MAXIMUM MUTUAL INFORMATION SPEAKER ADAPTED TRAINING WITH SEMI-TIED COVARIANCE MATRICES

John McDonough and Alex Waibel

Interactive Systems Laboratories
Institut für Logik, Komplexität, und Deduktionssysteme
Universität Karlsruhe
Am Fasanengarten 5
76128 Karlsruhe, Germany
jmcd@ira.uka.de, ahw@cs.cmu.edu

ABSTRACT

We present re-estimation formulae for semi-tied covariance (STC) transformation matrices based on a maximum mutual information (MMI) criterion. These re-estimation formulae are different from those that have appeared previously in the literature. Moreover, we present a positive definiteness criterion with which the regularization constant present in all MMI re-estimation formulae can be reliably set to provide both consistent improvements in the total mutual information of the training set, as well as fast convergence. We combine the STC re-estimation formulae with their like for speaker-independent means and variances, and update *all* parameters during MMI speaker adapted training (MMI-SAT). We present the results of two sets of speech recognition experiments conducted on the the 1998 Broadcast News evaluation set, as well as a corpus of Meeting Room data collected at the Interactive Systems Laboratories of the Carnegie Mellon University.

1. INTRODUCTION

Since Woodland [1] discovered that the word error rate (WER) reductions provided by discriminative training techniques over and above their maximum likelihood counterparts, could be greatly enhanced by scaling all acoustic log-likelihoods during training, MMI training has enjoyed a spate of renewed interest and a concomitant flurry of publications. Perhaps most noteable among these was the work by Gunawardana [2] which set forth a much simplified derivation of Normandin's [3] original continuous density re-estimation formulae, one which does not require the discrete density approximations Normandin used.

In [4] the current authors used Gunawardana's theorem to derive re-estimation formulae for the speaker-independent (SI) means and variances of a hidden Markov model (HMM) when speaker-adapted training [5] (SAT) is conducted on the latter under an MMI criterion. It would appear that since the publication of [2], and perhaps well before, research into MMI training schemes has followed similar if independent tracks at both the Johns Hopkins University and the University of Karlsruhe. Indeed, Byrne *et al* [6] presented a scheme for MMI-SATraining in which all parameters, including the speaker-dependent (SD) adaptation parameters, were estimated with an MMI criterion. The same authors also presented a technique for performing MMI estimation of a transformation matrix suitable for use with semi-tied covariance (STC) matrices proposed by Gales [7], but commented on the tendency of their technique to produce a transformation that was "effectively identity."

In this work, we also present re-estimation formulae for STC transformation matrices based on an MMI criterion. As will be shown, these re-estimation formulae are different from those in [6], and the transformation matrices therewith obtained are distinctly different from the identity. Moreover, we present a positive definiteness criterion, with which the regularization constant present in all MMI re-estimation formulae can be reliably set to provide both consistent improvements in the total mutual information of the training set, as well as fast convergence. We also combine the STC re-estimation formulae with their like for the SI means and variances, and update *all* parameters during MMI-SATraining. To demonstrate the effectiveness of the proposed techniques, we present the results of two sets of speech recognition experiments conducted on the the 1998 Broadcast News evaluation set, as well as a corpus of Meeting Room data collected at the Interactive Systems Laboratories of the Carnegie Mellon University.

The balance of this work is organized as follows. In Section 2 we briefly review the MMI-SAT mean and covariance re-estimation formulae previously derived in [4]. We also present our derivation of a MMI re-estimation formula for STC transformation matrices, along with a scheme for optimal regression class assignment. In Section 3 we present the results of our initial sets of experiments combining the re-estimation of all relevant parameters during MMI-SATraining. Finally, in Section 4 we summarize our efforts, and present plans for further work.

2. MAXIMUM MUTUAL INFORMATION ESTIMATION

Assume we wish to estimate the k^{th} mean μ_k and diagonal covariance matrix D_k of a continuous density hidden Markov model. Let s be an index over all speakers in the training set, and let $x_t^{(s)}$ denote the t^{th} observation from speaker s. Also let $c_{k,t}^{(s)}$ denote the *posterior probability* that $x_t^{(s)}$ was drawn from the k^{th} Gaussian in the HMM whose parameters we wish to re-estimate. Let us define the quantities

$$c_k^{(s)} = \sum_t c_{k,t}^{(s)} \qquad o_k^{(s)} = \sum_t c_{k,t}^{(s)} x_t^{(s)} \qquad s_k^{(s)} = \sum_t c_{k,t}^{(s)} x_t^{(s)2}$$

0-7803-7663-3/03/$17.00 ©2003 IEEE

which are typically accumulated during forward-backward training. In the sequel we let $\Lambda_k = \{\mu_k, D_k\}$ denote the parameters of the k^{th} Gaussian component and $\Lambda = \{\Lambda_k\}$ the speaker-independent parameters of the entire HMM.

The mean and covariance re-estimation formulae for maximum mutual information speaker-adapted training (MMI-SAT) have appeared in prior work by the current authors [4]. We summarize them here only to introduce the notation used in the current work. Let $x^{(s)}$, $n^{(s)}$, and $w^{(s)}$ respectively denote observation, Gaussian component and word *sequences* associated with an utterance of speaker s. Define *mutual information* as

$$I(W,O;\Lambda) = \sum_s \log \frac{p(w^{(s)}, x^{(s)}; A^{(s)}, \Lambda)}{p(w^{(s)})\, p(x^{(s)}; A^{(s)}, \Lambda)}$$

where $A^{(s)}$ is the matrix of maximum likelihood linear regression [8] parameters for speaker s. Let $\Lambda^0 = \{\Lambda_k^0\}$ denote the current set of parameter values and define the *auxiliary function*

$$\begin{aligned} Q(\Lambda|\Lambda^0) &= S^{(1)}(\Lambda|\Lambda^0) + E\cdot S^{(2)}(\Lambda|\Lambda^0) \qquad (1)\\ &= \sum_k \left[S_k^{(1)}(\Lambda_k|\Lambda_k^0) + E\cdot S_k^{(2)}(\Lambda_k|\Lambda_k^0)\right] \end{aligned}$$

where

$$S^{(1)}(\Lambda|\Lambda^0) = \sum_{t,s} c_{k,t}^{(s)} \log p(x_t^{(s)}; A^{(s)}, \Lambda_k)$$

$$S_k^{(2)}(\Lambda|\Lambda^0) = \sum_s d_k^{(s)} \int_x p(x; A^{(s)}, \Lambda_k^0) \log p(x; A^{(s)}, \Lambda_k)\, dx$$

In the above, it is necessary to define $c_{k,t}^{(s)}$ as the difference in posterior probabilities of $n^{(s)}$ that comes from knowledge of the correct word transcription

$$\begin{aligned} c_{k,t}^{(s)} = p(n_t^{(s)} = k|w^{(s)}, x^{(s)}; A^{(s)}, \Lambda^0)\\ - p(n_t^{(s)} = k|x^{(s)}; A^{(s)}, \Lambda^0) \qquad (2) \end{aligned}$$

and set

$$d_k^{(s)} = \sum_{t,s} p(n_t^{(s)} = k|x^{(s)}; A^{(s)}, \Lambda^0) \qquad (3)$$

The real constant $E > 0$ is typically chosen heuristically; for mean and variance re-estimation, good results have been obtained with $E = 1$.

Gunawardana [2] showed that $Q(\Lambda|\Lambda^0) > Q(\Lambda^0|\Lambda^0)$ implies $I(W,O;\Lambda) > I(W,O;\Lambda^0)$. In prior work [4], the current authors used (1) and Gunawardana's theorem to derive the maximum mutual information mean re-estimation formula $\mu_k = \mathbf{M}_k^{-1}\mathbf{v}_k$ where

$$\mathbf{M}_k = \sum_s \left(c_k^{(s)} + E\cdot d_k^{(s)}\right) A^{(s)\,T} D_k^{-1} A^{(s)} \qquad (4)$$

$$\mathbf{v}_k = \sum_s A^{(s)\,T} D_k^{-1} \left[\left(o_k^{(s)} - c_k^{(s)} b^{(s)}\right) + E\cdot d_k^{(s)} A^{(s)} \mu_k^0\right] \qquad (5)$$

The corresponding covariance re-estimation formula is

$$\begin{aligned} \sigma_{kn}^2 = \sum_s \Big\{ \left(s_k^{(s)} - 2\, o_k^{(s)} \hat{\mu}_{kn}^{(s)} + c_k^{(s)} \hat{\mu}_{kn}^{(s)2}\right)\\ + E\cdot d_k^{(s)} \left[\sigma_{kn}^{0\,2} + \left(\hat{\mu}_{kn}^{0(s)} - \hat{\mu}_{kn}^{(s)}\right)^2\right]\Big\} \Big/ \sum_s (c_k^{(s)} + E\cdot d_k^{(s)}) \end{aligned}$$

where $\hat{\mu}_{kn}^{0(s)}$ is the n^{th} component of $\hat{\mu}_k^0 = A^{(s)}\mu_k^0$ and $\sigma_{kn}^{0\,2}$ is the current value of the variance.

Semi-Tied Covariance Estimation

For reasons of brevity, we only summarize our derivation here; full details can be found in [9]. Gales [7] defines a *semi-tied covariance matrix* as $\Sigma_k = P\, D_k P^T$ where D_k is, as before, the diagonal covariance matrix for the k^{th} Gaussian component, and P is a transformation matrix shared by many Gaussian components. Both D_k and P can then be updated using a ML criterion. Rather than optimizing P directly, however, Gales defines $M = P^{-1}$ and then optimizes M instead. According to Gales' original formulation, this can be achieved by setting $o_t^{(s)} = x_t^{(s)} - \mu_k$ and defining the auxiliary function

$$\begin{aligned} S^{(1)} = \sum_k \{c_k \log|M| - \tfrac{1}{2} c_k \log|2\pi D_k|\\ - \tfrac{1}{2}\sum_{t,s} c_{k,t}^{(s)} \left[o_t^{(s)T} M^T D_k^{-1} M o_t^{(s)}\right]\} \qquad (6) \end{aligned}$$

where $c_k = \sum_{t,s} c_{k,t}^{(s)}$ and $c_{k,t}^{(s)}$ is the posterior probability that observation $x_t^{(s)}$ was drawn from Gaussian component k. If speaker adaptation is performed in the model space, Gales' development can be readily extend by setting

$$o_t^{(s)} = x_t^{(s)} - \left(A^{(s)}\mu_k + b^{(s)}\right) \qquad (7)$$

The components $\{m_{ij}\} = M$ can be updated using a recursive procedure [7].

In order to estimate M using an MMI criterion, we need only modify (6) slightly. Firstly, we redefine $c_{k,t}^{(s)}$ as the difference in posterior probabilities (2). Now assume that we wish only to optimize M. Hence the Gaussian normalization factor $-\frac{1}{2}\log|2\pi D_k|$ can be excluded as it does not depend on M; what remains is

$$S^{(1)} = \sum_k \left\{ c_k \log|M| - \tfrac{1}{2}\sum_{t,s} c_{k,t}^{(s)} \left(o_t^{(s)T} M^T D_k^{-1} M o_t^{(s)}\right)\right\}$$

Let m_i^T denote the i^{th} row of M, and let f_i^T denote the i^{th} row in the *cofactor* matrix of M. Hence (6) can be rewritten as

$$S^{(1)} = \tfrac{1}{2}\left\{ c \log(m_i^T f_i)^2 - \sum_j m_j^T W_j^{(1)} m_j \right\} \qquad (8)$$

where $c = \sum_k c_k$ and

$$W_j^{(1)} = \sum_k \frac{1}{\sigma_{k,j}^2} \sum_{t,s} c_{k,t}^{(s)} o_t^{(s)} o_t^{(s)T}$$

Turning now to the second term in (1), we must write

$$\begin{aligned} S^{(2)}(\Lambda|\Lambda^0) = \sum_k \sum_s d_k^{(s)} \int_x p(x; A^{(s)}, \Lambda_k^0, P^0)\cdot\\ \log p(x; A^{(s)}, \Lambda_k, P)\, dx \end{aligned}$$

It is straightforward to show that $S^{(2)}$ can be expressed as

$$S^{(2)} = \tfrac{1}{2}\left\{ d \log(m_i^T f_i)^2 - \sum_j m_j^T W_j^{(2)} m_j \right\} \qquad (9)$$

where $d = \sum_k d_k$ and

$$W_j^{(2)} = \sum_k \frac{1}{\sigma_{k,j}^2} \left(d_k \sum_i \sigma_{k,i}^{0\,2} p_i^0 p_i^{0\,T} \right)$$

In the above, p_i^0 denotes the i^{th} column of P^0. Substituting (8) and (9) into (1), we find

$$Q(\Lambda|\Lambda^0) = \tfrac{1}{2} \left\{ (c + E \cdot d) \log(m_i^T f_i)^2 - \sum_j m_j^T W_j m_j \right\}$$

where $W_j = W_j^{(1)} + E \cdot W_j^{(2)}$. Following the development in Gales [7], it can easily be shown that $Q(\Lambda|\Lambda^0)$ can be maximized by iteratively updating the rows of M according to

$$m_j^T = f_j^T W_j^{-1} \sqrt{\frac{c + E \cdot d}{f_j^T W_j^{-1} f_j}} \tag{10}$$

Equation (10) cannot be used indiscriminately; care must be taken to choose E large enough so that each W_j is positive definite. From the foregoing, it readily follows that

$$\frac{\partial^2 Q(\Lambda|\Lambda^0)}{\partial m_i \, \partial m_i^T} = -(c + E \cdot d) \frac{f_i f_i^T}{(m_i^T f_i)^2} - W_i$$

A maximum of $Q(\Lambda|\Lambda^0)$ requires that $\partial^2 Q / \partial m_i^T \, \partial m_i$ is negative definite. As $c + E \cdot d = E \cdot d > 0$ and $f_i f_i^T$ is a symmetric rank-one update, and hence positive definite, this will certainly hold if W_i is positive definite. It may be possible, however, to derive a weaker condition.

Regression Class Estimation

In order to obtain the largest possible reduction in word error rate, quite often we estimate not a single global transformation $(A^{(s)}, b^{(s)})$ for each speaker, but a set of transformations $\{(A_r^{(s)}, b_r^{(s)})\}$. In this case, it is of interest to reassign the k^{th} Gaussian to a regression class r so as to maximize the mutual information. This can be done with a slight extension of the prior analysis, to wit: it is necessary to calculate the actual value of the auxiliary function in addition to the optimal parameters Λ_k achieving its maximum. Letting

$$Q_k(\Lambda_k|\Lambda_k^0) = S_k^{(1)} + E \cdot S_k^{(2)}$$

it can be shown

$$Q_k(\Lambda_k|\Lambda_k^0) = a_k + \mu_k^T \mathbf{v}_k - \tfrac{1}{2} \mu_k^T \mathbf{M}_k \mu_k \tag{11}$$

where

$$a_k = -\frac{1}{2} \sum_s \Big[(c_k^{(s)} b_r^{(s)} - o_k^{(s)})^T D_k^{-1} (c_k^{(s)} b_r^{(s)} - o_k^{(s)}) / c_k^{(s)} + E \cdot d_k^{(s)} \mu_k^{0\,T} A_r^{(s)\,T} D_k^{-1} A_r^{(s)} \mu_k^0 \Big]$$

and $\mathbf{M}_k$ and $\mathbf{v}_k$ are as given in (4) and (5). With these definitions, it is possible to choose the optimal regression class r^* as that class which minimizes (11); see [10, §4.5].

3. SPEECH RECOGNITION EXPERIMENTS

The speech experiments described below were conducted with the Janus Recognition Toolkit (JRTk), which is developed and maintained jointly at Universität Karlsruhe, in Karlsruhe, Germany and at the Carnegie Mellon University in Pittsburgh, Pennsylvania, USA.

For the experiments reported below, HMM training was conducted on a combined training set consisting of the Broadcast News (BN) corpus, which totals approximately 64 hours of speech, along with the ESST set. The complete training set contains speech contributed by 2,989 speakers. Two test sets were used to determine system performance: the first was that set used for the 1998 Broadcast News evaluation which contains 15,310 words; the second Meeting Room (MR) test set was collected at the Interactive Systems Laboratories (ISL) of the Carnegie Mellon University. The MR test set contains 11,214 words spoken in discussions of various research projects currently underway at ISL. The speech therein is conversational and entirely spontaneous. Although the entire MR corpus is English, many of the speaker are non-native. As such, it makes for a very challenging automatic recognition task [11]. For these experiments, our baseline recognizer was comprised of 4,144 continuous density codebooks, each of which contained either 16 or 32 Gaussians.

All speech data was digitally sampled at a rate of 16 kHz. The speech features used for all experiments were obtained by first estimating 13 cepstral components, concatenating nine (9) successive features together, then performing linear discriminant analysis to obtain a final feature of length 40. Features were calculated every 10 ms using a 16 ms sliding window. Speaker-dependent frequency-domain vocal tract length normalization (VTLN) was used in calculating all speech feaures for both training and test.

The word lattices annotated with word start- and end-times used for discriminative training were written with the Ibis decoder [12]. Training was conducted by first performing a Viterbi alignment on the correct transcription for each utterance, and accumulating the appropriate *numerator* statistics [13]. The correct utterance together with its time markings was then inserted into the corresponding lattice, and Viterbi rescoring was performed on each link of this enhanced lattice based on the fixed start- and end-times. Using these new acoustic scores, the posterior probability of each link was calculated and used to accumulate the *denominator* statistics. As recommended in [1], the log acoustic scores were scaled by a factor of 1/15 during discriminative training, and a unigram language model was used in calculating the link posterior probabilities.

Shown in Figure 1 are the results of our initial speech recognition experiments on the BN and MR test sets. For these systems, a STC transformation matrix was estimated during conventional speaker-independent (SI) training, and held fixed thereafter for both ML- and MMI-SAT. To generate these results, we first did a complete decoding with the baseline MLE system, simultaneously writing both word lattices and errorful transcripts. The word lattices were then rescored with the appropriate acoustic models and, where necessary, adaptation parameters to generate the subsequent results. In our initial experiments, we used an acoustic with relatively few parameters: 16 Gaussian components for each of 4,144 codebooks. To determine the susceptibility of the MMI models to overtraining, we then trained a model with 32 Gaussians for each of 4,144 codebooks. As is clear from the tabulated results, the relative gains for both systems were comparable. MMI-SATraining gave a substantial improvements over MLE-SATraining, and best performance was obtained after two iterations of MMI-SAT for both large and small models.

System	% Word Error Rate			
	$4,144 \times 16$		$4,144 \times 32$	
	BN	MR	BN	MR
MLE Baseline	23.6	45.3	21.4	42.9
MLE-SAT	20.0	41.6	18.8	39.9
MMI-SAT-1	19.2	40.1	18.2	38.8
MMI-SAT-2	18.6	39.9	17.6	38.3
MMI-SAT-3	19.4	40.3	18.2	39.2

Fig. 1. Word error rate results on the 1998 Broadcast News evaluation set (BN) and the Interactive Systems Laboratories Meeting Room set (MR).

The next set of experiments was intended to determine if further reductions in word error rate could be achieved by re-estimating the STC transformation matrix during both ML- and MMI-SATraining. These experiments were conducted only on the large $4,144 \times 32$ model. Training under both criteria was conducted by making a first pass through the training set to re-estimate the STC transformation matrix, then making a second pass to re-estimate the SI means and variances. As a means of reducing computation, only the contribution of each frame to the *most likely* Gaussian in any given state was accumulated during STC estimation. The constant E in (3) was set as follows: First the matrices $\{W_j^{(1)}\}$ and $\{W_j^{(2)}\}$ were accumulated for all speakers in the training set. Thereafter E was set to the low value of 1/128 and successively doubled until all matrices $W_j = W_j^{(1)} + E \cdot W_j^{(2)}$ were positive definite. The values thereby obtained were $E = 1/16, 1/2$, and $1/2$ for the first, second, and third iterations respectively. This procedure proved very robust, providing both consistent improvements in the total mutual information of the training set, as well as rapid convergence.

For the sake of comparison, a second system was trained by holding the STC transformation matrix fixed after ML-SATraining, and re-estimating only the SI means and varianes during MMI-SATraining. The WER results obtained with both systems are tabulated in Table 2.

System	% Word Error Rate			
	ML-SAT STC		MMI-SAT STC	
	BN	MR	BN	MR
MLE-SAT	18.8	39.9	18.8	39.9
MMI-SAT-1	18.2	38.8	18.1	39.1
MMI-SAT-2	17.6	38.3	19.1	40.7
MMI-SAT-3	18.2	39.2	62.7	75.1

Fig. 2. WER results comparing STC transformation matrices estimated during ML- and MMI-SATraining.

Unfortunately, re-estimation of the STC transformation matrix provides no further reductions in WER, much the opposite in fact. At present it is not fully understood why this is so.

4. CONCLUSIONS

We have presented a practical technique for performing SAT on a continuous density hidden Markov model using an MMI criterion. In a set of experiments on two large vocabulary speech recognition tasks, we have demonstrated the effectiveness of MMI-SAT in reducing word error rate with respect to that obtained with MLE-SAT. We have also enhanced the basic MMI-SAT estimation with re-estimation formulae for the transformation matrices used with semi-tied covariances. To date no further reductions in WER have been obtained from STC estimation with an MMI criterion. This remains an area of active research, however.

Acknowledgments: The authors wish to acknowledge Hagen Soltau, Christian Fügen, and Florian Metze for their assistance in developing the lattice code used to perform MMI training as discussed in this work.

5. REFERENCES

[1] P. Woodland and D. Povey, "Large scale discriminative training for speech recognition," in *ISCA ITRW Automatic Speech Recognition: Challenges for the Millenium*, 2000, pp. 7–16.

[2] A. Gunawardana, "Maximum mutual information estimation of acoustic HMM emission densities," Tech. Rep. 40, Center for Language and Speech Processing, The Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218, USA, 2001.

[3] Y. Normandin, *Hidden Markov Models, Maximum Mutual Information, and the Speech Recognition Problem*, Ph.D. thesis, McGill University, 1991.

[4] J. McDonough, T. Schaaf, and A. Waibel, "On maximum mutual information speaker-adapted training," in *Proc. ICASSP*, 2002.

[5] T. Anastasakos, J. McDonough, R. Schwarz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP*, 1996, pp. 1137–1140.

[6] S. Tsakalidis, V. Doumpiotis, and W. Byrne, "Disciminative linear transforms for feature normalization and speaker adaptation in HMM estimation," in *Proc.ICSLP*; 2002, pp. 2585–2588.

[7] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.

[8] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, April 1995.

[9] J. McDonough, T. Schaaf, and A. Waibel, "On maximum mutual information speaker-adapted training," *Computer Speech and Language*, submitted for publication.

[10] J. McDonough, *Speaker Compensation with All-Pass Transforms*, Ph.D. thesis, The Johns Hopkins University, 2000.

[11] A. Waibel, H. Yu, H. Soltau, T. Schultz, T. Schaaf, Y. Pan, F. Metze, and M. Bett, "Advances in meeting recognition," in *Proc. Human Language Technology Conference, San Diego*, 2001.

[12] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A one pass decoder based on polymorphic linguistic context assignment," in *Proc. ASRU*, Trento, Italy, 2001.

[13] V. Valtchev, P. C. Woodland, and S. J. Young, "MMIE training of large vocabulary speech recognition systems," *Speech Communication*, vol. 22, pp. 303–314, 1997.