-, or vector- valued, does not designate the implementation of the dependencies (e.g., general regression, probability table, neural network, etc.), does not determine the model families for each of the variables (e.g., Gaussian, Laplace, etc.), does not force the underlying Markov chain to be time-homogeneous, and does not fix the parameters or any tying mechanism. Summing over $Q_{1:T}$, we get a marginal distribution over $X_{1:T}$ that forms a general discrete time stochastic process with, as we will see, great flexibility. In fact,

$$p(x_{1:T}) = \sum_{q_{1:T}} \prod_t p(x_t|q_t)p(q_t|q_{t-1}) \tag{3}$$

This holds regardless of the form used for $p(x|q)$. The factorization properties of an HMM makes for extremely efficient computation of $p(x_{1:T})$ [24], and which is a special case of statistical inference on the Bayesian network shown in Figure 1 ([26, 24]).

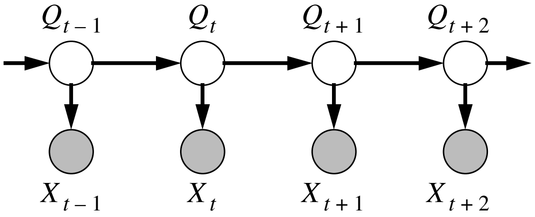

Fig. 1. A Hidden Markov Model

3. WHAT HMMS CAN DO

This HMM definition and Equations 1 and 2 can be used to better understand the capabilities of HMMs. In particular, it is possible to consider a particular quality in the context of conditional independence, in an effort to understand how and where that quality might apply, and its implications for using HMMs in a speech recognition system. The rest of this paper therefore compiles and then analyzes in detail a list of such qualities.

3.1. Observations i.i.d.

Given Equation 3, it can be seen that an HMM over $X_{1:T}$ is not i.i.d. since in general $p(x_{1:T})$ cannot factorize as $p(x_1)p(x_2)\ldots(x_T)$ unless only one state in the hidden Markov chain has non-zero probability for all times (which is never the case in practice).

3.2. Backwards-time influence

Equations (1) and (2) imply a large assortment of conditional independence statements including that the future is independent of the past given the present. The definition does not imply, given Q_{t-1}, that Q_t is unaffected by future variables. In fact, the distribution of Q_t could dramatically change, even given Q_{t-1}, when the variables X_τ or $Q_{\tau+1}$ change, for $\tau > t$.

3.3. Conditionally i.i.d. observations

HMMs *are* i.i.d. conditioned on certain state sequences since

$$p(x_{1:T}|q_{1:T}) = \prod_t p(x_t|q_t).$$

But this is not an inherent limitation. When sampling from an HMM, each sample will possess a different hidden Markov chain assignment. Unless one and only one state assignment has non-zero probability, the hidden state sequence will change with each sample. Therefore, the fact that HMMs are conditionally i.i.d. do not (necessarily) have repercussions when HMMs are actually used since HMM probabilities of a speech signal are obtained from the marginal distribution $p(x_{1:T})$, and not from the conditional distribution $p(X_{1:T}|Q_{1:T})$ where conditional i.i.d. holds.

3.4. Viterbi i.i.d.

HMMs are also not i.i.d. conditioned on the Viterbi path[24, 15], defined as follows:

$$q^*_{1:T} = \operatorname*{argmax}_{q_{1:T}} p(x_{1:T}, q_{1:T})$$

When using an HMM, often the joint distribution is taken as the effective Viterbi distribution:

$$p_{\text{vit}}(x_{1:T}) \propto p(x_{1:T}, q^*_{1:T})$$

Even under this approximation, however, the resulting distribution is not necessarily i.i.d. unless the Viterbi paths for all observation assignments are identical. Since the Viterbi path is typically different for each $x_{1:T}$, and the max operator does not commute with the product, $p_{\text{vit}}(x_{1:T})$ does not in general factorize.

3.5. Uncorrelated observations

Two observations at different times might be dependent, but are they correlated? If X_t and X_{t+h} are uncorrelated, then $E[X_t X'_{t+h}] = E[X_t]E[X_{t+h}]'$. Consider an HMM that has single component Gaussian observation distributions, i.e., $p(x|q) \sim \mathcal{N}(x|\mu_q, \Sigma_q)$ for all states q. Under these assumptions, and assuming $p(q)$ is currently at a stationary distribution π, and letting A be the matrix with $(i,j)^{th}$ entry $p(j|i)$, it can be shown [2] that cov(X_t, X_{t+h}) may be expressed as:

$$\sum_{ij} \mu_i \mu'_j (A^h)_{ij} \pi_i - \left(\sum_i \mu_i \pi_i\right)\left(\sum_i \mu_i \pi_i\right)'$$

a quantity that is not in general the zero matrix. Therefore HMMs, even as simple as ones that use single Gaussian observation distributions and under a stationary Markov chain, can capture correlation between feature vectors (see also [22]). To empirically demonstrate this correlation, the mutual information [7] in bits was computed between feature vectors from speech data sampled using 4-state per phone word HMMs trained from an isolated word task using MFCCs and their deltas [28]. Figure 2 compares the average pair-wise mutual information over time of this HMM with i.i.d. samples from a Gaussian mixture. The HMM clearly shows more correlation than the true i.i.d. process, since the HMM's hidden variables indirectly encode this information, and as the number of hidden states increases, so does the amount of information that can be encoded.

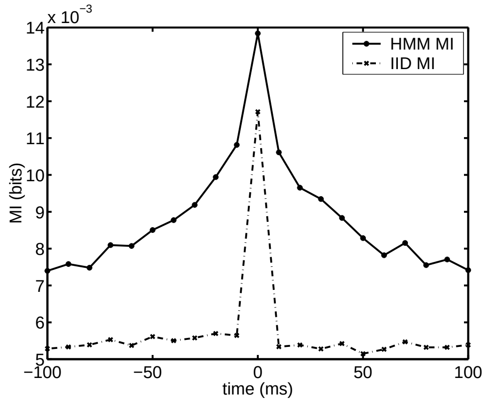

Fig. 2. HMM frame correlation vs. true i.i.d. process.

3.6. Piece-wise or segment-wise stationary

An HMM is stationary over $X_{1:T}$ whenever

$$p(X_{t_{1:n}+h} = x_{1:n}) = p(X_{t_{1:n}} = x_{1:n})$$

for all n, h, $t_{1:n}$, and $x_{1:n}$. It can be shown [2] that this conditional holds when $p(Q_{t_1+h} = q_1) = p(Q_{t_1} = q_1)$ for all h. Therefore, the HMM is stationary only when the underlying hidden Markov chain is stationary, even when the Markov chain is time-homogeneous. An HMM therefore does not necessarily correspond to a stationary distribution.

For ASR, HMMs commonly have non-ergodic left-to-right state-transition topologies where transition matrices are upper triangular ($P(j|i) = 0 \ \ \forall j < i$). More strongly, any left-to-right HMM is not stationary unless all non-final states have zero probability [2]. HMMs are also unlikely to be "piece-wise" stationary, in which an HMM is in a particular state for a time segment and where observations in that time segment are i.i.d. and therefore stationary. Recall, each HMM sample uses a separate sample from the hidden Markov chain. As a result, a segment (a sequence of identical state assignments to successive hidden variables) in the hidden chain of one HMM sample will not necessarily be a segment in the chain of a different sample. Therefore, HMMs are not stationary unless either 1) every HMM sample always results in the same hidden assignment for some fixed-time region, or 2) the hidden chain is always stationary over that region. With standard left-to-right HMMs, neither is true in practice.

3.7. Within-frame stationary

Speech is a band-limited continuous-time signal. A feature extraction process generates speech frames at regular time intervals (e.g., 10ms) over a window (e.g., 25ms). An HMM then characterizes the distribution over this discrete-time set of frame vectors. Might HMMs have trouble representing speech because information encoded by within-frame variation is lost via the framing of speech? This also is unlikely to produce problems since the properties of speech that convey any message are band-limited in the modulation domain, and if the rate of hidden state change is high enough, and if the frame-window width is small enough, framing of speech would not result in any information loss.

3.8. Geometric state distributions

In a Markov chain, the time duration D that a specific state i is active is a random variable distributed according to a geometric distribution. That is, D has distribution $P(D = d) = p^{d-1}(1 - p)$ where $d \geq 1$. It seems possible that HMMs might be deficient because their state duration distributions are inherently geometric, and geometric distributions do not accurately represent typical speech unit (e.g., phoneme or syllable) durations.[1]

HMMs, however, do not necessarily have such problems, and this occurs because of "state-tying", where multiple different states can share the same observation distribution. In general, a collection of HMM states sharing the same observation distribution may be combined in both series and parallel. If a sequence of n states using the same observation distribution are strung together in series, and each of the states has self transition probability α, then the resulting distribution (the sum of n independent geometrically distributed random variables) has a negative binomial distribution (a discrete version of the gamma distribution) [27]. Unlike a geometric distribution, a negative binomial distribution has a mode located away from zero. When combined in parallel, the resulting distribution is a weighted mixture of the individual distributions. This process can of course be combined (see Figure 3) and repeated recursively as well. Therefore, simply by increasing the hidden state space cardinality, an HMM can produce an broad class of speech-unit duration distributions.

[1] It has been suggested that a gamma distribution is a more appropriate speech-unit durational distribution[20].

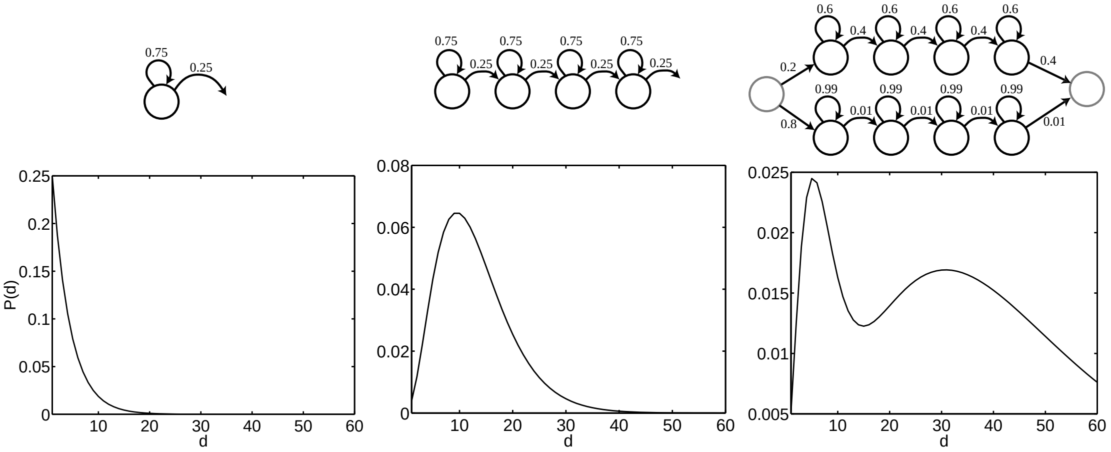

Fig. 3. Three speech-unit duration distributions with an HMM, and their respective Markov chain topologies.

3.9. First-order hidden Markov assumption

As described in [16], any n^{th}-order Markov chain may be transformed into a first-order chain. Therefore, assuming a first-order Markov chain possess a sufficient states, there is no inherent fidelity loss when using a first-order as opposed to an n^{th}-order HMM.

3.10. Conditions for HMM Accuracy

Suppose that $p(X_{1:T})$ is the true distribution of the observation variables $X_{1:T}$. If an HMM must represent this distribution accurately, necessary conditions on the number of hidden states and the necessary complexity of the observation distributions may be found. Let $p_h(X_{1:T})$ be the joint distribution over the observation variables under an HMM. HMM accuracy can be defined as KL-divergence [7] between the two distributions being zero. The following theorem is proven in [2].

Theorem 3.1. Necessary conditions for HMM accuracy. *An HMM with joint distribution $p_h(X_{1:T})$ will accurately model the true distribution $p(X_{1:T})$ (zero KL-divergence) only if the following three conditions hold for all t:*

- $I_h(X_{\neg t}; Q_t) \geq I(X_t; X_{\neg t})$,
- $I_h(Q_t; X_t) \geq I(X_t; X_{\neg t})$, *and*
- $|\mathcal{Q}| \geq 2^{I(X_t;X_{\neg t})}$

where $I_h(X_{\neg t}; Q_t)$ (resp. $I_h(Q_t; X_t)$) is the mutual information between $X_{\neg t}$ and Q_t (resp. Q_t and X_t) under an HMM, and $I(X_t; X_{\neg t})$ is the true mutual information between $I(X_t; X_{\neg t})$.

Since Q_t is the "channel" representation through which information about the past and future must travel, too few states can overburden the hidden variables and lead to inaccuracies. If there are enough states, and if the information in the surrounding acoustic context is appropriately encoded in the hidden states, the information may be compressed and represented by Q_t. To achieve high accuracy, it is likely that a finite number of states is required since signals representing natural real-world phenomena will have bounded mutual information. Can sufficient conditions for HMM accuracy be found? An initial attempt is given in [2], but more results are needed. It moreover remains to be seen if simultaneously necessary and sufficient conditions can be derived for HMM accuracy, if it is possible to derive sufficient conditions for continuous observation vector HMMs under reasonable conditions (e.g., finite power), and what conditions might exist for an HMM that is allowed to have a fixed upper-bound KL-divergence error.

3.11. Synthesis vs. Recognition

Given a speech utterance M, an HMM for M is a representation of the distribution $p(X_{1:T}|M)$ which can be viewed as a synthesis or generative model because sampling from this distribution should produce (or synthesize) an instance of the a speech sound. For ASR, however, one instead desires $P(M|X)$ which is a recognition or discriminative model since such a quantity achieves Bayes error. That HMMs inherently represent $p(X|M)$, however, is less restrictive than what might initially appear.

First, $p(M|X) = p(X|M)p(M)/p(X)$ so if an HMM accurately represents $p(X|M)$ and with accurate $P(M)$, an accurate posterior will ensue. Approximating a distribution such as $p(X|M)$ might require more effort (parameters, training data, and compute time) than necessary to achieve good classification accuracy. Representing the entire set of class conditional distributions $p(X|M)$, which includes regions between decision boundaries, is more difficult than necessary to achieve good performance. One may instead produce any approximating distribution $p_h(X|M)$ that achieves the same Bayes error, so classification accuracy will not be compromised. A sample from such a conditional distribution will not necessarily result in a "good" speech utterance, but this is of no consequence to classification accuracy. Moreover, using a simpler model $p_h(X|M)$ can have statistical parameter estimation benefits as well [13]. Under this view, an HMM is the temporal analog of the Naive Bayes classifier.

Moreover, even under a generative model, the degree to which decision boundary information is represented by an HMM depends on the parameter training method. Discriminative training methods have long ago been developed (and are currently usefully employed [25]) that adjust the parameters of each model to increase not the individual likelihood but rather approximate the posterior probability or Bayes decision rule [1, 6, 11, 12, 18, 17, 23]. Apart from the training method, the degree to which boundary information is represented can depend on each model's intrinsic ability to produce an accurate distribution at decision boundaries vs. its ability to represent the regions between boundaries. This is the inherent discriminability of the structure of the model for each class, independent of its parameters, a property that has been called structurally discriminative [5].

How structurally discriminative are HMMs when attempting to model the distinctive attributes of speech utterances? Certainly, the left-to-right Markov chain topology helps significantly, since there is much discriminative information in the hard sequencing properties of such HMMs. At the very least, HMMs are not structurally indiscriminate because, even when trained using maximum likelihood procedure, HMM-based speech recognition systems perform reasonably well. Moreover, the structure of an HMM might be further adjusted to improve discriminability [3]. Earlier sections of this paper suggested that HMM distributions are not lacking in their flexibility, but this section claims that for recognition, HMM need not even accurately represent the true likelihood $p(X|M)$ to achieve high classification accuracy. While HMMs are powerful, a fortunate consequence of the current discussion is that HMMs need not capture many of the nuances in a speech signal, and are thus allowed to be simpler still as a result. In other words, just because a particular HMM does not accurately represent speech does not mean it is poor at the recognition task.

4. WHAT HMMS CAN'T DO

From the above, there appears to be little an HMM can not do. One problem might be the way an HMM is used: a particular HMM trained in a particular way might be inaccurate due to too too few hidden states, weak observation distributions, or when it is trained non-discriminatively or without sufficient training data. These problems, of course, can be corrected still staying within the HMM framework. Considering also an HMM's extremely good computational properties, it is in fact be a difficult model to surpass. In our quest for a new model for speech recognition, therefore, we should be concerned not with what is wrong with HMMs, and rather seek a model that is inherently more parsimonious, more intrinsically discriminative, equally computationally tractable, and where knowledge-rich speech and language representations can be much more easily incorporated. It is believed by this author that graphical models [19, 4] satisfy all of the above requirements and are thus most likely, once computational frameworks are readily available, to eventually overtake the HMM.

5. REFERENCES

[1] L.R. Bahl, P.F. Brown, P.V. de Souza, and R.L. Mercer. Maximum mutual information estimation of HMM parameters for speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 49–52, Tokyo, Japan, December 1986.

[2] J. Bilmes. What HMMs can do. Technical Report UWEETR-2002-003, University of Washington, Dept. of EE, 2002.

[3] J. Bilmes. Buried markov models: A graphical modeling approach to automatic speech recognition. *Computer Speech and Language*, 17:213—231, April—July 2003.

[4] J. A. Bilmes. Graphical models and automatic speech recognition. In R. Rosenfeld, M. Ostendorf, S. Khudanpur, and M. Johnson, editors, *Mathematical Foundations of Speech and Language Processing*. Springer-Verlag, New York, 2003.

[5] J.A. Bilmes. Dynamic Bayesian Multinets. In *Proceedings of the 16th conf. on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 2000.

[6] P.F. Brown. *The Acoustic Modeling Problem in Automatic Speech Recognition*. PhD thesis, Carnegie Mellon University, 1987.

[7] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[8] S.J. Cox. Hidden Markov Models for automatic speech recognition: Theory an d application. In C. Wheddon and R. Linggard, editors, *Speech and Language Processing*, pages 209–230, 1990.

[9] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, Inc., 1973.

[10] Y. Ephraim. Hidden markov processes. *IEEE Trans. Info. Theory*, 48(6):1518–1569, June 2002.

[11] Y. Ephraim, A. Dembo, and L. Rabiner. A minimum discrimination information approach for HMM. *IEEE Trans. Info. Theory*, 35(5):1001–1013, September 1989.

[12] Y. Ephraim and L. Rabiner. On the relations between modeling approaches for speech recognition. *IEEE Trans. Info. Theory*, 36(2):372–380, September 1990.

[13] J.H. Friedman. On bias, variance, 0/1–loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997.

[14] K. Fukunaga. *Introduction to Statistical Pattern Recognition, 2nd Ed.* Academic Press, 1990.

[15] X.D. Huang, Y. Ariki, and M. Jack. *Hidden Markov Models for Speech Recognition*. Edinburgh University Press, 1990.

[16] F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1997.

[17] B.-H. Juang, W. Chou, and C.-H. Lee. Minimum classification error rate methods for speech recognition. *IEEE Trans. on Speech and Audio Signal Processing*, 5(3):257–265, May 1997.

[18] B-H Juang and S. Katagiri. Discriminative learning for minimum error classification. *IEEE Trans. on Signal Processing*, 40(12):3043–3054, December 1992.

[19] S.L. Lauritzen. *Graphical Models*. Oxford Science Publications, 1996.

[20] S.E. Levinson. Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*, I:29–45, 1986.

[21] S.E. Levinson, L.R. Rabiner, and M.M. Sondhi. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *The Bell System Technical Journal*, pages 1035–1073, 1983.

[22] I.L. MacDonald and W. Zucchini. *Hidden Markov and Other Models for Discrete-valued Time Series*. Chapman and Hall, 1997.

[23] D. Povey and P.C. Woodland. Minimum phone error and I-smoothing for improved discriminative training. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 2002.

[24] L.R. Rabiner and B.H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 1986.

[25] EARS rich transcription 2004 (RT-04) workshop, November 2004.

[26] P. Smyth, D. Heckerman, and M.I. Jordan. Probabilistic independence networks for hidden Markov probability models. Technical Report A.I. Memo No. 1565, C.B.C.L. Memo No. 132, MIT AI Lab and CBCL, 1996.

[27] D. Stirzaker. *Elementary Probability*. Cambridge, 1994.

[28] S. Young. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine*, 13(5):45–56, September 1996.