Large Vocabulary Conversational Speech Recognition with a Subspace Constraint on Inverse Covariance Matrices

Scott Axelrod, Vaibhava Goel, Brian Kingsbury, Karthik Visweswariah, Ramesh Gopinath

IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

{axelrod,vgoel,bedk,kv1,rameshg}@us.ibm.com

Abstract

This paper applies the recently proposed SPAM models for acoustic modeling in a Speaker Adaptive Training (SAT) context on large vocabulary conversational speech databases, including the Switchboard database. SPAM models are Gaussian mixture models in which a subspace constraint is placed on the precision and mean matrices (although this paper focuses on the case of unconstrained means). They include diagonal covariance, full covariance, MLLT, and EMLLT models as special cases. Adaptation is carried out with maximum likelihood estimation of the means and feature-space under the SPAM model. This paper shows the first experimental evidence that the SPAM models can achieve significant word-error-rate improvements over state-of-the-art diagonal covariance models, even when those diagonal models are given the benefit of choosing the optimal number of Gaussians (according to the Bayesian Information Criterion). This paper also is the first to apply SPAM models in a SAT context. All experiments are performed on the IBM "Superhuman" speech corpus which is a challenging and diverse conversational speech test set that includes the Switchboard portion of the 1998 Hub5e evaluation data set.

1. Introduction

SPAM models [1] are a class of acoustic models for speech recognition which are Gaussian mixture models in which subspace constraints are placed on the precision (inverse covariance) matrices and the means of the individual Gaussians. These models are generalization of diagonal, full covariance, MLLT [2, 3], and EMLLT [4, 5] models. They also provide a method for dimensional reduction which generalizes the LDA/HDA approach. Several papers [1, 6, 7], have previously presented algorithms for training these models and have shown that they can lead to significant improvements in error rate at small or even negative computational cost relative to MLLT or EMLLT models. In particular, it is demonstrated that the SPAM models allow one to obtain much of the gain possible when going from MLLT models to full covariance models at the same per Gaussian computational cost at decode time as that for MLLT models. However, all of the results reported in the references cited above were on an IBM internal database, the language models were grammar-based, no speaker adaptation was applied, and the acoustic models were smaller than optimal sized.

Despite the limitations of the experiments described above, one could fairly well expect ahead of time that the SPAM models can achieve gains over large, speaker adaptive, models on Switchboard and other large vocabulary tasks based on the facts that: (1) the SPAM models reduce error rate as compared to EMLLT models with the same per Gaussian computational cost (at least on the database used in [4, 1, 6, 7]); and (2) the EMLLT models outperform diagonal models in a VTL+SAT context with large models (150K Gaussians) on Switchboard [8]. However, even for those Switchboard experiments, the EMLLT models were compared to a large, but not necessarily *optimal* sized diagonal model.

In this paper we report results for SPAM models applied to IBM's "Superhuman" speech corpus [9] which consists of difficult large-vocabulary conversational speech in many domains recorded under a wide variety of noise conditions by speakers with a broad array of ages, accents, and speaking styles. We focus in particular on one well known and available subset of the Superhuman corpus, namely the Switchboard portion of the 1998 Hub5e evaluations set.

For baseline models to compare the SPAM models to, we will choose diagonal covariance models in an LDA+MLLT projected VTL+SAT feature space where the number of Gaussians per context-dependent state is selected using the Bayesian Information Criterion (BIC) [10], which is theoretically optimal. Further, we scan through the value of the BIC penalty parameter λ to allow several model sizes from which to choose the best diagonal baseline model. In this paper we show that the SPAM models lead to significant word error rate reduction even over this optimized baseline model. We also present new algorithms for FMLLR and MLLR transformation under a SPAM model and show that they lead to additional improvements.

2. SPAM Model and Adaptation

In this paper we will consider SPAM models in which the mean vectors are unconstrained. For such a model with acoustic feature vectors in dimension d, the means $\mu_g \in \mathbf{R}^d$ are independent for each Gaussian g and the precision (inverse covariance) matrices are all required to belong to a "tied" D-dimensional affine subspace of the space of $d \times d$ symmetric matrices. The precision matrices are written as:

$$P_g = S_0 + \sum_{k=1}^{D} \lambda_g^k S_k \quad . \tag{1}$$

Here the λ_g^k are "untied" (Gaussian-dependent) constants, and the $\{S_k\}$ are "tied" (Gaussian-independent) symmetric matrices. All parameters are trained so as to maximize likelihood [7] or approximately maximize likelihood [6, 1] on the training data, subject to the constraint that the P_g are all positive definite. The probability $p(x|s)$ of a context dependent state s is a

Gaussian mixture model:

$$p(x|s) = \sum_{g \in \mathcal{G}(s)} \pi_g \mathcal{N}(x|P_g^{-1}, \mu_g) \quad (2)$$

$$\mathcal{N}(x|P_g^{-1}, \mu_g) = det\left(\frac{P_g}{2\pi}\right)^{1/2} e^{-1/2(x-\mu_g)^T P_g (x-\mu_g)}, \quad (3)$$

where $\mathcal{G}(s)$ is the set of Gaussians for state s.

We will perform unsupervised speaker adaptation of a SPAM model so as to maximize the (EM bound for the) likelihood of the observed speaker specific acoustic data. We will consider FMLLR transforms [11], which are affine linear transformation of the data vector x of the form $x \mapsto Ax + c$; MLLR transforms [12], in which the Gaussian means μ_g are transformed by $\mu_g \mapsto B\mu_g - c$; and combined FMLLR and MLLR transforms in which both the data and the means are transformed as follows:

$$x \mapsto Ax + c \qquad \mu_g \mapsto B\mu_g \ . \quad (4)$$

For conciseness, it is convenient to view FMLLR transforms as the special case of (4) in which B is constrained to be the identity, and MLLR transforms as the special case of (4) in which A is constrained to be the identity.

Each kind of adaptation proceeds by optimizing the following total Q function of A, B, and c, with A and B constrained to be the identity for the case of pure MLLR and pure FMLLR, respectively:

$$Q(A,B,c) = -2\sum_{t,g} \gamma_t(g) \Big[\frac{d}{2}\log(2\pi) + \log\det(|A|) + \log \mathcal{N}(Ax + c; P_g^{-1}, B\mu_g) \Big] \quad (5)$$

where $\gamma_t(g)$ is the usual Gaussian posterior of the forward-backward algorithm. Q may be rewritten in terms of sufficient statistics as follows:

$$Q(A,B,c) = -2\beta \log\det|A| + \sum_k \mathrm{Tr}(S_k\ X^k) \quad (6)$$

$$X^k = A\ G_1^k\ A^T + B\ G_4^k\ B^T + c\ G_6^k\ c^T - 2\ A\ G_3^k\ B^T - 2\ c\ \left(G_5^k\right)^T B^T + 2\ c\ \left(G_2^k\right)^T A^T. \quad (7)$$

Associating to any function $f(x,\mu)$ of a data vector and a mean vector, it's total count:

$$< f(x_t, \mu_g) >_k = \sum_{t,g} \lambda_g^k \gamma_t(g) f(x_t, \mu_g) \ , \quad (8)$$

$G_1^k, ..., G_6^k$ are given as:

$$G_1^k = < x_t x_t^T >_k \text{ , a } d \times d \text{ symmetric matrix} \quad (9)$$
$$G_2^k = < x_t >_k \text{ , a column vector} \quad (10)$$
$$G_3^k = < x_t \mu_g^T >_k \text{ , a } d \times d \text{ matrix} \quad (11)$$
$$G_4^k = < \mu_g \mu_g^T >_k \text{ , a } d \times d \text{ symmetric matrix} \quad (12)$$
$$G_5^k = < \mu_g >_k \text{ , a column vector} \quad (13)$$
$$G_6^k = < 1 >_k \text{ , a constant} \quad (14)$$
$$\beta = \sum_{t,g} \gamma_t(g) \ . \quad (15)$$

As was the case for adaptation of EMLLT models [8], adaptation of SPAM models differs form standard MLLR/FMLLR adaptation in that the optimization problem for (A, B, c) does not break down into independent problems when one considers one row at a time. However, the optimization still proceeds readily by using any standard quasi-newton optimization package.

3. Experimental Setup

All of our experiments are in the context of the IBM "Superhuman speech recognition project" [9]. The IBM 2002 Superhuman speech recognition system uses a multi-pass decoding strategy with successive adaptation steps to achieve good performance across a range of acoustic conditions and speaker characteristics. The final recognition hypothesis is produced through combination of hypotheses from a number of recognition systems that employ different feature sets and acoustic models.

Our focus in this paper is on the process involved in training and decoding with the SPAM models and the baseline diagonal models they will be compared to. That process uses the following input data. All items apply on both the training and testing side, unless otherwise noted. The reader may note that the process of creation of this input data is fairly elaborate, involving many acoustic models in several feature spaces. An attempt was made to balance between making the optimal choice at each step and bowing to practical considerations.

I1 Training data.
This consists of 247 hours of Switchboard data and 17 hours of CallHome English, all released by the LDC.

I2 Test corpus.
The 2002 IBM Superhuman speech corpus consists of five sub-sets. The first subset, abbreviated *swb98* is simply the Switchboard part of the 1998 Hub5e evaluation set. The second subset, abbreviated *mtg* consists of meeting data, a version of which is to be made public by ICSI [13]. Two other data sets, referred to as *cc1* and *cc2* consist of IBM internal telephone call center data, and the final subset, *vm*, consists of voice mail data, reported on in [14].

I3 Vocal-tract length normalization (VTLN) factors.
These are created with a system that uses MFCC features with cepstral mean normalization and an acoustic model comprising 4078 context-dependent states and 171K Gaussian. The VTLN [15] warp factors are computed based on frames that align to vowels and semi-vowels. (For the test data the alignment is to an initial unsupervised decoding).

I4 Initial large word lattice for test data.
Initial word lattices are generated using a bigram language model and a system that uses VTLN PLP [16] features with cepstral mean normalization and a speaker-adaptively trained acoustic model comprising 3688 context-dependent states and 151K Gaussians. The bigram lattices are then rescored with a trigram language model and pruned.

I5 Forced alignment of training data.
This is created using the VTLN PLP model of I4.

I6 Mean normalized MFCC feature vectors.
These are 24-dimensional Mel-frequency cepstral coefficients (MFCCs) calculated using 25-ms frames with a 10-ms step size, with power spectra pre-processed by spectral flooring (performed by adding the equivalent of one bit of additive noise to the power spectra prior to Mel binning) and

periodogram averaging to smooth the binned power spectra. The VTL warping factors from I3 determine a linear scaling of the frequency axis prior to Mel binning. The MFCCs are computed using a bank of 24 Mel filters spanning 0–4 kHz. The MFCC features are all mean normalized.

I7 $LDA + MLLT$ projection of spliced MFCC features. Sixty dimensional acoustic feature vectors are computed by splicing together 9 consecutive frames of MFCC features from I6 and projecting to a 60 dimensional feature space using an LDA+MLLT projection.

I8 FMLLR transforms into the "SAT_0" feature space. Using an acoustic model on the feature space of I7 comprising 4440 context-dependent states and 163K diagonal Gaussians, a single FMLLR transform is computed per-speaker. On the testing side this uses the one-best hypothesis from the lattice of I4.

As mentioned above, the final recognition hypothesis of the IBM 2002 Superhuman speech recognition system is produced through combination of hypothesis created from a number of sub-systems. Each sub-system consists of a feature space, a model on that feature space, and an algorithm for computing a test speaker adaptive transformation of that model. For each such sub-system, the lattices of I4 are rescored using the acoustic weights from the speaker adaptively transformed model associated with the sub-system. These lattices are then further pruned and rescored using a 4-gram language model. The lattices for each of sub-systems are then processed to generate confusion networks [17] which are combined to produce the final overall system recognition output.

Further details on the full Superhuman system appear in [9]. In this paper we will content ourselves to focus on a single subsystem and will report word error rates for the one-best hypotheses of the lattices obtained by acoustically rescoring the lattices of I4 using the given sub-system.

4. Experiments and Results

In this section we describe the construction of several subsystems and report on their word error rates (computed as described in the previous paragraph). Specifically, we describe how various diagonal, full covariance, and SPAM acoustic models on the SAT_0 feature space I8 are constructed using the forced alignment I5 of the training data I1. Each of these models specifies distributions for 10133 left context dependent states. We use the technique described in Section 2 to construct FMLLR, MLLR, and combined FMLLR+MLLR transforms of the SPAM models. As described in I8 above, the feature space SAT_0 is trained based on a $163K$ Gaussian diagonal model from a previous stage of the overall system. As a check that the diagonal models we compare the SPAM models to have little room for improvement, we also construct a large $617K$ Gaussian diagonal model in it's own SAT feature space, which we refer to as the "SAT_1" feature space in the caption to Table 3.

Our first step was to train diagonal models of various sizes. This was dones using the Bayesian Information Criterion [10] (BIC) to choose the number of Gaussians for each context-dependent state. For a given BIC penalty parameter λ, the BIC criterion tells us to choose the model θ for state s which maximizes the following function:

$$F_s(\theta) \quad = \quad \log P(X_s|s,\theta) - \frac{\lambda}{2}\#(\theta)\log(N_s). \qquad (16)$$

Here the first term is just the total likelihood under the model θ of the data points X_s aligned to state s; N_s is the number of such data points; $\#(\theta)$ is the number of parameters in the model θ; and λ is called the penalty weight. Bayesian analysis tells us that, near the asymptotic limit, the best choice of penalty weight is $\lambda = 1$. In fact, we varied the penalty weight and generated models of various sizes and tested all of these models on the *swb98* test data in the SAT_0 feature space. The results are presented in Table 1. As can be seen, the model which performed best had $357K$ Gaussians and an error rate of 32.4%, with larger models becoming overtrained. The model with the theoretically preferred penalty weight of 1.0 had $217K$ Gaussian and performed nearly as well as the best performing model, so we chose this as the model size to use for creating SPAM models.

Penalty weight	# Gaussians	WER
0.51441	$617K$	32.7%
0.55000	$563K$	32.7%
0.67032	$357K$	32.4%
0.81873	$280K$	32.6%
1.00000	$217K$	32.6%
1.22140	$175K$	33.2%
1.49182	$149K$	33.3%
2.7183	$85K$	33.8%
4.0552	$61K$	35.3%

Table 1: *Error rates on swb98 for diagonal models in the SAT_0 feature space built with various BIC penalty weights.*

In order to explore the benefits of using full covariance modeling and to have a model to be used for training the SPAM basis $\{S_k\}$ for various D, we next created a full covariance model with $61K$ Gaussians in the SAT_0 feature space, seeded by the diagonal model in the last line of Table 1. (The choice of $61K$ Gaussians, rather than a larger model was made because of implementation constraints.) Next, we trained SPAM basis matrices $\{S_k\}$ using the modified Frobenius approximation technique [1]. That is, we chose the $\{S_k\}$ to be the tied parameters of a SPAM model which minimizes a certain approximation to the Kullback-Liebler distance from the SPAM model to the full covariance model. Finally, we used the EM algorithm for SPAM to train a SPAM model with $61K$ Gaussians and basis of size $D = d = 60$. Table 2 gives the error rate on the *swb98* test data for the $61K$ Gaussian full covariance, diagonal covariance, and SPAM models with $D = d = 60$. The table shows that 2% absolute error rate improvement is obtained in going from a diagonal covariance to a full covariance model, and that the SPAM model, which has equal computational burden to the diagonal model, is able to achieve half of the improvement that the full covariance model achieved.

MODEL	WER
full cov	33.4%
SPAM D=d	34.4%
diag	35.3%

Table 2: *Error rates on swb98 for full covariance, diagonal, and SPAM models (with comparable computational cost to the diagonal model) having $61K$ Gaussians.*

The results so far are consistent with previous experiments showing gains possible with SPAM models. Our main result, however, is to compare the SPAM models with the optimal sized diagonal model. To do this, we trained SPAM models in the SAT_0 feature space with $217K$ Gaussians whose tied models use the $\{S_k\}$ discussed above and which have basis size D equal to d, $2d$, and $4d$. The untied parameters of these models

were all trained by the EM algorithm for SPAM, seeded by the $217K$ diagonal model.

Table 3 presents our main results. It lists word error rates for the various sub-sets of the 2002 Superhuman corpus described in I1 above, as well as overall average error rates. Each line corresponds to a different choice of model as described in the caption to the table.

model	swb98	mtg	cc1	cc2	vm	all
diag-big	32.4	39.9	41.9	38.4	26.3	35.8
diag	32.6	41.3	41.3	38.5	27.0	36.1
SPAM d	32.3	40.9	41.0	37.9	26.2	35.6
SPAM 2d	31.7	40.3	42.6	37.6	25.8	35.6
SPAM 4d	31.5	39.5	43.8	37.9	25.4	35.6
SPAM 2d + F	31.6					
SPAM 2d + M	31.0					
SPAM 2d+FM	30.8	39.0	41.1	36.7	25.6	34.6

Table 3: *Comparison of word error rates on the Superhuman corpus for various models. The first model is a diagonal model with $617K$ Gaussians in the SAT_1 feature space, The remaining models all have $217K$ Gaussians and are distributions on the SAT_0 feature space. They are: a diagonal model, SPAM models with bases size $D = d$, $2d$, and $4d$, and the spam model with basis size $D = 2d$ with additional FMLLR, MLLR, and combined FMLLR and MLLR transforms.*

5. Analysis and Conclusion

In analyzing our main results in Table 3, we begin by suggesting the *swb98* task is the most significant one to focus on for the present work since it is the most studied in the literature and because the other tasks in the Superhuman corpus were chosen specifically because they have special difficulties, e.g. clicks in the acoustic waveforms, which we have not specifically addressed here.

The first two lines of Table 3 give results for diagonal models that we have made a sincere effort at making very close to optimal. The next three lines show that the SPAM models with $D = d$, $2d$, and $4d$ provide significant improvement on the diagonal models, with the $D = 4d$ model giving a 0.9% absolute improvement over the best diagonal model on the *swb98* task.

The final two lines of the table show that FMLLR and combined FMLLR+MLLR transforms of the $D = 2d$ SPAM model leads to significant improvements over the untransformed model.

It can also be seen that the FMLLR transform on top of the SPAM model leads to little improvement (on top of either the plain $D = 2d$ SPAM model or that model with an MLLR transform). This just verifies that the SAT_0 feature space (which is trained based on a $163K$ diagonal model) is close to what a true SAT space for the SPAM models would be. It also explains why we did not invest the time to compute the error rates on the test sets other than *swb98* for the "SPAM 2d+F" and "SPAM 2d+M" models.

As a final comment, we note that the SPAM results themselves are slightly sub-optimal and could possibly be improved by increasing the basis size, varying the number of Gaussians, or training the basis to maximize likelihood as in [7], rather than the quadratic approximation to likelihood of [1].

6. Acknowledgements

We are grateful to the members of the IBM "Superhuman team" for discussion, models, lattices, code, and all that, in particular to George Saon, Stan Chen, and Lidia Mangu. We would also like to thank Peder Olsen for helpful discussions.

7. References

[1] S. Axelrod, R. A. Gopinath, and P. Olsen, "Modeling with a subspace constraint on inverse covariance matrices," in *Proc. ICSLP*, 2002.

[2] R.A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *ICASSP 1998*.

[3] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, 1999.

[4] P. Olsen and R. A. Gopinath, "Modeling inverse covariance matrices by basis expansion," in *ICASSP*, 2002.

[5] P. Olsen and R. A. Gopinath, "Modeling inverse covariance matrices by basis expansion," *IEEE Transactions on Speech and Audio Processing*, submitted 2001.

[6] S. Axelrod, R. A. Gopinath, P. Olsen, and K. Visweswariah, "Dimensional reduction, covariance modeling, and computational complexity in ASR systems," in *Proc. ICASSP*, 2003.

[7] K. Visweswariah, P. Olsen, R. A. Gopinath, and S. Axelrod, "Maximum likelihood training of subspaces for inverse covariance modeling," in *Proc. ICASSP*, 2003.

[8] J. Huang, V. Goel, R. Gopinath, B. Kingsbury, P. Olsen, and K. Visweswariah, "Large vocabulary conversational speech recognition with the extended maximum likelihood linear transformation (EMLLT) model," in *ICSLP 2002*.

[9] B. Kingsbury, L. Mangu, G. Saon, G. Zweig, S. Axelrod, V. Goel, K. Visweswariah, and M. Picheny, "Toward domain-independent conversational speech recognition," submitted to *Eurospeech* 2003.

[10] S. Chen and R. A. Gopinath, "Model selection in acoustic modeling," in *Proc. Eurospeech*, September 1999.

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

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

[13] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," in *Proc. HLT 2001*, 2001.

[14] M. Padmanabhan, G. Saon, J. Huang, B. Kingsbury, and L. Mangu, "Automatic speech recognition performance on a voicemail transcription task," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 433–442, 2002.

[15] S. Wegman, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *Proc. ICASSP*, 1996.

[16] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, April 1990.

[17] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.