BURIED MARKOV MODELS FOR SPEECH RECOGNITION

Jeff A. Bilmes
<bilmes@cs.berkeley.edu>

International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA

CS Division, Department of EECS
University of California at Berkeley
Berkeley, CA 94720, USA

ABSTRACT

Good HMM-based speech recognition performance requires at most minimal inaccuracies to be introduced by HMM conditional independence assumptions. In this work, HMM conditional independence assumptions are relaxed in a principled way. For each hidden state value, additional dependencies are added between observation elements to increase both accuracy and discriminability. These additional dependencies are chosen according to natural statistical dependencies extant in training data that are not well modeled by an HMM. The result is called a *buried Markov model* (BMM) because the underlying Markov chain in an HMM is further hidden (buried) by specific cross-observation dependencies. Gaussian mixture HMMs are extended to represent BMM dependencies and new EM update equations are derived. On preliminary experiments with a large-vocabulary isolated-word speech database, BMMs are able to achieve an 11% improvement in WER with only a 9.5% increase in the number of parameters using a single state per mono-phone speech recognition system.

1. INTRODUCTION

Hidden Markov Models (HMMs) are the most common method used in automatic speech recognition systems to model the joint probability distribution of feature vectors for a given utterance model. Two conditional independence assumptions characterize HMMs: 1) observations are conditionally independent of other observations given the hidden state at the current time, and 2) the hidden state is conditionally independent of any preceding variables given the previous hidden state. In principle, an HMM can model a given probability distribution to an arbitrarily high degree of accuracy [3, 7]. The conditional independence assumptions associated with HMMs as they are used in practice, however, have not been demonstrably or provably shown to be sufficient for optimal speech recognition. Therefore, an open challenge is how to either increase the modeling power of or change the modeling assumptions made by an HMM such that, without an enormous increase in free parameters and complexity, speech-recognition error rates can improve.

One method to increase an HMM's ability is to increase the number of hidden states[7, 3] or to use a factored hidden-variable representation [6]. An alternative method adds explicit fixed dependencies from observations to other observations in the acoustic context, e.g., AR-HMMs [12, 8] and correlation models [13]. And in segmental models[10], a hidden state corresponds to a segment trajectory that defines local statistics.

In this paper, a new method is introduced that augments an HMM's modeling power in a systematic way. Starting with an existing HMM, statistical dependencies are added between observations that, for each hidden state value, provide both useful and discriminative information not already provided by the hidden variable. Previous results [3] showed WER improvements on a small-vocabulary isolated-word speech database using whole-word models. In this paper, preliminary reports show WER reductions for the PHONEBOOK[11] database relative to a baseline HMM. Section 2 motivates the addition of data-derived statistical dependencies to an HMM. Section 3 outlines a dependency selection algorithm based on conditional mutual information. Section 4 uses a graphical model[9] to describe the result. Section 5 introduces and derives EM update equations for Gaussian mixture extensions. Section 6 provides WER results on the PHONEBOOK database.

2. LEARNING BOTH STATISTICAL DEPENDENCIES AND MODEL PARAMETERS FROM DATA

In many statistical pattern classification tasks, a corpus of data is used to train a parametric probabilistic model. While the parameters of the model vary during training, the underlying model structure in terms of the attainable statistical dependencies stays fixed. Typically, a priori fixed statistical models are used without analyzing their ability to model crucial dependencies in the data. If such a model is unable to represent the important dependencies, performance might suffer regardless of how well the model is trained.

Natural sensory signals originating from the environment have idiosyncratic statistical dependencies. There have been successful attempts to predict neural processing using only environmental statistical properties [1]. Indeed, it is these statistical dependencies that separate *signal* from *noise* and can help distinguish one class of signals from another or one type of object within a class from another. A logical approach therefore is to "predict" a statistical model from the data. For a class of signals such as speech, benefits could be obtained by explicitly modeling speech's unique statistical dependencies.

Modern speech recognition systems use HMMs to represent probability distributions. A reasonable approach therefore is to augment an HMM with only those statistical dependencies that are found to be missing but useful according to training data. In the work presented here, the dependencies are between individual observation elements. The training data is then used as usual to adjust the resulting model's parameters. The result is called a *buried Markov model* (BMM) because the underlying Markov chain in an HMM is further hidden (buried) by specific cross-observation dependencies.

3. BUILDING BMM DEPENDENCIES

For a given number of hidden-variable states, the degree to which a hidden variable does not contain contextual information can be measured using conditional mutual information. The conditional mutual information $I(X_t; X_{<t}|Q_t) = \sum_q I(X_t; X_{<t}|Q_t = q)p(Q_t = q)$ represents[1] the quantity of additional information $X_{<t}$ provides about X_t not already provided by Q_t, where Q_t is the hidden random variable at time t. In particular, $I(X_t; X_{<t}|Q_t = q)$ represents the amount missing for a particular hidden state value q. This suggests that if $I(X_t; X_{<t}|Q_t = q) > 0$, the accuracy of

[1]The notation $X_{1:N}$ represents the set $\{X_1, \ldots, X_N\}$ and $X_{<t} = X_{1:(t-1)}$

To Appear IEEE ICASSP 1999.

an HMM can be improved without increasing the number of states by augmenting the observation models with dependencies directly on contextual data. It also suggests that dependencies should be added 1) only on the "relevant" contextual data, 2) that are potentially distinct for each value of Q_t, and 3) that are chosen to provide only new information not already provided by Q_t.

Using the hidden-variable first-order Markov assumption, the joint distribution of the observations can be written:

$$p(X_{1:T}) = \sum_{q_{1:T}} \prod_t p(X_t|X_1, \ldots, X_{t-1}, q_t) p(q_t|q_{t-1}).$$

In this form, the distribution of X_t depends on all previous time frames. As described in [3] (and motivated in [4]), the following slightly more general model is considered:

$$p(X_{1:T}) = \sum_{q_{1:T}} \prod_t p(X_t|X_{R_{q_t}}, q_t) p(q_t|q_{t-1})$$

where $X_{R_{q_t}} \subset \{X_1, \ldots, X_{t-1}, X_{t+1}, \ldots, X_T\}$ is a subset of X_t's surrounding context. Assume the number of elements in $X_{R_{q_t}}$ is fixed. If the elements of $X_{R_{q_t}}$ are chosen to maximize the conditional mutual information $I(X_t; X_{R_{q_t}}|Q_t = q_t)$ for each possible value of q_t, $X_{R_{q_t}}$ will be a vector consisting of relevant (i.e., entropy reducing) and non-redundant (i.e., containing information not already provided by Q_t) portions of X_t's context given $Q_t = q$.

To increase tractability, dependencies are considered and added individually for each feature element. Define the context of X_{ti} (the i^{th} element of X_t) as the set $\mathcal{Z}_{ti} = \{X_{t-\ell,j} : \forall \ell, j\} - \{X_{ti}\}$. The set of N variables $Z^i_{k_{1:N}} = \{Z^i_{k_1}, \ldots, Z^i_{k_N}\}$ providing the greatest entropy reduction of X_{ti} when $Q_t = q$ can be found by evaluating:

$$\underset{Z^i_{k_{1:N}} \subset \mathcal{Z}_{ti}}{\text{argmax}} \quad I(X_{ti}; Z^i_{k_{1:N}}|Q_t = q).$$

Alone, this selection method suffices to increase the descriptive power (i.e., lead to a higher likelihood) of the model for a particular state q but does not necessarily decrease classification error. A potential problem, therefore, is that the chosen dependencies might also reduce "entropy" in the context of a different and incorrect state. To increase the discriminability between different states, dependencies should be chosen that both 1) decrease entropy in the context of the correct state and 2) do not decrease the entropy (as much) in other contexts. This second concept can be represented with the following mutual information-like quantity:[2]

$$I_{\{Q_t=r\}}(X_{ti}; Z^i_{k_{1:N}}|Q_t = q) =$$

$$E_{p(X_{ti}, Z^i_{k_{1:N}}|Q_t=r)}\left[\log \frac{p(X_{ti}, Z^i_{k_{1:N}}|Q_t = q)}{p(X_{ti}|Q_t = q)p(Z^i_{k_{1:N}}|Q_t = q)}\right]$$

for $r \in C_q$ where C_q is the set of states confusable with q. The quantity $I_{\{Q_t=r\}}(X_{ti}; Z^i_{k_{1:N}}|Q_t = q)$ is similar to mutual information except that the individual event-wise entropy reductions are averaged using the probability distribution for the confusable context r rather than the original context q. Using this notation, $I(X_{ti}; Z^i_{k_{1:N}}|Q_t = q) = I_{\{Q_t=q\}}(X_{ti}; Z^i_{k_{1:N}}|Q_t = q)$. When $r \neq q$, it represents the situation in a classification task during evaluation of a model in an incorrect context.

A general dependency selection algorithm is as follows: for each q and i, choose the size N_q set of variables $Z^i_{k_{1:N_q}}$ for which

[2]Using the notation $E_{p(X)}[f(X)] = \int p(x)f(x)dx$.

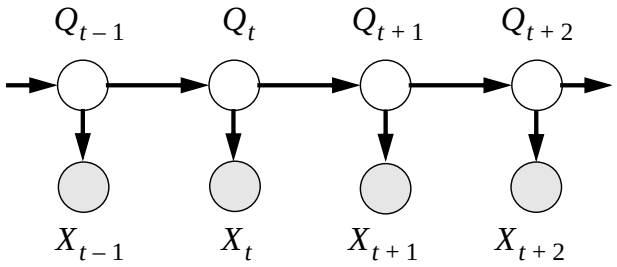

Figure 1: Graphical model of an HMM. Shaded (resp. unshaded) nodes represent observed (resp. hidden) variables.

$I(X_{ti}; Z^i_{k_{1:N_q}}|Q_t = q)$ is large and $I_{\{Q_t=r\}}(X_{ti}; Z^i_{k_{1:N_q}}|Q_t = q)$ is small for each $r \in C_q$.

In [3], it is empirically shown that $I(X_{ti}; Z^i_{k_{1:N_q}}|Q_t = q) > 0$. Moreover, several approximations are made to the above procedure that lead to a heuristic dependency selection algorithm. This algorithm chooses a set of dependencies $\mathbf{Z}_{qi}$ for each q and i:

Set $\mathbf{Z}_{qi} = \emptyset$
Sort $Z_j \in \mathcal{Z}_{ti}$ into an order decreasing by $U_{ti}(Z_j)$
Repeat over j until $U_{ti}(Z_j) < \tau_u$ or $|\mathbf{Z}_{qi}| = N_q$:
  If Z_j satisfies all the following criteria:
    1) $I(X_{ti}; Z_j|Q_t = q) > \tau_q$
    2) For each $Z \in \mathbf{Z}_{qi}$, $I(Z_j; Z|Q_t) < \tau_g I(Z_j; X_{ti}|Q_t = q)$
    3) $I(X_{ti}; Z_j|Q_t \in C_q) < \tau_c$
  then add Z_j to $\mathbf{Z}_{qi}$.

Z_j is a variable in X_{ti}'s context under consideration and $U_{ti}(Z_j)$ is the *utility* of Z_j approximated as:

$$\hat{U}_{ti}(Z_j) = I(X_{ti}; Z_j|Q = q) - I(X_{ti}; Z_j|Q \in C_q).$$

This algorithm requires only the computation of pairwise conditional mutual information for a given labeling scheme.

4. GRAPHICAL MODELS

A graphical model [9] (also called a Bayesian network) provides a natural way of depicting conditional independence assumptions about a collection of random variables. A directed graphical model is a graph where nodes represent random variables and directed edges represent conditional independence assumptions corresponding to directed separation (or d-separation[9]) properties. Both HMMs and BMMs can be described using a graphical model. Indeed, the preceding dependency selection procedure can be considered a form of graphical model structure-learning algorithm.

The graphical model for an HMM is described in Figure 1. Both types of HMM conditional independence assumptions are represented by this picture.

BMM conditional independence assumptions can also be described by such a graph, as shown in Figure 2. This graph shows those dependencies only for a particular assignment to the hidden variables. A different assignment will result in a different set of cross-observation dependencies. While it is possible to use a graphical model to describe the dependencies under all hidden-variable assignments, such a graph quickly becomes unwieldy (because of repeated burying).

In Figure 2, a distinction is made between two different observation vector streams $X_{1:T}$ and $Y_{1:T}$. For an HMM,

$$p(X_{1:T}, Y_{1:T}) = \sum_{q_{1:T}} \prod_t p(X_t, Y_t|q_t) p(q_t|q_{t-1}).$$

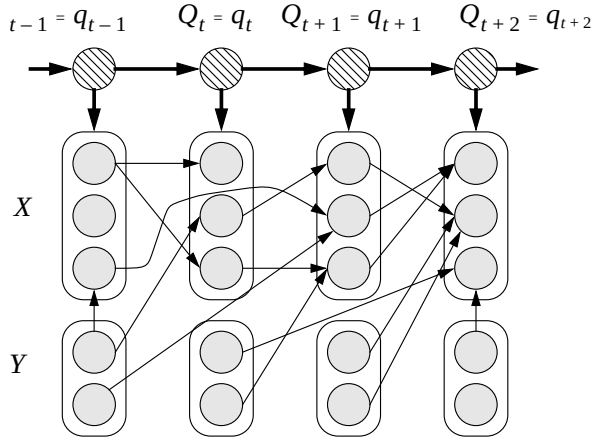

Figure 2: Graphical model of a BMM with a particular assignment to the hidden variables. Striped nodes indicate hidden variable binding.

and $p(X_t, Y_t|q_t) = p(X_t|Y_t, q_t)p(Y_t|q_t)$. If Y_t is marginally independent of q_t but Y_t not independent of q_t given X_t, then modeling $p(Y_t|q_t)$ is at best superfluous and at worst detrimental (i.e., it could add irrelevant and interfering parameters to the model). With the BMM described in the figure, $p(Y_t|q_t)$ is ignored but the dependency variables may consist of elements from both the X and Y streams. For speech recognition, X_t could be RASTA-PLP or MFCC features; Y_t could be acoustic features informative about vocal tract length, gender, speaking rate, noise condition, etc. all of which have little if any dependence on the hidden state (e.g., phone, syllable, etc.). With such features, a BMM is perhaps analogous to an online speaker adaptation procedure.

5. GAUSSIAN-MIXTURE BMMS

Gaussian mixture HMMs can be extended to include the cross-observation dependencies specified by a BMM. The observation models should allow their entropy to be affected by the additional dependencies while still leading to efficient EM update equations. To this end, hidden variables m and v are introduced to obtain the following:

$$p(x|z,q) = \sum_{m=1}^{M}\sum_{v=1}^{V} p(x,m,v|z,q)$$

where $x = (x_1,\ldots,x_d)'$ is an observation vector, $z = (z_1,\ldots,z_s,1)'$ is the entire collection of dependency variables any element of x might use (appended with the constant 1 to compute a fixed mean offset), m indicates a mixture component, and v indicates the class of z. m is assumed to be independent of other variables given v and q, and v is assumed to be independent of other variables given z resulting in:

$$p(x|z,q) = \sum_{m=1}^{M}\sum_{v=1}^{V} p(x|m,v,z,q)p(m|v,q)p(v|z)$$

where $p(m|v,q)$ is a discrete probability table, $p(v|z)$ is the probability of class v given continuous vector z, and

$$p(x|m,v,z,q) = \frac{1}{(2\pi)^{d/2}|\Sigma_{qmv}|^{1/2}} e^{-\frac{1}{2}(x-B_{qmv}z)'\Sigma_{qmv}^{-1}(x-B_{qmv}z)}$$

is a Gaussian distribution with mean $B_{qmv}z$ and covariance Σ_{qmv}. The $d\times(s+1)$-sized B_{qmv} matrices have a sparse structure determined by the BMM dependencies for state q.

With z containing observations only from x's past, these equations alone constitute a generalization of vector-valued autoregressive HMMs [8, 13] ($d > 1$, $M = 1$, $V = 1$). With $V > 1$ and $M > 1$, this model can be considered a mixture of mixtures. An important difference from previous work is that here the dependency *structure*, as represented by B_{qmv}, is sparse, data-derived, and hidden-variable dependent as described in Section 3. Furthermore, z can contain observations from x's past, present, future, or from a different feature stream.

Consider data matrices $\mathbf{x} = x_{1:T}$, $\mathbf{m} = m_{1:T}$, $\mathbf{v} = v_{1:T}$, and $\mathbf{z} = z_{1:T}$. The transition matrix update equations are the same as usual, so their derivation is skipped and the dependence on q is momentarily dropped. The EM auxiliary equation becomes:

$$Q(\Theta,\Theta^g) = \sum_{\mathbf{m}}\sum_{\mathbf{v}} \log p(\mathbf{x},\mathbf{m},\mathbf{v}|\mathbf{z},\Theta)p(\mathbf{m},\mathbf{v}|\mathbf{x},\mathbf{z},\Theta^g)$$

$$= \sum_{m=1}^{M}\sum_{v=1}^{V}\sum_{t=1}^{T} \log p(x_t,m,v|z_t,\Theta)\gamma_{mvt}$$

where $\gamma_{mvt} = p(m,v|x_t,z_t,\Theta^g)$, Θ^g are parameters from the previous iteration, and Θ are the parameters to optimize. This reduces to three equations which can be independently maximized.

$$Q_1(\Theta,\Theta^g) = \sum_{m,v,t} \log p(x_t|m,v,z_t,\Theta)\gamma_{mvt} \quad (1)$$

$$Q_2(\Theta,\Theta^g) = \sum_{m,v,t} \log p(m|v,\Theta)\gamma_{mvt} \quad (2)$$

$$Q_3(\Theta,\Theta^g) = \sum_{m,v,t} \log p(v|z_t\Theta)\gamma_{mvt} = \sum_{v,t} \log p(v|z_t,\Theta)\gamma_{vt} \quad (3)$$

where $\gamma_{vt} = p(v|x_t,z_t,\Theta^g)$.

Ignoring any constants, Equation 1 can be represented as:

$$\sum_{m,v,t} -\frac{1}{2}\left[\log(|\Sigma_{mv}|) + (x_t - B_{mv}z_t)'\Sigma_{mv}^{-1}(x_t - B_{mv}z_t)\right]\gamma_{mvt}$$

Taking the derivative with respect to B_{mv} and setting the result to zero gives:

$$\sum_{t=1}^{T}(x_t - B_{mv}z_t)z_t'\gamma_{mvt} = 0$$

which can easily be solved for B_{mv}. To find the update rule for Σ_{mv}, let $w_{mvt} = x_t - B_{mv}z_t$ and $\mu_m = 0$. Equation 1 becomes

$$\sum_{m,v,t} -\frac{1}{2}\left[\log(|\Sigma_{mv}|) + (w_{mvt} - \mu_m)'\Sigma_{mv}^{-1}(w_{mvt} - \mu_m)\right]\gamma_{mvt}$$

which has the same form as the usual Gaussian mixture case [2, 7].

Equation 2 can be optimized by introducing a Lagrange multiplier λ:

$$\sum_{m,v,i} \log p(m|v,\Theta)\gamma_{mvt} - \lambda\left(\sum_m p(m|v,\Theta) - 1\right)$$

Equation 3 can be optimized by noting that v is assumed independent of x given z:

$$\sum_{v,t} \log p(v|z_t,\Theta)p(v|z_t,\Theta^g)$$

This quantity is maximized when $D(p(v|z_t,\Theta)||p(v|z_t,\Theta^g))$ is minimized which occurs when $\Theta = \Theta^g$ for those portions of Θ that affect this distributions. Therefore, $p(v|z)$ does not change between EM iterations, so any (perhaps unsupervised) classification method can be used prior to EM BMM learning.

Reintroducing the q variable, the EM update equations for maximum-likelihood parameter estimation are as follows:

$$B_{qmv} = \left(\sum_{t=1}^{T} \gamma_{qmv}(t) x_t z_t'\right)\left(\sum_{t=1}^{T} \gamma_{qmv}(t) z_t z_t'\right)^{-1},$$

$$\Sigma_{qmv} = \frac{\sum_{t=1}^{T} \gamma_{qmv}(t)(x_t - B_{qmv} z_t)(x_t - B_{qmv} z_t)'}{\sum_{t=1}^{T} \gamma_{qmv}(t)},$$

and

$$p(m|v,q) = \frac{\sum_{t=1}^{T} \gamma_{qmv}(t)}{\sum_{t=1}^{T}\sum_{m=1}^{M} \gamma_{qmv}(t)}$$

where $\gamma_{qmv}(t) = p(q_t = q, m_t = m, v_t = v|\mathbf{x}, \mathbf{z})$.

6. PHONEBOOK **RESULTS**

Speech recognition results were obtained using PHONEBOOK, a large-vocabulary, phonetically-rich, isolated-word, telephone-speech database[11]. All data is represented using 12 MFCCs plus c_0 and deltas resulting in a $d = 26$ element feature vector sampled every 10ms.

The training and test sets are as defined in [5]. Test words do not occur in the training vocabulary, so test word models are constructed using phone-models learned during training. Strictly left-to-right transition matrices were used except for an optional beginning and ending silence model.

An HMM baseline system, bootstrapped using uniform segmental k-means, was developed using 42 phone models (41 monophones + silence), a single HMM state per mono-phone, and no durational modeling. Each phone model uses a mixture of 48 diagonal covariance Gaussians. While more complex context-dependent multi-state phone-models are known to be beneficial, a demonstrable improvement in this initial simpler case is desirable. The dictionary included with PHONEBOOK distribution was used for all pronunciations.

BMMs with $V = 1$ are bootstrapped with an HMM as described in [3]. Eight phonetically derived clusters were used to define the confusable classes C_q: silence, tongue fronting vowels, diphthongs, tongue retraction, retroflex, nasals, fricatives, and plosives.

Lex. Size	75	150	300	600	Params
HMM	5.7%	7.6%	9.8%	14.1%	105k
BMM	5.1%	7.1%	9.3%	13.4%	115k

Table 1: HMM and BMM comparison. The dependency selection parameters are $\tau_u = 0.0, \tau_q = 2.25 \times 10^{-3}$, $\tau_g = 75\%$, $\tau_c = 6.1 \times 10^{-3}$, $N_q = 20$, and C_q is phonetically derived.

Results are shown in Table 1. Lexicon sizes of 75 (averaged over 8 independent test cases), 150 (4 cases), 300 (2 cases), and 600 words are presented. In each test case for each lexicon, a BMM is never worse than the corresponding HMM, so the increase in average performance is due only to a BMM outperforming an HMM. In the above, dependency links were allowed to span a maximum of 100ms (10 frames) into the past. Threshold values were chosen using 10^{th} percentiles of the mutual information data values.

The baseline HMM system used 105k observation model parameters and the BMM used an additional 10k parameters resulting in 115k parameters. The 5.7% baseline HMM result for the 75 word vocabulary is better than the 7.1% result reported for a comparable 153k observation-parameter mono-phone system [5]. The results show that for the 75 word lexicon, an 11% BMM WER performance improvement is be obtained with only a modest 9.5% increase in the number of parameters. Improvements are also found for the larger lexicon sizes.

7. DISCUSSION

HMM conditional independence assumptions can be relaxed by including additional probabilistic state-specific dependencies only to the relevant and discriminative observation context. In this paper, a method has been provided that chooses this context using conditional mutual information. WER performance improvements have been demonstrated on a large-vocabulary isolated-word speech database.

The selection of good dependencies was found to be important for achieving good WER results. For example, if τ_c is set high enough to render it nonexistent, and if Z_j is chosen ordered not by utility but by maximum mutual information alone, the likelihoods of the resulting models increase dramatically but the WER results become worse. This could explain the mixed success of AR-HMMs in the past[8] where "dependencies" are fixed a priori without regard to their affect on information gain and discriminability.

This work has benefited from discussions with Geoff Zweig, Katrin Kirchhoff, Nelson Morgan, and Nir Friedman and has been partially sponsored by ONR URI Grant N00014-92-J-1617 and a DoD IDEA grant.

8. REFERENCES

[1] J.J. Atick. Could information theory provide an ecological theory of sensory processing? *Network*, 3, 1992.

[2] J.A. Bilmes. A gentle tutorial on the EM algorithm and application to Gaussian Mixtures and Baum-Welch. Technical Report TR-97-021, ICSI, 1997.

[3] J.A. Bilmes. Data-driven extensions to HMM statistical dependencies. In *Proc. ICSLP*, Sidney, Australia, December 1998.

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

[5] S. Dupont, H. Bourlard, O. Deroo, J.-M. Fontaine, and J.-M. Boite. Hybrid HMM/ANN systems for training independent tasks: Experiments on phonebook and related improvements. In *ICASSP*, 1997.

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

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

[8] P. Kenny, M. Lennig, and P. Mermelstein. A linear predictive HMM for vector-valued observations with applications to speech recognition. *IEEE Trans. ASSP*, 38(2), February 1990.

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

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

[11] J. Pitrelli, C. Fong, S.H. Wong, J.R. Spitz, and H.C. Lueng. PhoneBook: A phonetically-rich isolated-word telephone-speech database. In *ICASSP*, 1995.

[12] A.B. Poritz. Linear predictive hidden markov models and the speech signal. In *ICASSP*, 1982.

[13] C.J. Wellekens. Explicit time correlation in hidden markov models for speech recognition. In *ICASSP*, 1987.