ON MAXIMUM MUTUAL INFORMATION SPEAKER–ADAPTED TRAINING

John McDonough, Thomas Schaaf, and Alex Waibel

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

ABSTRACT

In this work, we combine maximum mutual information-based parameter estimation with speaker-adapted training (SAT). As will be shown, this can be achieved by performing *unsupervised* parameter estimation on the test data, a distinct advantage for many recognition tasks involving conversational speech. We also propose an approximation to the maximum likelihood and maximum mutual information SAT re-estimation formulae that greatly reduces the amount of disk space required to conduct training on corpora such as Broadcast News, which contains speech from thousands of speakers. We present the results of a set of speech recognition experiments on three test sets: the English Spontaneous Scheduling Task corpus, Broadcast News, and a new corpus of Meeting Room data collected at the Interactive Systems Laboratories of the Carnegie Mellon University.

1. INTRODUCTION

Since the publications of [1, 2] linear transform techniques have become extremely popular for performing speaker adaptation on continuous density hidden Markov models (HMMs). A good summary of many of the refinements of linear regression-based speaker adaptation is given by Gales [3]. Other work [4] proposes a set of linear transforms that are specified by very few free parameters, and hence able to be robustly estimated with little speaker-dependent enrollment data. Regardless of the precise formulation of the linear transform, these techniques, until very recently, have been based on maximum likelihood (ML) parameter estimation. This is also true of the technique Speaker-Adapted Training (SAT) [5], in which transform parameters are estimated for all speakers in a training set, and then used during the re-estimation of the speaker-independent means and variances.

Gopalakrishnan *et al* [6] first proposed a practical technique for performing maximum mutual information (MMI) training of hidden Markov models, and commented on the fact that MMI- is superior to ML-based parameter estimation given that the amount of available training data is always limited, and that the HMM is not the actual model of speech production. Gopalakrishnan's development was subsequently extended by Normandin [7] to the case of continuous density HMMs. While these initial works were of theoretical interest, for several years it was believed that the marginal performance gains that could be obtained with MMI did not justify the increase in computational effort it entailed with respect to ML training. This changed when Woodland [8] discovered that these gains could be greatly increased by scaling all acoustic log-likelihoods during training. Since the publication of [8], MMI training has enjoyed a spate of renewed interest and a concomitant flurry of publications, including [9] in which an MMI criterion is used for estimating linear regression parameters, and [10] in which different update formulae are proposed for the standard MMI-based mean and covariance re-estimation. Also noteworthy is the recent work by Gunawardana [11] which sets forth a much simplified derivation of Normandin's original continuous density re-estimation formulae, one which does not require the discrete density approximations Normandin used.

In this work, we take the next logical step of combining MMI-based parameter estimation with SAT. As will be shown, this can be achieved by performing *unsupervised* parameter estimation on the test data, a distinct advantage for many recognition tasks involving conversational speech. We also propose an approximation to the basic SAT re-estimation formulae that greatly reduces the amount of disk space required to conduct training. We present the results of three sets of speech recognition experiments conducted on the English Spontaneous Scheduling Task corpus, the 1998 Broadcast News evaluation set, as well as a new 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 describe the basic SAT mean and covariance re-estimation formula for ML-based training, and introduce a novel way to reduce the amount of hard disk space required to conduct this training. We then present re-estimation formulae for performing SAT under an MMI criterion. In Section 3 we present the results of our initial sets of experiments. Finally, in Section 4 we summarize our efforts, and present plans for further work.

2. MAXIMUM MUTUAL INFORMATION ESTIMATION

The re-estimation formula used in maximum likelihood speaker-adapted training (ML-SAT) have appeared previously in the literature [5]. We summarize them here only as an aid for our subsequent discussion. 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-

0-7803-7402-9/02/$17.00 ©2002 IEEE

estimate. Let us define the quantities

$$c_k^{(s)} = \sum_t c_{k,t}^{(s)}$$

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

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

which are typically accumulated during forward-backward training. The ML re-estimation formula for μ_k is then given by

$$\mu_k = \mathbf{M}_k^{-1} \mathbf{v}_k \tag{1}$$

where

$$\mathbf{M}_k = \sum_s c_k^{(s)} A^{(s)T} D_k^{-1} A^{(s)} \tag{2}$$

$$\mathbf{v}_k = \sum_s A^{(s)T} D_k^{-1} o_k^{(s)} \tag{3}$$

and $A^{(s)}$ is the matrix of maximum likelihood linear regression (MLLR) parameters [2] for speaker s. Moreover, the n^{th} diagonal covariance component of the k^{th} Gaussian can be re-estimated from

$$\sigma_{kn}^2 = \frac{1}{c_k} \sum_s \left(s_k^{(s)} - 2\, o_k^{(s)} \hat{\mu}_{kn}^{(s)} + c_k^{(s)} \hat{\mu}_{kn}^{(s)2} \right) \tag{4}$$

where $\hat{\mu}_{kn}^{(s)}$ is the n^{th} component of $\hat{\mu}_k^{(s)} = A^{(s)} \mu_k$.

A typical implementation of SAT requires writing out the quantities $c_k^{(s)}$, $o_k^{(s)}$, and $s_k^{(s)}$, in addition to $A^{(s)}$, for every speaker in the training set. A moments thought will reveal, however, that $\mathbf{M}_k$ can be accumulated from only $\{A^{(s)}\}$ and $\{c_k^{(s)}\}$. Moreover, the sum $\mathbf{v}_k$ can be accumulated for all speakers and written to disk just once, or else just once for each processor used in parallel forward-backward training, after which these partial sums can be added together. For a training corpus with several thousand speakers such as Broadcast News (BN), this results in a tremendous savings in the disk space required for SAT, as first noted in [12]. Although the sum in (4) requires that the *newly-updated* mean μ_k be used in calculating $\hat{\mu}_{kn}^{(s)}$, experience has proven that the re-estimation works just as well if the *prior* value of μ_k is used instead, in which case the partial sums in (4) can also be written out just once for each parallel processor. This novel, and useful, approximation has been dubbed *fast*[1] speaker-adapted training (FSAT).

The re-estimation formulae for maximum mutual information speaker-adapted training (MMI-SAT) are quite similar to their maximum likelihood counterparts. For the sake of brevity, we only outline their derivation here. Let $x^{(s)}$, $n^{(s)}$, and $w^{(s)}$ respectively denote observation, state and word *sequences* associated with an utterance of speaker s. Define *mutual information* as

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

Also define the *auxiliary function*

$$Q(\Lambda|\Lambda^0) = \sum_{s,n^{(s)}} c^{(s)} \log p(x^{(s)}|n^{(s)}; \Lambda)$$

$$+ \sum_{s,n^{(s)}} d'(n^{(s)}) \int_{x^{(s)}} p(x^{(s)}|n^{(s)}; \Lambda^0) \log p(x^{(s)}|n^{(s)}; \Lambda)\, dx^{(s)}$$

[1] *Fast* bedeutet "almost" auf Englisch.

where Λ^0 denotes the current set of parameter values, and

$$c^{(s)} = p(n^{(s)}|w^{(s)}, x^{(s)}; A^{(s)}, \Lambda^0) - p(n^{(s)}|x^{(s)}; A^{(s)}, \Lambda^0)$$

Gunawardana [11] shows that maximizing $Q(\Lambda|\Lambda^0)$ with respect to Λ, is sufficient to ensure $I(W, O; \Lambda) > I(W, O; \Lambda^0)$.

Let us set $\Lambda_k = \{\mu_k, D_k\}$, whereupon it is straightforward to show that

$$Q(\Lambda|\Lambda^0) = \sum_k Q_k(\Lambda_k|\Lambda_k^0)$$

where

$$Q_k(\Lambda_k|\Lambda_k^0) = \sum_{t,s} c_{k,t}^{(s)} \log p(x_t^{(s)}; A^{(s)}, \Lambda_k) \tag{5}$$

$$+ \sum_s d_k^{(s)} \int_x p(x; A^{(s)}, \Lambda_k^0) \log p(x; A^{(s)}, \Lambda_k)\, dx$$

and $d_k^{(s)}$ is a convergence constant, and we must redefine

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

Next observe that the acoustic log-likelihood for the k^{th} Gaussian can be expressed as

$$\log p(x|\Lambda_k) =$$

$$-\frac{1}{2}\left[\log|2\pi D_k| + (x - A^{(s)}\mu_k)^T D_k^{-1} (x - A^{(s)}\mu_k)\right] \tag{6}$$

Substituting (6) into (5), taking the derivative with respect to μ_k and equating the result to zero, we find that the new value of μ_k can be calculated from (1) as before, provided that we define

$$\mathbf{M}_k = \sum_s (c_k^{(s)} + d_k^{(s)})\, A^{(s)T} D_k^{-1} A^{(s)}$$

$$\mathbf{v}_k = \sum_s A^{(s)T} D_k^{-1} \left(o_k^{(s)} + d_k^{(s)} A^{(s)} \mu_k^0 \right)$$

where μ_k^0 is the current value of the k^{th} mean. It remains only to choose a value for $d_k^{(s)}$; good results have been obtained with

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

for $E = 1.0$ or 2.0 as recommended in [8]. Setting $E \geq 1.0$ also ensures that $\mathbf{M}_k$ is positive definite, which is necessary if $\mu_k = \mathbf{M}_k^{-1} \mathbf{v}_k$ is to be an optimal solution. Similar calculations reveal that the diagonal variance components can be obtained from

$$\sigma_{kn}^2 = \sum_s \left\{ \left(s_k^{(s)} - 2\, o_k^{(s)} \hat{\mu}_{kn}^{(s)} + c_k^{(s)} \hat{\mu}_{kn}^{(s)2} \right) \right.$$

$$\left. + d_k^{(s)} \left[\sigma_{kn}^{02} + \left(\hat{\mu}_{kn}^{0(s)} - \hat{\mu}_{kn}^{(s)} \right)^2 \right] \right\} \Big/ \sum_s (c_k^{(s)} + d_k^{(s)}) \tag{7}$$

where $\hat{\mu}_{kn}^{0(s)}$ is the n^{th} component of $\hat{\mu}_k^0 = A^{(s)} \mu_k^0$ and σ_{kn}^{02} is the current value of the variance. We can make the same FSAT approximation as before in accumulating the terms (4) and (7) have in common, which has proven to work well in practice.

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.

ESST Experiments

For our first set of recognition experiments, training was conducted on the English Spontaneous Scheduling Task (ESST) corpus, which contains approximately 35 hours of speech contributed by 242 speakers. For these experiments, we used a relatively small baseline model with eight (8) Gaussians for each of 2,340 codebooks, giving a total of 535 Gaussians per hour of training speech, which is well within the optimal range reported in [8]. These initial experiments were intended to establish the effectiveness of our algorithms on a tractably–sized training set, to provide for fast experimental turnaround.

All speech data was digitally sampled at a rate of 16 kHz. The speech features used for all experiments were obtained by estimating 13 cepstral components, along with their first and second differences, then performing linear discriminant analysis to obtain a final feature of length 32. 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.

Unsupervised speaker adaptation for all test conditions requiring it, was performed on the errorful test set transcriptions obtained with the baseline recognizer. MLLR parameter estimation was conducted by iterating twice over the test set. In all cases, only a single transformation was used.

Before beginning conventional MMI training, it was first necessary to write word lattices for all utterances in the training set [13]; this was accomplished with the baseline HMM, an 18,000 word vocabulary covering the training set, and a bigram language model trained on the ESST corpus. As suggested in [8], the acoustic log-likelihood scores calculated during MMI training were scaled by a factor of 1/15, before being combined with the unscaled log-likelihood returned by a *unigram* language model in order to calculate the posterior probability of any word appearing in the recognition lattice. Once the posterior probability of a given word in the lattice was known, the best state sequence based on fixed start and end times was calculated, and the statistics required for parameter re-estimation were accumulated for only those states lying on this Viterbi path.

ML-based speaker-adapted training was begun with the baseline conventional ML model described above and continued for three iterations over the training set. The final MLE-SAT acoustic model and speaker-dependent MLLR parameters were used as the starting point for MMI-SAT. The mean and variance parameters were updated during MMI-SAT as described in Section 2 using word lattices written with the adapted MLE-SAT acoustic model; the MLLR parameters for the training set speakers were *not* re-estimated during MMI-SAT. The same approach was used for unsupervised parameter estimation on the test set: speaker-dependent MLLR parameters were estimated with the MLE-SAT acoustic model, and then used with both the MLE-SAT and MMI-SAT acoustic models for lattice rescoring.

Shown in Figure 1 are the results of our initial experiments on the ESST test set, which contains 22,889 total words. The result labelled "MMI" was obtained with an conventional MMI-trained system without any speaker adaptation apart from VTLN. The "MLE with MLLR" result was obtained by using unsupervised MLLR adaptation on the MLE baseline model; the "MMI with MLLR" result was obtained by using unsupervised MLLR adaptation on the MMI model. The "MLE-SAT" result was obtained by performing MLLR adaptation on the MLE-SATrained model. Finally, the result labelled "MMI-SAT" was obtained by using the adaptation parameters estimated with the MLE-SATrained acoustic model to transform the means of the MMI-SATrained acoustic model.

System	% Word Error Rate
MLE Baseline	29.38
MMI	27.47
MLE with MLLR	25.53
MMI with MLLR	25.54
MLE-SAT	24.60
MMI-SAT	23.65

Fig. 1. Word error rates obtained with the small system on the ESST test set.

From these results, it is clear that the performance gain obtained using conventional MMI- instead of MLE-training does not carry over to the case when MLLR adapation is used, a fact that is rather surprising in light of what has been reported previously [8]. This gain does, however, carry over when an MMI-SATrained system is transformed with the MLLR parameters estimated with the MLE-SAT model.

On the small training set, we were also able to compare system performance after various numbers of training iterations and with various values of the constant E. These results are shown in Figure 2, from which it is clear that optimal performance with conventional MMI training is obtained after two iterations, whereas MMI-SAT requires three or more iterations. Moreover, a convergence value of $E = 1.0$ is adequate, a larger value only slows convergence.

System	% Word Error Rate		
	Training Iteration		
	1	2	3
MLE Baseline	29.38		
MMI, $E = 1.0$	27.93	27.53	27.89
MMI, $E = 2.0$	28.32	27.66	27.47
MMI-SAT, $E = 1.0$	24.09	23.87	23.65
MMI-SAT, $E = 2.0$	24.14	23.86	23.85

Fig. 2. Comparison of system performance versus number of training iterations.

Broadcast News and Meeting Room Experiments

In a second set of experiments, 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 [14]. For these experiments, our baseline recognizer was comprised of 4,144 continuous density codebooks, each of which contained 16 Gaussians, for a total of 670 Gaussians per hour of training speech.

For the systems trained on the combined BN-ESST set, we used the same feature extraction as before, save that the final feature length after LDA analysis was 40 instead of 32.

Once more three iterations of MLE-SAT were conducted on the training set. Using the FSAT approximation described in Section 2, less than 700 Mb of hard disk space was required to store all speaker-dependent forward-backward statistics needed for mean and covariance re-estimation.

Shown in Figure 3 are the results of our initial speech recognition experiments on the BN and MR test sets. 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. Apart from this, the labels for the several results have the same meanings as those in Figure 1. In these experiments, it was found that best performance was obtained after two iterations of MMI-SATraining.

System	% Word Error Rate	
	BN	MR
MLE Baseline	23.6	45.3
MMI	22.6	44.6
MLE with MLLR	20.8	43.6
MLE-SAT	19.9	42.2
MMI-SAT	18.9	40.2

Fig. 3. Word error rate results on the 1998 Broadcast News evaluation set (BN) and the Interactive Systems Laboratories Meeting Room set (MR).

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 three 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.

Further work is needed to refine the techniques presented here: The time required to conduct system training is fairly large, but could be greatly reduced through a more efficient implementation of the lattice forward-backward routines. Moreoever, it is possible that larger performance gains could be obtained by using multiple MLLR transforms for each speaker instead of the global transforms used for all experiments presented here, or by using a different linear transform for speaker adaptation. All of these issues are topics for further research.

Acknowledgments: The authors wish to acknowledge Hagen Soltau, Florian Metze, Christian Fügen, and Ivica Rogina for their aid in developing the source code and running the experiments described in this work, as well as Phil Woodland and Dan Povey for sharing their insight into MMI parameter re-estimation.

5. REFERENCES

[1] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained estimation of gaussian mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 357–366, 1995.

[2] 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.

[3] M. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, vol. 12, 1998.

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

[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] P. Gopalakrishnan, D. Kanevsky, A. Nádas, and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Transactions on Information Theory*, vol. 37, pp. 107–113, January 1991.

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

[8] 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.

[9] L. Uebel and P. Woodland, "Improvements in linear transform based speaker adaptation," in *Proc. ICASSP*, 2001.

[10] J. Zheng, J. Butzberger, H. Franco, and A. Stolcke, "Improved maximum mutual information estimation training of continuous density hmms," in *Proc. Eurospeech*, 2001.

[11] 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.

[12] S. Matsoukas, R. Schwartz, H. Jin, and L. Nguyen, "Practical implementations of speaker-adaptive training," in *Proc. ICASSP*, 1996.

[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.

[14] 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.