DATA-DRIVEN EXTENSIONS TO HMM STATISTICAL DEPENDENCIES

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

In this paper, a new technique is introduced that relaxes the HMM conditional independence assumption in a principled way. Without increasing the number of states, the modeling power of an HMM is increased by including only those additional probabilistic dependencies (to the surrounding observation context) that are believed to be both relevant and discriminative. Conditional mutual information is used to determine both relevance and discriminability. Extended Gaussian-mixture HMMs and new EM update equations are introduced. In an isolated word speech database, results show an average 34% word error improvement over an HMM with the same number of states, and a 15% improvement over an HMM with a comparable number of parameters.

1. INTRODUCTION

Hidden Markov Models (HMMs) are the most common statistical method used for automatic speech recognition where they model the joint probability distribution of a collection of random variables under certain statistical assumptions. Under the *first-order Markov assumption*, a set of "hidden" variables, one for each time point, form a discrete-valued first-order Markov chain. Under the *conditional independence assumption*, a set of observation variables, again one for each time point, are each conditionally independent of past variables given the corresponding hidden variable.[1] While HMMs can potentially represent rich probability distributions, these assumptions burden the hidden variables with the task of containing all relevant information about the observation variables' environment.

The conditional independence assumption can be further examined by observing how an HMM models $p(X_t|X_{<t})$ and comparing this with the "true" distribution; X_t is an observation vector at time t, and $X_{<t} = \{X_1, \ldots, X_{t-1}\}$ is the observed context preceding X_t. Without any modeling assumptions, X_t can be viewed as the output of a noisy channel with input $X_{<t}$ (Figure 1). For an accurate representation of $p(X_t|X_{<t})$, any channel model must have information transmission rate at least as big as $I(X_t; X_{<t})$ where $I(X;Y)$ is the mutual information between random vectors X and Y.

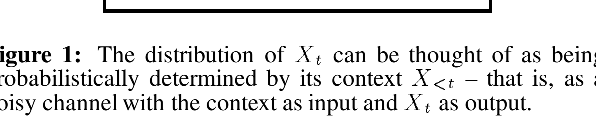

Figure 1: The distribution of X_t can be thought of as being probabilistically determined by its context $X_{<t}$ – that is, as a noisy channel with the context as input and X_t as output.

Under an HMM, $p(X_t|X_{<t}) = \sum_q p(X_t|Q_t = q)p(Q_t = q|X_{<t})$, where Q_t represents the random hidden state variable at time t. An HMM, therefore, attempts to compress the information about X_t contained in $X_{<t}$ into a single discrete variable Q_t (Figure 2). For an accurate representation, these two channels must be sufficiently powerful, i.e., $R_i \geq I(X_t; X_{<t})$ where R_i is the mutual information between the input and output of channel i. Furthermore, the number of hidden states must be large enough to accurately encode the information being transmitted. This is essentially a requirement that $|Q| \geq 2^{I(X_t;X_{<t})}$ where $|Q|$ is the number of hidden states. Assuming Q appropriately encodes the information contained in $X_{<t}$ relevant to X_t, an HMM's accuracy can be increased by increasing the number of states (as has been repeatedly noted in the past).

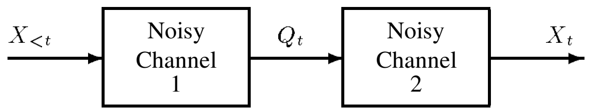

Figure 2: With an HMM, the information about X_t contained in $X_{<t}$ is "squeezed" through the hidden state variable Q_t. Depending on the number of hidden states, this can overburden Q_t and result in an inaccurate probabilistic model.

In this paper, a new technique is introduced that relaxes the conditional independence assumption in a principled way. Without increasing the number of states, the modeling power of an HMM is increased by including only those additional probabilistic dependencies believed to be useful according to training data. This can potentially lead to a more powerful statistical model without a large free-parameter increase. Section 2 introduces a data-driven method used to expand an HMM's probabilistic dependencies. Section 3 describes a heuristic approximation to the dependency selection algorithm given in Section 2. Section 4 describes an implementation of the extended HMMs and includes an EM training procedure, and Section 5 gives word-error results for an isolated-word digits data-base.

2. BURIED MARKOV MODELS

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 the quantity of additional information $X_{<t}$ provides about X_t not already provided by Q_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 an HMM can be improved without increasing the number of states by augmenting the probabilistic 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. This is depicted in Figure 3.

[1] Both are conditional independence assumptions; these names are used to distinguish the two assumptions later in the paper.

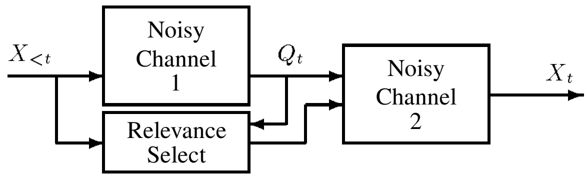

Figure 3: Improving an HMM by including additional direct dependencies on the relevant portions of $X_{<t}$ depending on the value of Q_t.

Using just the first-order Markov assumption, the joint distribution of the observations can be written:[2]

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

In this form, the model for the distribution of X_t depends only on previous time frames. While not necessary for subsequent analysis, the chain rule of probability can be violated[3] to get:

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

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. For a fixed size $X_{R_{q_t}}$, the problem becomes choosing the elements of $X_{R_{q_t}}$ to maximize the conditional mutual information $I(X_t; X_{R_{q_t}}|Q_t = q_t)$ for each q_t. In this case, $X_{R_{q_t}}$ is 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 $Q_t = q_t$.

Does additional information typically exist in the surrounding context given Q? Figure 4 shows a *conditional* mutual information density plot $I(\Delta f, \ell|Q) = avg_{i-j=\Delta f} I(X_{ti}; X_{t-\ell,j}|Q)$ in bits per unit area computed (as in [1]) from a 2 hour random selection of the Switchboard continuous-speech database where X_{ti} is the i^{th} element of the random vector X_t and ℓ is time-lag. Feature channels consist of cube root-compressed sub-band envelopes (so Δf is frequency difference) and Q represents decision-tree clustered triphones.[4] As can be seen, additional information is on average distributed throughout the acoustic context. Similar results have been found both for different labeling schemes (monophones and syllables) and feature sets (MFCCs, LPC and RASTA-PLP coefficients).

To increase tractability, dependencies are considered and added individually for each feature element. Define the context of X_{ti} 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:[5]

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

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

for $r \in C_q$ where C_q is the set of state values that could lead to a confusion with state q. Using this notation, $I(X_{ti}; Z^i_{k_{1:N}}|Q_t = q) = I_{\{Q_t=q\}}(X_{ti}; Z^i_{k_{1:N}}|Q_t = q)$. 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 under the probability distribution for the confusable context r rather than the original context q. When $r \neq q$, it represents the situation in a classification task during evaluation of a model in an incorrect context.

The dependency selection algorithm is therefore as follows: for each q and i, choose the size N_q set of variables $Z^i_{k_{1:N_q}}$ for which $I(X_{ti}; Z^i_{k_{1:N_q}}|Q_t = q)$ is large and $I_{\{Q_t=r\}}(X_{ti}; Z^i_{k_{1:N_q}}|Q_t = q)$ is small for each $r \in C_q$.

This approach is distinct from previous work [3, 8, 4, 9] in that the dependency structure may be sparse and may change for each value of Q_t rather than depending on an additional, fixed, and arbitrarily chosen sets of observations. And rather than depending on a location in a segment trajectory [5], the dependencies are data-derived; using conditional mutual information, the dependencies are chosen to provide new and discriminative information about X_t not already provided by the current value of Q_t. This potentially leads to a more accurate statistical model without a large free-parameter increase. 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.

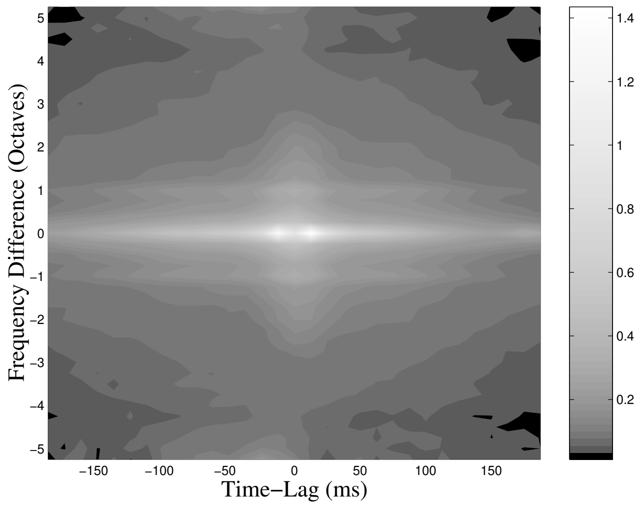

Figure 4: The conditional mutual information density of a randomly selected 2-hour section of the Switchboard continuous-speech database (in bits per unit area).

[2] The notation $X_{1:N}$ represents the set $\{X_1, \ldots, X_N\}$.

[3] This might sound like an egregious mistake but it is actually quite common and can be beneficial in practice, e.g., delta features, hybrid ANN/HMM systems[2], etc. The theoretical problems could potentially be eliminated if each probability distribution is considered a potential function (as in a Markov Random Field) and if appropriate normalization terms are used for each HMM. Such issues are not addressed further in this work.

[4] Thanks to Katrin Kirchhoff for these labels.

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

3. HEURISTIC DEPENDENCY SELECTION ALGORITHM

The general algorithm presented in the previous section involves the computation of mutual information between vectors evaluated under different probabilistic contexts. Computing of such quantities directly would involve more data and/or computation time than is typically available. In this section, a tractable heuristic algorithm is developed for selecting a good set of dependencies $\mathbf{Z}_{qi}$ for each q and feature position i.

To avoid potentially computing $\sum_q |C_q|$ values for each candidate dependency set, the quantity $I_{\{Q_t=r\}}(X_{ti}; Z^i_{k_{1:N}} | Q_t = q)$ is approximated using $I(X_{ti}; Z^i_{k_{1:N}} | Q_t = r)$, a reasonable guess at an upper bound. The difference between the two quantities is:

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

where $D(p_1||p_2)$ is the relative-entropy between distributions p_1 and p_2. While there is no guarantee that this difference is non-negative, intuitively it can be argued that additionally conditioning on $Z^i_{k_{1:N}}$ is not likely to decrease the relative-entropy between $p(X_{ti}|Q_t = r)$ and $p(X_{ti}|Q_t = q)$. This is because, for $r \in C_q$, the quantity $D(p(X_{ti}|Q_t = r)||p(X_{ti}|Q_t = q))$ is already small. And $Z^i_{k_{1:N}}$ is chosen in a sense to highlight rather than suppress differences between the distribution of X_{ti} given $Q_t = q$ and given $Q_t = r$. It is unlikely such a chosen $Z^i_{k_{1:N}}$ will cause a further decrease in relative-entropy, even if selected using a different probabilistic model as above. Therefore, the following relation is assumed typical for $r \in C_q$.

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

Using a liberal estimate for C_q (i.e., $\hat{C}_q \supseteq C_q$ as an estimate of C_q), results in a stronger constraint on the chosen $Z^i_{k_{1:N}}$. A liberal $\hat{C}_q$ potentially eliminates some useful dependencies, but any remaining dependencies will still be informative and discriminative for the confusable classes. C_q can therefore be approximated with a larger set, perhaps even the entire set of states (sans q).

A second difficulty stems from evaluating mutual information between vectors rather than scalars. The chain rule of mutual information says $I(X_{ti}; Z_{1:N}|Q) = \sum_j I(X_{ti}; Z_j | Z_{1:(j-1)}, Q)$. This can be approximated by first finding Z_1 so that $I(X_{ti}; Z_1|Q)$ is large, Z_2 so that $I(X_{ti}; Z_2|Z_1, Q)$ is large, and so on. Because of this approximation, earlier dependency selections can affect later ones. Each of the Z variables are therefore considered in order of decreasing utility choosing the most informative and discriminative variables first. Using an argument similar to the previous paragraph, utility is defined as $U_{ti}(Z_j) = I(X_{ti}; Z_j|Z_{1:(j-1)}, Q = q) - I(X_{ti}; Z_j|Z_{1:(j-1)}, Q \in C_q)$ where $I(X_{ti}; Z_j|Z_{1:(j-1)}, Q \in C_q) = \sum_{q \in C_q} I(X_{ti}; Z_j|Z_{1:(j-1)}, Q = q)p(Q = q)/\gamma$ with $\gamma = \sum_{q \in C_q} p(Q = q)$. The remaining difficulty is the evaluation of $I(X_{ti}; Z_j|Z_{1:(j-1)}, Q)$ which captures the notion that a variable should not be added to $\mathbf{Z}_{qi}$ if it contains only redundant information already provided by previously added variables (i.e., no variable in $\mathbf{Z}_{qi}$ should have a Markov blanket in $\mathbf{Z}_{qi}$ shielding it from X_{ti} given Q). To approximate this quality, $I(X_{ti}; Z_j|Z_{1:(j-1)}, Q)$ is considered large if both $I(X_{ti}; Z_j|Q)$ is large and $I(Z_j; Z_k|Q)$ is small for $k < j$ and is considered small if $I(X_{ti}; Z_j|Q)$ is small.

These approximations lead to the following heuristic dependency selection algorithm for choosing $\mathbf{Z}_{qi}$ 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 $\mathbf{Z}_{qi}$.

τ_u places a lower bound on utility. Criterion 1 ensures that any added dependency provides a significant amount of information (determined by the threshold τ_q) to the current model. Criterion 2 is a redundancy check, and puts an upper bound on the amount of information a dependency variable may have about previously added dependency variables. Criterion 3 places an upper bound τ_c on the prior-weighted cost of this dependency when evaluating the current model in other potentially confusable contexts. It is possible to end up with fewer than N_q (or even zero) dependencies if no satisfying Z exists for the current thresholds. This algorithm requires only the computation of pairwise conditional mutual information for a given labeling scheme.

4. GAUSSIAN-MIXTURE BMMS

In this section, Gaussian mixture HMMs are extended to include the cross-observation dependencies specified by a BMM. The dependencies affect only state specific observation models so modifications involve only Gaussian mixture models.

The observation models should allow their entropy to be affected by the additional dependencies. 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 independent of z given v and q and v is assumed independent of q 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

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

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 are similar to or generalize auto-regressive HMMs [6, 7] ($d = 1$, $M = 1$, $V = 1$), vector-valued auto-regressive HMMs [4, 9, 8][6] ($d > 1$, $M = 1$, $V = 1$), mixture auto-regressive HMMs [3] ($d = 1$, $M > 1$, $V = 1$), and the usual Gaussian mixture models ($d = 1$, $M > 1$, $V = 1$, $s = 0$). 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,

[6] [9] uses discriminative output distributions similar to state-specific LDA and also considers dependencies from future observations.

and hidden-variable dependent as described in Section 2. Furthermore, z is allowed to contain observations from x's past, present, and future.

By introducing an auxiliary function and taking its derivative, it can be shown that 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 $\gamma_{qmv}(t) = p(q_t = q, m_t = m, v_t = v | \mathbf{o}, \mathbf{z})$ and where $\mathbf{o}$ (resp. $\mathbf{z}$) is the set of training vectors (resp. context vectors). $p(v|z)$ does not change between EM iterations, so any (perhaps unsupervised) classification method can be used prior to EM BMM learning. The update equations for the transition probabilities are the same as usual.

5. RESULTS ON AN ISOLATED DIGITS DATABASE

Gaussian-mixture BMMs were tested with $d > 1$, $M = 5$, $V = 1$, and with diagonal covariance matrices on *digits+*, a telephone quality database of isolated digits and control words from Bellcore. The data is represented using 12 MFCCs plus c_0 and includes deltas resulting in a $d = 26$ element feature vector sampled every 10ms. Dependency links were allowed to span a maximum of 70ms (7 frames) on either side of t.

All word error rates (WER) reported are obtained using data from 200 speakers totaling 2600 examples from 4 jackknifed cuts – scores shown are the average of 4 tests in which 150 speakers were used for training and 50 different speakers used for testing. WER is computed using Viterbi probability evaluation.

Num. States	3	4	5	6	7
WER	1.73%	1.34%	1.15%	1.19%	1.19%
Num. Params.	10140	13520	16900	20280	23660

Num. States	8	9	10	11	12
WER	0.89%	1.35%	1.08%	1.08%	1.00%
Num. Params.	27040	30420	33800	37180	40560

Table 1: Results for a HMM with various number of states.

The following procedure is performed independently for each cut and number of states per word. Whole-word strictly left-to-right HMM models bootstrapped using a uniform segmental k-means procedure are created. Full EM training is performed until convergence is achieved and then HMM word error is calculated. Using the HMMs, the Viterbi path is computed for each word determining the state of each frame. Conditional mutual information is computed (as described in [1]) using the resulting labels. The BMM dependency selection algorithm of Section 3 is performed. The BMMs are trained starting with the means and covariances given by the corresponding HMM and with initial dependency link values set to zero. Forced-Viterbi training is performed on the BMMs using the labels derived from the HMM.

Table 1 shows the WER for normal HMMs with varying numbers of states per word along with the corresponding number of observation model parameters. Table 2 shows BMM WER. As can be seen, for a given number of states per word, the BMM error rate is always better than the corresponding HMM WER.

Num. States	3	4	5	6
WER	0.96%	0.85%	0.96%	0.73%
Num. Params.	19157	25511	32070	38521

Table 2: Results for a BMM with various number of states. The dependency selection parameters are $\tau_u = 5 \times 10^{-4}$, $\tau_q = 10^{-3}$, $\tau_g = 75\%$, $\tau_c = 5 \times 10^{-2}$, $N_q = 2$ for all q, and C_q is the set of all states except q.

The average percentage WER decrease[7] from an HMM to a BMM in this case is 34%. The table also shows that a BMM is always better than an HMM even when comparing with an HMM using a comparable number of parameters. The average percentage WER decrease from an HMM to a BMM in this case is 15% (BMMs with 3,4,5, and 6 states are compared with HMMs with 6, 8, 10, and 12 states respectively). The best WER achieved is 0.54% with a BMM using 6 states per word, 61877 parameters, and $N_q = 7$. The same procedure using JRASTA features shows comparable WER results and BMM advantages.

6. CONCLUSIONS

The HMM conditional independence assumption can be relaxed by including additional probabilistic 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. In an isolated word speech database, BMMs show improved performance over comparable HMMs.

The model building scheme presented above can be considered discriminative, but maximum likelihood training is currently being used. A discriminative training scheme such as MCE combined with these discriminatively built models might yield an additional advantage.

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

7. REFERENCES

1. J.A. Bilmes. Maximum mutual information based reduction strategies for cross-correlation based joint distributional modeling. In *Proc. IEEE ICASSP*, Seattle, WA, May 1998.
2. H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1994.
3. B.-H. Juang and L.R. Rabiner. Mixture autoregressive hidden markov models for speech signals. *IEEE Trans. ASSP*, 33(6):1404–1413, December 1985.
4. 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):220–225, February 1990.
5. 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):360–378, September 1996.
6. A.B. Poritz. Linear predictive hidden markov models and the speech signal. In *ICASSP*, pages 1291–1294, 1982.
7. A.B. Poritz. Hidden markov models: A guided tour. In *ICASSP*, pages 7–13, 1988.
8. C.J. Wellekens. Explicit time correlation in hidden markov models for speech recognition. In *ICASSP*, 1987.
9. P.C. Woodland. Hidden markov models using vector linear prediction and discriminative output distributions. In *ICASSP*, pages I–509–512, 1992.

[7] Defined as $avg_k(WER_{hmm_k} - WER_{bmm_k})/WER_{hmm_k}$