STRUCTURALLY DISCRIMINATIVE GRAPHICAL MODELS FOR AUTOMATIC SPEECH RECOGNITION – RESULTS FROM THE 2001 JOHNS HOPKINS SUMMER WORKSHOP

G. Zweig[1]*, J. Bilmes*[2]*, T. Richardson*[2]*, K. Filali*[2]*, K. Livescu*[3]*, P. Xu*[4]*, K. Jackson*[5]*, Y. Brandman*[6]*, E. Sandness*[7]*, E. Holtz*[8]*, J. Torres*[9]*, B. Byrne*[4]

[1]IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598
[3]Massachusetts Institute of Technology, Cambridge MA, 02139
[5]U.S. Department of Defense, Ft. Meade MD, 20755
[7]SpeechWorks, Boston MA, 02111
[9]Stanford University, Stanford CA, 94305

[2]University of Washington, Seattle, WA 98195
[4]Johns Hopkins University, Baltimore MD, 21218
[6]Phonetact Inc., Los Altos CA, 94002
[8]Harvard University, Cambridge MA, 02138

ABSTRACT

In recent years there has been growing interest in discriminative parameter training techniques, resulting from notable improvements in speech recognition performance on tasks ranging in size from digit recognition to Switchboard. Typified by Maximum Mutual Information training, these methods assume a fixed statistical modeling structure, and then optimize only the associated numerical parameters (such as means, variances, and transition matrices). In this paper, we explore the significantly different methodology of discriminative *structure* learning. Here, the fundamental dependency relationships between random variables in a probabilistic model are learned in a discriminative fashion, and are learned separately from the numerical parameters. In order to apply the principles of structural discriminability, we adopt the framework of graphical models, which allows an arbitrary set of variables with arbitrary conditional independence relationships to be modeled at each time frame. We present results using a new graphical modeling toolkit (described in a companion paper) from the recent 2001 Johns Hopkins Summer Workshop. These results indicate that significant gains result from discriminative structural analysis of both conventional MFCC and novel AM-FM features on the Aurora continuous digits task.

1. INTRODUCTION

Discriminative parameter learning techniques are becoming an important part of speech recognition technology, as indicated by recent advances in large vocabulary tasks such as Switchboard [16], which now complement well known improvements in small vocabulary tasks like digit recognition [12]. These techniques are exemplified by the *maximum mutual information* learning technique [1], which specifies a procedure for discriminatively optimizing HMM transition and observation probabilities. These methodologies adopt a fixed pre-specified model structure and optimize only the numeric parameters.

The technique of *structural* discriminability [4, 2, 5] stands in significant contrast to these methods because, in this case, the goal is to learn discriminatively the actual dependency structure between random variables in class-conditional probabilistic models. It is thus both orthogonal and complementary to the methods used for fixed-structure parameter optimization.

At the basis of all pattern classification problems is a set of K classes $C_1, \dots, C_K$, and a representation of each of these classes in terms of a set of T random variables $X_1, \dots, X_T$ (denoted as $X_{1:T}$). For each class, a probabilistic model $P(X_{1:T}|C_k)$ is used to represent the class-conditional distribution over the random variables, and these distributions are used to perform pattern classification. Bayes decision theory states that, modulo a 0/1-loss function, the optimal choice is the class with the highest posterior probability:

$$k^* = \underset{k}{\operatorname{argmax}}\, P(C_k|X_{1:T}) = \underset{k}{\operatorname{argmax}}\, P(X_{1:T}|C_k)P(C_k)$$

The aim of structural discriminability is to identify a minimal set of dependencies in class conditional distributions $P(X_{1:T}|C_k)$ such that there is little or no degradation in classification accuracy relative to the decision rule above. A measure that achieves this goal is described in Section 3.

In this paper, we will focus on class-conditional probabilistic models that can be expressed as Bayesian networks, a type of directed graphical model [13, 11]. A directed graphical model is a graph in which nodes represent random variables, and arcs encode conditional independence assumptions amongst the variables. If we denote the parents of a variable X_i as X_{π_i}, and a specific value of X_i as x_i, and specific values for its parents by x_{π_i}, then the joint distribution can be factored as

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_i P(X_i = x_i|X_{\pi_i} = x_{\pi_i})$$

In this work, we adopted the Graphical Models Toolkit (GMTK) - a newly developed open source toolkit described in detail in a companion paper [3]. The benefits of this framework include the ability to rapidly and easily express a wide variety of models, and use them in as efficient a way as possible for a given model structure.

In the remainder of this paper, we will present the work done at the 2001 Johns Hopkins Summer workshop, where we applied discriminative graphical models to the Aurora connected digits task, and demonstrated significant improvements from structure learning, both with standard MFCCs and with novel AM-FM features. In Section 2 we review the graphical model structures appropriate for speech recognition, followed in Section 3 by a summary of our structure learning algorithms. In Section 4, we present our experimental results, and conclude in Section 5.

2. BASE GRAPHICAL MODEL STRUCTURES

The overall goal of our project was to begin with HMM-equivalent graphical models, and then extend them in structurally discriminative ways; in this section, we briefly present the basic training and decoding structures that we used to emulate an HMM. Since the equivalence between certain graphical model structures and basic HMMs has been discussed previously [14], as have the methods that can be used to build a full speech recognition system [17, 18, 19], we present only a sketch.

The key to creating a graphical model that explicitly emulates an HMM is to create state and transition variables in each time frame, whose values refer to states and transitions in an underlying finite state HMM graph. Equivalence is proved by setting the

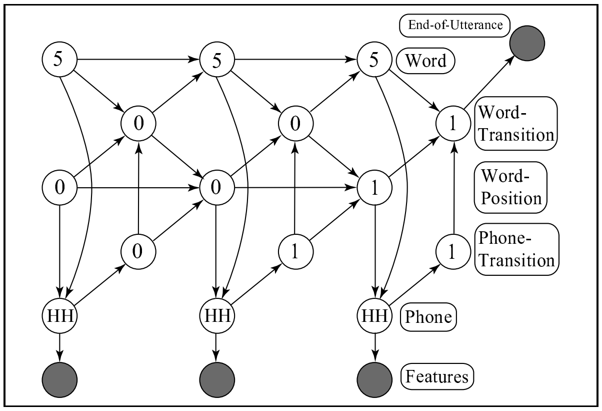

Fig. 1. Decoding network unrolled for 3 frames. Observed variables are shaded.

conditional probabilities in the graphical model so that each assignment of values to its variables corresponds to a path through the HMM graph, and has a probability equal to that of the HMM path (details may be found in [17]).

Figure 1 shows the graphical model structure used for decoding on the Aurora task. The variables in this network are:

1. Word: indicates which word is being spoken.
2. Word-position: within-word position.
3. Phone: acoustic state corresponding to word-position.
4. Phone-transition: 1 when phone transition occurs.
5. Word-transition. 1 when word ends.
6. Acoustics. The feature vector for a frame.
7. End-of-utterance. Has an assigned value of 1, and a probability distribution that assigns to that value 0 probability unless the final word-transition variable has value 1 (meaning, there is a transition out of the final state of a word).

The variables in Figure 1 have been assigned values corresponding to an occurrence of the word “hi”. In this case, the value of 5 for the word means that “hi” is the fifth word in the vocabulary. The word position sequences through 0 (corresponding to /HH/) and 1 (corresponding to /AY/). The word transition variable is 0 except in the last frame, when a phone transition in the last position of the word forces a value of 1. It should be noted that in decoding, the single likeliest assignment of variable values is found; Figure 1 shows just one of many possible. Details of our training structure will be described in forthcoming workshop documentation.

This basic graphical model structure was used as a skeleton onto which additional edges between observations were added discriminatively (see Figure 2). This corresponds to expanding the graph over observation feature vectors from Figure 1, and then adding edges between those variables.

3. DISCRIMINATIVE STRUCTURE LEARNING ALGORITHMS

In the past, there has been a significant amount of work devoted to the induction of probabilistic models, and much of this work involves identifying model structures that capture the underlying conditional independence relationships of the variables being modeled. Sometimes this is called statistical model selection [6] and more recently learning Bayesian networks [8, 7]. Similarly, there has been work in inducing HMM model topology, e.g. [15], though this is different in spirit as it does not focus on conditional independence relations. Most often, these methods work by attempting to find a model structure that maximizes the probability of some observed data, i.e. selection is performed according to the maximum likelihood principle.

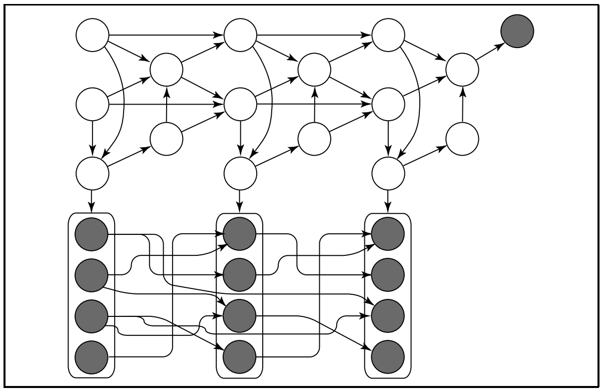

Fig. 2. Additional sparse discriminative structure added over observed variables.

The above approaches are significantly different from the focus of this work. When the task is pattern classification (or ASR), it is no longer the case that the class conditional distribution that maximizes the likelihood of observed data is required. Instead, we seek only discriminative representations that maximize classification accuracy. There are two orthogonal ways that these can be constructed; the first is via discriminative parameter training methods (discussed above), and the second is to have the underlying structure of each class-conditional model represent only that which helps for discrimination [2, 5].

In order to produce a discriminatively structured graphical model, it must be possible to measure the discriminative quality of an edge in a graph. In the work presented herein, we measure the quality of edges between observation variables. For example, suppose that X_{ti} is the i^{th} component of the t^{th} feature vector. In a typical HMM, a hidden variable representing a phone or some sub-phonetic unit is the parent of X_{ti}. Here, however, we consider adding additional between-observation edges, by allowing $X_{\tau j}$ to also be a parent of X_{ti}, where $\tau < t$.

In order to choose such edges discriminatively, we use the EAR (explaining-away residual) measure, an information-theoretic discriminative edge quality measurement first defined in [2, 5]. Assuming that Q is a class random variable, and that we are considering adding edges to all the elements of the random vectors X from all the elements of Y, the EAR measure is defined as follows:

$$\text{EAR}(X, Y) = I(X, Y|Q) - I(X, Y)$$

where $I(X, Y|Q)$ is the conditional and $I(X, Y)$ is the unconditional mutual information between vectors X and Y. It can be shown [5] that choosing edges which optimize the EAR measure is identical to minimizing the KL-divergence between the actual and an approximate class posterior probability distribution. Additional insight into the EAR measure follows from the fact that optimizing it is equivalent to optimizing $I(X, Q|Y)$. This last interpretation implies that a goal of the EAR measure is to choose additional parents of X to increase as much as possible the mutual information between X and Q.

The EAR measure in its most general form is difficult to compute, so in many cases one must instead use only an approximation. In this work, we assume that X and Y are scalars rather than vectors. This means that only the pair-wise quality of edges can be measured, and that the utility of multiple edges are not measured jointly. Single edges are then chosen in a greedy fashion. In other work [5], a form of switching EAR measure approximation was used where a class conditional model was designed for each $Q = q$ using $I(X, Y|Q = q) - I(X, Y)$. In the work reported herein, we designed a single global discriminative structure for all class conditional models using $I(X, Y|Q) - I(X, Y)$, represent-

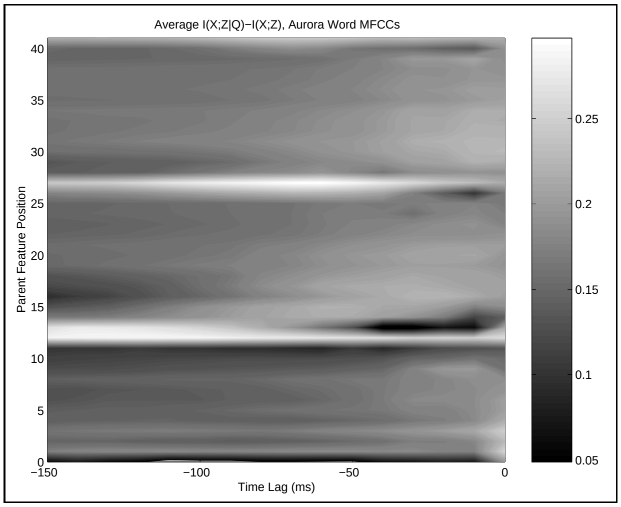

Fig. 3. Discriminative mutual information as a function of parent feature position j and time lag. Features $0, \ldots 12$ are $C_1, \ldots, C_{12}$; next is C_0 and then log-energy. The pattern repeats for deltas and double-deltas. Q ranged over word values.

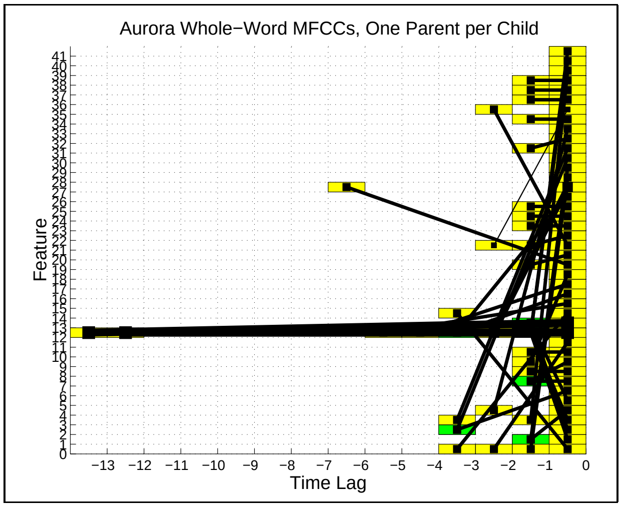

Fig. 4. Induced conditioning relationships for the system of Figure 3. The strength of the EAR measure is indicated by the thickness of the lines. The feature ordering is as in Figure 3, so C_0 is the 13th row. Each feature position was conditioned on one other entry.

ing the average across the possible values of Q. Depending on the desired level of granularity, Q may range over either words, or individual states within words — we present results for both.

4. EXPERIMENTAL RESULTS

4.1. Corpora

Our experimental results focus on the Aurora 2.0 continuous digit recognition task [9]. The Aurora database consists of TIDigits data, which has been additionally passed though telephone channel filters, and subjected to a variety of additive noises. There are eight different noise types ranging from restaurant to train-station noise, and SNRs from -5dB to 20dB. For training, we used the "multi-condition" set of 8440 utterances that reflect the variety of noise conditions. We present aggregate results for test sets A,B, and C, which total about 70,000 test sentences [9].

We processed the Aurora data in two significantly different ways. In the first, we used the standard front-end provided with the database to produce MFCCs, including log-energy and C_0. We then appended delta and double-delta features and performed cepstral mean subtraction, to form a 42 dimensional feature vector. In the second approach, we computed AM (Amplitude Modulation) and FM (Frequency Modulation) features [1]. These are computed by dividing the spectrum into 20 equally spaced bands using multiple complex quadrature band pass filters. For each neighboring pair of filters, the higher-band filter output is multiplied by the conjugate of the lower-band output. The result is low-pass filtered and sampled every 10ms. The FM features are the sine of the angle of the sampled output, and the AM feature is the log of the real component. Although we expect that these features could be improved by further processing (e.g. cosine transform, mean subtraction, derivative-concatenation) we used the raw features to provide the maximum contrast with MFCCs.

4.2. Mutual Information Measures

The first step of our analysis was a computation of the discriminative mutual information between all possible pairs of conditioning variables. Although we could compute this for hidden variables as well as observations, for expediency and simplicity we focused on conditioning between observation components alone. Thus, the structures we present later are essentially expanded views of conditioning relationships among the individual entries of the acoustic feature vectors (see Figure 2).

Consider the mutual information between observation components $X_{\tau j}$ and X_{ti}. We visualize the EAR measure by plotting it as a function of either i or j, and the lag $t - \tau$. In Figure 3, we present discriminative mutual information as a function of j and the time lag, for a system based on whole-word models. Interestingly, Figure 3 shows that discriminative mutual information is strongest when C_0 or log-energy are the parent features, and this is true on average for all children i (note that the features are ordered so that these appear after $C_1 - C_{12}$, i.e. in the middle of the plot). Moreover, information is strongest at a syllable-length lag of 100-150ms. The deltas of these quantities are also highly informative.

4.3. Induced Structures

Using the method of Section 3, we induced conditioning relationships using both MFCCs and AM-FM features. In Figure 4, we show the induced structure for an MFCC system based on whole-word models, and using Q-values corresponding to words in the EAR measure. As expected, there is conditioning between C_0 and its value more than 100 ms previously.

In a second set of experiments, we used the AM-FM features as possible conditioning parents for the MFCCs; the induced conditioning relationships are shown in Figure 5. The first 42 features are the MFCCs; these are followed by AM features, and finally the FM features. This graph indicates that FM features provide significant discriminative information about the MFCCs.

4.4. Word Error Rate Results

To validate our structure-learning methods, we built baseline systems (with GMTK emulating an HMM), and then enhanced them with discriminative structure. Although we built both whole-word and shared-phone systems, we describe only our best results here, which were from the whole-word system. Our results with phone-based systems were qualitatively identical.

[1] We thank Y. Brandman of Phonetact, Inc. for providing this technology.

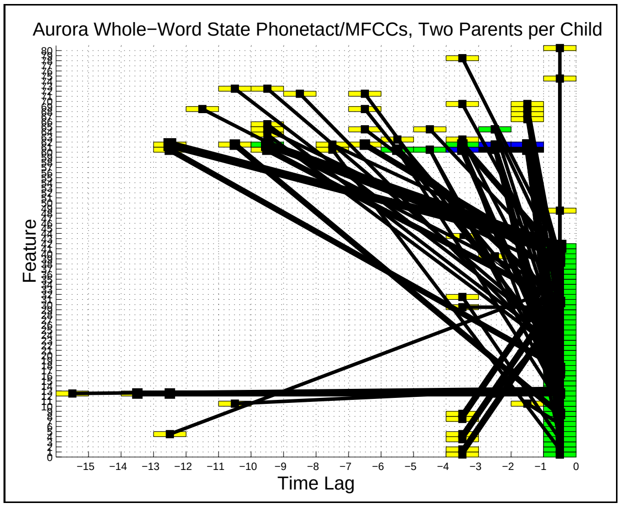

Fig. 5. Induced conditioning relationships for AM-FM features. Q ranged over word-state values. MFCC features are at the bottom, followed by AM, and then FM features. Each MFCC was conditioned on up to two parents.

	clean	20	15	10	5	0	-5
GMTK	99.2	98.5	97.8	96.0	89.2	66.4	21.5
HP	98.5	97.3	96.2	93.6	85.0	57.6	24.0

Table 1. Word recognition rates our baseline GMTK system as a function of SNR. HP is reproduced from [9].

The baseline whole-word system is quite similar to that specified in [9]: each of the 11 vocabulary words has 16 states, with no parameter tying between states. Additionally, silence and short-pause models were used, with three silence states and the middle state tied to short-pause. All models were strictly left-to-right, and used 4 Gaussians per state for a total of 715 Gaussians.

In Table 1 we show the absolute recognition rates for our baseline system as a function of SNR, averaged across all test conditions. Also presented is the published baseline result [9] with a system that had somewhat fewer (546) Gaussians; we see that GMTK performs competitively when it emulates an HMM. (We used 4 Gaussians per state rather than 3 because we used a splitting process that doubles the number after each split.)

Table 2 presents the relative improvement in word-error-rate for several structure-induced systems. There are several things to note. The first is that significant improvements were obtained in all cases. The second is that structure induction successfully identified the synergistic information present in the AM-FM features, and resulted in a significant improvement over raw MFCCs. The final point is that when we increased the size of a conventional system to the same number of parameters, performance was much worse in high noise conditions. Thus, structure induction improves performance in a robust way.

	clean	20	15	10	5	0	-5
WWS	16.3	19.3	14.2	10.5	9.85	19.0	12.6
AMFM	10.4	9.73	6.91	4.29	7.05	17.4	15.5
WW	7.16	7.02	5.51	5.93	5.05	16.0	15.0
EP	18.9	6.56	14.7	10.7	7.16	5.09	1.20

Table 2. Percent word-error-rate improvement for structure-induced systems. WWS is a system where Q ranges over states; AMFM conditions MFCCs on AM-FM features; In WW, Q ranges over words; and EP is a straight Gaussian system with twice as many Gaussians as the baseline. For the WW and WWS systems, one parent per feature was used; in the AMFM case, two parents. EP has the same number of parameters as WW and WWS.

5. CONCLUSION

In this paper we described the results of the 2001 Johns Hopkins CLSP workshop. Using a newly developed graphical models toolkit, GMTK [3], we implemented and tested a variety of structurally discriminative graphical models. We found significant improvements of 10-15% on the Aurora 2.0 recognition task, using both MFCCs and novel AM-FM features. We expect that discriminative *structure learning* techniques will be a good complement to traditional discriminative *parameter learning* methods.

This work was partially supported by NSF Grant IIS-0097467 and DARPA contracts N66001-99-2-8916 and N66001-99-2-892403.

6. REFERENCES

[1] L.R. Bahl, P.F. Brown, P.V. de Souza, and R.L. Mercer. Maximum mutual information estimation of HMM parameters for speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 49–52, Tokyo, Japan, December 1986.

[2] J. Bilmes. *Natural Statistical Models for Automatic Speech Recognition.* PhD thesis, U.C. Berkeley, Dept. of EECS, CS Division, 1999.

[3] J. Bilmes and G. Zweig. The graphical models toolkit: An open source software system for speech and time-series processing. *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 2002.

[4] J.A. Bilmes. Buried Markov models for speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, March 1999.

[5] J.A. Bilmes. Dynamic Bayesian Multinets. In *Proceedings of the 16th conf. on Uncertainty in Artificial Intelligence.* Morgan Kaufmann, 2000.

[6] K.P. Burnham and D.R. Anderson. *Model Selection and Inference : A Practical Information-Theoretic Approach.* Springer-Verlag, 1998.

[7] N. Friedman. The Bayesian structural EM algorithm. *14th Conf. on Uncertainty in Artificial Intelligence*, 1998.

[8] D. Heckerman, D. Geiger, and D.M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. Technical Report MSR-TR-94-09, Microsoft, 1994.

[9] H. G. Hirsch and D. Pearce. The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions. *ICSA ITRW ASR2000*, September 2000.

[10] K. G. Jvreskog. *Structural equation models in the social sciences*, chapter A general method for estimating a linear structural equation system. Seminar Press/Harcourt Brace, 1973.

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

[12] Y. Normandin. An improved mmie training algorithm for speaker indepedendent, small vocabulary, continuous speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 1991.

[13] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 2nd printing edition, 1988.

[14] P. Smyth, D. Heckerman, and M.I. Jordan. Probabilistic independence networks for hidden Markov probability models. Technical Report A.I. Memo No. 1565, C.B.C.L. Memo No. 132, MIT AI Lab and CBCL, 1996.

[15] A. Stolcke and S. Omohundro. Best-first model merging for hidden model induction. In *NIPS94*, 1994.

[16] P.C. Woodland and D. Povey. Large scale discriminative training for speech recognition. In *ICSA ITRW ASR2000*, 2000.

[17] G. Zweig. *Speech Recognition with Dynamic Bayesian Networks.* PhD thesis, U.C. Berkeley, 1998.

[18] G. Zweig and S. Russell. Speech recognition with dynamic Bayesian networks. *AAAI-98*, 1998.

[19] G. Zweig and S. Russell. Probabilistic modeling with bayesian networks for automatic speech recognition. *Australian Journal of Intelligent Information Processing*, 5(4):253–260, 1999.