LARGE VOCABULARY CONVERSATIONAL SPEECH RECOGNITION WITH THE EXTENDED MAXIMUM LIKELIHOOD LINEAR TRANSFORMATION (EMLLT) MODEL

Jing Huang, Vaibhava Goel, Ramesh Gopinath, Brian Kingsbury, Peder Olsen & Karthik Visweswariah

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598
{jghg,vgoel,rameshg,bedk,pederao,kv1}@us.ibm.com

ABSTRACT

This paper applies the recently proposed Extended Maximum Likelihood Linear Transformation (EMLLT) model in a Speaker Adaptive Training (SAT) context on the Switchboard database. Adaptation is carried out with maximum likelihood estimation of linear transforms for the means, precisions (inverse covariances) and the feature-space under the EMLLT model. This paper shows the first experimental evidence that significant word-error-rate improvements can be achieved with the EMLLT model (in both VTL and VTL+SAT training contexts) over a state-of-the-art diagonal covariance model in a difficult large-vocabulary conversational speech recognition task. The improvements were of the order of 1% absolute in multiple scenarios.

1. INTRODUCTION

The EMLLT model, proposed by Olsen and Gopinath [1, 2], estimates the inverse covariance (or precision) matrices of all the Gaussians of an HMM-based acoustic model as a linear combination of rank-one symmetric basis matrices. By choice of the number and nature of these rank one basis matrices, the EMLLT model allows one to gradually vary the complexity of the model from diagonal covariances on the one extreme to full-covariances on the other.

The EMLLT model has been shown to yield significant word error rate gains on several speech recognition tasks [1]. However, in all the tasks in that paper the databases used were IBM internal, the baseline acoustic models were of moderate size (ranging from $10K$–$50K$ Gaussians) and the language models were grammar-based with a wide range of perplexities, but with a relatively small vocabulary. Recently Gopinath *et al.* [3] applied the EMLLT model on the SPINE [4] database and demonstrated the usefulness of these models in a VTL+SAT context. While the SPINE task is based on an n-gram language model, the acoustic model is constrained (about 15K Gaussians) by the small size of the training corpus and the language model perplexity is about 30. This paper, in contrast, applies the EMLLT model in both VTL and VTL+SAT contexts on the much larger, more difficult and better understood Switchboard database where our best baseline performance is typically obtained with an n-gram language model of perplexity of about 80 and an acoustic model that has about $150K$ Gaussians.

The Switchboard training corpus comprises 4870 conversation sides with a total duration of 265 hours. Typical Switchboard acoustic models have between three and ten thousand context-dependent sub-phone units and $150K$–$180K$ Gaussian mixture components. Switchboard recognition vocabularies typically contain $30K$–$40K$ baseforms.

Our primary result is that the EMLLT model yields improvements in recognition accuracy over our best diagonal-covariance models on the large-scale Switchboard task as well as on the smaller tasks previously reported.

The rest of this paper is organized as follows. Section 2 briefly reviews the EMLLT modeling framework, and describes how feature-space and mean adaptation is carried out in conjunction with this model. Section 3 presents our experimental setup and comparisons of VTL and VTL+SAT EMLLT models with corresponding baseline VTL and VTL+SAT models.

2. EMLLT MODEL AND ADAPTATION

This section outlines the EMLLT modeling framework and the adaptation of an EMLLT model. It is essentially a review of the work presented in Olsen *et al.* [1] and Gopinath *et al.* [3].

In the EMLLT model the d-dimensional acoustic features (be they original MFCC/PLP or canonical VTL+SAT features) are modeled with Gaussians of a special form. Specifically the precision matrix of the j^{th} Gaussian, say $P_j (= \Sigma_j^{-1})$ is of the form

$$C\Lambda_j C^T, \quad C \in \mathbf{R}^{d\times D}, \Lambda_j \in \mathbf{R}^{D\times D}, \tag{1}$$

Λ_j diagonal and $d \leq D \leq d(d+1)/2$. The parameters of this model are C, Λ_j (and of course the means m_j and priors π_j). These parameters can be estimated in an ML fashion using a generalized EM algorithm [1].

The means of EMLLT model are adapted using an affine transformation, $\hat{\mu}_j^s = A^s \mu_j + b^s$, so as to maximize the likeli-

hood of the observed speaker specific acoustic data. Estimation of A^s and b^s is carried out by maximizing an auxiliary function which is similar to the auxiliary function used in the standard MLLR [5]. This function, Q_{MT}, is given as [3]

$$-1/2 \sum_{j,t} \gamma_j(t)(x_t - A^s\mu_j - b^s)^T C\Lambda_j C^T (x_t - A^s\mu_j - b^s).$$

$Q_{\mathrm{MT}}(A^s, b^s)$ is quadratic and hence strictly concave in the parameters (A^s, b^s). The difference between this case and standard MLLR is that the problem no longer breaks down into independent sub-problems when one considers one row at a time. For large d (say $d = 60$ as in one example in this paper) directly solving this quadratic in $d^2 + d$ variables is very slow. We use a conjugate-gradient algorithm with pre-conditioning to solve this optimization problem [6]. The pre-conditioner used is the matrix of diagonal elements of the positive-definite matrix associated with the quadratic form Q_{MT}. The cost of the algorithm is roughly the cost of computing the gradient once per iteration. A formula for the gradient of Q_{MT} explicitly in terms of the statistics is given in Gopinath *et al.* [3].

For adaptation of the feature-space an affine feature transform, $\hat{x}^s = A^s x + b^s$, is used. The auxiliary function $Q_{\mathrm{FT}}(A^s, b^s)$ in this case is also quite like the auxiliary function used in standard f-MLLR [7]; it is given as

$$-\sum_{j,t} \gamma_j(t)(A^s x_t + b^s - \mu_j)^T C\Lambda_j C^T (A^s x_t + b^s - \mu_j) + 2\beta \log|\det A^s|.$$

Because of the additional $\log|\det A^s|$ term Q_{FT} is not concave in (A^s, b^s). However, considered as a function of one row of A^s (with all other parameters fixed) and searching only in the region where $\det A^s$ has the same sign as the initial point Q_{MT} is a strictly concave function with a unique maximizer which can be found following a technique described in [7]. Once an A^s is fixed, a quadratic optimization is carried out for b^s. A^s and b^s are optimized in an alternating fashion until convergence is achieved.

The precision matrices of the acoustic model are transformed as $P_j^s = A^s P_j (A^s)^T$. The auxiliary function in this case is

$$-\sum_{j,t} \gamma_j(t)(x(t) - \mu_j)^T A^s C\Lambda_j C^T A^{sT} (x(t) - \mu_j) + 2\beta \log|\det A^s|.$$

Formally this optimization problem is equivalent to the optimization problem in the case of feature space transforms. Thus any method good for one will be good for the other. For the precision transforms we explicitly computed the function and the gradient and used conjugate gradient optimization function available in a numerical package.

The three types of transformations can be applied in any order iteratively to better match the EMLLT model to a test speaker's data. Notice that this does not require further recognition passes to get improved scripts. Also it is straightforward to apply these transformations in a class-dependent fashion by appropriately accumulating class dependent statistics.

3. EXPERIMENTS ON SWITCHBOARD

3.1. Experimental Setup

Our system architecture has two parallel branches corresponding to two different acoustic features: perceptual linear prediction (PLP) cepstral coefficients and Mel-frequency cepstral coefficients (MFCC). The 12th-order PLP features were computed from an 18-filter Mel filterbank that covered the 125 Hz - 3.8 kHz frequency range. The 24 MFCC features were computed from a 24-filter Mel filterbank spanning the 0 Hz - 4.0 kHz frequency range. In both cases, the power spectra were smoothed via periodogram averaging, and for the PLP features, the equivalent of 1 bit of noise was added to the power spectra prior to the Mel binning. Every 9 consecutive cepstral frames are spliced together and projected down to 60 dimensions using a discriminant feature space transformation – LDA for PLP [8] and heteroscedastic discriminant analysis (HDA [9]) for MFCC. The range of these transformations is further diagonalized by means of a maximum likelihood linear transform (MLLT [10]). Another difference between the feature streams is that the PLP cepstra are mean and variance normalized on a per side basis (except the energy term which is normalized on a per utterance basis) while the MFCC cepstra are mean normalized on a per utterance basis and does not have variance normalization.

For training we used the data released by LDC, which includes 248 hours of Switchboard and 17 hours of CallHome English. Tests are conducted on the Hub5 2000 evaluation data with 40 sides of Switchboard 1, and Hub5 2001 evaluation data with 40 sides of Switchboard 1, 40 sides of Switchboard 2, and 40 sides of Switchboard 2 cellular.

Recognition in the MFCC branch is performed with IBM's stack-search decoder under a trigram language model while recognition in the PLP branch is performed in a lattice rescoring setup where initial lattices are generated from the word-internal PLP models and a trigram LM. Two sets of acoustic rescoring experiments were performed: one on the initial trigram lattices and second on the lattices rescored with a four-gram language model.

3.2. Baseline System: VTL and VTL+SAT Models

The baseline VTL and VTL+SAT models are built from the following steps. The VTL warp factors are first determined for each training speaker. For this, a first-pass decoding

is performed using speaker independent MFCC models. A forced alignment is then carried out against this output to select the frames corresponding to vowels which are used together with a voicing GMM model to determine the speaker warp factor (out of 21 allowing for 20% scaling in either direction). Both PLP and MFCC features are then warped accordingly. VTL normalized acoustic models are built in both the MFCC and PLP branches. The next step uses these gender independent VTL models to generate statistics for feature-space transform estimation. The acoustic feature vectors of each speaker are then transformed and the VTL+SAT models are trained on the resulting features.

The MFCC SAT model has about three thousand context-dependent HMM states and about $150K$ Gaussians modeling these states. The PLP VTL+SAT model has about ten thousand context-dependent HMM states and about $180K$ Gaussians modeling the states. The number of Gaussians were chosen by BIC [11] when the baseline VTL+SAT acoustic models were built.

3.3. EMLLT System on VTL and VTL+SAT Space

EMLLT models were built on the VTL MFCC, VTL+SAT MFCC and VTL+SAT PLP feature spaces respectively. These models are in one-to-one correspondence with their associated diagonal covariance baseline models. In principle, for training the EMLLT VTL+SAT model each training speaker's SAT transform has to be estimated with an EMLLT model. However, in order to save on computation (and to be sure to have results in time for the paper deadline) we decided instead to use for each speaker the SAT transform estimated for the baseline diagonal covariance VTL+SAT model; at least for the MFCC based models, we were able to experimentally confirm that this choice had no significant effect on the final results as reported in Section 3.5. This also meant that at test time we had to use the baseline MFCC/PLP VTL+SAT model to estimate a speaker specific feature-space transform. All EMLLT adaptation experiments were then carried out on top of this feature-space transform.

The C-matrix in EMLLT was initialized as follows [1]: the HMM states were split into groups based on acoustic-phonetic knowledge, and an MLLT matrix was generated for each group. These matrices were then stacked on top of each other to give final C matrix; we did not carry out a likelihood based re-estimation. Using this C matrix, Λ_j parameters were updated using the generalized EM algorithm.

3.4. Results and Analysis

Table 1 compares the performance of the EMLLT model trained on MFCC features in both VTL and VTL+SAT spaces with the baseline model trained with diagonal Gaussians on the same feature space. Results in Table 1 are on the Hub5 2000 evaluation test set. The EMLLT model had 240 basis elements (i.e., C is a 60×240 matrix) corresponding to MLLT directions of four groups of HMM states. The first group includes vowels and diphthongs, the second group consists of voiced and unvoiced stops, the third group contains fricatives and affricates, and the fourth group includes liquids and glides. All EMLLT results reported in this paper had 240 basis elements. We also performed experiments with 120 and 480 basis elements. As expected the word error rate performance of the model with 240 basis elements is significantly better than that of the model with 120 basis elements. Going from 240 basis elements to 480 basis elements gives a marginal improvement in word error rate at double the number of parameters.

Adaptation is performed in the following steps:
a) Statistics for the initial feature space transform are obtained using scripts from an initial-pass decode of the test data with a speaker-independent VTL-normalized diagonal Gaussian acoustic model.
b) A feature space transform is obtained using the statistics in step a) and the test data is re-decoded with a SAT system
c) We generate statistics for mean transforms using the scripts obtained in step b).
d) We generate single or multiple mean transforms and (optionally) an additional precision transform using statistics from Step c). The threshold used for generating multiple transforms was 50 seconds.
Steps b), c) and d) are carried out separately for the EMLLT system and the baseline system.

	MFCC	
	Diagonal	EMLLT
1	26.8	25.2
2	24.6	23.1
3	24.2	22.6
4	23.6	22.6
5	-	22.3

Table 1. Word error rate comparison of diagonal Gaussian and EMLLT models on Switchboard 2000 eval testset: 1) VTL models; 2) VTL+SAT models; 3) VTL+SAT + 1 mean transform; 4) VTL+SAT + multiple mean transforms; 5) VTL+SAT + 1 mean transform + 1 precision transform.

Table 1 clearly shows the gains obtained by using the EMLLT model: at all stages we get a reduction of 1% or more absolute in word error rate. We note that although the EMLLT model has roughly 2 times more parameters than the diagonal Gaussian model, we did not see any gain in performance by building a diagonal covariance MLLT Gaussian model with $300K$ Gaussians. Table 2 shows results comparing diagonal and EMLLT models in the PLP VTL+SAT normalized feature space. The results reported in Table 2 are lattice rescoring numbers; the two models are used to rescore the same lat-

tices. The results reported here are for the Switchboard 2001 evaluation test set, which has 120 test speakers. Although the

	PLP	
	Diagonal	EMLLT
1	29.1	28.4
2	28.0	27.2

Table 2. Word error rate comparison of VTL+SAT diagonal Gaussian and EMLLT models on the Switchboard 2001 evaluation testset: 1) initial trigram lattices; 2) rescoring of lattices from 1) with 4-gram language model and acoustic model adapted with one mean transform.

gains on the 2001 evaluation test set are smaller than those we see on the 2000 test set, they are still significant. The reason for the smaller gain in rescoring may be due to the fact that the dynamic range of the likelihoods under the EMLLT model is quite different from that under the baseline model, which may necessitate a re-tuning of the acoustic model weight. This is not critical for the rank-based decoding scheme used in conjunction with the MFCC system.

3.5. SAT Transforms with EMLLT Model

As described in Section 3.3 to build the VTL+SAT EMLLT model we found it expedient to use the the SAT transform for each training speaker estimated for building the baseline VTL+SAT model. Since then, for the MFCC based system we have built an EMLLT VTL+SAT model where the SAT transform for each training speaker was estimated using the EMLLT VTL model. The performance of this VTL+SAT EMLLT model is similar to the one reported earlier in the paper. In fact the precise word error rates with this VTL+SAT EMLLT model on the test data with a SAT transform, SAT+mean transforms and SAT+mean+precision transforms are respectively 23.6%, 22.9% and 22.6%; from Table 1 the corresponding word error rates with the VTL+SAT EMLLT model described earlier in the paper are 23.1%, 22.6% and 22.3% respectively.

4. CONCLUSION

In this paper we have compared EMLLT models with corresponding diagonal covariance Gaussian models on the Switchboard database. The comparisons were made on VTL+SAT normalized MFCC and PLP features with two kinds of decoding strategies - single pass stack decoding and Viterbi lattice rescoring. Experiments were conducted on Hub5 2000 and 2001 evaluation data. Significant word error rate improvements were observed on both these tests over our best diagonal covariance models and these improvements are not solely attributable to the larger number of parameters in the EMLLT models.

Acknowledgment

The authors would like to thank George Saon for building the MFCC VTL+SAT baseline model and Lidia Mangu for providing the initial lattices.

5. REFERENCES

[1] P. A. Olsen and R. A. Gopinath, "Modeling Inverse Covariance Matrices by Basis Expansion," *Submitted to Transactions in Speech and Audio Processing*, 2001, http://www.research.ibm.com/people/r/rameshg/olsen-trans-sap2001-updated.ps.

[2] P. Olsen and R. A. Gopinath, "Modeling Inverse Covariance Matrices by Basis Expansion," in *ICASSP*, Orlando, Florida, 2002.

[3] R. A. Gopinath, V. Goel, K. Visweswariah, and P. Olsen, "Adaptation Experiments on the Spine Database with the EMLLT Model," in *ICASSP*, Orlando, Florida, 2002.

[4] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust Speech Recognition in Noisy Environments: The IBM SPINE-2 Evaluation System," in *ICASSP*, Orlando, Florida, 2002.

[5] C. J. Legetter, P. C. Woodland, "Maximum likelihood linear regression speaker adaptation of continuous density HMMs," *Computer speech and language*, 1997.

[6] J. Shewchuk, "An Introduction to the Conjugate Gradient Method without the Agonizing Pain," Technical Report CMU-CS-94-125, Carnegie Mellon University, March 1994.

[7] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Technical Report TR 291, Cambridge University, 1997.

[8] R. O. Duda and P. B. Hart, *Pattern classification and scene analysis*, Wiley, 1973.

[9] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum Likelihood Discriminant Feature Spaces," in *ICASSP*, Istanbul, Turkey, 2000.

[10] R. A. Gopinath, "Maximum Likelihood Modeling with Gaussian Distributions for Classification," in *Proceedings of ICASSP*, Seattle, USA, 1998, vol. II, pp. 661–664.

[11] S. Chen and P.S. Gopalakrishnan, "Clustering via the Bayesian Information Criterion with Applications in Speech Recognition," in *Proceedings of ICASSP*, Seattle, USA, 1998, vol. I.