# Mixed Bayesian Networks with Auxiliary Variables for Automatic Speech Recognition

Todd A. Stephenson Mathew Magimai-Doss
Hervé Bourlard
Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), CH-1920 Martigny, Switzerland
Swiss Federal Institute of Technology Lausanne (EPFL), CH-1015 Lausanne, Switzerland

## Abstract

*In standard automatic speech recognition (ASR), hidden Markov models (HMMs) calculate their emission probabilities by an artificial neural network (ANN) or a Gaussian distribution conditioned only upon the hidden state variable. Recent work [12] showed the benefit of conditioning the emission distributions also upon a discrete auxiliary variable, which is observed in training and hidden in recognition. Related work [3] has shown the utility of conditioning the emission distributions on a continuous auxiliary variable. We apply mixed Bayesian networks (BNs) to extend these works by introducing a continuous auxiliary variable that is observed in training but is hidden in recognition. We find that an auxiliary pitch variable conditioned itself upon the hidden state can degrade performance unless the auxiliary variable is also hidden. The performance, furthermore, can be improved by making the auxiliary pitch variable independent of the hidden state.*

## 1. Introduction

Hidden Markov models [8] calculate at each time $n$ the likelihood of the acoustic observation $x_n$ being produced, given that the hidden state variable $q_n$ has the discrete value of $k, 1 \leq k \leq K$:

$$p(x_n|q_n = k). \tag{1}$$

This is typically computed using an ANN or a Gaussian distribution, with mean $\mu_k$ and covariance $\Sigma_k$:

$$p(x_n|q_n = k) \sim \mathcal{N}(\mu_k, \Sigma_k). \tag{2}$$

There may be information not directly available in the acoustic observation $x_n$ that may be of use in enhancing the models. Such auxiliary information $a_n$, which can be continuous or discrete, may be derived from the acoustic signal or may be obtained from a secondary source [11]. $q_n$ and $a_n$ can then jointly condition the emission likelihoods, replacing (1) with:

$$p(x_n|q_n = k, a_n = z). \tag{3}$$

In [12], $a_n$ was defined as a discrete variable. It took a codebook of four values, each representing a pitch range. For this case, the performance was better when the pitch was hidden in recognition than when it was observed. However, some auxiliary information is more naturally used as continuous information than in reducing it to discrete values, as done above. In [3], an increase in recognition performance was observed when a continuous $a_n$ was introduced. For this case, the means of the Gaussian distributions (2) can then be shifted using the regression weights $B_k$ and the value of $a_n$, producing *conditional* Gaussians:

$$p(x_n|q_n = k, a_n = z) \sim \mathcal{N}(\mu_k + B_k^T\, z, \Sigma_k), \tag{4}$$

In this work we continue with these findings by using continuous $a_n$ in the framework of mixed BNs (BNs that have a *mixture* of continuous and discrete variables). The BN formalism has previously been presented as a statistical pattern recognition framework that is more generic than that of HMMs [10]. That is, while they are in the same family of models [9], BNs are more general in that they provide more *flexibility* in changing the topology of the model and, hence, the structure of the component distributions. With this flexibility, we address two questions:

1. Should the distribution for $a_n$ itself be conditioned upon $q_n$: $p(a_n|q_n)$, or be left independent: $p(a_n)$? That is, is $a_n \perp\!\!\!\perp q_n$ (read, "$a_n$ is independent of $q_n$")?
2. Should the distribution of $x_n$ be conditioned upon $q_n$ and $a_n$, as in (3), or only upon $q_n$, as in (1)? That is, is $x_n \perp\!\!\!\perp a_n \mid q_n$ (read, "$x_n$ is conditionally independent of $a_n$, given $q_n$")?

The contributions of this work, hence, are threefold. First, we introduce mixed BNs to ASR. To our knowledge, this has never been done before–at least not in the more complicated case where continuous variables can be hidden. Second, we look at an additional way to model the auxiliary information $a_n$ itself–that is, conditioning it upon the state variable $q_n$. Third, taking advantage of this general framework provided by mixed BNs, we show the effects of hiding the auxiliary information $a_n$.

We begin in Section 2 by introducing the emission probabilities of $x_n$ and $a_n$ that we will be modeling. Section 3 introduces mixed BNs as well as distributions conditioned upon both continuous and discrete variables. Section 4 then presents the incorporation of auxiliary information graphically in a BN. Section 5 then presents the experimental results followed by the conclusion in Section 6.

## 2. Introducing Auxiliary Information

Standard HMM-based pattern recognition models $p(X,Q)$, the evolution of the observed space $X = \{x_1, x_2, \ldots, x_N\}$ and the hidden state space $Q = \{q_1, q_2, \ldots, q_N\}$ for time $n = 1, \ldots, N$ as:

$$p(X,Q) \approx \prod_{n=1}^{N} p(x_n|q_n) \cdot P(q_n|q_{n-1}), \tag{5}$$

assuming time-independence for $x_n$ and a first-order Markov assumption for $q_n$ (specifically, that $q_n$ is independent of all previous variables given $q_{n-1}$).

For incorporating the auxiliary information $A = \{a_1, a_2, \ldots, a_N\}$ to the hidden or observed space, the modeling of $p(X,A,Q)$ factors as:

$$p(X,A,Q) \approx \prod_{t=1}^{N} p(x_n|a_n,q_n) \cdot p(a_n|q_n) \cdot P(q_n|q_{n-1}), \tag{6}$$

assuming time-independence for $x_n$ and $a_n$ and the first-order Markov assumption for $q_n$.

In our experiments, we present two separate ways to further relax the distribution in (6):

1. $a_n$ independent of $q_n$ ($a_n \perp\!\!\!\perp q_n$): $p(a_n|q_n) \rightarrow p(a_n)$. Similar to that done in [3], this assumes that the current hidden state $q_n$ does not influence the value of $a_n$. The only thing in common between $q_n$ and $a_n$ is that they jointly emit the acoustics $x_n$:

$$\prod_{t=1}^{N} p(x_n|a_n,q_n) \cdot p(a_n) \cdot P(q_n|q_{n-1}) \tag{7}$$

2. $x_n$ independent of $a_n$ ($x_n \perp\!\!\!\perp a_n \mid q_n$): $p(x_n|a_n,q_n) \rightarrow p(x_n|q_n)$. This assumes that $x_n$ and $a_n$ are two independent processes that are jointly emitted by $q_n$. This is equivalent to using a standard HMM with a single feature vector comprised of the concatenation of $x_n$ and $a_n$ (assuming a diagonal covariance matrix).

$$\prod_{t=1}^{N} p(x_n|q_n) \cdot p(a_n|q_n) \cdot P(q_n|q_{n-1}) \tag{8}$$

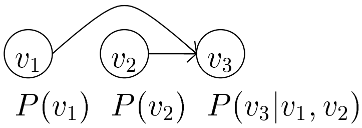

**Figure 1. Bayesian network modeling $P(v_1,v_2,v_3) = P(v_1) \cdot P(v_2) \cdot P(v_3|v_1,v_2)$.**

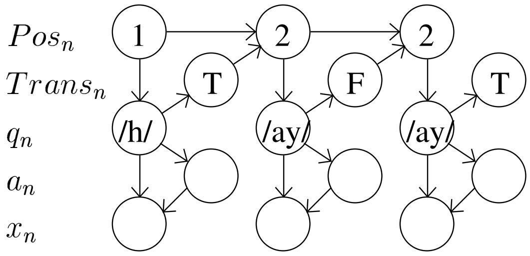


**Figure 2. BN for ASR (probabilities omitted) with auxiliary information [13], with $N = 3$. In System 3, $q_n$ and $a_n$ are not connected; in System 4, $a_n$ and $x_n$ are not connected. $a_n$ was not included in System 1.**

## 3. Mixed Bayesian Networks

A BN [1], or directed graphical model–see Figure 1, is a probabilistic model composed of three items:

1. a set of variables $V = \{v_1, \ldots, v_w, \ldots, v_W\}$
2. a directed acyclic graph (DAG), with a one-to-one mapping between each of its vertices and each $v_w \in V$
3. for each $v_w \in V$, a local probability distribution which is conditioned upon the values of its parents in the DAG: $P(v_w|\text{parents}(v_w))$.

The joint distribution of $V$ is then defined as the product of all the local probability distributions:

$$P(V) = \prod_{w=1}^{W} P(v_w|\text{parents}(v_w)) \tag{9}$$

The following are the forms that each local probabilities in (9) can take, depending on whether $v_w$ is continuous or discrete and on whether its parents are continuous, discrete, or mixed:

- Continuous $v_w$
  - Continuous parents $Z$ - conditional Gaussian:

$$p(v_w|Z=z) \sim \mathcal{N}(\mu_w + B_w^T z, \Sigma_w) \quad (10)$$

  - Discrete parents $J$ - Set of Gaussians:

$$\{p(v_w|J=j) \sim \mathcal{N}(\mu_{wj}, \Sigma_{wj})\}_J \quad (11)$$

  - Mixed parents - Set of conditional Gaussians:

$$\{p(v_w|J=j, Z=z) \sim \mathcal{N}(\mu_{wj} + B_{wj}^T z, \Sigma_{wj})\}_J \quad (12)$$

- Discrete $v_w$
  - Continuous or mixed parents - Not defined in [5]
  - Discrete parents - table of probabilities

Thus, the distribution for a discrete variable is only defined if all of its parents are discrete. A continuous variable can have continuous, discrete, or mixed parents.

We use the BN inference algorithm in [5] to compute $P(v_w|O)$, the posterior marginal distribution of $v_w$ given all of the observations $O$, as well as $P(O|V)$, the likelihood of the observations. Any variable can be observed, hidden, or partially observed, regardless of whether it is continuous or discrete valued. The computed posterior marginal distributions can be used for the expected counts in expectation-maximization (EM) training [4] for learning the discrete probabilities $P(\cdot)$, the means $\mu$, the regression weights $B$, and the covariances $\Sigma$.

# 4. Topologies

Figure 2 presents the BN, based on [13], for an isolated word recognition task. It contains the following variables:

- Deterministic variables
  - $Pos_n$ - The position (sub-model index) in the word model.
  - $q_n$ - The hidden phoneme state mapped to the given position.
- Random variables
  - $Trans_n$ - The presence of a change of sub-models (transition) between two time frames.
  - $a_n$ - The auxiliary information.
  - $x_n$ - The acoustics.

The upper three variables in Figure 2, $Pos_n$, $Trans_n$, and $q_n$, are referred to as the control layer as they "control" the permitted sequences of sub-models.

# 5. Experiments

## 5.1. Systems

Using the PhoneBook speech corpus [7] with the small training set defined in [2], we train four mixed BN systems to do speaker-independent, task-independent, isolated-word recognition.

**System 1** $x_n$ only, based on (5), as in a standard HMM

**System 2** $x_n$ & $a_n$, based on (6)

**System 3** $x_n$ & $a_n$, based on (7), with $a_n \perp\!\!\!\perp q_n$

**System 4** $x_n$ & $a_n$, based on (8), with $x_n \perp\!\!\!\perp a_n \mid q_n$, equivalent to a standard HMM with independent features $x_n$ and $a_n$

There are 41 context-independent phones in these systems, each modeled by three hidden phoneme states; with the initial silence model and end silence model, there are $41 * 3 + 2 = 125$ hidden state values for $q_n$. Both $x_n$ and $a_n$ are modeled using single (conditional) Gaussians for these initial tests; future extensions of the models would use multiple (conditional) Gaussians.

## 5.2. Features

Similarly to [13], $x_n$ is the mel-frequency cepstral coefficients (MFCCs), which are extracted from the speech signal, sampled at 8 kHz, using a window of 25 ms with a shift of 8.3 ms for each successive time frame. Cepstral mean subtraction and energy normalization are performed. Ten MFCCs plus $C_0$ (the energy coefficient) as well as the deltas (first-derivatives) of those eleven coefficients are computed for each time frame.

$a_n$ is defined only as pitch in this work and is estimated using the simple inverse filter tracking (SIFT) algorithm [6], which is based on an inverse filter formulation. This method retains the advantages of the autocorrelation and cepstral analysis techniques. The speech signal is prefiltered by a low pass filter with a cut-off frequency of 800 Hz, and the output of the filter is sampled at 2 kHz before computing the inverse filter coefficients using the Durbin algorithm.

## 5.3. Results

Training was done using expectation-maximization (EM) training, using a convergence criterion of stopping one iteration after the log-likelihood of the training data increased by less than $0.1\%$. As shown in Table 1, each system with auxiliary information was tested two times using the test set defined in [2]: (1) with both $X$ and $A$ observed and (2) with $X$ observed and $A$ hidden.

| | Observed $A$ | Hidden $A$ |
|---|---|---|
| System 1 | 19.0% | |
| System 2 | 49.0% | 21.0% |
| System 3 | 17.5% | 17.6% |
| System 4 | 54.2% | 19.1% |

**Table 1. Word error rate (WER) for small vocabulary (75 words) isolated word recognition using the systems in Section 5.1. Those trained with $A$ were tested twice: with observed and hidden $A$.**

## 6. Conclusion

First, $a_n$, such as the pitch used here, can be hurtful to the model when introduced with a dependency upon $q_n$. This is illustrated in Systems 2 & 4, which have very poor performance with observed $A$. However, these same systems perform almost the same as the baseline System 1 (statistically equivalent, in the case of System 4) when the $A$ are hidden and, therefore, marginalized out. This can potentially be extended to the actual elements within $x_n$. That is, if particular elements within $x_n$ are actually hampering recognition, perhaps they should be marginalized out as well in recognition.

Second, $a_n$, such as the pitch used here, can be beneficial to the model when introduced independent of $q_n$. This is illustrated in System 3, which performs significantly better than all of the other systems. Furthermore, in contrast to Systems 2 & 4, the performance of System 3 does not degrade with observed $A$. So, likewise, if an element of $x_n$ is found to be hurting recognition, perhaps the recognition would be better if the element were put into the conditional part of the emission distribution and made independent of the state.

Finally, modeling the distributions with single (conditional) Gaussians provides insights into the strengths and weaknesses of different ways to model auxiliary information. However, multiple (conditional) Gaussians will need to be incorporated into future models to make them more comparable to state-of-the-art ASR systems. Furthermore, although the performance improvement here is not dramatic, more significant improvement should be expected for the case of spontaneous speech and for other auxiliary variables.

## 7. Acknowledgments

Todd A. Stephenson and Mathew Magimai-Doss are supported by the Swiss National Science Foundation under grants # 2000-064172.00/1 and 2100-057245.99/1, respectively. This work benefited from discussions during the presentation of previous work at Eurospeech 2001.

## References

[1] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Statistics for Engineering and Information Science. Springer-Verlag New York, Inc., 1999.

[2] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J.-M. Boite. Hybrid HMM/ANN systems for training independent tasks: Experiments on phonebook and related improvements. In *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-97)*, volume 3, pages 1767–1770, Munich, April 1997.

[3] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama. Multiple-regression hidden Markov model. In *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-01)*, volume 1, pages 513–516, Salt Lake City, Utah, USA, May 2001.

[4] S. L. Lauritzen. The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*, 19:191–201, 1995.

[5] S. L. Lauritzen and F. Jensen. Stable local computations with conditional Gaussian distributions. *Statistics and Computing*, 11(2):191–203, April 2001.

[6] J. D. Markel. The SIFT algorithm for fundamental frequency estimation. *IEEE Trans. Audio and Electroacoustics*, 20:367–377, 1972.

[7] J. F. Pitrelli, C. Fong, S. H. Wong, J. R. Spitz, and H. C. Leung. PhoneBook: A phonetically-rich isolated-word telephone-speech database. In *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-95)*, volume 1, pages 101–104, Detroit, MI, May 1995.

[8] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[9] S. Roweis and Z. Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11(2), 1999.

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

[11] T. A. Stephenson, H. Bourlard, S. Bengio, and A. C. Morris. Automatic speech recognition using dynamic Bayesian networks with both acoustic and articulatory variables. In *6th International Conference on Spoken Language Processing: ICSLP 2000 (Interspeech 2000)*, Beijing, October 2000.

[12] T. A. Stephenson, M. Mathew, and H. Bourlard. Modeling auxiliary information in Bayesian network based ASR. In *7th European Conference on Speech Communication and Technology*, volume 4, pages 2765–2768, Aalborg, Denmark, September 2001.

[13] G. G. Zweig. *Speech Recognition with Dynamic Bayesian Networks*. PhD thesis, University of California, Berkeley, 1998.