DYNAMIC BAYESIAN NETWORK BASED SPEECH RECOGNITION WITH PITCH AND ENERGY AS AUXILIARY VARIABLES

Todd A. Stephenson* Jaume Escofet†
Mathew Magimai-Doss* Hervé Bourlard*

Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)
P. O. Box 592
rue du Simplon 4
CH-1920 Martigny
Switzerland

Phone: +41 27 721 77 11
Fax: +41 27 721 77 12

Abstract. Pitch and energy are two fundamental features describing speech, having importance in human speech recognition. However, when incorporated as features in automatic speech recognition (ASR), they usually result in a significant degradation on recognition performance due to the noise inherent in estimating or modeling them. In this paper, we show experimentally how this can be corrected by either conditioning the emission distributions upon these features or by marginalizing out these features in recognition. Since this is not obvious to do with standard hidden Markov models (HMMs), this work has been performed in the framework of dynamic Bayesian networks (DBNs), resulting in more flexibility in defining the topology of the emission distributions and in specifying whether variables should be marginalized out.

1 INTRODUCTION

The choice of the acoustic features has a large impact on ASR performance. Mel-frequency cepstral coefficients (MFCCs) are one type of acoustic feature that has proven to provide good recognition. While also being strongly

*also with The Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland

†with The Technical University of Catalonia (UPC), Barcelona, Spain, visiting IDIAP under the European Masters in Language and Speech

relevant to speech recognition, features for pitch and energy are not usually included with these standard MFCCs in the acoustic feature vector as they have been found to often degrade performance. This degradation could be explained by either difficulty in estimating them or falsely assuming what their underlying distribution is. Traditionally, the remedy to the performance degradations caused by using pitch and energy has been to not use them at all in any part of developing the system.

Acoustic modeling in ASR, therefore, considers for time frames $n = 1, \ldots, N$, the sequence of acoustic vectors $X = \{x_1, \ldots, x_n, \ldots, x_N\}$, associated with a sequence of hidden, discrete states, $Q = \{q_1, \ldots, q_n, \ldots, q_N\}$, where each state q_n can take one of K discrete values: $\{1, \ldots, k, \ldots, K\}$, each of these being associated with a specific probability distribution. Each distribution then models the emission of each x_n at time frame n:

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

Usually, attempts to use pitch or energy information in ASR were associating with each x_n an additional variable a_n, here referred to as an "auxiliary variable," yielding the sequence $A = \{a_1, \ldots, a_n, \ldots, a_N\}$. ASR was then performed by incorporating a_n in the emission distribution:

$$p(x_n, a_n|q_n). \tag{2}$$

However, this usually degraded the recognition.

While a discrete valued a_n is possible, we consider continuous valued a_n, which can have value $a_n = z$ and can be either pitch or energy values. We call a_n an auxiliary variable as it contains information that is not itself important for recognition but which has an impact on x_n. With these auxiliary variables, we investigate here two approaches to properly using them in ASR:

1. Conditioning the distribution of x_n upon a_n, as done in [3]. That is, using emission distributions of the form:

   $$p(x_n|q_n, a_n), \tag{3}$$

   where a_n appears as a continuous conditional variable.

2. Training with a_n but marginalizing it out (i.e. *hiding* it) in recognition, for example, in the case of (2) with continuous a_n:

   $$p(x_n|q_n) = \int p(x_n, a_n|q_n)\,\mathrm{d}a_n \tag{4}$$

   We note here that this is similar in spirit to work done in missing feature ASR [9], which marginalizes out features that are assumed to be corrupted by noise. In the simplest case, it ignores the noisy dimensions of the feature vector in calculating the emission likelihood.

These two approaches resemble what has already been done in the case of a discrete a_n representing gender [4]. One method of using gender modeling involves conditioning the distribution of x_n upon the gender–having a distribution for males and a distribution for females, based on (3). The distribution giving the highest likelihood when inserted into the ASR system is then used. Alternatively, the two distributions can be summed for each time frame, in a parallel manner to the integration in (4).

In this paper, we use DBNs as our framework for research into auxiliary variables with ASR. They are closely related to HMMs but are a more general framework that allows both the topology and the distributions to be easily modified (e.g., using (3) instead of (2)). Additionally, they allow the data to be arbitrarily hidden, thus marginalizing it out, as in (4). This work builds upon that of [11], which used the same training database and similar features but with single (conditional) Gaussians.

In Section 2 we will go into more detail about how a_n can be incorporated using the approaches proposed above. We do this work in the context of DBNs, which are explained in Section 3. Section 4 then gives more details of these pitch and energy auxiliary features, followed in Section 5 by the experimental results. We conclude in Section 6.

2 AUXILIARY INFORMATION

With both standard features x_n (MFCCs in this work) and auxiliary features a_n (either pitch or energy in this work) for time frame n, different statistical independence assumptions can be made between features. Here we propose that x_n needs to be dependent upon a_n; we then show how the resulting distributions might be modeled. We also propose an assumption that a_n needs to be marginalized out in recognition for certain cases.

Conditional Auxiliary Information.
In standard ASR, the distribution of x_n is dependent only on the discrete hidden state q_n, using a Gaussian distribution with mean vector μ_k^x and covariance matrix Σ_k^x for each state $q_n = k$:

$$p(x_n|q_n = k) \sim \mathcal{N}_x(\mu_k^x, \Sigma_k^x) \tag{5}$$

Σ_k^x is normally assumed to be a diagonal covariance matrix, thus containing non-zero elements only along the diagonal. This implies that there is no statistical correlation between the dimensions within the Gaussian and, thus, reduces the complexity of the system. This Gaussian distribution, as well as that of (6), (7), and (9) below, is based on Gaussian mixtures in our experiments, as is typically done in ASR. The exception here is that we always model a_n with a single Gaussian.

In attempting to add a_n to the ASR models, the simplest manner is to append it to the standard feature vector x_n, thus producing the Gaussian:

$$p(x_n, a_n | q_n = k) \sim \mathcal{N}_{x,a}(\mu_k^{x,a}, \Sigma_k^{x,a}) \tag{6}$$

With standard approaches, this would also assume a diagonal covariance in the expanded Gaussian, thus suggesting that there is no correlation between a_n and x_n. This is indeed the assumption, for example, between MFCCs and pitch/energy: the MFCCs are assumed to have pitch and energy removed (assuming that the zeroth coefficient is not used). However, these auxiliary features are such fundamental features of speech, that it may be a very erroneous to assume that they are uncorrelated with x_n. So, we propose that, conversely, there may be correlation between x_n and an a_n of either pitch or energy that needs to be modeled.

To model the correlation between x_n and a_n, we therefore propose that a_n should not be appended to x_n as above. Rather, the distribution for x_n should be conditioned upon the continuous value of a_n, as in (3). However, the modeling of $p(x_n | q_n = k, a_n = z)$ is not a straightforward task. Just as there are many approaches to modeling (1), such as Gaussians and artificial neural networks (ANNs), there may be many viable approaches to modeling (3). If we had been using a discrete valued a_n with Z discrete values, a straightforward way would have been to have $K \cdot Z$ Gaussians for each of the possible values of $(q_n = k, a_n = z)$, thus resembling the approach to gender modeling with ANNs in [4]. However, with a continuous valued a_n, we need a distribution for value $q_n = k$ which adapts itself to the continuous value $a_n = z$. This adaptation could involve linear methods (e.g., regression) or non-linear methods (e.g., ANN). Furthermore, in order to be incorporated into the full DBN framework, it should have the necessary operators for distributions in DBNs: marginalization to fewer dimensions, combination with other like distributions, etc.

We have chosen to represent (3) as *conditional* Gaussians, which have already been incorporated into the DBN framework [5] and have also been recently proposed by others in ASR research [3]:

$$p(x_n | q_n = k, a_n = z) \sim \mathcal{N}_x(u_k, \Sigma_k^x), \tag{7}$$

where x_n is modeled by a Gaussian whose mean is itself a regression on the mean of x_n and the value of a_n: $u_k = \mu_k^x + B_k^T z$. The weight on μ_k^x itself is 1 while B_k is the matrix containing the weights upon z, the value of a_n. A drawback of this distribution is that Σ_k^x itself is not dependent upon z; so, the same Σ_k^x will be used no matter what value of u_k is computed using the regression. Using only this distribution to calculate the emission likelihoods assumes that a_n itself is independent of q_n, that is, $p(a_n | q_n) = p(a_n)$. (In the implementation, (7) is actually multiplied by this value $p(a_n)$).

However, with (7) we do have the further possibility of whether a_n itself should be conditioned upon q_n, as was done in (6). This would be done if the evolution of A was assumed to be dependent upon that of Q. A simple way to model a_n would be to use a Gaussian for each $q_n = k$:

$$p(a_n | q_n = k) \sim \mathcal{N}_a(\mu_k^a, \Sigma_k^a). \tag{8}$$

Thus, the product of (7) and (8) would be used to compute the joint emission likelihood of x_n and a_n:

$$p(x_n|q_n = k, a_n) \cdot p(a_n|q_n = k) \sim \mathcal{N}_x(u_k, \Sigma_k^x) \otimes \mathcal{N}_a(\mu_k^a, \Sigma_k^a), \tag{9}$$

where $\otimes$ is the combination operator for (conditional) Gaussians, as defined in [7]. The difference between (9) and (6) is that we have here accounted for the correlation between x_n and a_n.

Marginalized Auxiliary Information.
Missing feature theory in ASR [9] has proposed to marginalize out those features which are noisy in recognition. Likewise, we propose a similar idea with auxiliary information. We still would want to use the auxiliary information in training so as to extract useful statistical information from it in order to better estimate the parameters in the models. While the data or its supposed model may be noisy, the training has the advantage of having a large amount of data over which it can extract relevant statistics. However, in recognition, there may be a lot of noise associated with the A presented for a single utterance. Using the estimated A (the "observed" A) in the emission distributions may produce a faulty likelihood. In such a case, it may be better to *hide* the A, which is accomplished by marginalizing it out of the emission distribution. In the case of the emission distribution (6), where a_n is appended to the feature vector, (4) illustrates this marginalization. After having been trained with conditional Gaussians, the emission distributions (7) and (9) may as well have problems with the noisy A. We can, therefore, obtain the distribution only for x_n by hiding, and, thus, integrating over a_n:

$$p(x_n|q_n) = \int p(x_n, a_n|q_n)\,\mathrm{d}a_n = \int p(x_n|q_n, a_n) \cdot p(a_n|q_n)\,\mathrm{d}a_n \tag{10}$$

$$\approx \int p(x_n|q_n, a_n) \cdot p(a_n)\,\mathrm{d}a_n. \tag{11}$$

where (10) applies to (9), where a_n is dependent upon q_n and (11) applies to the case of (7), where we assume that $p(a_n) = p(a_n|q_n)$.

3 DYNAMIC BAYESIAN NETWORKS

In our work, we incorporated auxiliary features in the DBN framework as it allows more flexibility in structuring the topology of the distributions and in allowing variables to arbitrarily be observed or hidden. HMMs can also model the same distributions and can have observed or hidden variables; however, they lack the generality in their algorithms that allows the topology of the distributions and the specification of hidden versus observed variables to be changed easily. So, we here outline what DBNs are and how they are visualized when using auxiliary information in ASR.

As illustrated in Figure 1, a DBN (a type of graphical model [1]) is a probabilistic model composed of three items:

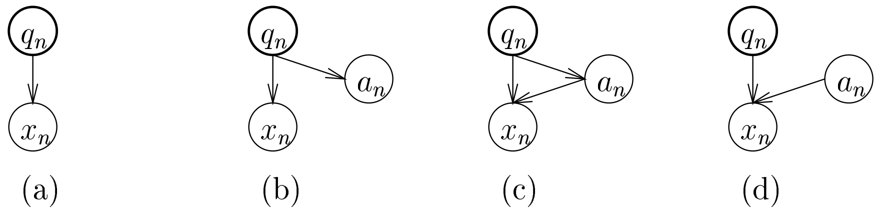

Figure 1: Portions of DBNs for time frame n (discrete variables having bold vertices), as initially proposed in [13]: (a) for standard HMM-based ASR; (b) for standard HMM-based ASR with concatenated a_n; (c) for ASR with x_n conditioned on a_n and a_n conditioned on q_n; (d) for ASR with x_n conditioned on a_n. Based on (5), (6), (9), and (7), respectively.

1. A set of variables $V = \{v_1^1, \ldots, v_1^w, \ldots, v_1^W, \ldots, v_N^1, \ldots, v_N^w, \ldots, v_N^W\}$. That is, there are W variables, each of which is modeled over the N time frames. The variables in the DBNs in Figure 1 are $\{q_1, a_1, x_1, \ldots, q_N, a_N, x_N\}$.

2. A directed acyclic graph (DAG), with a one-to-one mapping between each of its vertices and each $v_n^w \in V$.

3. For each $v_n^w \in V$, a local probability distribution which is conditioned upon the values of its parents in the DAG:

$$P(v_n^w|\text{parents}(v_n^w)). \tag{12}$$

   For example, the local probability distribution of x_n in Figure 1 (c) is $p(x_n|\text{parents}(x_n)) = p(x_n|q_n, a_n)$, which is the same as (7).

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

$$P(V) = \prod_{v_n^w \in V} P(v_n^w|\text{parents}(v_n^w)) \tag{13}$$

For a discrete v_n^w with zero or more discrete parents, its local probability distribution is defined by a table of discrete probabilities (it is not allowed to have any continuous parents in this framework). For a continuous v_n^w, its local probability distribution is defined by a Gaussian if it has no continuous parents or by a conditional Gaussian if it has continuous parents; if there are discrete parents, there will be a (conditional) Gaussian for each possible instantiation of the discrete parents. In the case of having continuous parents, the conditional Gaussian's mean is a regression on the mean of v_n^w itself and on the values of the continuous parents.

We use the inference algorithm in [7] to compute $P(v_n^w|O)$, the posterior marginal distribution of v_n^w given all of the observations O, as well as $P(O|V)$, the likelihood of the observations. For example, if in the DBN in Figure 1 (c), we have the observation $a_n = 2.5$, the inference algorithm would give the posterior marginals of $P(q_n|a_n = 2.5)$ and $p(x_n|a_n = 2.5)$ as well as the likelihood of the observation $p(a_n = 2.5)$. Any variable can be observed or hidden, 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 [6] for learning the discrete probabilities $P(\cdot)$, the means μ, the regression weights B, and the covariances Σ.

4 PITCH AND ENERGY AS AUXILIARY VARIABLES

In a first set of experiments, the auxiliary variable a_n was defined as the pitch value at time frame n. In our case, this pitch value, which we defined here as being the fundamental frequency F_0, was estimated using the simple inverse filter tracking (SIFT) algorithm [8], 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. While a fundamental property of the speech signal, it is a hard feature to estimate. Thus, any estimation of pitch will have inherent noise in it.

In a second set of experiments, the auxiliary variable a_n was defined as the short-term energy and was computed as follows:

$$a_n = \frac{1}{C} \sum_{t=1}^{T} s_n^2[t] \cdot w^2[t] \tag{14}$$

where $\{s_n[1], \ldots, s_n[t], \ldots, s_n[T]\}$ is the speech signal of T samples associated with time frame n, and $\{w[1], \ldots, w[t], \ldots, w[T]\}$ is a Hamming window, and C is a normalizing constant used to give manageable values for the short-term energy. It is straightforward to estimate in clean speech but harder to estimate in noisy speech.

5 EXPERIMENTS

Using the PhoneBook telephone speech corpus [10] with the small training set defined in [2], we train four types of DBN systems to do speaker-independent, task-independent, small vocabulary (75 words) isolated-word recognition. There are 41 context-independent, three-state phones in these systems, as well as initial silence and end silence models. Training was done using the EM algorithm, using a convergence criterion of stopping one iteration after the log-likelihood of the training data increased by less than 0.1% over that of the previous iteration. Each system with auxiliary information was tested two times on the test utterances defined in [2].

Similarly to [13], mel-frequency cepstral coefficients (MFCCs) are extracted from the speech signal, sampled at 8 kHz, using a window of 25 ms

DBN	Eq	Mix	Obs. Pitch	Hid. Pitch
Figure 1 (a) (Baseline)	(5)	4	5.9% (21k)	
Figure 1 (a) (Baseline)	(5)	6	4.3% (32k)	
Figure 1 (b)	(6)	4	60.5% (22k)	19.2% (21k)
Figure 1 (c)	(9)	4	48.9% (32k)	6.2% (21k)
Figure 1 (d)	(7)	4	5.3% (32k)	6.0% (21k)

Table 1: Pitch. Word error rates (WERs) (and number of parameters) using Pitch as an auxiliary variable. The labels of the underlying equations and the number of Gaussian mixtures for x_n (a_n has a single Gaussian) are also given. Equation (5) is equivalent to standard HMM-based ASR using only x_n while (6) is equivalent to standard HMM-based ASR using x_n and a_n in a single feature vector (except that a_n has a single Gaussian). Equations (9) and (7) use conditional Gaussians, with (7) treating a_n as independent of q_n. With "hidden" a_n, we are marginalizing it out of the emission distribution. Systems with a similar number of parameters are to be grouped together for performance comparisons against the respective baseline system.

with a shift of 8.3 ms for each successive frame. x_n is composed of the following MFCC elements: $C_1, \ldots, C_{10}, \Delta C_1, \ldots, \Delta C_{10}, \Delta C_0$, where C_i is the ith MFCC and ΔC_i is its approximate first derivative.

The recognition results where a_n is pitch and where a_n is short-term energy, as well as for the baseline x_n-only systems, are given in Tables 1 & 2, respectively. When marginalizing over a_n, its parameters are removed, having been merged into the parameters for x_n, as shown in (4), (10), and (11). Thus, the WERs with hidden (marginalized out) A show a lower effective number of parameters than when A is observed. Therefore, with A marginalized out in an auxiliary system, it has essentially the same structure and parameters as a baseline; the difference is that the parameters have been trained using an auxiliary variable. This is the reason for two baseline systems: for comparing against a baseline system, we use a system that has the same effective number of parameters. We note that it was not our intention to find the number of mixtures which gives each system its optimum performance. Rather, within each set of experiments, we wanted to have systems that were comparable in the number of parameters.

These results confirm the difficulty in incorporating auxiliary information in the traditional way, using (6), which provides a very poor recognition WER of 60.5% for pitch and 28.9% for energy. Furthermore, they show

DBN	Eq	Mix	Obs. Energy	Hid. Energy
Figure 1 (a) (Baseline)	(5)	4	5.9% (21k)	
Figure 1 (a) (Baseline)	(5)	6	4.3% (32k)	
Figure 1 (b)	(6)	4	28.9% (22k)	6.3% (21k)
Figure 1 (c)	(9)	4	27.4% (32k)	5.9% (21k)
Figure 1 (d)	(7)	4	5.9% (32k)	19.4% (21k)

Table 2: Energy. Word error rates (WERs) using short-term energy as an auxiliary variable, presented as in Table 1.

the great improvement we can achieve by letting a_n condition x_n's emission distribution. That is, by using a conditional Gaussian for x_n, as in (7), instead of (6), we decreased the WER by a relative 91% (60.5% to 5.3%) for pitch and 80% (28.9% to 5.9%) for energy. It is the system with (7) where observed pitch or energy auxiliary information provides its most promising results.

Marginalization (i.e., using hidden auxiliary information) dramatically increases the performance of the poorly performing systems, those using (6) or (9), with pitch or energy auxiliary information. Moreover, marginalizing out a_n on the systems using (9) "recovers" the performance of the baseline system with four mixtures. Marginalization of those using (7), however, has a negative effect on performance. As this is done using (11), the prior $p(a_n)$ is used, which was not learned in training but was just defined using the mean (and variance) of a_n across all of the training data. Using a global mean for a_n may have introduced problems in computing the marginals.

6 CONCLUSION

We have presented a new approach for properly including auxiliary variables in standard ASR. Although it is not yet perfect, the results reported here demonstrate the validity of this approach. While the results here do not improve over the baseline approach, earlier results showed how discretized pitch auxiliary information does bring improvement [12]. So continuous auxiliary information, as was used in the current work, still has the potential to improve over the baseline within the current framework.

More work is now required using continuous auxiliary variables. First, we need to improve the estimation of the auxiliary variables. For example, with energy, this could involve using the logarithm of the energy, using a longer-term energy, or in using the energy of a frequency sub-band (as done in [3]). Second, better distributions (e.g. Gaussian mixtures) may be needed to better model a_n instead of just single Gaussians. Finally, equivalence classes (a form of parameter tying [13]) to model a_n conditioned upon q_n may prove to be more robust; these could be used, for example, to have a_n conditioned on broad classes of q_n, such as vowels and consonants, thus having a hybrid between (9) and (7).

ACKNOWLEDGMENTS

Todd A. Stephenson and Mathew Magimai-Doss are supported by the Swiss National Science Foundation under grants FN 2000-064172.00/1 and FN 2100-057245.99/1, respectively. Andrew Morris provided useful insight into missing feature theory.

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)**, Munich, April 1997, vol. 3, pp. 1767–1770.

[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)**, Salt Lake City, Utah, USA, May 2001, vol. 1, pp. 513–516.

[4] Y. Konig and N. Morgan, "GDNN: A Gender-Dependent Neural Network for Continuous Speech Recognition," in **Proceedings of the 1992 International Joint Conference on Neural Networks (IJCNN)**, Baltimore, MD, June 1992, pp. 332–337.

[5] S. L. Lauritzen, "Propagation of Probabilities, Means, and Variances in Mixed Graphical Association Models," **Journal of the American Statistical Association**, vol. 87, no. 420, pp. 1098–1108, December 1992, Theory and Methods.

[6] S. L. Lauritzen, "The EM Algorithm for Graphical Association Models with Missing Data," **Computational Statistics & Data Analysis**, vol. 19, pp. 191–201, 1995.

[7] S. L. Lauritzen and F. Jensen, "Stable local computations with conditional Gaussian distributions," **Statistics and Computing**, vol. 11, no. 2, pp. 191–203, April 2001.

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

[9] A. C. Morris, M. P. Cooke and P. D. Green, "Some solutions to the missing feature problem in data classification, with application to noise robust ASR," in **Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-98)**, Seattle, Washington, USA, May 1998, vol. 2, pp. 737–740.

[10] 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)**, Detroit, MI, May 1995, vol. 1, pp. 101–104.

[11] T. A. Stephenson, M. Magimai-Doss and H. Bourlard, "Mixed Bayesian Networks with Auxiliary Variables for Automatic Speech Recognition," in **International Conference on Pattern Recognition (ICPR 2002)**, Quebec City, PQ, Canada, August 2002, to appear.

[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**, Aalborg, Denmark, September 2001, vol. 4, pp. 2765–2768.

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