Efficient Estimation of Speaker-specific Projecting Feature Transforms

Jonas Lööf, Ralf Schlüter, and Hermann Ney

Lehrstuhl für Informatik 6, Computer Science Department
RWTH Aachen University, 52056 Aachen, Germany
{loof,schluter,ney}@cs.rwth-aachen.de

Abstract

This paper introduces a new, efficient approach for estimating projecting feature transforms for speech recognition. It is based on the MMI′ criterion, a likelihood ratio criterion motivated by a simplification of the MMI criterion, and is shown to be closely related to HLDA. In comparison to current methods, the new method is faster, making it more suitable for speaker adaptive training, where the number of speakers, and therefore the number of transforms are substantial.

The proposed method was integrated into the RWTH parliamentary speeches transcription system. Experimental results are presented using speaker specific projecting transforms, both when used in recognition only and when used for speaker adaptive training, showing consistent improvements. Furthermore, the observed improvements are shown to be additive to the improvement of MLLR. Comparisons to DLT are presented, and results are presented for a new projecting DLT method.

Index Terms: Speech recognition, speaker adaptation, heteroscedastic linear discriminant analysis

1. Introduction

Several approaches to using projecting transforms for speaker adaptation have been proposed. This section gives an overview of the proposed methods.

The maximum likelihood (ML) estimation of linear transformations, as described in [1], requires computing the Jacobian of the transform. For projecting matrices, the algorithms presented are unsuitable in unmodified form.

One method, presented in [2], is derived from the criterion used for heteroscedastic linear discriminant analysis (HLDA) [3], an extension of linear discriminant analysis (LDA) that allows class specific covariances. In this method, as in the linear transform estimation algorithm presented in [1], the matrices are estimated from accumulated statistics using a row-wise iterative update algorithm. One drawback of the method is that it requires inverting an $n \times n$ matrix per matrix row and iteration, where n is the dimension of the untransformed features.

Another approach, presented in [4], involves estimating a speaker specific non-projecting matrix in a high dimensional feature space, before applying a global dimension reducing transform. The speaker specific matrices are estimated on a single Gaussian model with full covariances. Although substantial improvements were achieved, it was concluded that to a large part this is due to using the single Gaussian model in estimation. This conclusion is supported by the improvements obtained in [5] when using a single Gaussian model to estimate the matrices in (non-projecting) speaker adaptive training (SAT). To what extent the improvement is to attribute to the speaker dependent projection is not clear from the results presented. Furthermore, since the adaptation target model uses full covariances a general optimization method was used instead of the row-wise iteration of [1].

This work was partially funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738, http://www.tc-star.org)

2. Estimation of Projections Using MMI′

One possible way to estimate projecting feature transforms would be to use a standard discriminative criterion such as maximum mutual information (MMI). For non-projecting mean transformation estimation, results show that this requires interpolation with an ML estimated matrix to be useful [6]. See Sec. 3 for details on estimation of projecting feature transforms using this approach.

One criterion that has proven to be useful for optimizing parameters in the signal processing front-end [7], including projecting feature transforms is a likelihood ratio criterion motivated in [7] as a simplification of the MMI criterion. Some of the derivations below were first presented in a previous work [8], where the MMI′ criterion was used as an evaluation metric for VTLN linear transforms.

Starting with the MMI criterion, the competing model is exchanged with a single full covariance Gaussian model that is optimized (using maximum likelihood) on the same data as the transformation. Specifically for the case of a feature transform W the objective function is

$$g_{\text{MMI}'}(M, W) = \\ -\log P(Wx_1^T|\mu', \Sigma') + \log P(Wx_1^T|M, w_1^N). \quad (1)$$

where x_1^T is the sequence of untransformed feature vectors, w_1^N the sequence of words in the transcription, M the acoustic model parameters, and Σ' and μ' the parameters of the competing Gaussian.

The resulting criterion is called the MMI′ criterion. In [7] this criterion was used in a direct optimization framework, using multiple passes over the training data to compute the objective function and its derivative. On the other hand, the close formal similarity to the standard ML criterion allows for the use of the EM algorithm by defining an auxiliary function, in exact correspondence to the ML case.

Like in the ML derivation, start by forming the difference between $g_{\text{MMI}'}$ for two different transformations W and $\hat{W}$ keeping other parameters fixed.

$$g_{\text{MMI}'}(M, \hat{W}) - g_{\text{MMI}'}(M, W) = \\ -\log \frac{P(\hat{W}x_1^T|\mu', \Sigma')}{P(Wx_1^T|\mu', \Sigma')} + \log \frac{P(\hat{W}x_1^T|M, w_1^N)}{P(Wx_1^T|M, w_1^N)} \quad (2)$$

The second term is equivalent to the difference in log likelihoods formed in the proof for EM estimation of hidden Markov models (HMMs), and can be handled exactly as in that proof, while the term corresponding to the competing Gaussian lack state dependence. Combining all terms give

$$g_{\text{MMI}'}(W,\hat{W}) - g_{\text{MMI}'}(W,W) = \tag{3}$$
$$\sum_{s=1}^{S}\sum_{t=1}^{T}\gamma_s(t)\big[-\log\frac{P(\hat{W}x_t|\mu',\Sigma')}{P(Wx_t|\mu',\Sigma')} + \log\frac{P(\hat{W}x_t|M_s)}{P(Wx_t|M_s)}\big],$$

plus a term guaranteed to be larger than 0 from the EM proof. $\gamma_s(t)$ are state occupation probabilities for state s at time t that are dependant on W, and M_s denote the model parameters in state s. Further rearranging the expression makes it possible to rewrite Eq. (3) as $Q(W,\hat{W}) - Q(W,W)$, were Q is the auxiliary function. Since the competing Gaussian model is trained on the same data as the matrix, the resulting term can be simplified;

$$P(\hat{W}x_t|\mu',\Sigma') = \frac{1}{2}\log\big|\Sigma'\big|. \tag{4}$$

Thus, the auxiliary function is given by

$$Q(W,\hat{W}) =$$
$$\sum_{s=1}^{S}\sum_{t=1}^{T}\gamma_s(t)\big[\frac{1}{2}\log\big|\Sigma'\big| + \log P(\hat{W}x_t|s,M_s)\big]. \tag{5}$$

Finally, expressing Eq. (5) in terms of sufficient statistics leads to

$$Q'(W,\hat{W}) = \tag{6}$$
$$\frac{T}{2}\log\big|\hat{W}\Sigma\hat{W}^T\big| - \frac{1}{2}\sum_{d=1}^{D}\big(\hat{w}_d G^{(d)}\hat{w}_d^T - 2\hat{w}_d k^{(d)^T}\big),$$

where $\hat{w}_d$ is the dth row of $\hat{W}$, and terms constant with respect to the transform $\hat{W}$ have been omitted. Σ is the full covariance of the untransformed adaptation data, while G and k are statistics as defined in [1], i.e.

$$G^{(d)} = \sum_{s=1}^{S}\frac{1}{\sigma_d^{(s)2}}\sum_{t=1}^{T}\gamma_s(t)x_t x_t^T \tag{7}$$

$$k^{(d)} = \sum_{s=1}^{S}\frac{1}{\sigma_d^{(s)2}}\mu_d^{(s)}\sum_{t=1}^{T}\gamma_s(t)x_t^T, \tag{8}$$

Using these equations, EM optimization can be carried out in exactly the same way as for non-projecting linear transforms by iteratively optimizing $\gamma_s(t)$ using the forward backward algorithm, and $\hat{W}$ by accumulating the sufficient statistics and optimizing Q. It should be noted that Q is a strict auxiliary function, and not as in the case of extended Baum Welsh a generalized one, and thus improvement in Q guarantee improvement in the objective function. To be able to estimate an affine transform, as opposed to the linear case presented in the formulas, the feature vectors are extended with a constant dummy feature, set to one. Since the equations are derived for projecting transforms they are still valid.

To optimize Q the auxiliary function is differentiated with respect to $\hat{w}_i$ yielding

$$\frac{\partial Q'(W,\hat{W})}{\partial\hat{w}_i} = T\big[\Sigma\hat{W}^T\big(\hat{W}\Sigma\hat{W}^T\big)^{-1}\big]_i - \hat{w}_i G^{(i)} + k^{(i)}, \tag{9}$$

where the subscript i on the bracketed term indicates taking the ith *column* of the term.

Using this expression combined with Eq. (6) allows using a general optimization algorithm to calculate the matrix. The approach chosen in this work though, was to derive a row-wise iterative estimation algorithm similar to the one introduced in [1] for non-projecting transformations.

In [1] the row-wise update is guaranteed to increase the objective function, since a closed form solution for $\hat{w}_i$ is possible. For the current problem no closed form solution can be found, but convergence still occur in practice. An argument to why this should be the case can be outlined as followed.

Equating Eq. (9) to zero and rewriting leads to a row-wise update formula for $\hat{W}$,

$$\hat{w}_i = \big[T\big[\Sigma\hat{W}^T\big(\hat{W}\Sigma\hat{W}^T\big)^{-1}\big]_i + k^{(i)}\big]\big(G^{(i)}\big)^{-1}. \tag{10}$$

This describes a fixed point map for $\hat{w}_i$, and also for the complete $\hat{W}$ if all rows except i are kept fixed. In practice, a modified row update similar to the one used for CMLLR in [1] is used,

$$\hat{w}_i = \big[\alpha T\big[\Sigma\hat{W}^T\big(\hat{W}\Sigma\hat{W}^T\big)^{-1}\big]_i + k^{(i)}\big]\big(G^{(i)}\big)^{-1}, \tag{11}$$

where α is computed as in [1]. To show convergence, one would have to show that Eq. 11 describes a contracting map. Although no proof of convergence is known, in practice the iteration converges and provides a convenient way of finding the zeros of Eq. (9).

The MMI′ criterion is closely related to the HLDA criterion and other linear discriminant criteria. The first term of Eq. (6) is the total scatter matrix, with LDA/HLDA terminology. The second term is the within class term, and if the target model is trained on the data used for estimation, the same expression as in HLDA is reached for this term. Although the criteria are very similar, the total scatter term is not exactly the same. From the discussion in [9] we see that the original HLDA uses the determinant of the diagonal of the total scatter matrix, whereas the MMI′ criterion uses the determinant of the full scatter matrix. In practice there seems to be no noticeable difference. Tests during development have shown that matrices estimated using the presented method are identical, within numerical precision, to those produced using the FMLLR-P (projecting feature space maximum likelihood linear regression) adaptation method [2], that is based on the original HLDA criterion.

3. Discriminative Linear Transforms

One well known adaptation method is the discriminative linear transform (DLT) method [6], based on an H-smoothed MMI criterion,

$$g_{\text{H}}(M,W) = (1-H)g_{\text{ML}}(M,W) + Hg_{\text{MMI}}(M,W). \tag{12}$$

When $H = 1$, the criterion is equivalent to the MMI criterion, while $H = 0$ leads to the ML criterion. The smoothing introduced has two purposes, both as a general smoothing of the criterion, but most importantly to ensure that convergence of the auxiliary function leads to convergence in the objective function. In [6], the criterion is used to estimate Gaussian mean transforms, but the same approach can also be used for feature transforms.

In the case of CMLLR the smoothing will be done with the ML criterion, but for the projecting case MMI′ will be used.

The criterion can be rewritten (for both the CMLLR and the projecting case) as

$$g_{\text{H}'}(M,W) = g_{\text{num}}(M,W) - Hg_{\text{den}}(M,W), \qquad (13)$$

where g_{num} is the ML (or MMI′) objective function, and g_{den}, is the same objective function but accumulated over a denominator lattice. Note that also g_{den} includes the term from the Jacobian (or the MMI′ contrast model term); when H=1 it will cancel with the Jacobian from g_{num} so that the pure MMI case contains no Jacobian. Analogous to the original DLT the estimation of the matrices can be carried out exactly as in Sec. 2 but with the accumulators defined as

$$G^{(d)} = G^{(d)}_{\text{num}} - HG^{(d)}_{\text{den}} \qquad (14)$$

$$k^{(d)} = k^{(d)}_{\text{num}} - Hk^{(d)}_{\text{den}}, \qquad (15)$$

and T defined as $(1-H)T$.

In all experiments for this paper two-gram lattices were used, and the posteriors were computed using a one-gram language model as proposed in [6]. The posterior were smoothed with the inverse language model scale.

4. Implementation Considerations

The proposed method can be efficiently implemented, using the update rule defined in Sec. 2. In the following discussion n will denote the length of the original feature vector, and p the size of the projected one.

For the proposed method the cost of the row update is dominated by the inversion of $\hat{W}\Sigma\hat{W}^T$ in Eq. (11), a $p \times p$ matrix. For this to be the case, care must be taken to each iteration only update the portions of $\hat{W}\Sigma\hat{W}^T$ and $\Sigma\hat{W}^T$ that has changed. Thus the cost of one row update is $O(p^3)$, and the cost of one complete iteration over the rows is $O(p^4)$.

In contrast, for the FMLLR-P method [2], the matrix inverted is the extended transformation matrix; an $n \times n$ matrix. Additionally, the iterations are performed over the rows of the extended matrix. Combined this leads to a time complexity of $O(n^4)$ for one iteration over all the rows.

Since informal tests show that both methods require approximately the same number of iterations to converge, the new method clearly has an advantage, especially when n is much larger than p. For the system used in this paper, having $n = 153$ and $p = 45$, a comparison showed a significant improvement in run time per matrix estimation; from 8 minutes for FMLLR-P to 20 seconds for the proposed method when using 200 iterations. This makes a large difference when performing speaker adaptive training.

5. Experimental Results

All recognition experiments were performed on the TC-STAR project English EPPS corpora as used in the 2006 evaluation, and the experiments were performed with systems developed for the 2006 evaluation as baseline [10]. The training material includes 88 hours of manually transcribed recordings. The tests were performed on the development and evaluation sets, each consisting of 3.2 hours. The system used a MFCC front-end augmented with a single voicedness feature, and the 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 to estimate the warping factor, was used. For the following experiments, maximum likelihood trained models were used, although discriminatively trained models were used in the evaluation. For further details of the baseline systems used, see [10].

Table 1: Recognition performance

CMLLR	MLLR	SAT	Projection	EPPS 2006 WER [%] Dev	Eval
no	no	no	no	16.4	13.5
yes				15.1	11.9
			yes	15.0	11.7
		yes	no	14.4	11.0
			yes	14.4	10.8
	yes	no	no	14.0	11.0
			yes	13.8	11.0
		yes	no	13.6	10.6
			yes	13.3	10.4

The baseline systems used LDA to map from a higher dimensional feature space to lower dimension; since the system utilizes a globally pooled covariance, HLDA has no advantage over LDA. The dimension of the LDA matrix was 45×153. When standard CMLLR was used, it was applied after the LDA transform. When using the new projecting method, the speaker-specific matrix was applied instead of the LDA matrix. The method were used as an affine transform, i.e. a dummy feature was added to the input dimension of the transform. When estimating the projecting transforms, the LDA was used as the starting point for the iteration.

Experiments were performed comparing the performance of the projecting adaptation matrices to standard non-projecting feature transformations, with and without using speaker adaptive training, and with and without combining with maximum likelihood linear regression (MLLR) model adaptation. All adaptation was *unsupervised*, i.e. performed with the first pass recognition output as ground truth. The speaker identity information on the recognition corpora was provided by a segment clustering algorithm. For the speaker adaptive training, the approach suggested in [5] was used.

Table 1 summarizes the results of these experiments. The baseline is the system with no adaptation except VTLN added. *CMLLR* denotes that a single affine feature transform was utilized per speaker. *Projection* indicates that the transform is projecting, estimated using the new method introduced in this paper. *SAT* means that feature transforms were estimated also on the training set and models estimated on the speaker-normalized features. In the case of using SAT, for both the training- and recognition set speakers, the matrices were estimated using a single Gaussian acoustic model. *MLLR* indicates that the model was further adapted using MLLR speaker wise.

Further experiments were performed comparing the performance of the projecting transform adaptation method to CMLLR based DLT. This included experiments using the projecting DLT as presented in Sec. 3. Table 2 shows the results of the CMLLR based DLT for different number of iterations and different values of the constant H, on the 2006 development set. Table 3 shows in the same way the experiments using projecting DLT. The column for $H = 0$ represents ML (or MMI′ for projecting matrices.) The results show that the improvements of DLT, both for the CMLLR case, but especially for the projecting transforms are very small, and also that the performance vary depending on the parameters. It seems unfeasible

Table 2: CMLLR based DLT

I	$H=$ 0	0.33	0.50	0.66	0.83	0.95
1	15.2	15.1	15.1	15.0	14.9	15.0
2	15.0	15.0	15.0	14.9	15.0	14.8
3	15.0	15.0	14.9	14.9	14.9	14.9
4	15.0	15.1	14.9	15.0	15.0	15.0
5	15.0	15.1	15.0	15.0	14.8	15.1

Table 3: Projecting DLT

I	$H=$ 0	0.33	0.50	0.66	0.83
1	15.0	15.0	15.0	15.0	14.9
2	14.8	14.8	14.8	14.8	15.0
3	14.9	14.8	14.8	14.7	14.9
4	14.7	14.8	14.7	14.8	14.9
5	14.8	14.7	14.7	14.7	14.8

to use DLT with success for unsupervised adaptation. It should be noted that this second set of experiments were performed using a slightly different setup compared to the first experiment, and the results are not directly comparable to those of Table 1.

As can be seen, in all cases the use of projecting matrices gives an improvement of about 0.2% absolute, or about 2% relative, when compared to the non-projecting case. Although not large, the improvements are consistent between the different experiments and corpora and in line with the results reported in [2]. When compared to the results in [4] the improvements seem small, but it must be remembered that those improvements to a large extent were due to the effects of using a zero split model as target model, an effect that is already included in the non-projecting SAT baseline in the current work.

It may be argued that the improvement of the method is mainly due to increased number of parameters, and not the use of the additional information in the input features. On the other hand, the fact that an improvement is observed in combination with (regression tree based) MLLR, shows that even with a very high number of parameters the projecting matrix still brings an improvement. Note that in this case, the use of regression classes for the feature transforms would be redundant in the RWTH system, due to the use of a globally pooled covariance.

6. Conclusions

This paper introduced a novel variant of feature space projections for adaptation, based on the MMI′ criterion. The criterion was shown to be closely related to the criterion used in HLDA, and the resulting matrices equal to the matrices produced by the FMLLR-P method, within numerical errors. An efficient estimation method for the matrices was presented, more efficient than FMLLR-P estimation, making the proposed method suitable for use with speaker adaptive training. Finally recognition results were presented, showing performance improvements using projecting transforms, both for recognition only, and when used for speaker adaptive training, and both with and without additional MLLR adaptation. Results were presented for CMLLR based, as well as projecting DLT, but no significant and practical improvements could be reached.

The improvements of the projecting transforms were consistent across corpora and condition, and when compared to DLT usable in practice for unsupervised adaptation. As a final note the method introduced in this paper was successfully applied for SAT in the 2007 RWTH TC-STAR Evaluation system, see [11] for details.

7. References

[1] 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.

[2] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Linear feature space projections for speaker adaptation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Salt Lake City, Utah, USA, May 2001, pp. 325 – 328.

[3] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 26, no. 4, pp. 283 – 297, Dec. 1998.

[4] S. Matsoukas and R. Schwartz, "Improved speaker adaptation using speaker dependent feature projections," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, St. Thomas, U.S. Virgin Islands, Dec. 2003, pp. 273 – 278.

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

[6] L. F. Uebel and P. C. Woodland, "Discriminative linear transforms for speaker adaptation," in *ISCA ITR-Workshop on Adaptation Methods in Speech Recognition*, Sophia Antipolis, France, Aug. 2001, pp. 61 – 64.

[7] K. Visweswariah and R. Gopinath, "Adaptation of front end parameters in a speech recognizer," in *Proc. Int. Conf. on Spoken Language Processing*, Jeju Island, Korea, Oct. 2004, pp. 21 – 24.

[8] J. Lööf, H. Ney, and S. Umesh, "VTLN warping factor estimation using accumulation of sufficient statistics," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006, pp. 1201 – 1204.

[9] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, June 2000, pp. 1129 – 1132.

[10] J. Lööf, M. Bisani, Ch. Gollan, G. Heigold, B. Hoffmeister, Ch. Plahl, R. Schlüter, and H. Ney, "The 2006 RWTH parliamentary speeches transcription system," in *Proc. Int. Conf. on Spoken Language Processing*, Pittsburgh, PA, USA, Sept. 2006, pp. 105 – 108.

[11] J. Lööf, Ch. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, Ch. 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.