Automatic Speech Recognition: the New Millennium

Khalid Daoudi

Automatic speech recognition: the new millennium

Khalid Daoudi

INRIA-LORIA
Speech Group
B.P. 101 - 54602 Villers les Nancy. France.
email: daoudi@loria.fr

Abstract. We present a new approach to automatic speech recognition (ASR) based on the formalism of Bayesian networks. We put the foundations of new ASR systems for which the robustness relies on the fidelity in speech modeling and on the information contained in training data.

1 Introduction

Sate-of-the-art automatic speech recognition (ASR) systems are based on probabilistic modeling of the speech signal using Hidden Markov Models (HMM). These models lead to the best recognition performances in ideal "lab" conditions or for easy tasks. However, in real word conditions of speech processing (noisy environment, spontaneous speech, non-native speakers...), the performances of HMM-based ASR systems can decrease drastically and their use becomes very limited. For this reason, the conception of robust and viable ASR systems has been a tremendous scientific and technological challenge in the field of ASR for the last decades.

The speech research community has been addressing this challenge for many years. The most commonly proposed solutions to deal with real world conditions are *adaptation* techniques (in the wide sense of the term) of HMM-based systems. Namely, the speech signals or/and the acoustic models are adapted to compensate for the miss-match between training and test conditions. While the ideas behind adaptation techniques are attractive and justified, the capability of HMM-based ASR systems to seriously address the challenge seems to be out of reach (at least in our opinion).

During the last two years, we[1] have been conducting research dedicated to address the challenge from another perspective. We focus on what we believe is the core of the problem: the *robustness* in speech modeling. Precisely, our strategy is to conceive ASR systems for which robustness relies on:

- the fidelity and the flexibility in speech modeling rather than (ad-hoc) tuning of HMMs,

[1] The Speech Group members involved in this project are: C. Antoine, M. Deviren, D. Fohr and myself (www.loria.fr/equipes/parole).

– a better exploitation of the information contained in the available statistical data.

This is motivated by the fact that the discrepancy of HMMs in real conditions is mainly due to their weakness in capturing some acoustic and phonetic phenomena which are specific to speech, and to their "limited" processing of the training databases. In order to hope obtaining robust ASR systems, it is then crucial to develop new probabilistic models capable of capturing all the speech features and of exploiting at best the available data.

A family of models which seems to be an ideal candidate to achieve this goal is the one of *probabilistic graphical models* (PGMs). Indeed, in last decade, PGMs have emerged as a powerful formalism unifying many concepts of probabilistic modeling widely used in statistics, artificial intelligence, signal processing and other fields. For example, HMMs, mixture models, factorial analysis, Kalman filters and Ising models are all particular instances of the more general PGMs formalism. However, the use of PGMs in automatic speech recognition has gained attention only very recently [2, 14, 21].

In this paper, we present an overview of our recent research in the field of ASR using probabilistic graphical models. The scope of this paper is to bring this new concept (from our perspective) to the attention of researchers, engineers and industrials who are interested in the conception of robust ASR. We develop the main ideas behind our perspective and argue that PGMs are a very promising framework in ASR and could be the foundation of a new generation of ASR systems. We do not provide full algorithmic and implementation details, but we give all the necessary references to readers interested in such details[2].

2 Probabilistic graphical models

During the last decade, PGMs have become very popular in artificial intelligence (and other fields) due to many breakthroughs in many aspects of inference and learning. The literature is now extremely rich in papers and books dealing with PGMs in artificial intelligence among of which we refer to [4] for a very good introduction. The formalism of probabilistic graphical models (PGMs) is well summarized in the following quotation by M. Jordan [18]:

"Graphical models are a marriage between probability theory and graph theory. They provide a natural tool for dealing with two problems that occur throughout applied mathematics and engineering - uncertainty and complexity - and in particular they are playing an increasingly important role in the design of machine learning algorithms. Fundamental to the idea of a graphical model is the notion of modularity - a complex system is built by combining simpler parts. Probability theory provides the glue whereby the parts are combined, ensuring that the system as whole is consistent, and providing ways to interface models to

[2] All author's papers can be down-loaded from *www.loria.fr/~daoudi* or *www.loria.fr/equipes/parole*

data. The graph theoretic side of graphical models provides both an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure that lends itself naturally to the design of efficient general-purpose algorithms."

More precisely, given a system of random variables (r.v.), a PGM consists in associating a graphical structure to the joint probability distribution of this system. The nodes of this graph represent the r.v., while the edges encode the (in)dependencies which exist between these variables. One distinguishes three types of graphs: directed, undirected and those for which the edges are a mixture of both. In first case, one talks about *Bayesian networks*, in the second case, one talks about *Markov random fields*, and in the third case one talks about *chain networks*. PGMs have two major advantages:

- They provide a natural and intuitive tool to illustrate the dependencies which exist between variables. In particular, the graphical structure of a PGM clarifies the conditional independencies embedded in the associated joint probability distribution.
- By exploiting these conditional independencies, they provide a powerful setting to specify efficient inference algorithms. Moreover, these algorithms may be specified automatically once the initial structure of the graph is determined.

So far, the conditional independencies semantics (or Markov properties) embedded in a PGM are well-understood for Bayesian networks and Markov random fields. For chain networks, these are still not well-understood. In our current research, given the causal and dynamic aspects of speech, Bayesian networks (BNs) are of particular interest to us. Indeed, thanks to their structure and Markov properties, BNs are well-adapted to interpret causality between variables and to model temporal data and dynamic systems. In addition, not only HMMs are a particular instance of BNs, but also the Viterbi and Forward-Backward algorithms (which made the success of HMMs in speech) are particular instances of generic inference algorithms associated to BNs [20]. This shows that BNs provide a more general and flexible framework than the HMMs paradigm which has ruled ASR for the last three decades.

Formally, a BN has two components: a directed acyclic graph S and a numerical parameterization Θ. Given a set of random variables $X = \{X_1, ..., X_N\}$ and $P(X)$ its joint probability distribution (JPD), the graph S encodes the conditional independencies (CI) which (are supposed to) exist in the JPD. The parameterization Θ is given in term of conditional probabilities of variables given their parents. Once S and Θ are specified, the JPD can be expressed in a factored way as[3][4]

$$P(x) = \prod_{i=1}^{N} P(x_i|pa(x_i)) \tag{1}$$

[3] This factorization can not be obtained if the graph is cyclic.

[4] In the whole paper, upper-case (resp. lower-case) letters are used for random variables (resp. outcomes).

where $pa(x_i)$ denotes an outcome of the parents of X_i.

The Markov properties of a BN imply that, conditioned on its parents, a variable is independent of all the other variables except its descendants. Thus, it is obvious to represent a HMM as a particular BN, as shown in Figure 1. Contrarily to the usual state transition diagram, in the BN representation each node H_t (resp. O_t) is a random variable whose outcome indicates the state occupied (resp. the observation vector) at time t. Time is thus made explicit and arrows linking the H_t must be understood as "causal influences" (not as state transitions).

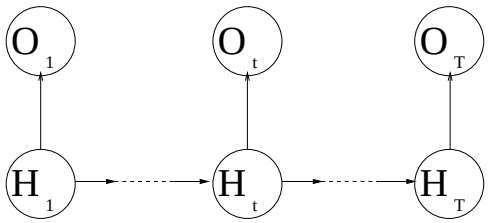

Fig. 1. a HMM represented as a Bayesian network

3 The language of data

When conceiving an ASR system, the only "real" material available is the training speech databases. It is then extremely important to exploit at best the information contained in these databases to build the speech acoustic models. The conception of state-of-the-art ASR systems can be decomposed in three major steps. First a front-end is decided for the parameterization of the speech signals, in general Mel Frequency Cepstral Coefficients (MFCC) are used to extract acoustic features vectors. Second, system variables are designed, in general continuous multi-Gaussian variables are used to represent the observed acoustic vectors and discrete variables are used to capture the hidden dynamics of the system. Finally, dependency relationships between system variables are *imposed* once for all in order to define the speech probabilistic model (a HMM in general), then training is done using this model. In our opinion, this last point can be a serious weakness in such systems. Indeed, these dependency relationships are motivated only by prior knowledge, no examination of data is done in order to check if they are really consistent with data. It is then more realistic to *combine* prior knowledge and data information to decide which dependency relationships are more appropriate in the modeling of each acoustic unit in the vocabulary. In the next subsections, we explain how to do so using the BNs framework.

3.1 Structural learning of BNs

As mentioned above, the Markov properties of a BN (and a PGM in general) determine the dependency relationships between system variables. Thus, the problem of finding the appropriate dependency relationships which are consistent with data comes back to finding the graphical structure (and its numerical parameterization) which best explain data. This problem is known as *structural learning* in the BN literature.

The general principle of structural learning [15] is to introduce a scoring metric that evaluates the degree to which a structure fits the observed data, then a search procedure (over all possible structures) is performed to determine the structure which yields the best score according to this metric. If S is a given structure and D is the observed data, the scoring metric has the following form in general :

$$Score(S, D) = \log P(D|\hat{\Theta}, S) - Pen(S) \tag{2}$$

where $\hat{\Theta}$ is the estimated numerical parameterization given the structure S. The scoring metric has two terms. The first one represents the (log)likelihood of data. The second one is a penalty term introduced to penalize overly complex structures. The consideration of this term is very important because otherwise complex structures would be always favored, resulting in untractable networks.

There are basically two approaches for penalizing complex structures. In the first one, a prior probability distribution on structures is used. This scoring metric, known as the Bayesian Dirichlet metric, gives high probabilities to more realistic structures [16]. The second approach, which is the one we use, is known as the Bayesian Information Criterion (BIC) score or the Minimum Description Length (MDL) score [17]. The BIC score, which is based on universal coding, defines a penalty term proportional to the number of parameters used to encode data:

$$Pen(S) = \frac{\log N}{2} \sum_{i=1}^{n} ||X_i, pa(X_i)|| \tag{3}$$

where N is the number of examples (realizations) in D and $||X, Y||$ is defined as the number of parameters required to encode the conditional probability $P(X|Y)$.

The evaluation of the scoring metric for all possible structures is generally not feasible. Therefore, many algorithms have been proposed to search the structure space so as to achieve a maximum scoring structure. In [13], a structural Expectation-Maximization (SEM) algorithm is proposed to find the optimal structure and parameters for a BN, in the case of incomplete data. This algorithm starts with random structure and parameters. At each step, first a parametric Expectation-Maximization algorithm is performed to find the optimal parameters for the current structure. Second, a structural search is performed to increase the scoring. The evaluation of the scoring metric for the next possible structure is performed using the parameters of the previous structure. This algorithm guarantees an increase in the scoring metric at each iteration and converges to a local maximum.

3.2 Application to speech recognition

In HMM-based systems, the observed process is assumed to be governed by a hidden (unobserved) dynamic one, under some dependency assumptions. The latter state that the hidden process is first-order Markov, and that each observation depends only on the current hidden variable. There is however a fundamental question regarding these dependency assumptions: are they consistent with data? In case the answer is no: what (more plausible) dependency assumptions should we consider? We have applied the structural learning framework to learn (from training data) the appropriate dependency relationships between the observed and hidden process variables, instead of imposing them as HMM-based systems do. We have also introduced a slight but important modification in the initialization of the SEM algorithm. Namely, instead of starting from a random structure, we initialize the algorithm using the HMM structure (see Figure 1). This way we exploit prior knowledge, namely that HMMs are good "initial" models for speech recognition. More importantly, we thus guaranty that the SEM algorithm will converge to a BN which models speech data with higher (or equal) fidelity than a HMM. We refer the reader to [8] for details on the application of this strategy to an isolated speech recognition task. In the latter, the decoding algorithm is readily given by the inference algorithms associated to BNs. However, in a continuous speech recognition task, decoding requires more attention. Indeed, the SEM algorithm yields in general different structures for different words in the vocabulary. This leads in turn to an *asymmetry* in the representation of dependency in the decoding process. In [9], we have developed a decoding algorithm which deals with such asymmetry and, consequently, allows recognition in a continuous speech recognition task.

Our approach to build acoustic speech models described above has many advantages:

- It leads to speech models which are consistent with training databases.
- It guarantees improvement in modeling fidelity w.r.t. to HMMs.
- It allows capturing phonetic features, such as the *anticipation* phenomena, which can not be modeled by any Markov process (not only HMMs).
- It is implicitly discriminative because two words in the vocabulary may be modeled using completely different networks (in term of Markov properties and the number of parameters).
- It is technically attractive because all the computational effort is made in the training phase.
- It allows the user to make a control on the trade-off between modeling accuracy and model complexity.

The experiments carried out in [9, 8] and the results obtained show that, indeed, this approach leads to significant improvement in recognition accuracy w.r.t. to a classical system.

4 Multi-band speech recognition

In the previous section we argued that data should be combined with prior knowledge in order to build speech acoustic models *a posteriori*. We applied this principle in the setting where speech is assumed to be the superposition of an observed process with a hidden one. We showed that (w.r.t. HMMs) substantial gain in recognition accuracy can be obtained using this methodology. In this section, we address the problem of modeling *robustness* from the multi-band principle perspective.

The multi-band principle was originally introduced by Harvey Fletcher [12] who conducted (during the first half of the 20th century) extensive research on the way humans process and recognize speech. He "showed" that the human auditory system recognizes speech using partial information across frequency, probably in the form of speech features that are localized in the frequency domain. However, Fletcher's work has been miss-known until 1994 when Jont B. Allen published a paper [1] in which he summarized the work of Fletcher and also proposed to adapt the multi-band paradigm to automatic speech recognition. Allen's work has then inspired researchers to develop a multi-band approach to speech recognition in order to overcome the limitations of classical HMM modeling. Indeed, in many applications (spontaneous speech, non-native speakers...) the performances of HMMs can be very low. One of the major reasons for this discrepancy (from the acoustic point of view) is the fact that the frequency dynamics are weakly modeled by classical MFCC parameterization. Another application where HMMs present a serious limitation is when the system is trained on clean speech but tested in noisy conditions (particularly additive noise). Even when the noise covers only a small frequency sub-band, HMMs yield bad performances since the MFCC coefficients are calculated on the whole spectrum and are then all corrupted.

In the classical multi-band (CMB) approach, the frequency axis is divided into several sub-bands, then each sub-band is independently modeled by a HMM. The recognition scores in the sub-bands are then fusioned with some recombination module. The introduction of multi-band speech recognition [3, 11] has been essentially motivated by two desires. The first one is to mime the behavior of the human auditory system (which decomposes the speech signal into different sub-bands before recognition [1]). The second one is to improve the robustness to band-limited noise. While the ideas leading to multi-band speech recognition are attractive, the CMB approach has many drawbacks. For instance, the sub-bands are assumed mutually independent which is an unrealistic hypothesis. Moreover, the information contained in one sub-band is not discriminative in general. In addition, it is not easy to deal with asynchrony between sub-bands, particularly in continuous speech recognition. As a consequence, the recombination step can be a very difficult task. Using the BNs formalism, we present in the next subsection an alternative approach to perform multi-band speech recognition which has the advantage to overcome *all* the limitations (mentioned above) of the CMB approach.

4.1 A multi-band Bayesian network

The basic idea behind our approach is the following: instead of considering an independent HMM for each sub-band (as in the CMB approach), we build a more complex but uniform BN on the time-frequency domain by "coupling" all the HMMs associated with the different sub-bands. By coupling we mean adding (directed) links between the variables in order to capture the dependency between sub-bands. A natural question is: what are the "appropriate" links to add?. Following our reasoning of the previous section, the best answer is to learn the graphical structure from training data. Meanwhile, it is also logical to first impose a "reasonable" structure in order to see if this new approach is promising. If the answer is yes, then this "reasonable" structure could play the role of the initial structure (prior knowledge) in a structural learning procedure (as we did with HMMs in the classical full-band case). We build such "reasonable" structure using the following computational and physical criteria. We want a model where no continuous variable has discrete children in order to apply an exact inference algorithm (see [19]). We also want a model with a relatively small number of parameters and for which the inference algorithms are tractable. Finally, we want to have links between the hidden variables along the frequency axis in order to capture the asynchrony between sub-bands. A simple model which satisfies these criteria is the one shown in Figure 2. In this BN, the hidden variables of sub-band n are linked to those of sub-band $n+1$ in such way that the state of a hidden variable in sub-band $n+1$ at time t is conditioned by the state of two hidden variables: at time $t-1$ in the same sub-band and at time t in sub-band n. Each $H_t^{(n)}$ is a discrete variable taking its values in $\{1, ..., m\}$, for

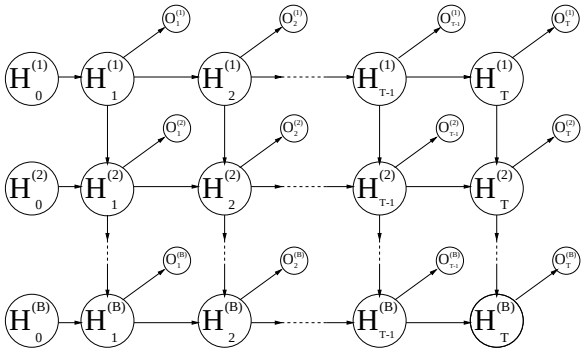

Fig. 2. B-band Bayesian network

some integer $m \geq 1$. Each $O_t^{(n)}$ is a continuous variable with a multi-Gaussian distribution representing the observation vector at time t in sub-band n ($n = 1, ..., B$), B is the number of sub-bands. We impose a left-to-right topology on each sub-band and assume that model parameters are stationary. The numerical

parameterization of our model is then:

$$a_{ij} \triangleq P(H_t^{(1)} = j | H_{t-1}^{(1)} = i) \ ;$$

$$u_{ijk}^{(n)} \triangleq P(H_t^{(n)} = k | H_t^{(n-1)} = i, H_{t-1}^{(n)} = j) \ ;$$

$$b_i^{(n)}(o_t^{(n)}) \triangleq P(O_t^{(n)} = o_t^{(n)} | H_t^{(n)} = i)$$

where $b_i^{(n)}$ is a multi-Gaussian density. The asynchrony between sub-bands is taken into account by allowing all the $u_{ijk}^{(n)}$ to be non-zero, except when $k < j$ or $k > j + 1$ because of the left-to-right topology.

We now stretch the advantages of such approach to multi-band ASR. Contrarily to HMMs, our multi-band BN provides "a" modeling of the frequency dynamics of speech. As opposed to the CMB approach, this BN allows interaction between sub-bands and the possible asynchrony between them is taken into account. Moreover, our model uses the information contained in all sub-bands and no recombination step is needed. A related work has been proposed in [10] where a multi-band Markov random field is analyzed by mean of Gibbs distributions. This approach (contrarily to ours) does not lead however to exact nor fast inference algorithms and assumes a linear model for asynchrony between sub-bands. In our approach, the asynchrony is learned from data. In the next subsection, we present some experiments which illustrate the power of this new approach.

4.2 Experiments

Implementation details and experiments in an isolated speech recognition task can be found in [5] and [6]. The experiments has been carried out in clean and noisy speech conditions and has led to three main results. First, in clean conditions, our approach not only outperforms the CMB one, but also HMMs. To the best of our knowledge, the only multi-band systems which out-perform HMMs in clean conditions use the full-band parameterization as an additional "sub-band". Second, through comparison to a synchronous multi-band BN, the results show the importance of asynchrony in multi-band ASR. Finally and more importantly, our approach largely outperforms HMMs and the CMB approach in noisy speech conditions.

Implementation details of our approach in a continuous speech recognition task can be found in [7]. A preliminary experiment has been carried out but only in clean speech conditions. The results obtained are analogous to those obtain in the isolated speech setting. We now present some of our latest experiments in noisy continuous speech conditions.

Our experiments are carried out on the Tidigits database. In learning we only use the isolated part of the training database where each speaker utters 11 digits twice. In test, we use the full (test) database in which 8636 sentences are uttered, each sentence contains between 1 and 7 digits. We show comparisons of the performances of a 2-band BN to HMMs (it is well-known that the performances of

the CMB approach in the continuous setting are generally lower than in the isolated setting). For every digit and the silence model, the number of hidden states is seven ($m = 7$) and we have a single Gaussian per state with a diagonal covariance matrix. We use a uniform language model, i.e., $P(v|v') = \frac{1}{12}$ (eleven digits + silence). The parameterization of the classical full-band HMM is done as follows: 25ms frames with a frame shift of 10ms, each frame is passed through a set of 24 triangular filters resulting in a vector of 35 features, namely, 11 static MFCC (the energy is dropped), 12 Δ and 12 $\Delta\Delta$. The parameterization of the 2-band BN is done as follows: each frame is passed through the 16 first (resp. last 8) filters resulting in the acoustic vector of sub-band 1 (resp. sub-band 2). Each vector contains 17 features: 5 static MFCC, 6 Δ and 6 $\Delta\Delta$. The resulting bandwidths of sub-bands 1 and 2 are $[0..2152Hz]$ and $[1777Hz..10000Hz]$ respectively. The training of all models is done on clean speech only. The test however is performed on clean and noisy speech. The latter is obtained by adding, at different SNRs, a band-pass filtered white noise with a bandwidth of $[3000Hz..6000Hz]$. Table 1 shows the digit recognition rates we obtain using both models. These results illustrate the potential of our approach in exploiting the information contained in the non-corrupted sub-band.

Table 1. Digit recognition scores

Noise SNR	HMM	2-band BN
26 db	89.95%	**96.16%**
20 db	82.17%	**94.89%**
14 db	73.27%	**90.81%**
8 db	62.57%	**82.27%**
2 db	58.86%	**75.51%**

In summary, our new approach to multi-band ASR seems to be more robust than the classical approach and HMMs, both in clean and noisy conditions. The next step (which we did not carry out yet) will be to perform a multi-band structural learning using our B-band BN as an initial structure. Such procedure should increase the robustness of the resulting multi-band system.

5 Conclusion

Based on the PGMs formalism, we presented a new methodology to ASR which seems to be very promising. While the results we obtained are from the most "striking" in the literature (to the best of our knowledge), we do not claim that our perspective of applying PGMs to automatic speech recognition is the "best" one. As we have mentioned, research in this field is still new and PGMs can be applied in many different ways. Our only claim is that, at the beginning of this new millennium, PGMs seem to have a bright future in the field of ASR.

Acknowledgments

The author would like to thank C. Antoine, M. Deviren and D. Fohr for their major contributions in the implementation of the ideas presented in this paper.

References

1. J. Allen. How do humans process and recognize speech. *IEEE Trans. Speech and Audio Processing*, 2(4):567–576, 1994.
2. Jeff A. Bilmes. *Natural Statistical Models for Automatic Speech Recognition.* PhD thesis, International Compute Science Institute, Berkeley, California, 1999.
3. H. Bourlard and S. Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. ICSLP'96.
4. Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems.* Springer, 1999.
5. K. Daoudi, D. Fohr, and C. Antoine. A new approach for multi-band speech recognition based on probabilistic graphical models. In *ICSLP*, 2000.
6. K. Daoudi, D. Fohr, and C. Antoine. A Bayesian network for time-frequency speech modeling and recognition. In *International Conference on Artificial Intelligence and Soft Computing*, 2001.
7. K. Daoudi, D. Fohr, and C. Antoine. Continuous Multi-Band Speech Recognition using Bayesian Networks. In *IEEE ASRU Workshop*, 2001.
8. M. Deviren and K. Daoudi. Structural learning of dynamic bayesian networks in speech recognition. In *Eurospeech*, 2001.
9. M. Deviren and K. Daoudi. Continuous speech recognition using structural learning of dynamic bayesian networks. Technical report, 2002.
10. G. Gravier et al. A markov random field based multi-band model. ICASSP'2000.
11. H. Hermansky et al. Towards ASR on partially corrupted speech. ICSLP'96.
12. H. Fletcher. *Speech and hearing in communication.* Krieger, New-York, 1953.
13. N. Friedman. Learning belief networks in the presence of missing values and hidden variables. In *Int. Conf. Machine Learning*, 1997.
14. G. Gravier. *Analyse statistique deux dimensions pour la modlisation segmentale du signal de parole: Application la reconnaissance.* PhD thesis, ENST Paris, 2000.
15. D. Heckerman. A tutorial on learning with bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division, March 1995.
16. D. Heckerman, D. Geiger, and D.M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.
17. W. Lam and F. Bacchus. Learning bayesian belief networks an approach based on the mdl principle. *Computational Intelligence*, 10(4):269–293, 1994.
18. M. Jordan, editor. Learning in graphical models. *MIT Press*, 1999.
19. S.L. Lauritzen. Propagation of probabilities, means, and variances in mixed graphical association models. *Jour. Amer. Stat. Ass.*, 87(420):1098–1108, 1992.
20. P. Smyth, D. Heckerman, and M. Jordan. Probabilistic independence networks for hidden markov probability models. *Neural Computation*, 9(2):227–269, 1997.
21. G.G. Zweig. *Speech Recognition with Dynamic Bayesian Networks.* PhD thesis, University of California, Berkeley, 1998.