THE 2001 GMTK-BASED SPINE ASR SYSTEM

Özgür Çetin, Harriet J. Nock, Katrin Kirchhoff, Jeff Bilmes, Mari Ostendorf

University of Washington - SSLI Lab
Department of Electrical Engineering
Box 352500 Seattle, WA 98195-2500

ABSTRACT

This paper provides a detailed description of the University of Washington automatic speech recognition (ASR) system for the 2001 DARPA SPeech In Noisy Environments (SPINE) task. Our system makes heavy use of the graphical modeling toolkit (GMTK), a general purpose graphical modeling-based ASR system that allows arbitrary parameter tying, flexible deterministic and stochastic dependencies between variables, and a generalized maximum likelihood parameter estimation algorithm. In our SPINE system, GMTK was used for acoustic model training whereas feature extraction, speaker adaptation, and first-pass decoding were performed by HTK. Our integrated GMTK/HTK system demonstrates the relative merits provided by each tool. Novel aspects of our SPINE system include the capturing of correlations among feature vectors via a globally-shared factored sparse inverse covariance matrix and generalized EM training.

1. INTRODUCTION

The performance of state-of-the-art speech recognition systems in carefully controlled acoustic environments has become more and more impressive over time. When the test conditions are noisy, however, or when training and test conditions are mismatched, even the best systems show a drastic decrease in performance. Human speech recognition performance, however, degrades only mildly under these conditions [10]. Additive background noise and convolutional channel distortion are two of the acoustic conditions that often have crippling effects on system performance. A number of techniques have been developed to improve the noise robustness of ASR systems [6] and a number of standard databases have been created in order to evaluate these techniques. The DARPA SPINE task is one such database that provides an excellent environment for testing recognition systems under various real-world noise and channel conditions. SPINE speech data is recorded in controlled conditions where two speakers play a battleship-type game. Speakers are placed in separate sound booths, and previously recorded background noise is produced while recording the speaker. Background noise affects speech recognition performance not only because of the contamination of the speech signal, but also because speakers adjust their articulation to compensate for the environmental disturbance, a phenomenon known as the Lombard effect [7]. Unlike some speech databases which simulate noisy conditions by adding noise to speech recorded in a quiet environment, SPINE does include some amount of Lombard speech.

Previous approaches to improving noise robustness include standard feature processing techniques such as cepstral mean subtraction and variance normalization, which are an attempt to compensate for what are typically minimal differences in channel effects, owing for example to different microphones or room acoustics. Another example is maximum likelihood linear regression (MLLR) [9], an adaptation methodology that can account for and adapt to adverse acoustic conditions to some degree. Ideally, however, models should be used that can truly distinguish the properties of speech from the properties of noise without making any assumptions about either speech or noise. We believe that these models require richer and more sophisticated structures than the simple hidden Markov model (HMM) structures commonly used. A *graphical model* [8] is a natural extension to HMMs in that it allows a much larger set of variables and dependencies between those variables to be incorporated. We have recently developed a toolkit for the development of graphical model-based ASR systems (GMTK) [3], which facilitates the exploration of this research direction. This paper describes our GMTK-based DARPA 2001 SPINE system. One goal of this work was to begin to demonstrate the feasibility of building a large state-of-the-art graphical model-based speech recognition system. The second goal was to explore a novel technique for noise robustness based on modeling the correlation between speech feature vectors.

The organization of the paper is as follows: In Section 2, we provide a brief overview of GMTK, followed in Section 3 by a discussion of the covariance models used in our system. In Section 4 we describe our baseline system that uses diagonal covariance matrices, and in Section 5, we describe a system that uses globally tied sparse inverse covariance matrices. Lastly, in Section 6, we provide experimental results on the SPINE evaluations, and discuss future work in Section 7.

2. GMTK OVERVIEW

GMTK is an open source, (soon to be) publicly available software package for developing graphical-model based speech recognition systems and general time-series applications [3]. Graphical models are graphs where nodes represent random variables, and edges represent sets of *conditional* independence relationships between these variables. These representations provide an explicit and visual account of the underlying assumptions associated with a statistical model. Inference in graphical models can be done efficiently by taking advantage of probabilistic factorizations that occur according to separation properties in a graph.

Graphical representations are useful because they provide a common framework for rapidly comparing different models and their underlying modeling assumptions. Thus, similarities between seemingly unrelated models can be made more explicit. For example, the graphs for a Kalman filter and an HMM are identical, differing only in the discrete vs. continuous representation of random variables. Inference in Kalman filters and HMMs is also “identical”, as seen in the more abstract space of operations associated with graphical inference.

Many time-series prediction, regression, and classification models can be represented with graphical models. Furthermore, many existing ASR techniques are subsumed by graphical models and their inference algorithms. This includes the forward-backward algorithm used in HMM training, which is a special case of the junction tree algorithm of graphical models inference. Multivariate Gaussians themselves may also be seen as either directed or undirected graphical models [2]. Nevertheless, graphical models have only recently begun to be used for ASR [12, 2].

The probabilistic semantics used in the current version of

GMTK are dynamic switching Bayesian networks, a type of directed acyclic graphical model that specifies the joint probability distribution over random variables in factored form — each factor represents the conditional probability of a variable given its parents according to the graph. In GMTK, the conditional probability tables can have dense, sparse, or deterministic forms. For continuous nodes with discrete and continuous parents, Gaussian mixture distributions with linear conditional dependencies on continuous parents can be used. In addition to the usual diagonal and full covariance matrices, banded and factored sparse inverse covariance matrices [4] (see Section 3) provide the advantages of full covariance matrices while alleviating the associated storage and computational costs. Moreover, *switching* parents [2] can be used to change a variable's parents conditioned on other values of parents. GMTK allows extensive parameter tying between similar kinds of model parameters, i.e. tying of probability mass functions, Gaussian means, diagonal covariance matrices, and full covariance matrix factorizations. Such parameter-tying can be useful for large models that otherwise cannot be estimated reliably using only a limited amount of training data. Parameter estimation in GMTK can be done using standard Expectation-Maximization (EM) for maximum likelihood parameter estimation [5], and also a generalized EM (GEM) algorithm when arbitrary parameter tying is used [1]. In addition to the standard mixture *splitting* algorithm to increase the number of components in a Gaussian mixture, GMTK also allows *vanishing* of improbable mixture components.

3. SPARSE INVERSE COVARIANCE MATRICES

The usual choice of distribution for state conditional observation model of phone HMMs is a mixture of multivariate Gaussians, i.e.

$$p(Y_t|Q_t = i) = \sum_{l=1}^{M_i} \alpha_{il} \mathcal{N}(Y_t; \mu_{il}, \Sigma_{il}) \tag{1}$$

where Y_t and Q_t refer to, respectively, observation and state of HMM at time t, M_i is the number of Gaussian components associated with state i, $\mathcal{N}(Y; \mu, \Sigma)$ denotes a multivariate Gaussian distribution with mean vector μ and covariance matrix Σ, and α_{il}, μ_{il} and Σ_{il} are the weight, mean vector and covariance matrix of the lth Gaussian component associated with state i. Because of the high cost of full-covariance Gaussians, covariance matrices are often diagonal in speech recognition systems. This is especially true in systems with large numbers of Gaussian components.[1] Viewing Gaussians as directed graphical models [2] and encoding only the conditional independence relations useful for discrimination can alleviate costs associated with computation, storage, and estimation without sacrificing the advantages associated with full covariance modeling.

The directed graphical model representation of a multivariate Gaussian is obtained through reformulating the exponent in the Gaussian probability density function [8]

$$f(X = x) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\} \tag{2}$$

as

$$f(x) = |2\pi\Lambda|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x - Bx - \tilde{\mu})^T \Lambda^{-1} (x - Bx - \tilde{\mu})\right\} \tag{3}$$

where $\Sigma \equiv U^T \Lambda U$ is the Cholesky decomposition with diagonal Λ and unit upper-triangular U, $B \equiv I - U$ where I is the identity matrix and $\tilde{\mu} \equiv U\mu$. Equations 2 and 3 are equivalent representations of a Gaussian. Further insight can be obtained by factoring the exponent, and making use of the fact that B is a strictly upper triangular matrix with zeros along the diagonal:

$$f(X = x) \quad = \quad \prod_{k=1}^{d} (2\pi\tilde{\sigma}_k^2)^{-\frac{1}{2}} \exp\left\{-\frac{(x_k - b_k^T z_k - \tilde{\mu}_k)^2}{2\tilde{\sigma}_k^2}\right\}$$

where $\tilde{\sigma}_k^2 \equiv [\Lambda]_{ii}$, $b_k \equiv [B_{k,k+1} \cdots B_{kd}]^T$ is the kth row of B in the upper triangular portion (note that $[B]_{ki} \equiv 0$ for $i \leq k$ by construction) of B and $z_k \equiv [x_{k+1} \cdots x_d]^T$ are the parents of child x_k where x_k refers to k-th element of vector x and $[A]_{ij}$ refers (i, j)-th element of matrix A. In this form it can be seen that the conditional distribution of x_k given $\{x_{k+1,\ldots,d}\}$ is another Gaussian with conditional mean $b_k^T z_k + \tilde{\mu}_k$ and conditional variance $\tilde{\sigma}_k^2$. Further details may be found in [4, 2].

The zeroes in the upper triangular portion of B encodes the conditional independence relationships among the elements of X. If $[B]_{ki} = 0$ for $i > k$, then x_k is conditionally independent of x_i given $\{x_{k+1}, \ldots, x_d\} \backslash \{x_i\}$. Note that even if B is sparse, i.e. it has only a few non-zero elements, the corresponding Σ could still be a full matrix. This fact can be put to good use in Equation 1 to replace diagonal covariance matrices with full covariance matrices with sparse inverses [4].

It is straightforward to derive closed form parameter update equations for conditional means, $\tilde{\mu}_k$, and variances, $\tilde{\sigma}_k^2$, and b_k for learning with hidden variables using the EM algorithm. Moreover, it is possible to share the same B matrix globally across a variety of Gaussian components in the system. This can further lessen the storage requirements and estimation problems, but the EM update equations are unfortunately no longer analytical. In this case a GEM training procedure [1] can be used, as implemented in GMTK and used for training the globally-shared sparse B system below (Section 5).

4. BASELINE SYSTEM

In the baseline system we use GMTK for acoustic model training. HTK is used for feature extraction, decision tree clustering, adaptation, and N-best list generation. The features are per side mean-subtracted and variance-normalized 12 mel-frequency cepstral coefficients (MFCCs) plus logarithm of energy with their associated first- and second-order differences.[2]

First, a monophone system with 45 distinct phones is trained using GMTK. Each phone is modeled by a 3-state HMM with left-to-right topology and single Gaussian component output distributions and diagonal covariance matrices. In addition to between-word silence phones, single state beginning and end of utterance silence models are used. At the start of monophone training, all Gaussian means and variances are set to 0 and 10, respectively, while all state transition matrices are set to be uniform, allowing only left-to-right state transitions without any skips. The transition matrices of the HMMs, Gaussian means and covariances are trained with GMTK for 8 EM iterations, which are then loaded into HTK for decision tree clustering of triphone states. The decision tree state-clustered triphone HMM system with 2743 distinct states is then re-trained using GMTK, using a splitting/vanishing scheme for 30 EM iterations to obtain roughly 12 Gaussian mixtures per state.

For recognition, the GMTK trained parameters are transferred back to HTK to perform adaptation and N-best list generation. After the segmentation of raw acoustic waveforms (described in Section 6.1) a 1-best decoding is performed with a bigram language model (LM), which is then used as the basis for the two-class MLLR adaptation of the phone and silence models. The 1-best decoding and MLLR adaptation steps are repeated and the

[1] There are approximately 30,000 Gaussian components in even a moderate-sized SPINE system, e.g. the one in Section 4.

[2] We thank Bill Byrne for his per side mean-subtraction and variance-normalization code.

resulting acoustic models are used to obtain a 200-best list which is rescored with a trigram language model and final adapted acoustic models. A diagram showing the various steps of training and recognition is depicted in Figure 1 for this system — also shown is a second system that uses a shared sparse inverse covariance matrix, described next. The experimental details and word error rates are reported in Section 6.

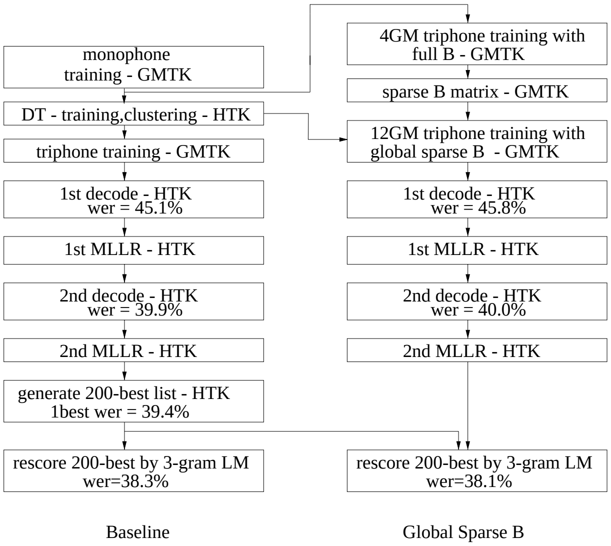
monophone training - GMTK
DT - training,clustering - HTK
triphone training - GMTK
1st decode - HTK wer = 45.1%
1st MLLR - HTK
2nd decode - HTK wer = 39.9%
2nd MLLR - HTK
generate 200-best list - HTK 1best wer = 39.4%
rescore 200-best by 3-gram LM wer=38.3%
4GM triphone training with full B - GMTK
sparse B matrix - GMTK
12GM triphone training with global sparse B - GMTK
1st decode - HTK wer = 45.8%
1st MLLR - HTK
2nd decode - HTK wer = 40.0%
2nd MLLR - HTK
rescore 200-best by 3-gram LM wer=38.1%
Baseline
Global Sparse B

Fig. 1. The training and testing of baseline and globally-shared sparse B matrix systems.

5. GLOBALLY-SHARED SPARSE INVERSE COVARIANCE SYSTEM

Instead of just diagonal covariance matrices, our second system uses a single globally shared sparse B matrix to mimic full covariances with virtually no increase in the number of parameters. Note that while a single B has been tied globally among the Gaussian components, the conditional means, $\tilde{\mu}_k$'s, and variances, $\tilde{\sigma}_k$'s, are still unique to each mixture and state. Hence, each Gaussian component has a distinct full covariance matrix.

The decision tree used for state clustering for the baseline system of Section 4 is re-used here. Initially, all conditional Gaussian means and variances are set to 0 and 10 with a globally shared strictly upper triangular full B matrix whose elements are all set to 0. After training a 4 component Gaussian mixture per state system, we "sparsified" the B matrix by keeping only those entries whose absolute values were greater than a threshold (0.01 in our case). The dense B matrix and the structure of the sparse B matrix is depicted in Figures 2 and 3 respectively.

Assessing the importance of a B matrix "link" between a parent and a child via the absolute value of the corresponding B element is justifiable for the per-side mean subtracted and variance normalized features since each element of the feature vectors has approximately the same dynamic range. The structure of the sparse B matrix in Figure 3 reveals that MFCCs depend on themselves plus their second-order differences, while first-order and second-order differences of coefficients only depend on themselves. This is expected given that the linear regression formulas used for first-order differences do not include the current feature vector whereas the second-order regression formula does. Using this sparse B matrix, further Gaussian splitting yielded our final 12 Gaussian component per mixture system. At all stages of training, only the elements of the globally-shared B matrix corresponding to the structure depicted in Figure 3 are trained, all other elements are fixed at 0.

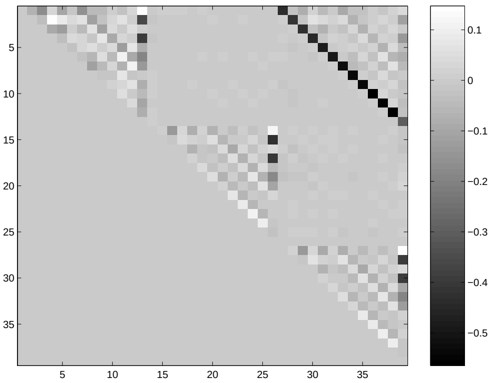

Fig. 2. The full non-sparse B matrix of 4 Gaussian mixture per state system. The features are ordered as 12 MFCCs and log energy, along with their first-order and second-order differences. Horizontal axis should be read as parents while the vertical axis as children.

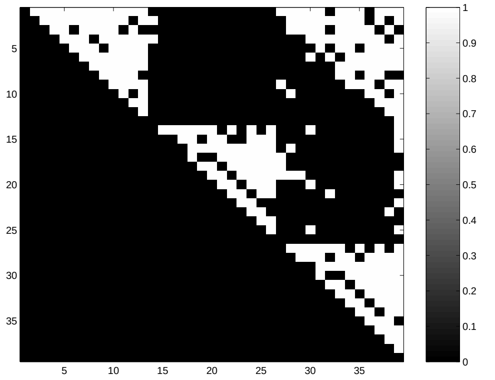

Fig. 3. The resulting structure (a binary mask) of the global sparse B matrix.

During recognition two passes of MLLR adaptation and 200-best rescoring with a trigram LM are performed with steps identical to the baseline. The rescored 200-best list is the 200-best list generated by the final adapted acoustic models of the baseline system. The word error rates are reported in Section 6.

6. EXPERIMENTS

All reported results that follow are from systems trained using the official SPINE-1 training and evaluation sets and the SPINE-2 training and development sets, totaling 17 hours of data[3]. The test set is the official SPINE-2 evaluation set, 3.2 hours of data. All recognition experiments use the standard SPINE-2 bigram and trigram language models provided to the participating sites by CMU [11].

6.1. Speech/Non-Speech Segmentations

Given the high computational and memory cost associated with decoding long segments of acoustic waveforms, the incoming acous-

[3] SPINE-1 and SPINE-2 refer to year 2000 and 2001 SPINE evaluations respectively.

Table 1. *WERs of baseline and sparse B systems before and after adaptation. "no adapt." results use no adaptation, and MLLR-1 and MLLR-2 refer to the results obtained after first and second MLLR adaptations with the bigram LM. 200-best refers to the trigram results obtained by rescoring a 200-best list generated using the MLLR-2 acoustic models and the bigram LM.*

Model	no adapt.	MLLR-1	MLLR-2	200-best
baseline	45.1	39.9	39.4	38.3
sparse B	45.8	40.0	39.6	38.1

tic speech signals needed to be segmented to find the "islands" of speech. Otherwise, recognition of the unsegmented waveforms might result in many insertions on the non-speech segments. An energy-only approach, where a segment is hypothesized as speech only if the energy is above some threshold, might not work reliably on the SPINE data due to the variety of background noise types. In our system, we therefore segmented the waveforms using a binary classifier consisting of an ensemble of neural networks. The procedure consists of three steps. First, speech features (MFCCs) are extracted from the unsegmented acoustic waveforms. Second, frame-level "speech" probabilities are obtained from the neural-network ensemble. Lastly, the frame level probabilities are post-processed to determine speech segments satisfying a minimum duration constraint.

The neural-network ensemble consists of two multi-layer perceptrons trained on different subsets of training data. Their outputs are combined by taking the product and renormalizing. The input to each network consists of speech features from a 9-frame window centered at the current frame. The training targets for the networks are the speech/non-speech designations as determined by the SPINE hand transcriptions. Using this criteria, between-word silences might also be treated as "speech" targets. The first neural network is trained using the SPINE-1 training set and the SPINE-2 training and development sets, and uses SPINE-1 evaluation set for cross validation. The second neural network is trained using the SPINE-1 training and evaluation sets and the SPINE-2 training set, and uses the SPINE-2 development set for cross-validation. The number of hidden units are 50 and 75 for the two networks respectively. A frame is classified as "speech" if the ensemble's output for that case is greater than 0.5. The consecutive frame level hypotheses whose length are less than 200 msec are modified to be consistent with the decisions in their context. 200 msec was chosen as the minimum duration length based on the fact that it corresponds roughly to minimum syllable length and was found to be the shortest segment in the training transcriptions identified as speech. On the official SPINE-2 evaluations data, the frame level error rate of the segmenter is 11.2% with 12.9% miss detection rate of "speech" segments as "non-speech" and 6.8% false alarm rate of "non-speech" segments as "speech". All experiments reported below use these segmentations.

6.2. Results

The word error rates (WERs) of the two systems at different stages of adaptation and for different noise conditions are reported in Tables 1 and 2 respectively. All decodings used the bigram language model except the final rescoring of 200-best list with the trigram language model. In Table 1, even though the initial WER of the globally-shared sparse B system (2nd column) is worse than the baseline system, the final WER of this system (5th column) is at least as good as that of the baseline. The WERs of both systems change drastically under different noise conditions, and there appears to be no significant systematic difference between the two systems. The results are competitive with other single (i.e., non-combined) systems on this task. Moreover, these results are quite preliminary, and we plan to further investigate the effect of a global B in future systems, and their utility for system combination.

Table 2. *WERs of baseline and sparse B systems under different noise conditions corresponding to the overall WERs of 38.3% and 38.1% (the last column of Table 1).*

noise	baseline	sparse B
quiet	40.8	41.4
office	35.0	34.9
street	41.0	41.0
car	35.0	34.5
f16	43.1	43.2
carrier	30.4	30.7
bradley	42.2	41.8

7. DISCUSSION

There are natural extensions to this work. First of all, converting the frame-based "speech" vs. "non-speech" decisions to segment level decisions can be made more systematic by using a left-to-right HMM on the frame level "speech" decisions of the ensemble to incorporate minimum duration constraints. Second, our simple threshold-based algorithm for determining the sparse B matrix should be replaced by a more statistically sound criteria such as minimum decrease in likelihood, or the better yet a discrimination measure. Third, instead of using a single globally shared B across all speech units, tying could be done more finely such as at the mixture or phonetic-category level.

This work was supported by NSF Grant 0093430, and by DARPA contract N66001-99-2-892403.

8. REFERENCES

[1] J. Bilmes. Gaussian parameter sharing and training using a GEM algorithm. In Preparation. University of Washington, Dept. of EE.

[2] J. Bilmes. Graphical models and automatic speech recognition. Technical Report UWEETR-2001-005, University of Washington, Dept. of EE, 2001.

[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. Factored sparse inverse covariance matrices. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, 2000.

[5] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society Series B*, 39:185–197, 1977.

[6] Y. Gong. Speech recognition in noisy environments: A survey. *Speech Communications*, 16:261–291, 1995.

[7] H. Lane and B. Tranel. The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research*, 14:677–709, 1971.

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

[9] C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9:171–185, 1995.

[10] R.P. Lippmann. Speech recognition by machines and humans. *Speech Communication*, 1(22):1–16, 1997.

[11] E. Marsh, A. Schmidt-Nielsen, and T. H. Crystal. NRL presentation. 2001 SPINE Workshop, http://elazar.itd.nrl.navy.mil/spine/nrl2/presentation/nrl2001.html.

[12] G. Zweig, J. Bilmes, T. Richardson, K. Filali, K. Livescu, P. Xu, K. Jackson, Y. Brandman, E. Sandness, E. Holtz, J. Torres, and B. Byrne. Structurally discriminative graphical models for automatic speech recognition: Results from the 2001 Johns Hopkins summer workshop. In *Proc. of Intl. Conf. on Acoustics, Speech and Signal Processing*, 2002.