interspeech 2008 incorporating SST 08

Speaker Adaptive Training Using Shift-MLLR

Jonas Lööf, Christian Gollan, Hermann Ney

Lehrstuhl für Informatik 6 - Computer Science Dept.
RWTH Aachen University, Aachen, Germany
{loof,gollan,ney}@cs.rwth-aachen.de

Abstract

In this paper a novel method for speaker adaptive training (SAT), based on Gaussian mean offset adaptation, so called Shift-MLLR, is presented. The method differs from previous SAT methods, where linear transformations of Gaussian means or features are utilized, in that only an offset vector is used for adaptation, but instead the number of regression classes is increased. This is shown to allow an efficient implementation. Furthermore, the use of word posterior confidence measures for Shift-MLLR is investigated, also in combination with the proposed SAT method. The presented methods are integrated into a state of the art speech recognition system, and performance is contrasted with Shift-MLLR without SAT, as well as with MLLR. Large and consistent improvements in word error rate are observed from the new SAT method, as well as from confidence based Shift-MLLR. The combination of the new speaker adaptive training method with confidence based estimation show consistent improvements.

Index Terms: speech recognition, speaker adaptation, speaker adaptive training

1. Introduction

Speaker adaptation is an important method for improving acoustic models in speech recognition. For each speaker encountered in recognition, the acoustic model is refined by taking into account acoustic data from that particular speaker. Widely used is affine transform based maximum likelihood (ML) model adaptation, especially the so called maximum likelihood linear regression (MLLR) method [1], where the acoustic model is adapted by applying affine transforms to the means of the Gaussian emission models.

Speaker adaptive training (SAT) is an important method to maximize the performance gains from speaker adaptation. While speaker adaptation already compensates for speaker differences during recognition, the idea in SAT is to do the same during acoustic model training. This is especially important when using training data with large diversity in speakers and recording conditions. Speaker adaptive training has been showed to yield important improvements to the quality of the acoustic model, see [2] for an overview of the topic.

MLLR based speaker adaptive training was originally proposed in [3], and shown to yield improvements in recognition performance. In its original form it requires a large amount of memory compared to standard acoustic model training, and is not straightforward to combine with discriminative training. Due to these issues, the use of MLLR based SAT is not common. In contrast to this, SAT using so called constrained MLLR (CMLLR), while delivering approximately the same performance improvement in equivalent conditions [4], is straightforward to combine with standard (ML), and discriminative training, especially in the case of a single global transform (per speaker) where it can be performed completely in feature space. Due to this, its inclusion in state of the art systems is common, while MLLR based SAT is hardly used to our knowledge.

Applying CMLLR adaptation with multiple (state dependent) regression classes can not be done completely in feature space, since the choice of transformation matrix depends on the actual state used in the acoustic model. Thus, to make use of such a setup in speaker adaptive training, the actual core training algorithm needs to be slightly modified. Probably due to this, multiple regression classes are not normally used with CMLLR SAT in state of the art systems. In [4], multiple regression classes were used only in recognition.

2. Shift-MLLR

In MLLR adaptation, the model means are adapted using an affine transform, that is a combination of a linear transform (deformation and rotation) and an additive offset. The means of the acoustic model are organized into (state dependent) regression classes, where the means in the same regression class share an adaptation matrix. The use of multiple regression classes is in effect a nonlinear aspect of the model, although limited since the number of classes are typically low.

In [5], this nonlinear modeling aspect of regression classes was further investigated. Instead of using an affine transform per regression class, only a simple offset was utilized. Since this is more robustly estimable, the number of regression classes could be dramatically increased. The number of regression classes was chosen dynamically as in tree based MLLR [6]. The authors of [5] call their approach Shift-MLLR, and this term is used in this paper for the combination of mean offset adaptation with regression classes chosen as in tree MLLR. The recognition performance of this setup was shown to be practically identical to that of tree-based MLLR, when both methods were combined with CMLLR based speaker adaptive training. The use of offsets (biases) on features or mean vectors for adaptation is not new, and has been utilized in [7] for condition adaptation.

In this work the approach suggested in [5] is used. The offsets are formulated as a model transform, a speaker dependent additive offset $b_{r,c}$ applied to each mean vector μ_s,

$$\mu'_s = \mu_s + b_{r,c}, \tag{1}$$

where s the state, r is the speaker, and c is a state dependent regression class.

The estimation of the offset vectors are done with maximum likelihood, using expectation maximization (EM). The auxiliary function with respect to both adaptation- and acoustic model

parameters is given by

$$Q(M,\hat{M}) = -\frac{1}{2}\sum_{t=1}^{T}\sum_{s=1}^{S}\gamma_s(t)\Big[\log|\hat{\Sigma}_s| + (x_t-\hat{\mu}_s-\hat{b}_{r,c})^T\hat{\Sigma}_s^{-1}(x_t-\hat{\mu}_s-\hat{b}_{r,c})\Big]. \quad (2)$$

Keeping acoustic model parameters fixed results in a closed form re-estimation update formula for the offset $b_{r,c}$;

$$\hat{b}_{r,c} = \Bigg(\sum_{t=1}^{T}\sum_{s\in c}\gamma_s(t)\Sigma_s^{-1}\Bigg)^{-1}\sum_{t=1}^{T}\sum_{s\in c}\gamma_s(t)\Big[\Sigma_s^{-1}(x_t-\mu_s)\Big], \quad (3)$$

where the summation over states go over all states belonging to regression class c. Using diagonal covariance matrices, this expression reduces to a component-wise form. With a single globally pooled diagonal covariance, as used in the present work, the formula simplifies to

$$\hat{b}_{r,c} = \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in c}\gamma_s(t)(x_t-\mu_s). \quad (4)$$

Furthermore, in the system used in the present work, the Viterbi approximation for EM estimation is utilized, and the estimation equation is further simplified by the fact that only one state is active at each time frame.

The number of regression classes are chosen dynamically using the same method as for tree MLLR [6], but with much fewer required observations per regression class, since offsets are more robust to estimate than affine transforms. The number of regression classes per speaker vary according to the amount of adaptation data; in the experiments performed as part of the present work the number mostly fall in the range between 500 and 2000 classes.

2.1. Use for Speaker Adaptive Training

Since Shift-MLLR shows performance equivalent to MLLR when used in recognition, it is especially attractive due to the simple form of the Shift-MLLR transforms to consider using it for speaker adaptive training. The hope is to gain the improvements of MLLR based SAT, while allowing for efficient implementation and combination with discriminative training methods.

To use Shift-MLLR for speaker adaptive training, we use Eq. (2) to derive re-estimation equations for the model parameters, with adaptation offsets kept fixed. This can be done in complete analogy to the standard ML acoustic model case, and result in

$$\hat{\mu}_s = \frac{1}{T}\sum_{t=1}^{T}\gamma_s(t)(x_t-b_{r,c}), \quad (5)$$

$$\hat{\Sigma}_s = \frac{1}{T}\sum_{t=1}^{T}\gamma_s(t)(x_t-b_{r,c}-\hat{\mu}_s)(x_t-b_{r,c}-\hat{\mu}_s)^T. \quad (6)$$

As can be seen, the form of the re-estimation equations are similar to those of the non SAT case, with the difference that the offset $b_{r,c}$ is subtracted from the feature vector x_t.

Due to this similarity, the needed modifications to the training software are limited. When using Viterbi approximation, as is the case in the RWTH system, only one state is active for each time frame, and the regression class can be chosen from the frame state alignment, and the offset directly applied to the feature vector. This allows isolating the difference between the SAT training and the standard acoustic model training to the feature extraction, thus putting the complete implementation in the feature extraction front end, leaving the actual training software unmodified.

2.2. Target Model for SAT

In classical SAT training, as presented in [3], acoustic model parameters, and adaptation parameters are jointly estimated on the training set, using interleaved re-estimation. In [8] alternatives to this approach are investigated, and a variant where the adaptation parameters are estimated once, followed by a complete re-training of the acoustic model parameters from scratch was demonstrated to yield better performance. It was also demonstrated that better performance was achieved by not using the real, optimal (up until that time) acoustic model in the estimation of the adaptation parameters, but instead use a coarser model, a so called simple target model.

As a tentative explanation or at least motivation for this result it is suggested that since a large, fully trained, acoustic model already captures much of the speaker variation, less room is left for improvements from adaptation. Since this results in adaptation of lower quality, the subsequent re-estimation of the model based on this adaptation also is of lower quality. When starting from a less complex model, on the other hand, almost no speaker specific information is contained in the model, leading to adaptation of better quality, and thus a better final model.

In this work, the approach using complete acoustic model retraining after estimating the adaptation offsets is used for all experiments. Different target models of varying complexity are used.

2.3. Estimation Using State Posterior Confidence Measures

In unsupervised adaptation, as typically used in state of the art transcription systems, the adaptation parameters are estimated using the output from a previous unadapted recognition step as ground truth. This means that the data on which the parameters are optimized will contain errors. One strategy to counter this is to apply confidence measures to select (or to weight) what portions of the automatic first pass transcription should be used for the estimation.

Many publications have shown that the application of confidence scores for adaptation estimation can improve recognition results. Small improvements for confidence based CMLLR adaptation is reported in [9]. In [10] the authors have investigated lattice-based MLLR applying a confidence threshold and report 2% relative improvement in word error rate (WER) over the 1-best transcription. In [11] 5% relative improvement is reported for MLLR adaptation by performing word confidence selection from the 1-best transcription.

In automatic speech recognition confidence scores can be developed and optimized for different units like utterances, words, phonemes or states. For acoustic model adaptation it makes sense to focus on the tied state label since distributions are associated with these units. Instead of rejecting an entire utterance or word, the system can use state confidence scores to select state-dependent data. State confidence scores are obtained from computing arc posteriors from the lattice output from the decoder. The arc posterior probabilities can be computed efficiently using the forward-backward algorithm as, for example, described in [12, 13]. For further details on the meth-

ods used for computation of state posterior confidence measures, see [14].

2.4. Combination with Discriminative Training

Due to the simple structure of the Shift-MLLR transformation, the combination of Shift-MLLR SAT with discriminative training is relatively straightforward. Instead of using an expectation maximization auxiliary function, the generalized auxiliary function as in normal discriminative training can be used. The difference compared to the normal re-estimation equations using common discriminative criteria, such as maximum mutual information (MMI) and minimum phone error (MPE) consist in the same subtraction of the offset $b_{r,c}$ from the feature vector as in the ML case.

While in the maximum likelihood case the use of Viterbi training allows to perform this subtraction in the front end, the discriminative criteria, even when using Viterbi approximation, require competing states to be active in the same time frame. Due to this, the implementation of Shift-MLLR SAT for discriminative criteria requires modifying the core discriminative training software to a limited extent. Experimental work on Shift-MLLR SAT with discriminative criteria is not included in the present work, and will be part of future work.

In CMLLR SAT, using discriminative criteria often only the model is re-estimated using the discriminative criterion, while the CMLLR matrix continues to be estimated using maximum likelihood - especially when unsupervised adaptation is to be performed in recognition; results presented in [15] show no advantage to using the discriminative criteria for adaptation estimation. To what extent this approach is preferable for Shift-MLLR SAT compared to also estimating the offsets using discriminative criteria, must be investigated experimentally.

3. Experimental Results

The recognition experiments were performed using one of the systems developed for the TC-STAR 2007 evaluation as baseline [16]. All experiments were performed on the English TC-STAR 2006 development and evaluation data sets. The acoustic training material includes 88 hours of manually transcribed recordings. The development and evaluation sets each consist of 3.2 hours of recordings. The system used a MFCC front-end augmented with a single voicedness feature, and the acoustic models used consisted of roughly 900k Gaussians sharing a single globally pooled covariance. Furthermore, in all experiments a one pass VTLN method, using a classifier for warping factor estimating, was used.

The baseline system furthermore included speaker adaptive training using dimension reducing feature transforms, as presented in [17], called FMLLRP here. FMLLRP SAT is similar to CMLLR SAT using one regression class, but was shown to give consistent improvements in error rate. Adaptation was performed unsupervised, using two or three recognition passes. For two-pass recognition both FMLLRP and Shift-MLLR or MLLR adaptors were estimated on the first pass output, while in three-pass recognition, only FMLLRP matrices were estimated using first pass output, and the output of a second pass FMLLRP SAT recognition was used to estimate Shift-MLLR or MLLR adaptors for a third recognition pass. Both MLLR and Shift-MLLR were performed using dynamic tree based regression classes.

Table 1 show word error rate (WER) for MLLR, Shift-MLLR and Shift-MLLR SAT, both using two pass and three pass setups. As can be seen there is a notable improvement due to speaker adaptive training using Shift-MLLR, especially when doing three pass recognition. Note that for MLLR, as well as for Shift-MLLR without SAT, there is no improvement from three pass recognition.

Table 1: *SAT shift results.*

	Dev06	Eval06
1st pass	16.3	13.2
FMLLRP SAT	14.2	10.5
MLLR	13.1	10.0
MLLR 3-pass	13.2	10.1
Shift	13.0	10.1
Shift 3-pass	13.2	10.2
SAT-Shift	12.5	9.9
SAT-Shift 3-pass	12.4	9.6

Table 2: *Target model influence.*

	Dev06	Eval06
Simple	12.4	9.6
Full	12.6	9.6

In Table 2, the influence of different types of target models for Shift-MLLR SAT, as discussed in Sec. 2.2 is presented. As expected from experiences with VTLN and CMLLR SAT, the simpler single Gaussian target model show slightly better performance, compared to the full model.

Table 3: *Results using state posterior confidences.*

	Dev06	Eval06
MLLR	13.2	10.1
MLLR Conf.	12.9	9.8
Shift	13.2	10.2
Shift Conf.	12.7	9.6
MAP	14.1	10.5
MAP Conf.	13.0	9.8
SAT-Shift Full	12.6	9.6
SAT-Shift Full Conf.	12.5	9.4
SAT-Shift Reduced	12.4	9.5
SAT-Shift Reduced Conf.	12.2	9.4
SAT-Shift Simple	12.5	9.5
SAT-Shift Simple Conf.	12.3	9.3

Table 3 summarizes the results (WER) when applying state confidence measures as discussed in Sec. 2.3 to adaptation estimation; Shift-MLLR with and without SAT, as well as MLLR. Since results in [14] show competitive results for confidence based maximum a posteriori adaptation (MAP), this is also included for comparison. Note that the experiments were all done using three pass adaptation. Here further experiments where conducted to investigate the influence of the target model. Three models were used: A single Gaussian unadapted model (Simple), a CMLLR SAT estimated model with approximately 70k Gaussians (Reduced), and the full CMLLR SAT model with about 900k Gaussians (Full). Due to the implementation of confidence measures in the system used for this experiment, the frame state alignments must be computed with the same acoustic model as used to compute confidence measures, where the

full model should be used. To remain consistent, for all results in Table 3 the full model was used for the alignment in Shift-MLLR estimation, both in training and recognition, while the actual target model was used in accumulation. This explains the slight differences in results compared to the previous experiment.

Significant improvements from confidence measures can be seen for the cases of MLLR as well as for Shift-MLLR without SAT. The improvements from confidence measures are consistently larger in the case of Shift-MLLR compared to MLLR; this is consistent with results from [14], where MAP adaptation was shown to benefit more from confidence measures than MLLR. When combining confidence based estimation with Shift-MLLR SAT, the improvements are small but consistent. The effect of using different target models remains limited, but it appears that the use of the full acoustic model as target model for SAT is not optimal.

4. Discussion

It is clear from the presented results that Shift-MLLR SAT represents an attractive method for improving a state of the art speech recognition system. The large improvements achieved - in the same range as improvements from MPE discriminative training (c.f. [16], the system used in the current work is subsystem S1) - combined with the ease of implementation, makes the presented method a good candidate for inclusion into many transcription systems.

About equally large improvements as those of Shift-MLLR SAT, was achieved by using confidence measures for Shift-MLLR estimation. The combination of the two methods has yet to deliver more than small, though consistent, improvements. The large improvements achieved in the non SAT case leads to the hope that further investigations of modified approaches will show advantages.

Although the combination of Shift-MLLR SAT and discriminative training was not experimentally investigated in this work, such a combination poses no theoretical or practical difficulties (c.f. Sec. 2.4). Furthermore, experience from the combination of CMLLR SAT and discriminative training, where the improvements are known to be essentially additive, encourages further investigation in this direction.

5. Conclusions

In this work, a novel method for speaker adaptive training using Shift-MLLR was presented. Re-estimation equations for the adaptation and for the acoustic model based on expectation maximization were presented, and an efficient implementation was described. The use of state posterior confidence measures in combination with Shift-MLLR and the proposed method was described. Results were presented combining Shift-MLLR SAT with CMLLR SAT, and its performance was contrasted with that of Shift-MLLR without SAT, as well as with MLLR, and large improvements in word error rate were observed. Results using confidence measures for Shift-MLLR also showed large improvements. When combining confidence measures with Shift-MLLR SAT consistent improvements were still observed.

Future work includes further investigations into the combination of Shift-MLLR SAT with confidence measures. Furthermore the combination of Shift-MLLR SAT with discriminative training needs to be investigated in detail. Though the influence of the target model remains inconclusive, further improvements might be achieved by choosing or estimating a target model that is optimal for SAT estimation. One limitation of the present work is that no comparison or combination is performed between Shift-MLLR SAT and CMLLR SAT using multiple regression classes. This could be rectified in future work.

6. References

[1] 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, no. 2, pp. 171 – 185, Apr. 1995.

[2] M. J. F. Gales, "Adaptive training for robust ASR," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Italy, Dec. 2001, pp. 15 – 20.

[3] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. Int. Conf. on Spoken Language Processing*, vol. 2, Philadelphia, PA, USA, Oct. 1996, pp. 1137 – 1140.

[4] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75 – 98, Apr. 1998.

[5] D. Giuliani and F. Brugnara, "Acoustic model adaptation with multiple supervisions," in *Proc.* TC-STAR *Workshop on Speech-to-Speech Translation*, Barcelona, Spain, Jun. 2006, pp. 151–154.

[6] C. Leggetter and P. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," in *Proc. ARPA Spoken Language Technology Workshop*, Austin, TX, USA, Jan. 1995, pp. 104 – 109.

[7] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 3, pp. 190–202, May 1996.

[8] G. Stemmer, F. Brugnara, and D. Giuliani, "Adaptive training using simple target models," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Philadelphia, PA, USA, Mar. 2005, pp. 997 – 1000.

[9] T. Anastasakos and S. V. Balakrishnan, "The use of confidence measures in unsupervised adaptation of speech recognizers," in *Proc. Int. Conf. on Spoken Language Processing*, vol. 6, Sydney, NSW, Australia, Dec. 1998, pp. 2303–2306.

[10] M. Padmanabhan, G. Saon, and G. Zweig, "Lattice-based unsupervised MLLR for speaker adaptation," in *Proc. ISCA Automatic Speech Recognition Workshop*, Paris, Sep. 2000, pp. 128–132.

[11] M. Pitz, F. Wessel, and H. Ney, "Improved MLLR speaker adaptation using confidence measures for conversational speech recognition," in *Proc. Int. Conf. on Spoken Language Processing*, vol. 4, Beijing, China, Oct. 2000, pp. 548 – 551.

[12] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 3, pp. 288 – 298, Mar. 2001.

[13] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," in *Proc. European Conf. on Speech Communication and Technology*, vol. 2, Rhodes, Greece, Sep. 1997, pp. 827 – 830.

[14] C. Gollan and M. Bacchiani, "Confidence scores for acoustic model adaptation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Las Vegas, NV, USA, Mar. 2008, pp. 4289 – 4292.

[15] L. Wang and P. C. Woodland, "Discriminative adaptive training using the MPE criterion," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, St. Thomas, U.S. Virgin Islands, Dec. 2003, pp. 279 – 284.

[16] J. Lööf, C. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, C. Plahl, D. Rybach, R. Schlüter, and H. Ney, "The RWTH 2007 TC-STAR evaluation system for European English and Spanish," in *Proc. Int. Conf. on Spoken Language Processing*, Antwerp, Belgium, Aug. 2007, pp. 2145 – 2148.

[17] J. Lööf, R. Schlüter, and H. Ney, "Efficient estimation of speaker-specific projecting feature transforms," in *Proc. Int. Conf. on Spoken Language Processing*, Antwerp, Belgium, Aug. 2007, pp. 1557 – 1560.