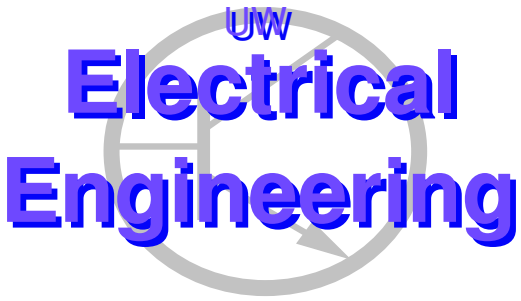
What HMMs Can Do

Jeff Bilmes
bilmes@ee.washington.edu

Dept of EE, University of Washington
Seattle WA, 98195-2500



UWEE Technical Report
Number UWEETR-2002-0003
January 2002

Department of Electrical Engineering
University of Washington
Box 352500
Seattle, Washington 98195-2500
PHN: (206) 543-2150
FAX: (206) 543-3842
URL: http://www.ee.washington.edu

What HMMs Can Do

Jeff Bilmes
bilmes@ee.washington.edu

Dept of EE, University of Washington
Seattle WA, 98195-2500

University of Washington, Dept. of EE, UWEETR-2002-0003
January 2002

Abstract

Since their inception over thirty years ago, hidden Markov models (HMMs) have have become the predominant methodology for automatic speech recognition (ASR) systems — today, most state-of-the-art speech systems are HMM-based. There have been a number of ways to explain HMMs and to list their capabilities, each of these ways having both advantages and disadvantages. In an effort to better understand what HMMs can do, this tutorial analyzes HMMs by exploring a novel way in which an HMM can be defined, namely in terms of random variables and conditional independence assumptions. We prefer this definition as it allows us to reason more throughly about the capabilities of HMMs. In particular, it is possible to deduce that there are, in theory at least, no theoretical limitations to the class of probability distributions representable by HMMs. This paper concludes that, in search of a model to supersede the HMM for ASR, we should rather than trying to correct for HMM limitations in the general case, new models should be found based on their potential for better parsimony, computational requirements, and noise insensitivity.

1 Introduction

By and large, automatic speech recognition (ASR) has been approached using statistical pattern classification [29, 24, 36], mathematical methodology readily available in 1968, and summarized as follows: given data presumably representing an unknown speech signal, a statistical model of one possible spoken utterance (out of a potentially very large set) is chosen that most probably explains this data. This requires, for each possible speech utterance, a model governing the set of likely acoustic conditions that could realize each utterance.

More than any other statistical technique, the Hidden Markov model (HMM) has been most successfully applied to the ASR problem. There have been many HMM tutorials [69, 18, 53]. In the widely read and now classic paper [86], an HMM is introduced as a collection of urns each containing a different proportion of colored balls. Sampling (generating data) from an HMM occurs by choosing a new urn based on only the previously chosen urn, and then choosing with replacement a ball from this new urn. The sequence of urn choices are not made public (and are said to be "hidden") but the ball choices are known (and are said to be "observed"). Along this line of reasoning, an HMM can be defined in such a generative way, where one first generates a sequence of hidden (urn) choices, and then generates a sequence of observed (ball) choices.

For statistical speech recognition, one is not only worried in how HMMs generate data, but also, and more importantly, in an HMMs distributions over observations, and how those distributions for different utterances compare with each other. An alternative view of HMMs, therefore and as presented in this paper, can provide additional insight into what the capabilities of HMMs are, both in how they generate data and in how they might recognize and distinquish between patterns.

This paper therefore provides an up-to-date HMM tutorial. It gives a precise HMM definition, where an HMM is defined as a variable-size collection of random variables with an appropriate set of conditional independence properties. In an effort to better understand what HMMs can do, this paper also considers a list of properties, and discusses how they each might or might not apply to an HMM. In particular, it will be argued that, at least within the paradigm

offered by statistical pattern classification [29, 36], there is no general 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 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.

This paper does not argue, however, that HMMs should be the final technology for speech recognition. On the contrary, a main hope of this paper is to offer a better understanding of what HMMs can do, and consequently, a better understanding of their limitations so they may ultimately be abandoned in favor of a superior model. Indeed, HMMs are extremely flexible and might remain the preferred ASR method for quite some time. For speech recognition research, however, a main thrust should be searching for inherently more parsimonious models, ones that incorporate only the distinct properties of speech utterances relative to competing speech utterances. This later property is termed structural discriminability [8], and refers to a generative model's inherent inability to represent the properties of data common to every class, even when trained using a maximum likelihood parameter estimation procedure. This means that even if a generative model only poorly represents speech, leading to low probability scores, it may still properly classify different speech utterances. These models are to be called discriminative generative models.

Section 2 reviews random variables, conditional independence, and graphical models (Section 2.1), stochastic processes (Section 2.2), and discrete-time Markov chains (Section 2.3). Section 3 provides a formal definition of an HMM, that has both a generative and an "acceptive" point of view. Section 4 compiles a list of properties, and discusses how they might or might not apply to HMMs. Section 5 derives conditions for HMM accuracy in a Kullback-Leibler distance sense, proving a lower bound on the necessary number of hidden states. The section derives sufficient conditions as well. Section 6 reviews several alternatives to HMMs, and concludes by presenting an intuitive criterion one might use when researching HMM alternatives

1.1 Notation

Measure theoretic principles are avoided in this paper, and discrete and continuous random variables are distinguished only where necessary. Capital letters (e.g., X, Q) will refer to random variables, lower case letters (e.g., x, q) will refer to values of those random variables, and script letters (e.g., $\mathcal{X}$, $\mathcal{Q}$) will refer to possible values so that $x \in \mathcal{X}$, $q \in \mathcal{Q}$. If X is distributed according to p, it will be written $X \sim p(X)$. Probabilities are denoted $p_X(X = x)$, $p(X = x)$, or $p(x)$ which are equivalent. For notational simplicity, $p(x)$ will at different times symbolize a continuous probability density or a discrete probability mass function. The distinction will be unambiguous when needed.

It will be necessary to refer to sets of integer indexed random variables. Let $A \triangleq \{a_1, a_2, \ldots, a_N\}$ be a set of T integers. Then $X_A \triangleq \{X_{a_1}, X_{a_2}, \ldots, X_{a_T}\}$. If $B \subset A$ then $X_B \subset X_A$. It will also be useful to define sets of integers using matlab-like ranges. As such, $X_{i:j}$ with $i < j$ will refer to the variables $X_i, X_{i+1}, \ldots, X_j$. $X_{<i} \triangleq \{X_1, X_2, \ldots, X_{i-1}\}$, and $X_{\neg t} \triangleq X_{1:T} \setminus X_t = \{X_1, X_2, \ldots, X_{t-1}, X_{t+1}, X_{t+2}, \ldots, X_T\}$ where T will be clear from the context, and $\setminus$ is the set difference operator. When referring to sets of T random variable, it will also be useful to define $X \triangleq X_{1:T}$ and $x \triangleq x_{1:T}$. Additional notation will be defined when needed.

2 Preliminaries

Because within an HMM lies a hidden Markov chain which in turn contains a sequence of random variables, it is useful to review a few noteworthy prerequisite topics before beginning an HMM analysis. Some readers may wish to skip directly to Section 3. Information theory, while necessary for a later section of this paper, is not reviewed and the reader is referred to the texts [16, 42].

2.1 Random Variables, Conditional Independence, and Graphical Models

A random variable takes on values (or in the continuous case, a range of values) with certain probabilities.[1] Different random variables might or might not have the ability to influence each other, a notion quantified by statistical independence. Two random variables X and Y are said to be (marginally) statistically independent if and only if

[1] In this paper, explanations often use discrete random variables to avoid measure theoretic notation needed in the continuous case. See [47, 103, 2] for a precise treatment of continuous random variables. Note also that random variables may be either scalar or vector valued.

$p(X = x, Y = y) = p(X = x)p(Y = y)$ for every value of x and y. This is written $X \perp\!\!\!\perp Y$. Independence implies that regardless of the outcome of one random variable, the probabilities of the outcomes of the other random variable stay the same.

Two random variables might or might not be independent of each other depending on knowledge of a third random variable, a concept captured by conditional independence. A random variable X is conditionally independent of a different random variable Y given a third random variable Z under a given probability distribution $p(\cdot)$, if the following relation holds:

$$p(X = x, Y = y|Z = z) = p(X = x|Z = z)p(Y = y|Z = z)$$

for all x, y, and z. This is written $X \perp\!\!\!\perp Y|Z$ and it is said that "X is independent of Y given Z under $p(\cdot)$". An equivalent definition is $p(X = x|Y = y, Z = z) = p(X = x|Z = z)$. The conditional independence of X and Y given Z has the following intuitive interpretation: if one has knowledge of Z, then knowledge of Y does not change one's knowledge of X and vice versa. Conditional independence is different from unconditional (or marginal) independence. Therefore, it might be true that $X \perp\!\!\!\perp Y$ but not true that $X \perp\!\!\!\perp Y|Z$. One valuable property of conditional independence follows: if $X_A \perp\!\!\!\perp Y_B|Z_C$, and subsets $A' \subset A$ and $B' \subset B$ are formed, then it follows that $X_{A'} \perp\!\!\!\perp Y_{B'}|Z_C$. Conditional independence is a powerful concept — when assumptions are made, a statistical model can undergo enormous simplifications. Additional properties of conditional independence are presented in [64, 81].

When reasoning about conditional independence among collections of random variables, graphical models [102, 64, 17, 81, 56] are very useful. Graphical models are an abstraction that encompasses an extremely large set of statistical ideas. Specifically, a graphical model is a graph $\mathcal{G} = (V, E)$ where V is a set of vertices and the set of edges E is a subset of the set $V \times V$. A particular graphical model is associated with a collection of random variables and a family of probability distributions over that collection. The vertex set V is in one-to-one correspondence with the set of random variables. In general, a vertex can correspond either to a scalar- or a vector-valued random variable. In the latter case, the vertex implicitly corresponds to a sub-graphical model over the individual elements of the vector. The edge set E of the model in one way or another specifies a set of conditional independence properties of the random variables that are true for every the member of the associated family. There are different types of graphical models. The set of conditional independence assumptions specified by a graphical model, and therefore the family of probability distributions it constitutes, depends on its type.

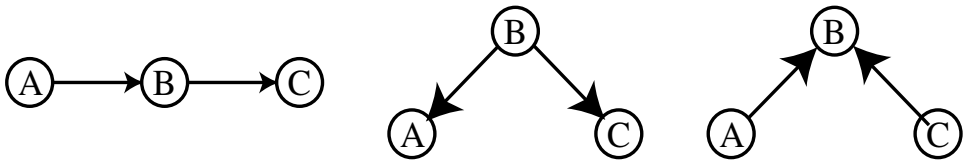


Figure 1: Like any graphical model, the edges in a DGM determine the conditional independence properties over the corresponding variables. For a DGM, however, the arrow directions make a big difference. The figure shows three networks with different arrow directions over the same random variables, A, B, and C. On the left side, the variables form a three-variable first-order Markov chain $A \rightarrow B \rightarrow C$ (see Section 2.3). In the middle graph, the same conditional independence property is realized although one of the arrows is pointing in the opposite direction. Both these networks correspond the property $A \perp\!\!\!\perp C|B$. These two networks do not, however, insist that A and B are not independent. The right network corresponds to the property $A \perp\!\!\!\perp C$ but it does not imply that $A \perp\!\!\!\perp C|B$.

A directed graphical model (DGM) [81, 56, 48], also called a Bayesian network, is only one type of graphical model. In this case, the graph is directed and acyclic. In a DGM, if an edges is directed from node A towards node B, then A is a parent of B and B is a child of A. One may also discuss ancestors, descendants, etc. of a node. A Dynamic Bayesian Network (DBN) [43, 108, 34] is one type of DGM containing edges pointing in the direction of time. There are several equivalent schemas that may serve to formally define the conditional independence relationships implied by a DGM[64]. This includes d-separation [81, 56], the directed local Markov property [64] (which states that a variable is conditionally independent of its non-descendants given its parents), and the Bayes-ball procedure [93] (which perhaps the easiest to understand and is therefore described in Figure 2).

An undirected graphical model (often called a Markov random field [23]) is one where conditional independence among the nodes is determined simply by graph separation, and therefore has a easier semantics than DGMs. The family of distributions associated with DGMs is different from the family associated with undirected models, but the

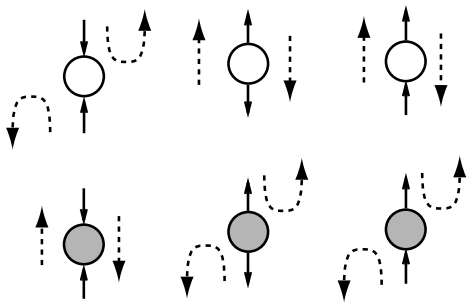


Figure 2: The Bayes-ball procedure makes it easy to answer questions about a DGM such as "is $X_A \perp\!\!\!\perp X_B | X_C$?", where A, B, and C are disjoint sets of node indices. First, shade every node having indices in C and imagine a ball bouncing from node to node along the edges in a graph. The answer to the above question is true if and only if a ball starting at some node in A can reach a node in B, when the ball bounces according to the rules depicted in the figure. The dashed arrows depict whether a ball, when attempting to bounce through a given node, may bounce through that node or if it must bounce back.

intersection of the two families is known as the decomposable models [64]. Other types of graphical models include causal models [82], chain graphs [64], and dependency networks [49].

Nodes in a graphical model can be either *hidden*, which means they have unknown value and signify a true random variable, or they can be *observed*, which means that the values are known. In fact, HMMs are so named because they possess a Markov chain that is hidden. A node may at different times be either hidden or observed, and for different reasons. For example, if one asks "what is the probability $p(C = c|A = a)$?" for the left graph in Figure 1, then B is hidden and A is observed. If instead one asks "what is the probability $p(C = c|B = b)$ or $p(A = a|B = b)$?" then B is observed. A node may be hidden because of missing values of certain random variables in samples from a database. Moreover, when the query "is $A \perp\!\!\!\perp B|C$?" is asked of a graphical model, it is implicitly assumed that A and B are hidden and C is observed. In general, if the value is known (i.e., if "evidence" has been supplied) for a node, then it is considered observed — otherwise, it is considered hidden.

A key problem with graphical models is that of computing the probability of one subset of nodes given values of some other subset, a procedure called probabilistic inference. Inference using a network containing hidden variables must "marginalize" them away. For example, given $p(A, B, C)$, the computation of $p(a|c)$ may be performed as:

$$p(a|c) = \frac{p(a, c)}{p(c)} = \frac{\sum_b p(a, b, c)}{\sum_{a,b} p(a, b, c)}$$

in which b has been marginalized (or integrated) away in the numerator. Inference is essential both to make predictions and to learn the network parameters with, say, the EM algorithm [20].

In this paper, graphical models will help explicate the HMM conditional independence properties. An additional important property of graphical models, however, is that they supply more efficient inference procedures [56] than just, ignoring conditional independence, marginalizing away all unneeded and hidden variables. Inference can be either exact, as in the popular junction tree algorithm [56] (of which the Forward-Backward or Baum-Welch algorithm [85, 53] is an example [94]), or can be approximate [91, 54, 57, 72, 100] since in the general case inference is NP-Hard [15].

Examples of graphical models include mixture models (e.g., mixtures of Gaussians), decision trees, factor analysis, principle component analysis, linear discriminant analysis, turbo codes, dynamic Bayesian networks, multi-layered perceptrons (MLP), Kalman filters, and (as will be seen) HMMs.

2.2 Stochastic Processes, Discrete-time Markov Chains, and Correlation

A discrete-time stochastic process is a collection $\{X_t\}$ for $t \in 1{:}T$ of random variables ordered by the discrete time index t. In general, the distribution for each of the variables X_t can be arbitrary and different for each t. There may also be arbitrary conditional independence relationships between different subsets of variables of the process — this corresponds to a graphical model with edges between all or most nodes.

Certain types of stochastic processes are common because of their analytical and computational simplicity. One example follows:

Definition 2.1. Independent and Identically Distributed (i.i.d.) *The stochastic process is said to be i.i.d.[16, 80, 26] if the following condition holds:*

$$p(X_t = x_t, X_{t+1} = x_{t+1}, \ldots, X_{t+h} = x_{t+h}) = \prod_{i=0}^{h} p(X = x_{t+i}) \tag{1}$$

for all t, for all $h \geq 0$, for all $x_{t:t+h}$, and for some distribution $p(\cdot)$ that is independent of the index t.

An i.i.d. process therefore comprises an ordered collection of independent random variables each one having exactly the same distribution. A graphical model of an i.i.d process contains no edges at all.

If the statistical properties of variables within a time-window of a stochastic process do not evolve over time, the process is said to be stationary.

Definition 2.2. Stationary Stochastic Process *The stochastic process $\{X_t : t \geq 1\}$ is said to be (strongly) stationary [47] if the two collections of random variables*

$$\{X_{t_1}, X_{t_2}, \ldots, X_{t_n}\}$$

and

$$\{X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_n+h}\}$$

have the same joint probability distributions for all n and h.

In the continuous case, stationarity means that $F_{X_{t_{1:n}}}(a) = F_{X_{t_{1:n}+h}}(a)$ for all a where $F(\cdot)$ is the cumulative distribution and a is a valid vector-valued constant of length n. In the discrete case, stationarity is equivalent to the condition

$$P(X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_n} = x_n) = P(X_{t_1+h} = x_1, X_{t_2+h} = x_2, \ldots, X_{t_n+h} = x_n)$$

for all $t_1, t_2, \ldots, t_n$, for all $n > 0$, for all $h > 0$, and for all x_i. Every i.i.d. processes is stationary.

The covariance between two random vectors X and Y is defined as:

$$\text{cov}(X, Y) = E[(X - EX)(Y - EY)'] = E(XY') - E(X)E(Y)'$$

It is said that X and Y are uncorrelated if $\text{cov}(X, Y) = \vec{0}$ (equivalently, if $E(XY') = E(X)E(Y)'$) where $\vec{0}$ is the zero matrix. If X and Y are independent, then they are uncorrelated, but not vice versa unless they are jointly Gaussian [47].

2.3 Markov Chains

A collection of discrete-valued random variables $\{Q_t :\geq 1\}$ forms an n^{th}-order Markov chain [47] if

$$\begin{aligned} &P(Q_t = q_t | Q_{t-1} = q_{t-1}, Q_{t-2} = q_{t-2}, \ldots, Q_1 = q_1) \\ &= \quad P(Q_t = q_t | Q_{t-1} = q_{t-1}, Q_{t-2} = q_{t-2}, \ldots, Q_{t-n} = q_{t-n}) \end{aligned}$$

for all $t \geq 1$, and all $q_1, q_2, \ldots, q_t$. In other words, given the previous n random variables, the current variable is conditionally independent of every variable earlier than the previous n. A first order Markov chain is depicted using the left network in Figure 1.

One often views the event $\{Q_t = i\}$ as if the chain is "in state i at time t" and the event $\{Q_t = i, Q_{t+1} = j\}$ as a transition from state i to state j starting at time t. This notion arises by viewing a Markov chain as a finite-state automata (FSA) [52] with probabilistic state transitions. In this case, the number of states corresponds to the cardinality of each random variable. In general, a Markov chain may have infinitely many states, but chain variables in this paper are assumed to have only finite cardinality.

An n^{th}-order Markov chain may always be converted into an equivalent first-order Markov chain [55] using the following procedure:

$$Q'_t \triangleq \{Q_t, Q_{t-1}, \ldots, Q_{t-n}\}$$

where Q_t is an n^{th}-order Markov chain. Then Q'_t is a first-order Markov chain because

$$\begin{aligned} & P(Q'_t = q'_t | Q'_{t-1} = q'_{t-1}, Q'_{t-2} = q'_{t-2}, \ldots, Q'_1 = q'_1) \\ = \;& P(Q_{t-n:t} = q_{t-n:t} | Q_{1:t} = q_{1:t}) \\ = \;& P(Q_{t-n:t} = q_{t-n:t} | Q_{t-n-1:t} = q_{t-n-1:t}) \\ = \;& P(Q'_t = q'_t | Q'_{t-1} = q'_{t-1}) \end{aligned}$$

This transformation implies that, given a large enough state space, a first-order Markov chain may represent any n^{th}-order Markov chain.

The statistical evolution of a Markov chain is determined by the state transition probabilities $a_{ij}(t) \triangleq P(Q_t = j | Q_{t-1} = i)$. In general, the transition probabilities can be a function both of the states at successive time steps and of the current time t. In many cases, it is assumed that there is no such dependence on t. Such a time-independent chain is called time-homogeneous (or just homogeneous) because $a_{ij}(t) = a_{ij}$ for all t.

The transition probabilities in a homogeneous Markov chain are determined by a transition matrix A where $a_{ij} \triangleq (A)_{ij}$. The rows of A form potentially different probability mass functions over the states of the chain. For this reason, A is also called a stochastic transition matrix (or just a transition matrix).

A state of a Markov chain may be categorized into one of three distinct categories [47]. A state i is said to be *transient* if, after visiting the state, it is possible for it never to be visited again, i.e.,:

$$p(Q_n = i \text{ for some } n > t | Q_t = i) < 1.$$

A state i is said to be *null-recurrent* if it is not transient but the expected return time is infinite (i.e., $E[\min\{n > t : Q_n = i\} | Q_t = i] = \infty$). Finally, a state is *positive-recurrent* if it is not transient and the expected return time to that state is finite. For a Markov chain with a finite number of states, a state can only be either transient or positive-recurrent.

Like any stochastic process, an individual Markov chain might or might not be a stationary process. The stationarity condition of a Markov chain, however, depends on 1) if the Markov chain transition matrix has (or "admits") a stationary distribution or not, and 2) if the current distribution over states is one of those stationary distributions.

If Q_t is a time-homogeneous stationary Markov chain then:

$$P(Q_{t_1} = q_1, Q_{t_2} = q_2, \ldots, Q_{t_n} = q_n) = P(Q_{t_1+h} = q_1, Q_{t_2+h} = q_2, \ldots, Q_{t_n+h} = q_n)$$

for all t_i, h, n, and q_i. Using the first order Markov property, the above can be written as:

$$\begin{aligned} & P(Q_{t_n} = q_n | Q_{t_{n-1}} = q_{n-1}) P(Q_{t_{n-1}} = q_{n-1} | Q_{t_{n-2}} = q_{n-2}) \ldots \\ & \quad P(Q_{t_2} = q_2 | Q_{t_1} = q_1) P(Q_{t_1} = q_1) \\ = \;& P(Q_{t_n+h} = q_n | Q_{t_{n-1}+h} = q_{n-1}) P(Q_{t_{n-1}+h} = q_{n-1} | Q_{t_{n-2}+h} = q_{n-2}) \ldots \\ & \quad P(Q_{t_2+h} = q_2 | Q_{t_1+h} = q_1) P(Q_{t_1+h} = q_1) \end{aligned}$$

Therefore, a homogeneous Markov chain is stationary only when $P(Q_{t_1} = q) = P(Q_{t_1+h} = q) = P(Q_t = q)$ for all $q \in \mathcal{Q}$. This is called a stationary distribution of the Markov chain and will be designated by ξ with $\xi_i = P(Q_t = i)$.[2]

According to the definition of the transition matrix, a stationary distribution has the property that $\xi A = \xi$ implying that ξ must be a left eigenvector of the transition matrix A. For example, let $p_1 = [.5, .5]$ be the current distribution over a 2-state Markov chain (using matlab notation). Let $A_1 = [.3, .7; .7, .3]$ be the transition matrix. The Markov chain is stationary since $p_1 A_1 = p_1$. If the current distribution is $p_2 = [.4, .6]$, however, then $p_2 A_1 \neq p_2$, so the chain is no longer stationary.

In general, there can be more than one stationary distribution for a given Markov chain (as there can be more than one eigenvector of a matrix). The condition of stationarity for the chain, however, depends on if the chain "admits" a stationary distribution, and if it does, whether the current marginal distribution over the states is one of the stationary

[2] This is typically designated using π, but that will be reserved for initial HMM distributions.

distributions. If a chain does admit a stationary distribution ξ, then $\xi_j = 0$ for all j that are transient and null-recurrent [47]; i.e., a stationary distribution has positive probability only for positive-recurrent states (states that are assuredly re-visited).

The time-homogeneous property of a Markov chain is distinct from the stationarity property. Stationarity, however, does implies time-homogeneity. To see this, note that if the process is stationary then $P(Q_t = i, Q_{t-1} = j) = P(Q_{t-1} = i, Q_{t-2} = j)$ and $P(Q_t = i) = P(Q_{t-1} = i)$. Therefore, $a_{ij}(t) = P(Q_t = i, Q_{t-1} = j)/P(Q_{t-1} = j) = P(Q_{t-1} = i, Q_{t-2} = j)/P(Q_{t-2} = j) = a_{ij}(t-1)$, so by induction $a_{ij}(t) = a_{ij}(t+\tau)$ for all τ, and the chain is time-homogeneous. On the other hand, a time-homogeneous Markov chain might not admit a stationary distribution and therefore never correspond to a stationary random process.

The idea of "probability flow" may help to determine if a Markov chain admits a stationary distribution. Stationary, or $\xi A = \xi$, implies that for all i

$$\xi_i = \sum_j \xi_j a_{ji}$$

or equivalently,

$$\xi_i(1 - a_{ii}) = \sum_{j \neq i} \xi_j a_{ji}$$

which is the same as

$$\sum_{j \neq i} \xi_i a_{ij} = \sum_{j \neq i} \xi_j a_{ji}$$

The left side of this equation can be interpreted as the probability flow out of state i and the right side can be interpreted as the flow into state i. A stationary distribution requires that the inflow and outflow cancel each other out for every state.

3 Hidden Markov Models

We at last arrive at the main topic of this paper. As will be seen, an HMM is a statistical model for a sequence of data items called the observation vectors. Rather than wet our toes with HMM general properties and analogies, we dive right in by providing a formal definition.

Definition 3.1. Hidden Markov Model *A hidden Markov model (HMM) is collection of random variables consisting of a set of T discrete scalar variables $Q_{1:T}$ and a set of T other variables $X_{1:T}$ which may be either discrete or continuous (and either scalar- or vector-valued). These variables, collectively, possess the following conditional independence properties:*

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

and

$$X_t \perp\!\!\!\perp \{Q_{\neg t}, X_{\neg t}\} | Q_t \tag{3}$$

for each $t \in 1 : T$. No other conditional independence properties are true in general, unless they follow from Equations 2 and 3. The length T of these sequences is itself an integer-valued random variable having a complex distribution (see Section 4.7).

Let us suppose that each Q_t may take values in a finite set, so $Q_t \in \mathcal{Q}$ where $\mathcal{Q}$ is called the state space which has cardinality $|\mathcal{Q}|$. A number of HMM properties may immediately be deduced from this definition.

Equations (2) and (3) imply a large assortment of conditional independence statements. Equation 2 states that the future is conditionally independent of the past given the present. One implication[3] is that $Q_t \perp\!\!\!\perp Q_{1:t-2} | Q_{t-1}$ which means the variables $Q_{1:T}$ form a discrete-time, discrete-valued, first-order Markov chain. Another implication of Equation 2 is $Q_t \perp\!\!\!\perp \{Q_{1:t-2}, X_{1:t-1}\} | Q_{t-1}$ which means that X_τ is unable, given Q_{t-1}, to affect Q_t for $\tau < t$. This 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 $\tau > t$.

The other variables $X_{1:T}$ form a general discrete time stochastic process with, as we will see, great flexibility. Equation 3 states that given an assignment to Q_t, the distribution of X_t is independent of every other variable

[3] Recall Section 2.1.

(both in the future *and* in the past) in the HMM. One implication is that $X_t \perp\!\!\!\perp X_{t+1} | \{Q_t, Q_{t+1}\}$ which follows since $X_t \perp\!\!\!\perp \{X_{t+1}, Q_{t+1}\} | Q_t$ and $X_t \perp\!\!\!\perp X_{t+1} | Q_{t+1}$.

Definition 3.1 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.

Any joint probability distribution over an appropriately typed set of random variables that obeys the above set of conditional independence rules is then an HMM. The two above conditional independence properties imply that, for a given T, the joint distribution over all the variables may be expanded as follows:

$$
\begin{aligned}
p(x_{1:T}, q_{1:T}) &= p(x_T, q_T | x_{1:T-1}, q_{1:T-1}) p(x_{1:T-1}, q_{1:T-1}) && \text{Chain Rule of probability.} \\
&= p(x_T | q_T, x_{1:T-1}, q_{1:T-1}) p(q_T | x_{1:T-1}, q_{1:T-1}) p(x_{1:T-1}, q_{1:T-1}) && \text{Again, chain rule.} \\
&= p(x_T | q_T) p(q_T | q_{T-1}) p(x_{1:T-1}, q_{1:T-1}) && \text{Since } X_T \perp\!\!\!\perp \{X_{1:T-1}, Q_{1:T-1}\} | Q_T. \\
& && \text{and } Q_T \perp\!\!\!\perp \{X_{1:T-1}, Q_{1:T-2}\} | Q_{T-1} \\
& && \text{which follow from Definition 3.1} \\
&= \dots \\
&= p(q_1) \prod_{t=2}^{T} p(q_t | q_{t-1}) \prod_{t=1}^{T} p(x_t | q_t)
\end{aligned}
$$

To parameterize an HMM, one therefore needs the following quantities: 1) the distribution over the initial chain variable $p(q_1)$, 2) the conditional "transition" distributions for the first-order Markov chain $p(q_t | q_{t-1})$, and 3) the conditional distribution for the other variables $p(x_t | q_t)$. It can be seen that these quantities correspond to the classic HMM definition [85]. Specifically, the initial (not necessarily stationary) distribution is labeled π which is a vector of length $|\mathcal{Q}|$. Then, $p(Q_1 = i) = \pi_i$. where π_i is the i^{th} element of π. The observation probability distributions are notated $b_j(x) = p(X_t = x | Q_t = j)$ and the associated parameters depend on $b_j(x)$'s family of distributions. Also, the Markov chain is typically assumed to be time-homogeneous, with stochastic matrix A where $(A)_{ij} = p(Q_t = j | Q_{t-1} = i)$ for all t. HMM parameters are often symbolized collectively as $\lambda \triangleq (\pi, A, B)$ where B represents the parameters corresponding to all the observation distributions.

For speech recognition, the Markov chain $Q_{1:T}$ is typically hidden, which naturally results in the name *hidden* Markov model. The variables $X_{1:T}$ are typically observed. These are the conventional variable designations but need not always hold. For example, X_τ could be missing or hidden, for some or all τ. In some tasks, $Q_{1:T}$ might be known and $X_{1:T}$ might be hidden. The name "HMM" applies in any case, even if $Q_{1:T}$ are not hidden and $X_{1:T}$ are not observed. Regardless, $Q_{1:T}$ will henceforth refer to the hidden variables and $X_{1:T}$ the observations.

With the above definition, an HMM can be simultaneously viewed as a generator and a stochastic acceptor. Like any random variable, say Y, one may obtain a sample from that random variable (e.g., flip a coin), or given a sample, say y, one may compute the probability of that sample $p(Y = y)$ (e.g., the probability of heads). One way to sample from an HMM is to first obtain a complete sample from the hidden Markov chain (i.e., sample from all the random variables $Q_{1:T}$ by first sampling Q_1, then Q_2 given Q_1, and so on.), and then at each time point t produce a sample of X_t using $p(X_t | q_t)$, the observation distribution according to the hidden variable value at time t. This is the same as choosing first a sequence of urns and then a sequence of balls from each urn as described in [85]. To sample just from $X_{1:T}$, one follows the same procedure but then throws away the Markov chain $Q_{1:T}$.

It is important to realize that each sample of $X_{1:T}$ requires a new and different sample of $Q_{1:T}$. In other words, two different HMM observation samples typically originate from two different state assignments to the hidden Markov chain. Put yet another way, an HMM observation sample is obtained using the marginal distribution $p(X_{1:T}) = \sum_{q_{1:T}} p(X_{1:T}, q_{1:T})$ and not from the conditional distribution $p(X_{1:T} | q_{1:t})$ for some fixed hidden variable assignment $q_{1:T}$. As will be seen, this marginal distribution $p(X_{1:T})$ can be quite general.

Correspondingly, when one observes only the collection of values $x_{1:T}$, they have presumably been produced according to some specific but unknown assignment to the hidden variables. A given $x_{1:T}$, however, could have been produced from one of many different assignments to the hidden variables. To compute the probability $p(x_{1:T})$, one

must therefore marginalize away all possible assignments to $Q_{1:T}$ as follows:

$$\begin{aligned} p(x_{1:T}) &= \sum_{q_{1:T}} p(x_{1:T}, q_{1:T}) \\ &= \sum_{q_{1:T}} p(q_1) \prod_{t=2}^{T} p(q_t|q_{t-1}) \prod_{t=1}^{T} p(x_t|q_t) \end{aligned}$$

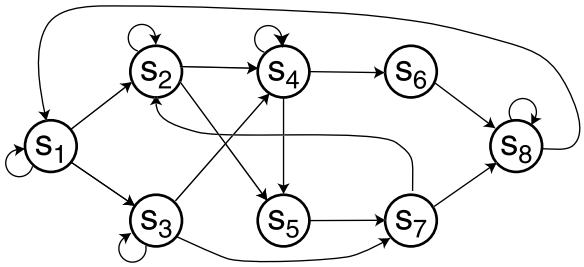


Figure 3: Stochastic finite-state automaton view of an HMM. In this case, only the possible (i.e., non-zero probability) hidden Markov chain state transitions are shown.

An HMM may be graphically depicted in three ways. The first view portrays only a directed state-transition graph as in Figure 3. It is important to realize that this view neither depicts the HMM's output distributions nor the conditional independence properties. The graph depicts only the allowable transitions in the HMM's underlying Markov chain. Each node corresponds to one of the states in $\mathcal{Q}$, where an edge going from node i to node j indicates that $a_{ij} > 0$, and the lack of such an edge indicates that $a_{ij} = 0$. The transition matrix associated with Figure 3 is as follows:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & 0 & 0 & 0 & 0 & 0 \\ 0 & a_{22} & 0 & a_{24} & a_{25} & 0 & 0 & 0 \\ 0 & 0 & a_{33} & a_{34} & 0 & 0 & a_{37} & 0 \\ 0 & 0 & 0 & a_{44} & a_{45} & a_{46} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & a_{57} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_{68} \\ 0 & a_{72} & 0 & 0 & 0 & 0 & 0 & a_{78} \\ a_{81} & 0 & 0 & 0 & 0 & 0 & 0 & a_{88} \end{pmatrix}$$

where it is assumed that the explicitly mentioned a_{ij} are non-zero. In this view, an HMM is seen as an extended stochastic FSA [73]. One can envisage being in a particular state j at a certain time, producing an observation sample from the observation distribution corresponding to that state $b_j(x)$, and then advancing to the next state according to the non-zero transitions.

A second view of HMMs (Figure 4) shows the collection of states and the set of possible transitions between states at each successive time step. This view also depicts only the transition structure of the underlying Markov chain. In this portrayal, the transitions may change at different times and therefore a non-homogeneous Markov chain can be pictured unlike in Figure 3. This view is often useful to display the HMM search space [55, 89] in a recognition or decoding task.

A third HMM view, displayed in Figure 5, shows how HMMs are one instance of a DGM. In this case, the hidden Markov-chain topology is unspecified — only the HMM conditional independence properties are shown, corresponding precisely to our HMM definition. That is, using any of the equivalent schemas such as the directed local Markov property (Section 2.1) or the Bayes ball procedure (Figure 2), the conditional independence properties implied by Figure 5 are identical to those expressed in Definition 3.1. For example, the variable X_t does not depend on any of X_t's non-descendants ($\{Q_{\neg t}, X_{\neg t}\}$) given X_t's parent Q_t. The DGM view is preferable when discussing the HMM statistical dependencies (or lack thereof). The stochastic FSA view in Figure 3 is useful primarily to analyze the underlying hidden Markov chain topology. It should be very clear that Figure 3 and Figure 5 display entirely different HMM properties.

There are many possible state-conditioned observation distributions [71, 85]. When the observations are discrete, the distributions $b_j(x)$ are mass functions and when the observations are continuous, the distributions are typically specified using a parametric model family. The most common family is the Gaussian mixture where

$$b_j(x) = \sum_{k=1}^{N_j} c_{jk} \mathcal{N}(x|\mu_{jk}, \Sigma_{jk})$$

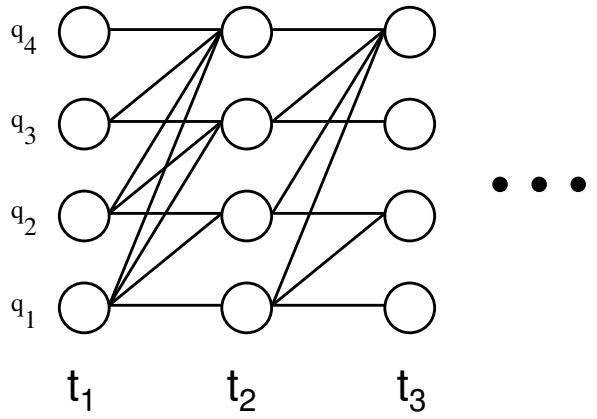


Figure 4: Time-slice view of a Hidden Markov Model's state transitions.

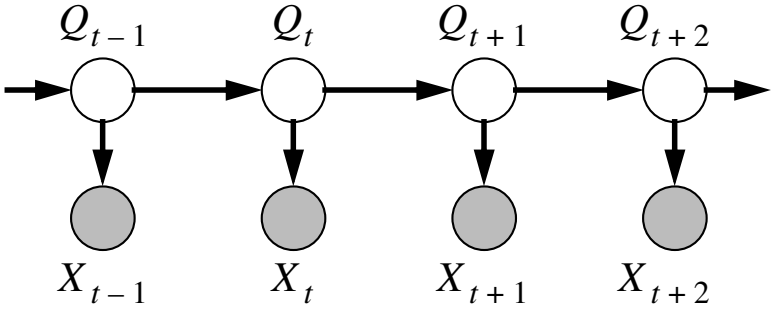


Figure 5: A Hidden Markov Model

and where $\mathcal{N}(x|\mu_{jk}, \Sigma_{jk})$ is a Gaussian distribution [74, 64] with mean vector μ_{jk} and covariance matrix Σ_{jk}. The values c_{jk} are mixing coefficients for hidden state j with $c_{jk} \geq 0$ and $\sum_k c_{jk} = 1$. Often referred to as a Gaussian Mixture HMM (GMHMM), this HMM has DGM depicted in Figure 6. Other observation distribution choices include discrete probability tables [85], neural networks (i.e., hybrid systems) [11, 75], auto-regressive distributions [83, 84] or mixtures thereof [60], and the standard set of named distributions [71].

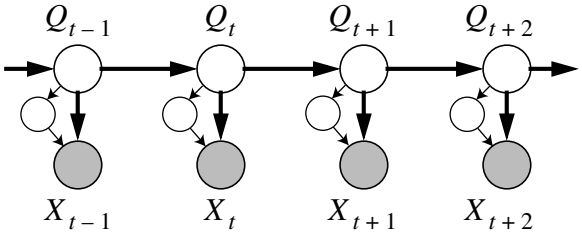


Figure 6: A Mixture-Observation Hidden Markov Model

One is often interested in computing $p(x_{1:T})$ for a given set of observations. Blindly computing $\sum_{q_{1:T}} p(x_{1:T}, q_{1:T})$ is hopelessly intractable, requiring $O(|\mathcal{Q}|^T)$ operations. Fortunately, the conditional independence properties allow for efficient computation of this quantity. First the joint distribution can be expressed as $p(x_{1:t}) = \sum_{q_t, q_{t-1}} p(x_{1:t}, q_t, q_{t-1})$,

the summand of which can be expanded as follows:

$$\begin{aligned}
p(x_{1:t}, q_t, q_{t-1}) &= p(x_{1:t-1}, q_{t-1}, x_t, q_t) \\
&= p(x_t, q_t | x_{1:t-1}, q_{t-1}) p(x_{1:t-1}, q_{t-1}) && \text{Chain rule of probability.} \\
&= p(x_t | q_t, x_{1:t-1}, q_{t-1}) p(q_t | x_{1:t-1}, q_{t-1}) p(x_{1:t-1}, q_{t-1}) \\
&= p(x_t | q_t) p(q_t | q_{t-1}) p(x_{1:t-1}, q_{t-1}) && \text{Since } X_t \perp\!\!\!\perp \{X_{1:t-1}, Q_{1:t-1}\} | Q_t \\
& && \text{and } Q_t \perp\!\!\!\perp \{X_{1:t-1}, Q_{1:t-2}\} | Q_{t-1} \\
& && \text{which follow from Definition 3.1.}
\end{aligned}$$

This yields,

$$p(x_{1:t}, q_t) = \sum_{q_{t-1}} p(x_{1:t}, q_t, q_{t-1}) \tag{4}$$

$$= \sum_{q_{t-1}} p(x_t | q_t) p(q_t | q_{t-1}) p(x_{1:t-1}, q_{t-1}) \tag{5}$$

If the following quantity is defined $\alpha_q(t) \triangleq p(x_{1:t}, Q_t = q)$, then the preceding equations imply that $\alpha_q(t) = p(x_t | Q_t = q) \sum_r p(Q_t = q | Q_{t-1} = r) \alpha_r(t-1)$. This is just the alpha, or forward, recursion [85]. Then $p(x_{1:T}) = \sum_q \alpha_q(T)$, and the entire computation requires only $O(|\mathcal{Q}|^2 T)$ operations. To derive this recursion, it was necessary to use only the fact that X_t was independent of its past given Q_t — X_t is also independent of the future given Q_t, but this was not needed. This later assumption, however, is obligatory for the beta or backward recursion.

$$\begin{aligned}
p(x_{t+1,T} | q_t) &= \sum_{q_{t+1}} p(q_{t+1}, x_{t+1}, x_{t+2:T} | q_t) \\
&= \sum_{q_{t+1}} p(x_{t+2:T} | q_{t+1}, x_{t+1}, q_t) p(x_{t+1} | q_{t+1}, q_t) p(q_{t+1} | q_t) && \text{Chain rule of probability.} \\
&= \sum_{q_{t+1}} p(x_{t+2:T} | q_{t+1}) p(x_{t+1} | q_{t+1}) p(q_{t+1} | q_t) && \text{Since } X_{t+2:T} \perp\!\!\!\perp \{X_{t+1}, Q_t\} | Q_{t+1} \\
& && \text{and } X_{t+1} \perp\!\!\!\perp Q_t | Q_{t+1} \text{ which follow} \\
& && \text{from Definition 3.1.}
\end{aligned}$$

Using the definition $\beta_q(t) \triangleq p(x_{t+1:T} | Q_t = q)$, the above equations imply the beta-recursion $\beta_q(t) = \sum_r \beta_r(t+1) p(x_{t+1} | Q_{t+1} = r) p(Q_{t+1} = r | Q_t = q)$, and another expression for the full probability $p(x_{1:T}) = \sum_q \beta_q(1) p(q) p(x_1 | q)$. Furthermore, this complete probability may be computed using a combination of the alpha and beta values at any t since

$$\begin{aligned}
p(x_{1:T}) &= \sum_{q_t} p(q_t, x_{1:t}, x_{t+1:T}) \\
&= \sum_{q_t} p(x_{t+1:T} | q_t, x_{1:t}) p(q_t, x_{1:t}) \\
&= \sum_{q_t} p(x_{t+1:T} | q_t) p(q_t, x_{1:t}) && \text{Since } X_{t+1:T} \perp\!\!\!\perp X_{1:t} | Q_t. \\
&= \sum_{q_t} \beta_{q_t}(t) \alpha_{q_t}(t)
\end{aligned}$$

Together, the alpha- and beta- recursions are the key to learning the HMM parameters using the Baum-Welch procedure (which is really the EM algorithm for HMMs [94, 3]) as described in [85, 3]. It may seem natural at this point to provide EM parameter update equations for HMM training. Rather than repeat what has already been provided in a variety of sources [94, 85, 3], we are at this point equipped with the machinery sufficient to move on and describe what HMMs can do.

4 What HMMs Can Do

The HMM conditional independence properties (Equations 2 and 3), can be used to better understand the general 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. This section therefore compiles and then analyzes in detail a list of such qualities as follows:

- 4.1 observation variables are i.i.d.
- 4.2 observation variables are i.i.d. conditioned on the state sequence or are "locally" i.i.d.
- 4.3 observation variables are i.i.d. under the most likely hidden variable assignment (i.e., the Viterbi path)
- 4.4 observation variables are uncorrelated over time and do not capture acoustic context
- 4.5 HMMs correspond to segmented or piece-wise stationary distributions (the "beads-on-a-string" phenomena)
- 4.6 when using an HMM, speech is represented as a sequence of feature vectors, or "frames", within which the speech signal is assumed to be stationary
- 4.7 when sampling from an HMM, the active duration of an observation distribution is a geometric distribution
- 4.8 a first-order Markov chain is less powerful than an n^{th} order chain
- 4.9 an HMM represents $p(X|M)$ (a synthesis model) but to minimize Bayes error, a model should represent $p(M|X)$ (a production model)

4.1 Observations i.i.d.

Given definition 2.1, it can be seen that an HMM is not i.i.d. Consider the following joint probability under an HMM:

$$p(X_{t:t+h} = x_{t:t+h}) = \sum_{q_{t:t+h}} \prod_{j=t}^{t+h} p(X_j = x_j|Q_j = q_j)a_{q_j q_{j-1}}.$$

Unless only one state in the hidden Markov chain has non-zero probability for all times in the segment $t\!:\!t+h$, this quantity can not in general be factored into the form $\prod_{j=t}^{t+h} p(x_j)$ for some time-independent distribution $p(\cdot)$ as would be required for an i.i.d. process.

4.2 Conditionally i.i.d. observations

HMMs are i.i.d. conditioned on certain state sequences. This is because

$$p(X_{t:t+h} = x_{t:t+h}|Q_{t:t+h} = q_{t:t+h}) = \prod_{\tau=t}^{t+h} p(X_\tau = x_\tau|Q_\tau = q_\tau).$$

and if for $t \leq \tau \leq t+h$, $q_\tau = j$ for some fixed j then

$$p(X_{t:t+h} = x_{t:t+h}|Q_{t:t+h} = q_{t:t+h}) = \prod_{\tau=t}^{t+h} b_j(x_\tau)$$

which is i.i.d. for this specific state assignment over this time segment $t\!:\!t+h$.

While this is true, recall that each HMM sample requires a potentially different assignment to the hidden Markov chain. Unless one and only one state assignment during the segment $t\!:\!t+h$ has non-zero probability, the hidden state sequence will change for each HMM sample and there will be no i.i.d. property. The fact that an HMM is i.i.d. conditioned on a state sequence does not necessarily have repercussions when HMMs are actually used. An HMM represents the joint distribution of feature vectors $p(X_{1:T})$ which is obtained by marginalizing away (summing over) the hidden variables. HMM probability "scores" (say, for a classification task) are obtained from that joint distribution, and are not obtained from the distribution of feature vectors $p(X_{1:T}|Q_{1:T})$ conditioned on one and only one state sequence.

4.3 Viterbi i.i.d.

The Viterbi (maximum likelihood) path [85, 53] of an HMM is defined as follows:

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

where $p(X_{1:T} = x_{1:T}, q_{1:T})$ is the joint probability of an observation sequence $x_{1:T}$ and hidden state assignment $q_{1:T}$ for an HMM.

When using an HMM, it is often the case that the joint probability distribution of features is taken according to the Viterbi path:

$$\begin{aligned} & p_{\text{vit}}(X_{1:T} = x_{1:T}) \\ &= c\, p(X_{1:T} = x_{1:T}, Q_{1:T} = q^*_{1:T}) \\ &= c \max_{q_{1:T}} p(X_{1:T} = x_{1:T}, Q_{1:T} = q_{1:T}) \\ &= c \max_{q_{1:T}} \prod_{t=1}^{T} p(X_t = x_t | Q_t = q_t) p(Q_t = q_t | Q_{t-1} = q_{t-1}) \end{aligned} \tag{6}$$

where c is some normalizing constant. This can be different than the complete probability distribution:

$$p(X_{1:T} = x_{1:T}) = \sum_{q_{1:T}} p(X_{1:T} = x_{1:T}, Q_{1:T} = q_{1:T}).$$

Even under a Viterbi approximation, however, the resulting distribution is not necessarily i.i.d. unless the Viterbi paths for all observation assignments are identical. The Viterbi path is different for each observation sequence, and the max operator does not in general commute with the product operator in Equation 6, the product form required for an i.i.d. process is unattainable in general.

4.4 Uncorrelated observations

Two observations at different times might be dependent, but are they correlated? If X_t and X_{t+h} are uncorrelated, then $E[X_t X'_{t+h}] = E[X_t]E[X_{t+h}]'$. For simplicity, consider an HMM that has single component Gaussian observation distributions, i.e., $b_j(x) \sim \mathcal{N}(x|\mu_j, \Sigma_j)$ for all states j. Also assume that the hidden Markov chain of the HMM is currently a stationary process with some stationary distribution π. For such an HMM, the covariance can be computed explicitly. In this case, the mean value of each observation is a weighted sum of the Gaussian means:

$$\begin{aligned} E[X_t] &= \int x p(X_t = x) dx \\ &= \int x \sum_i p(X_t = x | Q_t = i) \pi_i dx \\ &= \sum_i E[X_t | Q_t = i] \pi_i \\ &= \sum_i \mu_i \pi_i \end{aligned}$$

Similarly,

$$\begin{aligned} E[X_t X'_{t+h}] &= \int x y' p(X_t = x, X_{t+h} = y) dx dy \\ &= \int x y' \sum_{ij} p(X_t = x, X_{t+h} = y | Q_t = i, Q_{t+h} = j) p(Q_{t+h} = j | Q_t = i) \pi_i dx dy \\ &= \sum_{ij} E[X_t X'_{t+h} | Q_t = i, Q_{t+h} = j] (A^h)_{ij} \pi_i dx dy \\ &= \sum_{ij} E[X_t X'_{t+h} | Q_t = i, Q_{t+h} = j] (A^h)_{ij} \pi_i dx dy \end{aligned}$$

The above equations follow from $p(Q_{t+h} = j|Q_t = i) = (A^h)_{ij}$ (i.e., the Chapman-Kolmogorov equations [47]) where $(A^h)_{ij}$ is the i, j^{th} element of the matrix A raised to the h power. Because of the conditional independence properties, it follows that:

$$E[X_t X'_{t+h}|Q_t = i, Q_{t+h} = j] = E[X_t|Q_t = i]E[X'_{t+h}|Q_{t+h} = j] = \mu_i \mu'_j$$

yielding

$$E[X_t X'_{t+h}] = \sum_{ij} \mu_i \mu'_j (A^h)_{ij} \pi_i$$

The covariance between feature vectors may therefore be expressed as:

$$\text{cov}(X_t, X_{t+h}) = \sum_{ij} \mu_i \mu'_j (A^h)_{ij} \pi_i - \left(\sum_i \mu_i \pi_i \right) \left(\sum_i \mu_i \pi_i \right)'$$

It can be seen that this quantity is not in general the zero matrix and therefore HMMs, even with a simple Gaussian observation distribution and a stationary Markov chain, can capture correlation between feature vectors. Results for other observation distributions have been derived in [71].

To empirically demonstrate such correlation, the mutual information [6, 16] in bits was computed between feature vectors from speech data that was sampled using 4-state per phone word HMMs trained from an isolated word task using MFCCs and their deltas [107]. As shown on the left of Figure 7, the HMM samples do exhibit inter-frame dependence, especially between the same feature elements at different time positions. The right of Figure 7 compares the average pair-wise mutual information over time of this HMM with i.i.d. samples from a Gaussian mixture.

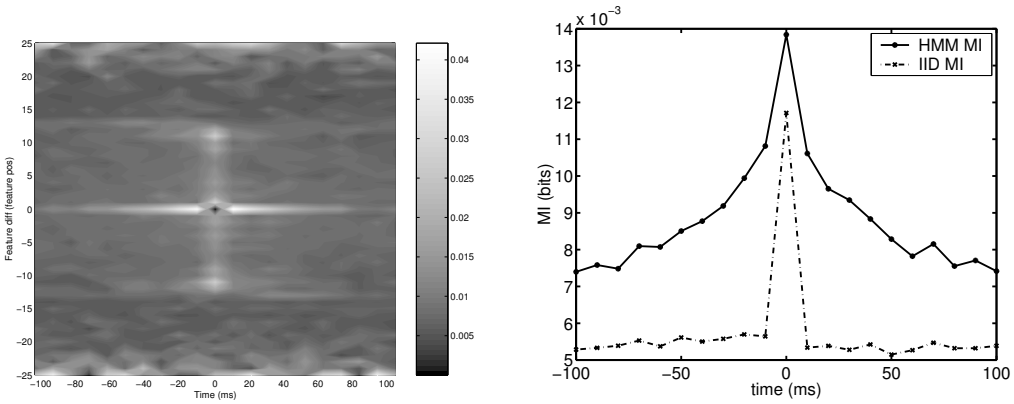


Figure 7: Left: The mutual information between features that were sampled from a collection of about 1500 word HMMs using 4 states each per context independent phone model. Right: A comparison of the average pair-wise mutual information over time between all observation vector elements of such an HMM with that of i.i.d. samples from a Gaussian mixture. The HMM shows significantly more correlation than the noise-floor of the i.i.d. process. The high values in the center reflect correlation between scalar elements within the vector-valued Gaussian mixture.

HMMs indeed represent dependency information between temporally disparate observation variables. The hidden variables indirectly encode this information, and as the number of hidden states increases, so does the amount of information that can be encoded. This point is explored further in Section 5.

4.5 Piece-wise or segment-wise stationary

A HMM's stationarity condition may be discovered by finding the conditions that must hold for an HMM to be a stationary process. In the following analysis, it is assumed that the Markov chain is time-homogeneous – if non-stationary can be shown in this case, it certainly can be shown for the more general time-inhomogeneous case.

According to Definition 2.2, an HMM is stationary when:

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

for all n, h, $t_{1:n}$, and $x_{1:n}$. The quantity $P(X_{t_{1:n}+h} = x_{1:n})$ can be expanded as follows:

$$
\begin{aligned}
&p(X_{t_{1:n}+h} = x_{1:n}) \\
&= \sum_{q_{1:n}} p(X_{t_{1:n}+h} = x_{1:n}, Q_{t_{1:n}+h} = q_{1:n}) \\
&= \sum_{q_{1:n}} p(Q_{t_1+h} = q_1) p(X_{t_1+h} = x_1 | Q_{t_1+h} = q_1) \prod_{i=2}^{n} p(X_{t_i+h} = x_i | Q_{t_i+h} = q_i) p(Q_{t_i+h} = q_i | Q_{t_{i-1}+h} = q_{i-1}) \\
&= \sum_{q_1} p(Q_{t_1+h} = q_1) p(X_{t_1+h} = x_1 | Q_{t_1+h} = q_1) \sum_{q_{2:T}} \prod_{i=2}^{n} p(X_{t_i+h} = x_i | Q_{t_i+h} = q_i) p(Q_{t_i+h} = q_i | Q_{t_{i-1}+h} = q_{i-1}) \\
&= \sum_{q_1} p(Q_{t_1+h} = q_1) p(X_{t_1+h} = x_1 | Q_{t_1+h} = q_1) \sum_{q_{2:T}} \prod_{i=2}^{n} p(X_{t_i} = x_i | Q_{t_i} = q_i) p(Q_{t_i} = q_i | Q_{t_{i-1}} = q_{i-1}) \\
&= \sum_{q_1} p(Q_{t_1+h} = q_1) p(X_{t_1} = x_1 | Q_{t_1} = q_1) f(x_{2:n}, q_1)
\end{aligned}
$$

where $f(x_{2:n}, q_1)$ is a function that is independent of the variable h. For HMM stationarity to hold, it is required that $p(Q_{t_1+h} = q_1) = p(Q_{t_1} = q_1)$ 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 stochastic process.

For speech recognition, HMMs commonly have left-to-right state-transition topologies where transition matrices are upper triangular ($a_{ij} = 0 \ \ \forall j > i$). The transition graph is thus a directed acyclic graph (DAG) that also allows self loops. In such graphs, all states with successors (i.e., non-zero exit transition probabilities) have decreasing occupancy probability over time. This can be seen inductively. First consider the start states, those without any predecessors. Such states have decreasing occupancy probability over time because input transitions are unavailable to create inflow. Consequently, these states have decreasing outflow over time. Next, consider any state having only predecessors with decreasing outflow. Such a state has decreasing inflow, a decreasing occupancy probability, and decreasing outflow as well. Only the final states, those with only predecessors and no successors, may retain their occupancy probability over time. Since under a stationary distribution, every state must have zero net probability flow, a stationary distribution for a DAG topology must have zero occupancy probability for any states with successors. All states with children in a DAG topology have less than unity return probability, and so are transient. This proves that a stationary distribution must bestow zero probability to every transient state. Therefore, any left-to-right HMM (e.g., the HMMs typically found in speech recognition systems) is not stationary unless all non-final states have zero probability.

Note that HMMs are also unlikely to be "piece-wise" stationary, in which an HMM is in a particular state for a time and where observations in that time 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 result in the same hidden assignment for some fixed-time region, or 2) the hidden chain is always stationary over that region. In the general case, however, an HMM does not produce samples from such piece-wise stationary segments.

The notions of stationarity and i.i.d. are properties of a random processes, or equivalently, of the complete ensemble of process samples. The concepts of stationarity and i.i.d. do not apply to a single HMM sample. A more appropriate characteristic that might apply to a single sequence (possibly an HMM sample) is that of "steady state," where the short-time spectrum of a signal is constant over a region of time. Clearly, human speech is not steady state.

It has been known for some time that the information in a speech signal necessary to convey an intelligent message to the listener is contained in the spectral sub-band modulation envelopes [30, 28, 27, 45, 46] and that the spectral energy in this domain is temporally band-limited. A liberal estimate of the high-frequency cutoff 50Hz. By band-pass filtering the sub-band modulation envelopes, this trait is deliberately used by speech coding algorithms which achieve significant compression ratios with little or no intelligibility loss. Similarly, any stochastic process representing the message-containing information in a speech signal need only possess dynamic properties at rates no higher than this rate. The Nyquist sampling theorem states that any band-limited signal may be precisely represented with a discrete-time signal sampled at a sufficiently high rate (at least twice the highest frequency in the signal). The statistical properties of speech may therefore be accurately represented with a discrete time signal sampled at a suitably high rate.

Might HMMs be a poor speech model because HMM samples are piece-wise steady-state and natural speech does not contain steady-state segments. An HMM's Markov chain establishes the temporal evolution of the process's statistical properties. Therefore, any band-limited non-stationary or non-steady-state signal can be represented by an HMM with a Markov chain having a fast enough average state change and having enough states to capture all the inherent signal variability. As argued below, only a finite number of states are needed for real-world signals.

The arguments above also apply to time inhomogeneous processes since they are a generalization of the homogeneous case.

4.6 Within-frame stationary

Speech is a continuous time signal. A feature extraction process generates speech frames at regular time intervals (such as 10ms) each with some window width (usually 20ms). 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. Because the properties of speech that convey any message are band-limited in the modulation domain, if the rate of hidden state change is high enough, and if the frame-window width is small enough, a framing of speech would not result in information loss about the actual message.

4.7 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 with parameter a_{ii}. That is, D has distribution $P(D = d) = p^{d-1}(1 - p)$ where $d \geq 1$ is an integer and $p = a_{ii}$. It seems possible that HMMs might be deficient because their state duration distributions are inherently geometric, and geometric distributions can not accurately represent typical speech unit (e.g., phoneme or syllable) durations [4]

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. 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 is equivalent to that of a random variable consisting of the sum of n independent geometrically distributed random variables. The distribution of such a sum has a negative binomial distribution (which is a discrete version of the gamma distribution) [95]. Unlike a geometric distribution, a negative binomial distribution has a mode located away from zero.

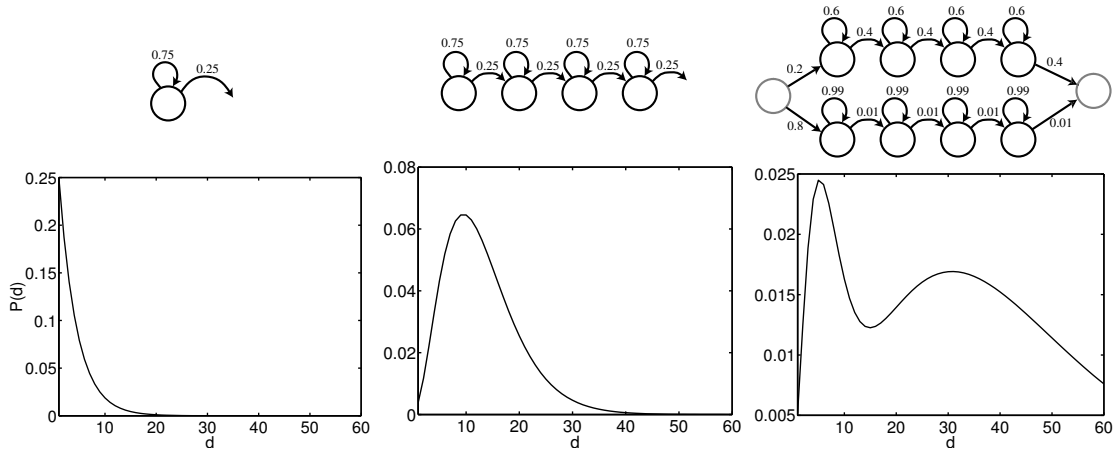


Figure 8: Three possible active observation duration distributions with an HMM, and their respective Markov chain topologies.

In general, a collection of HMM states sharing the same observation distribution may be combined in a variety of serial and parallel fashions. When combined in series, the resulting distribution is a convolution of the individual

[4] It has been suggested that a gamma distribution is a more appropriate speech-unit durational distribution[68].

distributions (resulting in a negative binomial from a series of geometric random variables). When combined in parallel, the resulting distribution is a weighted mixture of the individual distributions. This process can of course be repeated at higher levels as well. In fact, one needs a recursive definition to define the resulting set of possible distributions. Supposing D is such a random variable, one might say that D has a distribution equal to that of a sum of random variables, each one having a distribution equal to a mixture model, with each mixture component coming from the set of possible distributions for D. The base case is that D has a geometric distribution. In fact, the random variable T in Definition 3.1 has such a distribution. This is illustrated for a geometric, a sum of geometric, and a mixture of sums of geometric distributions in Figure 8. As can be seen, by simply increasing the hidden state space cardinality, this procedure can produce an broad class of distributions that can represent the time during which a specific observation distribution is active.

4.8 First-order hidden Markov assumption

As was demonstrated in Section 2.3 and as described in [55], 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. [5]

4.9 Synthesis vs. Recognition

HMMs represent only the distribution of feature vectors for a given model, i.e., the likelihood $p(X|M)$. This can viewed as a synthesis or a generative model because sampling from this distribution should produce (or synthesize) an instance of the object M (e.g., a synthesized speech utterance). To achieve Bayes error, however, one should use the posterior $p(M|X)$. This can be viewed as a recognition or a discriminative model since, given an instance of X, a sample from $p(M|X)$ produces a class identifier (e.g., a string of words), the goal of a recognition system. Even though HMMs inherently represent $p(X|M)$, there are several reasons why this property might be less severe than expected.

First, by Bayes rule, $p(M|X) = p(X|M)p(M)/p(X)$ so if an HMM accurately represents $p(X|M)$ and given accurate priors $P(M)$, an accurate posterior will ensue. Maximum-likelihood training adjusts model parameters so that the resulting distribution best matches the empirical distribution specified by training-data. Maximum-likelihood training is asymptotically optimal, so given enough training data and a rich enough model, an accurate estimate of the posterior will be found just by producing an accurate likelihood $p(X|M)$ and prior $p(M)$.

On the other hand, 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. In a classification task, one of a set of different models M_i is chosen as the target class for a given X. In this case, only the decision boundaries, that is the sub-spaces $\{x : p(M_i|x)p(M_i) = p(M_j|x)p(M_j)\}$ for all $i \neq j$, affect classification performance [29]. 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.

The use of generative conditional distributions, as supplied by an HMM, is not necessarily a limitation, since for classification $p(X|M)$ need not be found. Instead, one of the many functions that achieve Bayes error can be approximated. Of course, one member of the class is the likelihood itself, but there are many others. Such a class can be described as follows:

$$\mathcal{F} = \{f(x,m) : \underset{m}{\text{argmax}}\ p(X = x|M = m)p(M = m) = \underset{m}{\text{argmax}}\ f(x,m)p(M = m) \quad \forall x, m\}.$$

The members of $\mathcal{F}$ can be arbitrary functions, can be valid conditional distributions, but need not be approximations of $p(x|m)$. A sample from these distributions will not necessarily result in an accurate object instance (or synthesized speech utterance in the case of speech HMMs). Instead, members of $\mathcal{F}$ might be accurate only at decision boundaries. In other words, statistical consistency of a decision function does not require consistency of any internal likelihood functions.

There are two ways that other members of such a class can be approximated. First, the degree to which boundary information is represented by an HMM (or any likelihood model) depends on the parameter training method. Discriminative training methods have been developed which adjust the parameters of each model to increase not the individual likelihood but rather approximate the posterior probability or Bayes decision rule. Methods such as maximum mutual

[5] In speech recognition systems, hidden state "meanings" might change when moving to a higher-order Markov chain.

information (MMI) [1, 13], minimum discrimination information (MDI) [32, 33], minimum classification error (MCE) [59, 58], and more generally risk minimization [29, 97] essentially attempt to optimize $p(M|X)$ by adjusting whatever model parameters are available, be they the likelihoods $p(X|M)$, posteriors, or something else.

Second, the degree to which boundary information is represented depends on each model's intrinsic ability to produce a probability distribution at decision boundaries vs. its ability to produce a distribution between boundaries. This is the inherent discriminability of the structure of the model for each class, independent of its parameters. Models with this property have been called structurally discriminative [8].

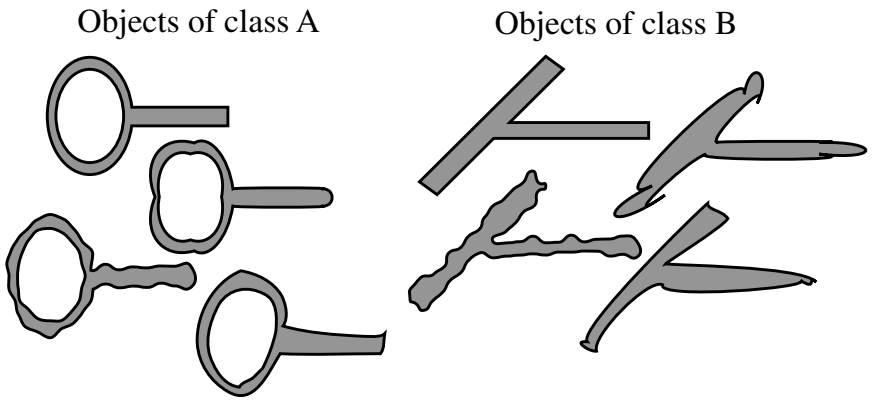


Figure 9: Two types of objects that share a common attribute, a horizontal bar on the right of each object. This attribute need not be represented in the classification task.

This idea can be motivated using a simple example. Consider two classes of objects as shown in Figure 9. Objects of class A consist of an annulus with an extruding horizontal bar on the right. Objects of class B consist of a diagonal bar also with an extruding horizontal bar on the right. Consider a probability distribution family in this space that is accurate only at representing horizontal bars — the average length, width, smoothness, etc. could be parameters that determine a particular distribution. When members of this family are used, the resulting class specific models will be blind to any differences between objects of class A and class B, regardless of the quality and type (discriminative or not) of training method. These models are structurally indiscriminant.

Consider instead two families of probability distributions in this 2D space. The first family accurately represents only annuli of various radii and distortions, and the second family accurately represents only diagonal bars. When each family represents objects of their respective class, the resulting models can easily differentiate between objects of the two classes. These models are inherently blind to the commonalities between the two classes regardless of the training method. The resulting models are capable of representing only the distinctive features of each class. In other words, even if each model is trained using a maximum likelihood procedure using positive-example samples from only its own class, the models will not represent the commonalities between the classes because they are incapable of doing so. The model families are structurally discriminative. Sampling from a model of one class produces an object containing attributes only that distinguish it from samples of the other class's model. The sample will not necessarily resemble the class of objects its model represents. This, however, is of no consequence to classification accuracy. This idea, of course, can be generalized to multiple classes each with their own distinctive attributes.

An HMM could be seen as deficient because it does not synthesize a valid (or even recognizable) spoken utterance. But synthesis is not the goal of classification. A valid synthesized speech utterance should correspond to something that could be uttered by an identifiable speaker. When used for speech recognition, HMMs attempt to describe probability distributions of speech in general, a distribution which corresponds to the average over many different speakers (or at the very least, many different instances of an utterance spoken by the same speaker). Ideally, any idiosyncratic speaker-specific information, which might result in a more accurate synthesis, but not more accurate discrimination, should not be represented by a probabilistic model — representing such additional information can only require a parameter increase without providing a classification accuracy increase. As mentioned above, an HMM should represent distinctive properties of a specific speech utterance relative to other rival speech utterances. Such a model would not necessarily produce high quality synthesized speech.

The question then becomes, how structurally discriminative are HMMs when attempting to model the distinctive attributes of speech utterances? With HMMs, different Markov chains represent each speech utterance. A reasonable

assumption is that HMMs are not structurally indiscriminant because, even when trained using a simple maximum likelihood procedure, HMM-based speech recognition systems perform reasonably well. Sampling from such an HMM might produce an unrealistic speech utterance, but the underlying distribution might be accurate at decision boundaries. Such an approach was taken in [8], where HMM dependencies were augmented to increase structural discriminability.

Earlier sections of this paper suggested that HMM distributions are not destitute in their flexibility, but this section claimed that for the recognition task an HMM need not accurately represent the true likelihood $p(X|M)$ to achieve high classification accuracy. While HMMs are powerful, a fortunate consequence of the above discussion is that HMMs need not capture many nuances in a speech signal and may be simpler as a result. In any event, just because a particular HMM does not represent speech utterances does not mean it is poor at the recognition task.

5 Conditions for HMM Accuracy

Suppose that $p(X_{1:T})$ is the true distribution of the observation variables $X_{1:T}$. In this section, it is shown that if an HMM represents 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 is defined as KL-distance between the two distributions being zero, i.e.:

$$D(p(X_{1:T})||p_h(X_{1:T})) = 0$$

If this condition is true, the mutual information between any subset of variables under each distribution will be equal. That is,

$$I(X_{S_1}; X_{S_2}) = I_h(X_{S_1}; X_{S_2})$$

where $I(\cdot;\cdot)$ is the mutual information between two random vectors under the true distribution, $I_h(\cdot;\cdot)$ is the mutual information under the HMM, and S_i is any subset of $1:T$.

Consider the two sets of variables X_t, the observation at time t, and $X_{\neg t}$, the collection of observations at all times other than t. X_t may be viewed as the output of a noisy channel that has input $X_{\neg t}$ as shown in Figure 10. The information transmission rate between $X_{\neg t}$ and X_t is therefore equal to the mutual information $I(X_{\neg t}; X_t)$ between the two.

Figure 10: A noisy channel view of X_t's dependence on $X_{\neg t}$.

Implied by the KL-distance equality condition, for an HMM to mirror the true distribution $p(X_t|X_{\neg t})$ its corresponding noisy channel representation must have the same transmission rate. Because of the conditional independence properties, an HMM's hidden variable Q_t separates X_t from its context $X_{\neg t}$ and the conditional distribution becomes

$$p_h(X_t|X_{\neg t}) = \sum_q p_h(X_t|Q_t = q)p_h(Q_t = q|X_{\neg t})$$

An HMM, therefore, attempts to compress the information about X_t contained in $X_{\neg t}$ into a single discrete variable Q_t. A noisy channel HMM view is depicted in Figure 11.

For an accurate HMM representation, the composite channel in Figure 11 must have at least the same information transmission rate as that of Figure 10. Note that $I_h(X_{\neg t}; Q_t)$ is the transmission rate between $X_{\neg t}$ and Q_t, and $I_h(Q_t; X_t)$ is the transmission rate between Q_t and X_t. The maximum transmission rate through the HMM composite channel is no greater than to the minimum of $I_h(X_{\neg t}; Q_t)$ and $I_h(Q_t; X_t)$. Intuitively, HMM accuracy requires $I_h(X_{\neg t}; Q_t) \geq I(X_t; X_{\neg t})$ and $I_h(Q_t; X_t) \geq I(X_t; X_{\neg t})$ since if one of these inequalities does not hold, then channel A and/or channel B in Figure 11 will become a bottle-neck. This would restrict the composite channel's transmission rate to be less than the true rate of Figure 10. An additional requirement is that the variable Q_t have enough storage capacity (i.e., states) to encode the information flowing between the two channels. This last condition must be a lower bound on the number of hidden states. This is formalized by the following theorem.

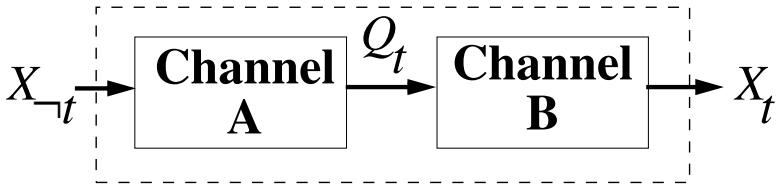


Figure 11: A noisy channel view of one of the HMM conditional independence property.

Theorem 5.1. Necessary conditions for HMM accuracy. *An HMM as defined above (Definition 3.1) with joint observation distribution $p_h(X_{1:T})$ will accurately model the true distribution $p(X_{1:T})$ 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 information transmission rate 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 information transmission rate between $I(X_t; X_{\neg t})$.

Proof. If an HMM is accurate (i.e., has zero KL-distance from the true distribution), then $I(X_{\neg t}; X_t) = I_h(X_{\neg t}; X_t)$. As with the data-processing inequality [16], the quantity $I_h(X_{\neg t}; Q_t, X_t)$ can be expanded in two ways using the chain rule of mutual information:

$$\begin{aligned} & I_h(X_{\neg t}; Q_t, X_t) && (7)\\ = \;& I_h(X_{\neg t}; Q_t) + I_h(X_{\neg t}; X_t | Q_t) && (8)\\ = \;& I_h(X_{\neg t}; X_t) + I_h(X_{\neg t}; Q_t | X_t) && (9)\\ = \;& I(X_{\neg t}; X_t) + I_h(X_{\neg t}; Q_t | X_t) && (10) \end{aligned}$$

The HMM conditional independence properties say that $I_h(X_{\neg t}; X_t | Q_t) = 0$, implying

$$I_h(X_{\neg t}; Q_t) = I(X_{\neg t}; X_t) + I_h(X_{\neg t}; Q_t | X_t)$$

or that

$$I_h(X_{\neg t}; Q_t) \geq I(X_{\neg t}; X_t)$$

since $I_h(X_{\neg t}; Q_t | X_t) \geq 0$. This is the first condition. Similarly, the quantity $I_h(X_t; Q_t, X_{\neg t})$ may be expanded as follows:

$$\begin{aligned} & I_h(X_t; Q_t, X_{\neg t}) && (11)\\ = \;& I_h(X_t; Q_t) + I_h(X_t; X_{\neg t} | Q_t) && (12)\\ = \;& I(X_t; X_{\neg t}) + I_h(X_t; Q_t | X_{\neg t}) && (13) \end{aligned}$$

Reasoning as above, this leads to

$$I_h(X_t; Q_t) \geq I(X_t; X_{\neg t}),$$

the second condition. A sequence of inequalities establishes the third condition:

$$\log |\mathcal{Q}| \geq H(Q_t) \geq H(Q_t) - H(Q_t | X_t) = I_h(Q_t; X_t) \geq I(X_t; X_{\neg t})$$

so $|\mathcal{Q}| \geq 2^{I(X_t; X_{\neg t})}$. □

A similar procedure leads to the requirement that $I_h(X_{1:t}; Q_t) \geq I(X_{1:t}; X_{t+1:T})$, $I_h(Q_t; X_{t+1:T}) \geq I(X_{1:t}; X_{t+1:T})$, and $|\mathcal{Q}| \geq 2^{I(X_{1:t}; X_{t+1:T})}$ for all t.

There are two implications of this theorem. First, an insufficient number of hidden states can lead to an inaccurate model. This has been known for some time in the speech recognition community, but a lower bound on the required

number of states has not been established. With an HMM, the information about X_t contained in $X_{<t}$ is squeezed through the hidden state variable Q_t. Depending on the number of hidden states, this can overburden Q_t and result in an inaccurate probabilistic model. But if there are enough states, and if the information in the surrounding acoustic context is appropriately encoded in the hidden states, the required information may be compressed and represented by Q_t. An appropriate encoding of the contextual information is essential since just adding states does not guarantee accuracy will increase.

To achieve high accuracy, it is likely that a finite number of states is required for any real task since signals representing natural objects will have bounded mutual information. Recall that the first order Markov assumption in the hidden Markov chain is not necessarily a problem since a first-order chain may represent an n^{th} order chain (see Section 2.3 and [55]).

The second implication of this theorem is that each of the two channels in Figure 11 must be sufficiently powerful. HMM inaccuracy can result from using a poor observation distribution family which corresponds to using a channel with too small a capacity. The capacity of an observation distribution is, for example, determined by the number of Gaussian components or covariance type in a Gaussian mixture HMM [107], or the number of hidden units in an HMM with MLP [9] observation distributions [11, 75].

In any event, just increasing the number of components in a Gaussian mixture system or increasing the number of hidden units in an MLP system does not necessarily improve HMM accuracy because the bottle-neck ultimately becomes the fixed number of hidden states (i.e., value of $|\mathcal{Q}|$). Alternatively, simply increasing the number of HMM hidden states might not increase accuracy if the observation model is too weak. Of course, any increase in the number of model parameters must accompany a training data increase to yield reliable low-variance parameter estimates.

Can sufficient conditions for HMM accuracy be found? Assume for the moment that X_t is a discrete random variable with finite cardinality. Recall that $X_{<t} \stackrel{\Delta}{=} X_{1:t-1}$. Suppose that $H_h(Q_t|X_{<t}) = 0$ for all t (a worst case HMM condition to achieve this property is when every observation sequence has its own unique Markov chain state assignment). This implies that Q_t is a deterministic function of $X_{<t}$ (i.e., $Q_t = f(X_{<t})$ for some $f(\cdot)$). Consider the HMM approximation:

$$p_h(x_t|x_{<t}) = \sum_{q_t} p_h(x_t|q_t)p_h(q_t|x_{<t}) \tag{14}$$

but because $H(Q_t|X_{<t}) = 0$, the approximation becomes

$$p_h(x_t|x_{<t}) = p_h(x_t|q_{x_{<t}})$$

where $q_{x_{<t}} = f(x_{<t})$ since every other term in the sum in Equation 14 is zero. The variable X_t is discrete, so for each value of x_t and for each hidden state assignment $q_{x_{<t}}$, the distribution $p_h(X_t = x_t|q_{x_{<t}})$ can be set as follows:

$$p_h(X_t = x_t|q_{x_{<t}}) = p(X_t = x_t|X_{<t} = x_{<t})$$

This last condition might require a number of hidden states equal to the cardinality of the discrete observation space, i.e., $|X_{1:T}|$ which can be very large. In any event, it follows that for all t:

$$\begin{aligned}
& D(p(X_t|X_{<t})||p_h(X_t|X_{<t})) \\
&= \sum_{x_{1:t}} p(x_{1:t}) \log \frac{p(x_t|x_{<t})}{p_h(x_t|x_{<t})} \\
&= \sum_{x_{1:t}} p(x_{1:t}) \log \frac{p(x_t|x_{<t})}{\sum_{q_t} p_h(x_t|q_t)p_h(q_t|x_{<t})} \\
&= \sum_{x_{1:t}} p(x_{1:t}) \log \frac{p(x_t|x_{<t})}{p_h(x_t|q_{x_{<t}})} \\
&= \sum_{x_{1:t}} p(x_{1:t}) \log \frac{p(x_t|x_{<t})}{p(x_t|x_{<t})} \\
&= 0
\end{aligned}$$

It then follows, using the above equation, that:

$$\begin{aligned}
0 &= \sum_t D(p(X_t|X_{<t})||p_h(X_t|X_{<t})) \\
&= \sum_t \sum_{x_{1:t}} p(x_{1:t}) \log \frac{p(x_t|x_{<t})}{p_h(x_t|x_{<t})} \\
&= \sum_t \sum_{x_{1:T}} p(x_{1:T}) \log \frac{p(x_t|x_{<t})}{p_h(x_t|x_{<t})} \\
&= \sum_{x_{1:T}} p(x_{1:T}) \log \frac{\prod_t p(x_t|x_{<t})}{\prod_t p_h(x_t|x_{<t})} \\
&= \sum_{x_{1:T}} p(x_{1:T}) \log \frac{p(x_{1:T})}{p_h(x_{1:T})} \\
&= D(p(X_{1:T})||p_h(X_{1:T}))
\end{aligned}$$

In other words, the HMM is a perfect representation of the true distribution, proving the following theorem.

Theorem 5.2. Sufficient conditions for HMM accuracy. *An HMM as defined above (Definition 3.1) with a joint discrete distribution $p_h(X_{1:T})$ will accurately represent a true discrete distribution $p(X_{1:T})$ if the following conditions hold for all t:*

- $H(Q_t|X_{<t}) = 0$
- $p_h(X_t = x_t|q_{x_{<t}}) = p(X_t = x_t|X_{<t} = x_{<t})$.

It remains to be seen if simultaneously necessary and sufficient conditions can be derived to achieve HMM accuracy, if it is possible to derive sufficient conditions for continuous observation vector HMMs under some reasonable conditions (e.g., finite power, etc.), and what conditions might exist for an HMM that is allowed to have a fixed upper-bound KL-distance error.

6 What HMMs Can't Do

From the previous sections, there appears to be little an HMM can't do. If under the true probability distribution, two random variables possess extremely large mutual information, an HMM approximation might fail because of the required number of states required. This is unlikely, however, for distributions representing objects contained in the natural world.

One problem with HMMs is how they are used; the conditional independence properties are inaccurate when there are too few hidden states, or when the observation distributions are inadequate. Moreover, a demonstration of HMM generality acquaints us not with other inherently more parsimonious models which could be superior. This is explored in the next section.

6.1 How to Improve an HMM

The conceptually easiest way to increase an HMM's accuracy is by increasing the number of hidden states and the capacity of the observation distributions. Indeed, this approach is very effective. In speech recognition systems, it is common to use multiple states per phoneme and to use collections of states corresponding to tri-phones, quad-phones, or even penta-phones. State-of-the-art speech recognition systems have achieved their performance on difficult speech corpora partially by increasing the number of hidden states. For example, in the 1999 DARPA Broadcast News Workshop [19], the best performing systems used penta-phones (a phoneme in the context of two preceding and two succeeding phonemes) and multiple hidden states for each penta-phone. At the time of this writing, some advanced systems condition on both the preceding and succeeding five phonemes leading to what could be called "unodeca-phones." Given limits of training data size, such systems must use methods to reduce what otherwise would be an enormous number of parameters — this is done by automatically tying parameters of different states together [107].

How many hidden states are needed? From the previous section, HMM accuracy might require a very large number. The computations associated with HMMs grow quadratically $O(TN^2)$ with N the number of states, so while increasing the number of states is simple, there is an appreciable associated computational cost (not to mention the need for more training data).

In general, given enough hidden states and a sufficiently rich class of observation distributions, an HMM can accurately model any real-world probability distribution. HMMs therefore constitute a very powerful class of probabilistic model families. In theory, at least, there is no limit to their ability to model a distribution over signals representing natural scenes.

Any attempt to advance beyond HMMs, rather than striving to correct intrinsic HMM deficiencies, should instead start with the following question: is there a class of models that inherently leads to more parsimonious representations (i.e., fewer parameters, lower complexity, or both) of the relevant aspects of speech, and that also provides the same or better speech recognition (or more generally, classification) performance, better generalizability, or better robustness to noise? Many alternatives have been proposed, some of which are discussed in subsequent paragraphs.

One HMM alternative, similar to adding more hidden states, factors the hidden representation into multiple independent Markov chains. This type of representation is shown as a graphical model in Figure 12. Factored hidden state representations have been called HMM decomposition [98, 99], and factorial HMMs [44, 92]. A related method that estimates the parameters of a composite HMM given a collection of separate, independent, and already trained HMMs is called parallel model combination [41]. A factorial HMM can represent the combination of multiple signals produced independently, the characteristics of each described by a distinct Markov chain. For example, one chain might represent speech and another could represent some dynamic noise source [61] or background speech [99]. Alternatively, the two chains might each represent two underlying concurrent sub-processes governing the realization of the observation vectors [70] such as separate articulatory configurations [87, 88]. A modified factorial HMMs couples each Markov chain using a cross-chain dependency at each time step [108, 110, 109, 92]. In this case, the first chain represents the typical phonetic constituents of speech and the second chain is encouraged to represent articulatory attributes of the speaker (e.g., the voicing condition).

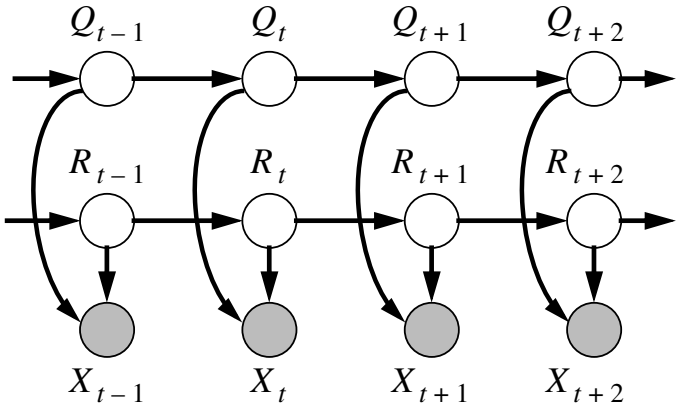


Figure 12: A factorial HMM with two underlying Markov chains Q_t and R_t governing the temporal evolution of the statistics of the observation vectors X_t.

The factorial HMMs described above are all special cases of HMMs. That is, they are HMMs with tied parameters and state transition restrictions made according to the factorization. Starting with a factorial HMM consisting of two hidden chains Q_t and R_t, an equivalent HMM may be constructed using $|\mathcal{Q}||\mathcal{R}|$ states and by restricting the set of state transitions and parameter assignments to be those only allowed by the factorial model. A factorial HMM using M hidden Markov chains each with K states that all span over T time steps has complexity $O(TMK^{M+1})$ [44]. If one translates the factorial HMM into an HMM having K^M states, the complexity becomes $O(TK^{2M})$. The underlying complexity of an factorial HMM therefore is significantly smaller than that of an equivalent HMM. An unrestricted HMM with K^M states, however, has more expressive power than a factorial HMM with M chains each with K states because in the HMM there can be more transition restrictions via the dependence represented between the separate chains.

More generally, dynamic Bayesian networks (DBNs) are Bayesian networks consisting of a sequence of DGMs

strung together with arrows pointing in the direction of time (or space). Factorial HMMs are an example of DBNs. Certain types of DBNs have been investigated for speech recognition [8, 108].

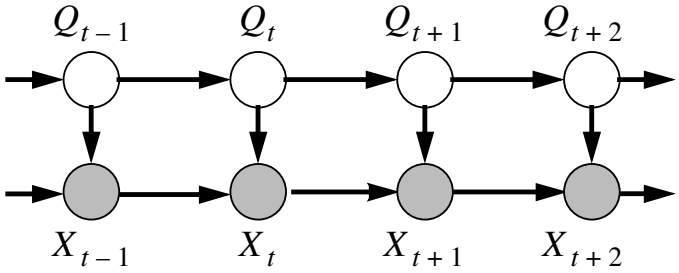


Figure 13: An HMM augmented with dependencies between neighboring observations.

Some HMMs use neural networks as discriminatively trained phonetic posterior probability estimators [11, 75]. By normalizing with prior probabilities $p(q)$, posterior probabilities $p(q|x)$ are converted to scaled likelihoods $p(x|q)/p(x)$. The scaled likelihoods are then substituted for HMM observation distribution evaluations. Multi-layered perceptrons (MLP) or recurrent neural networks [9] are the usual posterior estimator. The size of the MLP hidden-layer determines the capacity of the observation distributions. The input layer of the network typically spans, both into the past and the future, a number of temporal frames. Extensions to this approach have also been developed [63, 35].

A remark that can be made about a specific HMM is that additional information might exist about an observation X_t in an adjacent frame (say X_{t-1}) that is not supplied by the hidden variable Q_t. This is equivalent to the statement that the conditional independence property $X_t \perp\!\!\!\perp X_{t-1}|Q_t$ is inaccurate. As a consequence, one may define correlation [101] or conditionally Gaussian [77] HMMs, where an additional dependence is added between adjacent observation vectors. In general, the variable X_t might have as a parent not only the variable Q_t but also the variables X_{t-l} for $l = 1, 2, \ldots, K$ for some K. The case where $K = 1$ is shown in Figure 13.

A K^{th}-order Gaussian vector auto-regressive (AR) process [47] may be exemplified using control-theoretic state space equations such as:

$$x_t = \sum_{k=1}^{K} A_k x_{t-k} + \epsilon$$

where A_k is a matrix that controls the dependence of x_t on the k^{th} previous observation, and ϵ is a Gaussian random variable with some mean and variance. As described in Section 3, a Gaussian mixture HMM may also be described using similar notation. Using this scheme, a general K^{th} order conditionally mixture-Gaussian HMM may be described as follows:

$$\begin{aligned} q_t &= i \text{ with probability } p(Q_t = i|q_{t-1}) \\ x_t &\sim \sum_{k=1}^{K} A_k^{q_t n} x_{t-k} + \mathcal{N}(\mu_{q_t n}, \Sigma_{q_t n}) \text{ with prob. } c_{q_t n} \text{ for } i = \{1, 2, \ldots, N\} \end{aligned}$$

where K is the auto-regression order, A_k^{in} is the regression matrix and c_{in} is the mixture coefficient for state i and mixture n (with $\sum_n c_{in} = 1$ for all i), and N is the number of mixture components per state. In this case, the mean of the variable X_t is determined using previous observations and the mean of the randomly chosen Gaussian component $\mu_{q_t n}$.

Although these models are sometimes called vector-valued auto-regressive HMMs, they are not to be confused with auto-regressive, linear predictive, or hidden filter HMMs [83, 84, 60, 85] which are HMMs that, inspired from linear-predictive coefficients for speech [85], use the observation distribution that arises from coloring a random source with a hidden-state conditioned AR filter.

Gaussian vector auto-regressive processes have been attempted for speech recognition with $K = 1$ and $N = 1$. This was presented in [101] along with EM update equations for maximum-likelihood parameter estimation. Speech recognition results were missing from that work, although an implementation apparently was tested [10] and found not to improve on the case without the additional dependencies. Both [13] and [62] tested implementations of such models with mixed success. Namely, improvements were found only when "delta features" (to be described shortly)

were excluded. Similar results were found by [25] but for segment models (also described below). In [79], the dependency structure in Figure 13 used discrete rather than Gaussian observation densities. And in [76], a parallel algorithm was presented that can efficiently perform inference with such models.

The use of dynamic or delta features [31, 37, 38, 39] has become standard in state-of-the-art speech recognition systems. While incorporating delta features does not correspond to a new model per se, they can also be viewed as an HMM model augmentation. Similar to conditionally Gaussian HMMs, dynamic features also represent dependencies in the feature streams. Such information is gathered by computing an estimate of the time derivative of each feature $\frac{d}{dt}X_t = \dot{X}_t$ and then augmenting the feature stream with those estimates, i.e., $X'_t = \{X_t, \frac{d}{dt}X_t\}$. Acceleration, or delta-delta, features are defined similarly and are sometimes found to be additionally beneficial [104, 65].

Most often, estimates of the feature derivative are obtained [85] using linear regression, i.e.,

$$\dot{x_t} = \frac{\sum_{k=-K}^{K} kx_t}{\sum_{k=-K}^{K} k^2}$$

where K is the number of points to fit the regression. Delta (or delta-delta) features are therefore similar to auto-regression, but where the regression is over samples not just from the past but also from the future. That is, consider a hypothetical process defined by

$$x_t = \sum_{k=-K}^{K} a_k x_{t-k} + \epsilon$$

where the fixed regression coefficients a_k are defined by $a_k = -k/\sum_{l=-K}^{K} l^2$ for $k \neq 0$ and $a_0 = 1$. This is equivalent to

$$x_t - \sum_{k=-K}^{K} a_k x_{t-k} = \frac{\sum_{k=-K}^{K} kx_{t-k}}{\sum_{l=-K}^{K} l^2} = \epsilon$$

which is the same as modeling delta features with a single Gaussian component.

The addition of delta features to a feature stream is therefore similar to additionally using a separate conditionally Gaussian observation model. Observing the HMM DGM (Figure 5), delta features add dependencies between observation nodes and their neighbors from both the past and the future (the maximum range determined by K). Of course, this would create a directed cycle in a DGM violating its semantics. To be theoretically accurate, one must perform a global re-normalization as is done with a Markov random field [23]. Nevertheless, it can be seen that the use of delta features corresponds in some sense to a relaxation of the HMM conditional independence properties.

As mentioned above, conditionally Gaussian HMMs often do not supply an improvement when delta features are included in the feature stream. Improvements were reported with delta features in [106] using discriminative output distributions [105]. In [66, 67], successful results were obtained using delta features but where the conditional mean, rather than being linear, was non-linear and was implemented using a neural network. Also, in [96], benefits were obtained using mixtures of discrete distributions. In a similar model, improvements when using delta features were also reported when sparse dependencies were chosen individually between feature vector elements, and according to an data-driven hidden-variable dependent information-theoretic criteria [8, 7, 5].

In general, one can consider the model

$$\begin{aligned} q_t &= i \text{ with prob. } p(Q_t = i|q_{t-1}) \\ x_t &= F_t(x_{t-1}, x_{t-1}, \ldots, x_{t-k}) \end{aligned}$$

where F_t is an arbitrary random function of the previous k observations. In [21, 22], the model becomes

$$x_t = \sum_{k=1}^{K} \phi_{q_t,t,k} x_{t-k} + g_{q_t,t} + \epsilon_{q_t}$$

where $\phi_{i,t,k}$ is a dependency matrix for state i and time lag k and is a polynomial function of t, $g_{i,t}$ is a fixed mean for state i and time t, and ϵ_i is a state dependent Gaussian. Improvements using this model were also found with feature streams that included delta features.

Another general class of models that extend HMMs are called segment or trajectory models [77]. In a segment model, the underlying hidden Markov chain governs the statistical evolution not of the individual observation vectors. Instead, it governs the evolution of sequences (or segments) of observation vectors where each sequence may be described using an arbitrary distribution. More specifically, a segment model uses the joint distribution of a variable length segment of observations conditioned on the hidden state for that segment. In a segment model, the joint distribution of features can be described as follows:

$$\begin{aligned} & p(X_{1:T} = x_{1:T}) \qquad (15) \\ = & \sum_{\tau} \sum_{q_{1:\tau}} \sum_{\ell_{1:\tau}} \prod_{i=1}^{\tau} p(x_{t(q_{1:\tau},\ell_{1:\tau},i,1)}, x_{t(q_{1:\tau},\ell_{1:\tau},i,2)}, \ldots, x_{t(q_{1:\tau},\ell_{1:\tau},i,\ell_i)}, \ell_i | q_i, \tau) p(q_i | q_{i-1}, \tau) p(\tau) \end{aligned}$$

There are T time frames and τ segments where the i^{th} segment has hypothesized length ℓ_i. The collection of lengths are constrained so that $\sum_{i=1}^{\tau} \ell_i = T$. For a hypothesized segmentation and set of lengths, the i^{th} segment starts at time frame $t(q_{1:\tau}, \ell_{1:\tau}, i, 1)$ and ends at time frame $t(q_{1:\tau}, \ell_{1:\tau}, i, \ell_i)$. In this general case, the time variable t could be a function of the complete Markov chain assignment $q_{1:\tau}$, the complete set of currently hypothesized segment lengths $\ell_{1:\tau}$, the segment number i, and the frame position within that segment 1 through ℓ_i. It is assumed that $t(q_{1:\tau}, \ell_{1:\tau}, i, \ell_i) = t(q_{1:\tau}, \ell_{1:\tau}, i+1, 1) - 1$ for all values of every quantity.

Renumbering the time sequence for a hypothesized segment starting at one, the joint distribution over the observations of a segment is given by:

$$p(x_1, x_2, \ldots, x_\ell, \ell | q) = p(x_1, x_2, \ldots, x_\ell | \ell, q) p(\ell | q)$$

where $p(x_1, x_2, \ldots, x_\ell | \ell, q)$ is the joint segment probability for length ℓ and for hidden Markov state q, and where $p(\ell|q)$ is the explicit duration model for state q.

An HMM occurs in this framework if $p(\ell|q)$ is a geometric distribution in ℓ and if

$$p(x_1, x_2, \ldots, x_\ell | \ell, q) = \prod_{j=1}^{\ell} p(x_j | q)$$

for a state specific distribution $p(x|q)$. The stochastic segment model [78] is a generalization which allows observations in a segment to be additionally dependent on a region within a segment

$$p(x_1, x_2, \ldots, x_\ell | \ell, q) = \prod_{j=1}^{\ell} p(x_j | r_j, q)$$

where r_j is one of a set of fixed regions within the segment. A slightly more general model is called a segmental hidden Markov model [40]

$$p(x_1, x_2, \ldots, x_\ell | \ell, q) = \int p(\mu | q) \prod_{j=1}^{\ell} p(x_j | \mu, q) d\mu$$

where μ is the multi-dimensional conditional mean of the segment and where the resulting distribution is obtained by integrating over all possible state-conditioned means in a Bayesian setting. More general still, in trended hidden Markov models [21, 22], the mean trajectory within a segment is described by a polynomial function over time. Equation 15 generalizes many models including the conditional Gaussian methods discussed above. An excellent summary of segment models, their learning equations, and a complete bibliography is given in [77].

Markov Processes on Curves [90] is a recently proposed dynamic model that may represent speech at various speaking rates. Certain measures on continuous trajectories are invariant to some transformations, such as monotonic non-linear time warpings. The arc-length, for example, of a trajectory $x(t)$ from time t_1 to time t_2 is given by:

$$\ell = \int_{t_1}^{t_2} [\dot{x}(t) g(x(t)) \dot{x}(t)]^{1/2} \, dt$$

where $\dot{x}(t) = \frac{d}{dt} x(t)$ is the time derivative of $x(t)$, and $g(x)$ is an arc-length metric. The entire trajectory $x(t)$ is segmented into a collection of discrete segments. Associated with each segment of the trajectory is a particular state

of a hidden Markov chain. The probability of staying in each Markov state is controlled by the arc-length of the observation trajectory. The resulting Markov process on curves is set up by defining a differential equation on $p_i(t)$ which is the probability of being in state i at time t. This equation takes the form:

$$\frac{dp_i}{dt} = -\lambda_i p_i \left[\dot{x}(t) g_i(x(t)) \dot{x}(t)\right]^{1/2} + \sum_{j \neq i} \lambda_j p_j a_{ji} \left[\dot{x}(t) g_j(x(t)) \dot{x}(t)\right]^{1/2}$$

where λ_i is the rate at which the probability of staying in state i declines, a_{ji} is the transition probability of the underlying Markov chain, and $g_j(x)$ is the length metric for state j. From this equation, a maximum likelihood update equations and segmentation procedures can be obtained [90].

The hidden dynamic model (HDM) [12] is another recent approach to speech recognition. In this case, the hidden space is extended so that it can simultaneously capture both the discrete events that ultimately are needed for words and sentences, and also continuous variables such as formant frequencies (or something learned in an unsupervised fashion). This model attempts to explicitly capture coarticulatory phenomena [14], where neighboring speech sounds can influence each other. In an HDM, the mapping between the hidden continuous and the observed continuous acoustic space is performed using an MLP. This model is therefore similar to a switching Kalman filter, but with non-linear hidden to observed mapping between continuous spaces rather than a Gaussian regressive process.

A Buried Markov model (BMM) [8, 7, 5] is another recent approach to speech recognition. A BMM is based on the idea that one can quantitatively measure where a specific HMM is failing on a particular corpus, and extend it accordingly. For a BMM, the accuracy is measured of the HMM conditional independence properties themselves. The model is augmented to include only those data-derived, sparse, and hidden-variable specific dependencies (between observation vectors) that are most lacking in the original model. In general, the degree to which $X_{t-1} \perp\!\!\!\perp X_t | Q_t$ is true can be measured using conditional mutual information $I(X_{t-1}; X_t | Q_t)$ [16]. If this quantity is zero, the model is perfect and needs no extension. The quantity indicates a modeling inaccuracy if it is greater than zero. Augmentations based on conditional mutual information alone is likely to improve only synthesis and not recognition, which requires a more discriminative model. Therefore, a quantity called discriminative conditional mutual information (derivable from the posterior probability) determines new dependencies. Since it attempts to minimally correct only those measured deficiencies in a particular HMM, and since it does so discriminatively, this approach has the potential to produce better performing and more parsimonious models for speech recognition.

All the models described above are interesting in different ways. They each have a natural mode where, for a given number of parameters, they succinctly describe a certain class of signals. It is apparent that Gaussian mixture HMMs are extremely well suited to speech as embodied by MFCC [107] features. It may be the case that other features [50, 51, 46, 4] are more appropriate under these models. As described in Section 5, however, since HMMs are so flexible, and since structurally discriminative but not necessarily descriptive models are required for speech recognition, it is uncertain how much additional capacity these models supply. Nevertheless, they all provide interesting and auspicious alternatives when attempting to move beyond HMMs.

7 Conclusion

This paper has presented a tutorial on hidden Markov models. Herein, a list of properties was subjected to a new HMM definition, and it was found that HMMs are extremely powerful, given enough hidden states and sufficiently rich observation distributions. Moreover, even though HMMs encompass a rich class of variable length probability distributions, for the purposes of classification, they need not precisely represent the true conditional distribution — even if a specific HMM only crudely reflects the nature of a speech signal, there might not be any detriment to their use in the recognition task, where a model need only internalize the distinct attributes of its class. This later concept has been termed structural discriminability, and refers to how inherently discriminative a model is, irrespective of the parameter training method. In our quest for a new model for speech recognition, therefore, we should be concerned less with what is wrong with HMMs, and rather seek models leading to inherently more parsimonious representations of only those most relevant aspects of the speech signal.

8 Acknowledgements

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] P. Billingsley. *Probability and Measure*. Wiley, 1995.

[3] J.A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical Report TR-97-021, ICSI, 1997.

[4] J.A. Bilmes. Joint distributional modeling with cross-correlation based features. In *Proc. IEEE ASRU*, Santa Barbara, December 1997.

[5] J.A. Bilmes. Data-driven extensions to HMM statistical dependencies. In *Proc. Int. Conf. on Spoken Language Processing*, Sidney, Australia, December 1998.

[6] J.A. Bilmes. Maximum mutual information based reduction strategies for cross-correlation based joint distributional modeling. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Seattle, WA, May 1998.

[7] J.A. Bilmes. Buried Markov models for speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, March 1999.

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

[9] C. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[10] H. Bourlard. Personal communication, 1999.

[11] H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[12] J. Bridle, L. Deng, J. Picone, H. Richards, J. Ma, T. Kamm, M. Schuster, S. Pike, and R. Reagan. An investigation fo segmental hidden dynamic models of speech coarticulation for automatic speech recognition. *Final Report for 1998 Workshop on Langauge Engineering, CLSP, Johns Hopkins*, 1998.

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

[14] J. Clark and C. Yallop. *An Introduction to Phonetics and Phonology*. Blackwell, 1995.

[15] G. Cooper and E. Herskovits. Computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42:393–405, 1990.

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

[17] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, 1999.

[18] 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.

[19] Darpa 1999 broadcast news workshop. DARPA Notebooks and Proceedings, Feb 1999. Hilton at Washington Dulles Airport.

[20] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Ser. B.*, 39, 1977.

[21] L. Deng, M. Aksmanovic, D. Sun, and J. Wu. Speech recognition using hidden Markov models with polynomial regression functions as non-stationary states. *IEEE Trans. on Speech and Audio Proc.*, 2(4):101–119, 1994.

[22] L. Deng and C. Rathinavelu. A Markov model containing state-conditioned second-order non-stationarity: application to speech recognition. *Computer Speech and Language*, 9(1):63–86, January 1995.

[23] H. Derin and P. A. Kelley. Discrete-index Markov-type random processes. *Proc. of the IEEE*, 77(10):1485–1510, October 1989.

[24] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Applications of Mathematics. Springer, 1996.

[25] V. Digalakis, M. Ostendorf, and J.R. Rohlicek. Improvements in the stochastic segment model for phoneme recognition. *Proc. DARPA Workshop on Speech and Natural Language*, 1989.

[26] J. L. Doob. *Stochastic Processes*. Wiley, 1953.

[27] R. Drullman, J. M. Festen, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *JASA*, 95(5):2670–2680, May 1994.

[28] R. Drullman, J.M. Festen, and R. Plomp. Effect of temporal envelope smearing on speech reception. *Journal of the Acoustical Society of America*, 95(2):1053–1064, February 1994.

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

[30] H. Dudley. Remaking speech. *Journal of the Acoustical Society of America*, 11(2):169–177, October 1939.

[31] K. Elenius and M. Blomberg. Effects of emphasizing transitional or stationary parts of the speech signal in a discrete utterance recognition system. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 535–538, 1982.

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

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

[34] N. Friedman, K. Murphy, and S. Russell. Learning the structure of dynamic probabilistic networks. *14th Conf. on Uncertainty in Artificial Intelligence*, 1998.

[35] J. Fritsch. ACID/HNN: A framework for hierarchical connectionist acoustic modeling. In *Proc. IEEE ASRU*, Santa Barbara, December 1997.

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

[37] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2):254–272, April 1981.

[38] S. Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(1):52–59, February 1986.

[39] Sadaoki Furui. On the role of spectral transition for speech perception. *Journal of the Acoustical Society of America*, 80(4):1016–1025, October 1986.

[40] M.J.F. Gales and S.J. Young. Segmental hidden Markov models. In *European Conf. on Speech Communication and Technology (Eurospeech)*, 3rd, pages 1579–1582, 1993.

[41] M.J.F. Gales and S.J. Young. Robust speech recognition in additive and convolutional noise using parallel model combination. *Computer Speech and Language*, 9:289–307, 1995.

[42] R.G. Gallager. *Information Theory and Reliable Communication*. Wiley, 1968.

[43] Z. Ghahramani. *Lecture Notes in Artificial Intelligence*, chapter Learning Dynamic Bayesian Networks. Springer-Verlag, 1998.

[44] Z. Ghahramani and M. Jordan. Factorial hidden Markov models. *Machine Learning*, 29, 1997.

[45] S. Greenberg. Understanding speech understanding: Towards a unified theory of speech perception. In William Ainsworth and Steven Greenberg, editors, *Workshop on the Auditory Basis of Speech Perception*, pages 1–8, Keele University, UK, July 1996.

[46] S. Greenberg and B. Kingsbury. The modulation spectrogram: in pursuit of an invariant representation of speech. In *Proceedings ICASSP-97*, pages 1647–1650, 1997.

[47] G.R. Grimmett and D.R. Stirzaker. *Probability and Random Processes*. Oxford Science Publications, 1991.

[48] D. Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft, 1995.

[49] D. Heckerman, Max Chickering, Chris Meek, Robert Rounthwaite, and Carl Kadie. Dependency networks for density estimation, collaborative filtering, and data visualization. In *Proceedings of the 16th conf. on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 2000.

[50] H. Hermansky. Int. Conf. on Spoken Language Processing, 1998. Panel Discussion.

[51] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[52] J. Hopcroft and J. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison Wesley, 1979.

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

[54] T.S. Jaakkola and M.I. Jordan. *Learning in Graphical Models*, chapter Improving the Mean Field Approximations via the use of Mixture Distributions. Kluwer Academic Publishers, 1998.

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

[56] F.V. Jensen. *An Introduction to Bayesian Networks*. Springer, 1996.

[57] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. *Learning in Graphical Models*, chapter An Introduction to Variational Methods for Graphical Models. Kluwer Academic Publishers, 1998.

[58] 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.

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

[60] B.-H. Juang and L.R. Rabiner. Mixture autoregressive hidden Markov models for speech signals. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 33(6):1404–1413, December 1985.

[61] M. Kadirkamanathan and A.P. Varga. Simultaneous model re-estimation from contaminated data by composed hidden Markov modeling. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 897–900, 1991.

[62] P. Kenny, M. Lennig, and P. Mermelstein. A linear predictive HMM for vector-valued observations with applications to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(2):220–225, February 1990.

[63] Y. Konig. *REMAP: Recursive Estimation and Maximization of A Posterior Probabilities in Transition-based Speech Recognition*. PhD thesis, U.C. Berkeley, 1996.

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

[65] C.-H. Lee, E. Giachin, L.R. Rabiner, R. Pieraccini, and A.E. Rosenberg. Improved acoustic modeling for speaker independent large vocabulary continuous speech recognition. *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 1991.

[66] E. Levin. Word recognition using hidden control neural architecture. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 433–436. IEEE, 1990.

[67] E. Levin. Hidden control neural architecture modeling of nonlinear time varying systems and its applications. *IEEE Trans. on Neural Networks*, 4(1):109–116, January 1992.

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

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

[70] B.T. Logan and P.J. Moreno. Factorial HMMs for acoustic modeling. *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 1998.

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

[72] D.J.C. MacKay. *Learning in Graphical Models*, chapter Introduction to Monte Carlo Methods. Kluwer Academic Publishers, 1998.

[73] C. D. Manning and H. Sch utze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[74] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[75] N. Morgan and H. Bourlard. Continuous speech recognition. *IEEE Signal Processing Magazine*, 12(3), May 1995.

[76] H. Noda and M.N. Shirazi. A MRF-based parallel processing algorithm for speech recognition using linear predictive HMM. *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 1994.

[77] M. Ostendorf, V. Digalakis, and O. Kimball. From HMM's to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Trans. Speech and Audio Proc.*, 4(5), September 1996.

[78] M. Ostendorf, A. Kannan, O. Kimball, and J. Rohlicek. Continuous word recognition based on the stochastic segment model. *Proc. DARPA Workshop CSR*, 1992.

[79] K.K. Paliwal. Use of temporal correlations between successive frames in a hidden Markov model based speech recognizer. *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages II–215/18, 1993.

[80] A. Papoulis. *Probability, Random Variables, and Stochastic Processes, 3rd Edition*. McGraw Hill, 1991.

[81] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 2nd printing edition, 1988.

[82] J. Pearl. *Causality*. Cambridge, 2000.

[83] A.B. Poritz. Linear predictive hidden Markov models and the speech signal. *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 1291–1294, 1982.

[84] A.B. Poritz. Hidden Markov models: A guided tour. *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 7–13, 1988.

[85] L.R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing Series, 1993.

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

[87] M. Richardson, J. Bilmes, and C. Diorio. Hidden-articulator markov models for speech recognition. In *Proc. of the ISCA ITRW ASR2000 Workshop*, Paris, France, 2000. LIMSI-CNRS.

[88] M. Richardson, J. Bilmes, and C. Diorio. Hidden-articulator markov models: Performance improvements and robustness to noise. In *Proc. Int. Conf. on Spoken Language Processing*, Beijing, China, 2000.

[89] S.J. Russel and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 1995.

[90] L. Saul and M. Rahim. Markov processes on curves for automatic speech recognition. *NIPS*, 11, 1998.

[91] L.K. Saul, T. Jaakkola, and M.I. Jordan. Mean field theory for sigmoid belief networks. *JAIR*, 4:61–76, 1996.

[92] L.K. Saul and M.I. Jordan. Mixed memory markov models: Decomposing complex stochastic processes as mixtures of simpler ones. *Machine Learning*, 1999.

[93] R.D. Shachter. Bayes-ball: The rational pastime for determining irrelevance and requisite information in belief networks and influence diagrams. In *Uncertainty in Artificial Intelligence*, 1998.

[94] 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.

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

[96] S. Takahashi, T. Matsuoka, Y. Minami, and K. Shikano. Phoneme HMMs constrained by frame correlations. *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 1993.

[97] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[98] A.P. Varga and R.K. Moore. Hidden Markov model decomposition of speech and noise. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 845–848, Alburquerque, April 1990.

[99] A.P. Varga and R.K. Moore. Simultaneous recognition of concurrent speech signals using hidden makov model decomposition. In *European Conf. on Speech Communication and Technology (Eurospeech)*, 2nd, 1991.

[100] Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation*, 12(1):1–41, 2000.

[101] C.J. Wellekens. Explicit time correlation in hidden Markov models for speech recognition. *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 384–386, 1987.

[102] J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. John Wiley and Son Ltd., 1990.

[103] D. Williams. *Probability with Martingales*. Cambridge Mathematical Textbooks, 1991.

[104] J.G. Wilpon, C.-H. Lee, and L.R. Rabiner. Improvements in connected digit recognition using higher order spectral and energy features. *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 1991.

[105] P.C. Woodland. Optimizing hidden Markov models using discriminative output distributions. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 1991.

[106] P.C. Woodland. Hidden Markov models using vector linear prediction and discriminative output distributions. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages I–509–512, 1992.

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

[108] G. Zweig. *Speech Recognition with Dynamic Bayesian Networks*. PhD thesis, U.C. Berkeley, 1998.

[109] G. Zweig and M. Padmanabhan. Dependency modeling with bayesian networks in a voicemail transcription system. In *European Conf. on Speech Communication and Technology (Eurospeech)*, 6th, 1999.

[110] G. Zweig and S. Russell. Probabilistic modeling with Bayesian networks for automatic speech recognition. In *Int. Conf. on Spoken Language Processing*, 1998.