EXPRESSIVE PROBABILITY MODELS FOR SPEECH RECOGNITION AND UNDERSTANDING

Stuart Russell

Computer Science Division
University of California
Berkeley, CA 94720, USA
russell@cs.berkeley.edu

ABSTRACT

The paper is a brief summary of an invited talk given at the ASRU 99 conference. The principal points are as follows: first, that the expressive power of the probability models available for use in speech recognition and understanding has expanded significantly; second, that using expressive models such as dynamic Bayesian networks can result in improved learning rates and recognition results; and finally that still further expansion is required to tackle many problems of interest in the area of speech understanding. This further expansion should combine probability theory with the expressive power of first-order logical languages. The paper sketches an approximate inference method for representation systems of this kind.

1. EXPRESSIVE POWER

Probability models (hypotheses) typically specify probability distributions over *events*; for example, a hidden Markov model specifies the probabilities of events defined as sequences of hidden states and observations. Roughly speaking, a model family $\mathcal{L}_2$ is at least as expressive as $\mathcal{L}_1$ iff, for every e_1 in $\mathcal{L}_1$, there is some e_2 in $\mathcal{L}_2$ such that e_1 and e_2 denote "the same" distribution. Crude expressive power in this sense is obviously important, but more important is *efficient* representation: we prefer desirable distributions to be represented by subfamilies with "fewer parameters," enabling faster learning.

The simplest probability model is the "atomic" distribution that specifies a probability explicitly for each event. For example, a model for a coin flip specifies 0.5 probability of heads and 0.5 probability of tails.

Bayesian networks, generalized linear models, and, to some extent, neural networks provide compact descriptions of *multivariate* distributions. These models are all analogues of *propositional* logic representations. A Bayesian network, for example, defines a joint distribution on variables $X_1, \ldots, X_n$ as the product

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i | Parents(X_i))$$

where $Parents(X_i)$ are the parent nodes of X_i in the directed network. If the indegree of each node in the network is bounded by a constant d, then the total number of parameters required to specify the joint distribution as a Bayesian network is $O(2^d n)$, as opposed to $O(2^d)$ for the atomic distribution. This has at least two consequences: first, it becomes possible to represent models with many hundreds of variables; second, the amount of data required to learn such models grows roughly linearly with the number of variables [3].

Additional expressive power is needed to represent temporal processes. A temporal process involves an infinite number of random variables, hence temporal modelling languages involve some kind of universal quantification over time. Hidden Markov models (HMMs), for example, are temporal models in which the transition model $\mathbf{P}(X_{t+1}|X_t)$ and the sensor or emission model $\mathbf{P}(E_t|X_t)$ are assumed constant for all t. HMMs are atomic models in the sense that the state of the modelled system is represented by a single state variable.

Propositional temporal models, in which the state is represented by a set of state variables, include Kalman filters, recurrent neural networks, and dynamic Bayesian networks (DBNs). In this paper, we will focus on DBNs because of their natural relationship to HMMs. In a (first-order) DBN, each state variable may have parents in the current or preceding time step; sensor variables have parents in the same time step; and all conditional distributions are assumed constant over time [5]. Every HMM can be represented as a DBN (see Figure 1), and every DBN can be represented as an HMM by constructing a single state variable whose domain is the cross-product of the state variables of the DBN.

The efficiency advantage of multivariate over atomic models in the static case carries over to the temporal case. For example, an HMM representing a system with n bits of state requires a state variable with 2^n values and a transition model

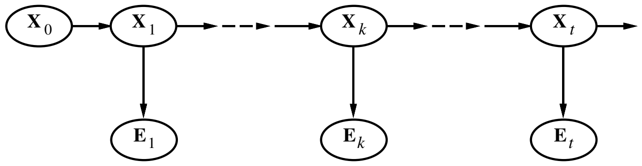

Figure 1: A DBN representation of an HMM.

with 2^{2n} parameters. On the other hand, a DBN requires $O(2^d n)$ parameters *provided the temporal process is sparse* in the sense that each variable depends on at most d predecessors. Note that sparseness for DBNs is *not* the same as sparseness for HMMs. A DBN representing n independently evolving processes has n separate chains, whereas the transition matrix of the equivalent HMM contains no zeroes. Section 2 suggests that this representational advantage results in improved performance on speech recognition tasks. DBNs seem promising for modelling a variety of other complex systems, such as economic systems, metabolic systems, and gene regulation networks.

Exact inference algorithms for DBNs are well-understood [15] and the analogies to HMM inference and Kalman filter inference have been explored in depth [21, 20]. Parametric learning for DBNs works in much the same way as for HMMs, with both EM and gradient-based methods leading to simple parameter update formulae that are additive over the sequence [21, 1]. Structural learning to create the topology of the network is less well developed [7] and methods for introducing new hidden variables are sorely needed.

2. DBNS FOR SPEECH RECOGNITION

Speech recognition is an obvious target for Bayesian network learning: large data sets are available, trainable probabilistic models (HMMs) are known to be effective, and the effectiveness of those models appears to be limited by the explosion in the number of parameters as one attempts to model speech more faithfully. In recent work carried out primarily by Geoff Zweig [22, 23], it has been shown that DBNs do in fact yield somewhat better results than HMMs of equivalent complexity.

Like HMMs, our DBN speech models include a pronunciation model (phonetic transitions) and an acoustic model (acoustic signal given phonetic state). However, our acoustic model includes additional state variables that we will call “articulatory context” variables; the intent is that these may capture the state of the articulatory apparatus of the speaker—for example, voicing, lips open/closed and stretched/rounded, jaw open/closed, nasalization, tongue position and shape. These variables can depend on both the current phonetic state and the previous articulatory context. Figure 2 illustrates a DBN structured for speech recognition in this manner. Our hypothesis is that the transition models learned for the articulatory context variables will represent physical constraints on the motion of the articulatory apparatus that are important in the production of coarticulation effects and other important speech phenomena. Ultimately, a such a model can be viewed also as a generative model that can be used for text-to-speech systems as well as recognition.

States per Phone	Number of Context Variables	Ctxt Var. Arity	Total System Params.	Word Error Rate
3	0(HMM)	-	96k	5.4%
3	1	2	191k	4.1%
3	1	3	287k	4.0%
3	1	4	383k	3.8%
3	2	2	383k	3.6%
4	0(HMM)	-	127k	4.8%
4	1	2	254k	3.6%
4	1	3	381k	3.5%
4	1	4	508k	3.2%
4	2	2	508k	3.2%

Table 1: Results for networks with one and two context variables per frame; $\sigma \approx 0.25\%$.

Network	Parameters	Error Rate
CDA-HMM	257k	3.2%
CDA-ChainDBN	515k	2.6%
CDA-HMM	510k	3.1%

Table 2: Test results with a context-dependent alphabet (CDA). The first two CDA results used 336 phonemes; the last CDA result used less frequently occurring phonemes and had a size of 666. The CDA-ChainDBN has the topology of Figure 2, with a binary context variable. $\sigma \approx 0.2\%$.

2.1. Initial results

Our initial experiments have been limited to minimal representations of articulatory state using just one or two variables with two to four values each. As a test-bed, we selected the Phonebook database, a large-vocabulary, isolated-word database compiled by researchers at NYNEX [18]. Some results from [23] are shown in Tables 1 and 2. The latter paper also includes results showing that the simple mechanism of adding a single atemporal variable to the DBN topology, with links to all the state variables, provides a method for automatic clustering of utterances. The learned models involving this variable appear to to be highly correlated with speaker gender, although some word clustering also occurs.

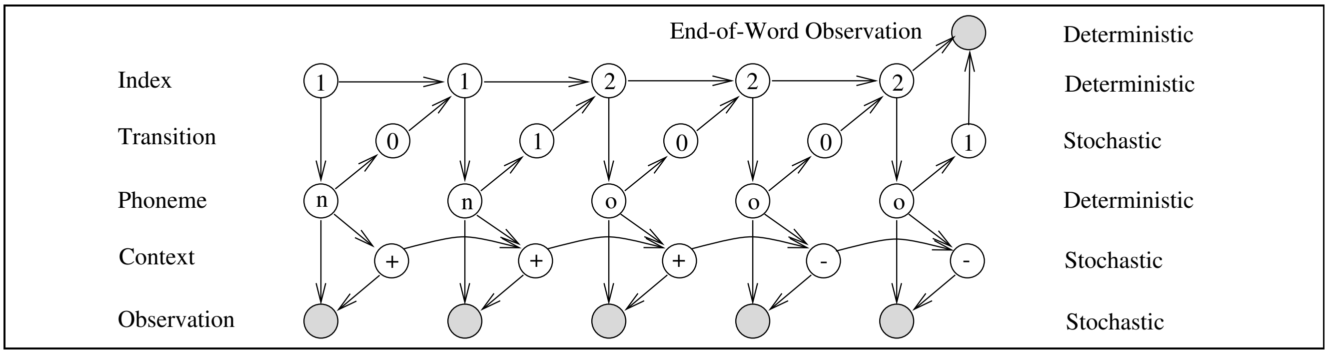

Figure 2: A DBN for speech recognition. The index, transition, phoneme, and end-of-word variables encode a probability distribution over phonetic sequences. The context and observation variables encode a distribution over observations, conditioned on phonetic sequence. A valid set of variable assignments is indicated for the word "no." In this picture, the context variable represents nasalization. The vowel $/o/$ is not usually nasalized, but in this case coarticulation causes nasalization of its first occurrence.

2.2. Immediate challenges

A number of problems must be addressed in order to achieve our goal of effective modelling of the speech apparatus.

- *Efficient inference*: as with HMMs, the cost of exact inference in DBNs scales with the square of the number of hidden states and hence exponentially in the number of state variables. A speech model with, say, a dozen articulatory context variables is probably infeasible for exact inference in the near term. Methods for approximate inference include *variational* algorithms [8] and Monte Carlo sampling algorithms, particularly the *particle filtering* method [11, 14, 4].
- *Structure learning*: once we have a reasonable number of context variables, it will become infeasible to explore all possible topologies for connecting them. Some recent work on structure learning for DBNs has suggested some heuristics that could be useful [2].
- *Acoustic modelling*: our current models use a quantized representation of the acoustic signal. It is likely that better results can be obtained using mixtures of Gaussians; this may also be better for speech generation.
- *Acoustic environment*: the greater expressive power of DBNs should facilitate an attempt to cope with more complex acoustic environments including background noise and multiple speakers. For example, multiple speakers can be represented by multiple, independently evolving DBNs with shared acoustic variables modelled additively. Possibly the addition of variables to represent speech rate and pitch would assist in separating the signals. In all cases, no additional algorithmic work is needed—all these issues are handled by refining the DBN model topology.

3. EXPRESSIVENESS AND UNDERSTANDING

There are a variety of scientific problems involving data interpretation for which none of the standard multivariate models suffice. For example, research on phylogenetic trees [6] and genetic linkage analysis [19] uses models with repeating parameter patterns and different tree structures for each data set; researchers in these areas have traditionally used special-purpose algorithms rather than a general modelling language. This is because standard tools lack the ability to handle distributions over attributes and relations of multiple objects. The same failing arises in any attempt to apply standard tools to speech *understanding*, because the content of utterances usually concerns objects and relations, where the number, type, and dependency structures of the objects are no known *a priori*. These capabilities are considered to require the expressive power of first-order languages (see, for example, [10].

First-order probabilistic logics (FOPL) knowledge bases specify probability distributions over events that are models of the first-order language [13]. The ability to handle objects and relations gives such languages a huge advantage in representational efficiency in certain situations: for example, the rules of chess can be written in about one page of Prolog but require perhaps millions of pages in a propositional language. Recent work by Koller and Pfeffer [16] has succeeded in devising restricted sublanguages that make it relatively easy to specify complete distributions over events (see also [17]). Current inference methods, however, are somewhat restricted in their ability to handle uncertainty in the relational structure of the domain or in the number of objects it contains. It is possible that algorithms based on MCMC [9] will yield more flexibility. The basic idea is to construct a Markov chain on the set of logical models of the FOPL theory. In each model, any query can be answered simply; the

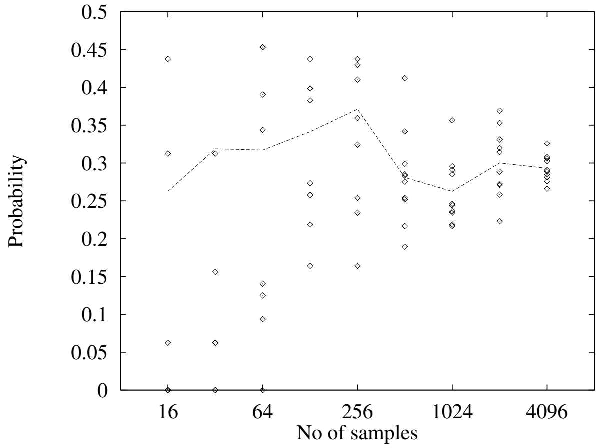

Figure 3: Solution of a first-order query by MCMC, showing the average probability over 10 runs as a function of sample size (*solid line*)

MCMC method samples models so that estimates of the query probability converge to the true posterior. For example, suppose we have the following simple knowledge base:

$$\begin{aligned}
&\forall x, y\ P(CoveredBy(x, y)) = 0.3 \\
&\forall x\ P(Safe(x) \mid \exists y\ CoveredBy(x, y)) = 0.8 \\
&\forall x\ P(Safe(x) \mid \neg\exists y\ CoveredBy(x, y)) = 0.5
\end{aligned}$$

and we want to know the probability that everyone who is safe is covered only by safe people, given certain data on objects a, b, c, d:

$$\begin{aligned}
P(\forall x\ &Safe(x) \Rightarrow \\
&\quad (\forall y\ CoveredBy(x, y) \Rightarrow Safe(y)) \\
&\mid Safe(a), Safe(b), \neg Safe(d))
\end{aligned}$$

Figure 3 shows the results of running MCMC for this query. Reasonable convergence is achieved after 4,000 samples even though the state space contains about 1,000,000 states.

Much work remains to be done, including convergence proofs for MCMC on infinite model spaces, analyses of the complexity of approximate inference for various sublanguages, and incremental query answering between models. The approach has much in common with the *general pattern theory* of Grenander [12], but should benefit from the vast research base dealing with representation, inference, and learning in first-order logic. Once inference (and, by the usual EM extension, learning) tools are available for flexible first-order probabilistic languages, it may be possible to address questions of speech recognition and understanding within a single, effective probabilistic framework.

4. REFERENCES

[1] John Binder, Daphne Koller, Stuart Russell, and Keiji Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29:213–244, 1997.

[2] Xavier Boyen, Nir Friedman, and Daphne Koller. Discovering the hidden structure of complex dynamic systems. In *Proc. 15th Annual Conference on Uncertainty in AI (UAI)*, Stockholm, 1999.

[3] Sanjoy Dasgupta. The sample complexity of learning fixed-structure Bayesian networks. *Machine Learning*, 29:165–180, 1997.

[4] J.F.G. de Freitas, M. Niranjan, and A.H. Gee. Sequential Monte Carlo methods to train neural network models. Neural Computation, to appear, 1999.

[5] Thomas Dean and Keiji Kanazawa. Probabilistic temporal reasoning. In *Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI-88)*, pages 524–528, St. Paul, Minnesota, 1988. American Association for Artificial Intelligence.

[6] J. Felsenstein. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods in Enzymology*, 266:418–427, 1996.

[7] Nir Friedman, Kevin Murphy, and Stuart Russell. Learning the structure of dynamic probabilistic networks. In *Uncertainty in Artificial Intelligence: Proceedings of the Fourteenth Conference*, Madison, Wisconsin, July 1998. Morgan Kaufmann.

[8] Z. Ghahramani and M. I. Jordan. Factorial hidden Markov models. *Machine Learning*, 29:245–274, 1997.

[9] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors. *Markov chain Monte Carlo in practice*. Chapman and Hall, London, 1996.

[10] R. P. Goldman and E. Charniak. Probabilistic text understanding. *Statistics and Computing*, 2(2):105–114, June 1992.

[11] Neil J. Gordon. *Bayesian methods for tracking*. PhD thesis, Imperial College, University of London, 1994.

[12] Ulf Grenander. *General pattern theory*. Oxford University Press, Oxford, 1993.

[13] J. Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46(3):311–350, 1990.

[14] Keiji Kanazawa, Daphne Koller, and Stuart Russell. Stochastic simulation algorithms for dynamic probabilistic networks. In *Uncertainty in Artificial Intelligence: Proceedings of the Eleventh Conference*, pages 346–351, Montreal, Canada, August 1995. Morgan Kaufmann.

[15] U. Kjaerulff. A computational scheme for reasoning in dynamic probabilistic networks. In *Uncertainty in Artificial Intelligence: Proceedings of the Eighth Conference*, pages 121–129, Stanford, California, July 1992. Morgan Kaufmann.

[16] D. Koller and A. Pfeffer. Probabilistic frame-based systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, Madison, Wisconsin, July 1998. AAAI Press.

[17] S. Muggleton. Stochastic logic programs. Journal of Logic Programming, to appear, 1999.

[18] J. Pitrelli, C. Fong, S. Wong, J. Spitz, , and H. Leung. Phonebook: A phonetically-rich isolated-word telephone-speech database. In *ICASSP-95: 1995 International Conference on Acoustics, Speech, and Signal Processing*, pages 101–104. IEEE, 1995.

[19] M. Polymeropoulos, J. Higgins, L. Golbe, et al. Mapping of a gene for Parkinson's disease to chromosome 4q21-q23. *Science*, 274(5290):1197–1199, 1996.

[20] S.T. Roweis and Z. Ghahramani. A unifying review of linear gaussian models. *Neural Computation*, 11(2):305–345, 1999.

[21] P. Smyth, D. Heckerman, and M. Jordan. Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, 9(2):227–269, 1997.

[22] Geoff Zweig and Stuart J. Russell. Speech recognition with dynamic Bayesian networks. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, Madison, Wisconsin, July 1998. AAAI Press.

[23] Geoff Zweig and Stuart J. Russell. Probabilistic modeling with Bayesian networks for automatic speech recognition. *Australian Journal of Intelligent Information Processing Systems*, 5(4):253–260, 1999.