LINEAR FEATURE SPACE PROJECTIONS FOR SPEAKER ADAPTATION

George Saon, Geoffrey Zweig and Mukund Padmanabhan

IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598
E-mail: {gsaon,gzweig,mukund}@us.ibm.com, Phone: (914)-945-2985

ABSTRACT

We extend the well-known technique of constrained Maximum Likelihood Linear Regression (MLLR) to compute a projection (instead of a full rank transformation) on the feature vectors of the adaptation data. We model the projected features with phone-dependent Gaussian distributions and also model the complement of the projected space with a single class-independent, speaker-specific Gaussian distribution. Subsequently, we compute the projection and its complement using maximum likelihood techniques. The resulting ML transformation is shown to be equivalent to performing a speaker-dependent heteroscedastic discriminant (or HDA) projection. Our method is in contrast to traditional approaches which use a single speaker-independent projection, and do speaker adaptation in the resulting subspace. Experimental results on Switchboard show a 3% relative improvement in the word error rate over constrained MLLR in the projected subspace only.

1. INTRODUCTION

In many modern speech recognition systems, speaker adaptation and discriminant transformations play key roles in acoustic modeling and/or front-end design. For the latter, the signal processing typically starts by computing Mel-frequency warped cepstral coefficients (MFCCs) for each speech frame, and by concatenating a window of several adjacent frames to form high-dimensional feature vectors. Unfortunately, these vectors, which have typically 100 to 200 dimensions, are difficult to model accurately, and must be projected down to a lower dimensional space. Early systems did this by extracting the first and second cepstral time derivatives from this context window, but more recently, it has been found that improved performance results from using linear discriminant analysis or related procedures to find the transformation to a low dimensional subspace [6, 9, 10]. As a final signal-processing step, the resulting features may be subjected to a diagonalizing transformation such as the maximum likelihood linear transform (MLLT) if diagonal covariance Gaussian models are to be employed [5, 6].

We note that the projection from a high dimensional space to a low dimensional one can be viewed as discarding or "rejecting" some of the dimensions as uninformative. That is, the information carried by the rejected dimensions is not further considered, and both the models and the data now "live" in the low dimensional space. In the case of LDA or HDA, one may argue that the projected subspace has been designed to carry most of the discriminant information useful for classification, and that the contribution of the rejected dimensions is minor and can be neglected. This is certainly true if one is constrained to use a single subspace across all speakers and channel conditions, but it is questionable in the context of speaker adaptation.

Speaker adaptation, as exemplified by MLLR, is a second key technique that is used in most state-of-the art systems. In this step, a linear transform is found such that, when it is applied to either the Gaussian means [7] or, as in constrained MLLR, to the feature vectors themselves [4], the likelihood of the acoustic data associated with an utterance is maximized with respect to an initial word hypothesis. The utterance is then re-decoded after applying the transform. Regardless of whether the models or feature vectors are transformed, this step is applied in the *reduced* subspace determined by the initial projection.

This paper is motivated by the following observations:

1. It may be desirable to make the subspace in which the classification is performed speaker dependent. Intuitively, we want to be able to "borrow" dimensions from the rejected subspace if those dimensions carry discriminant information for a particular speaker.

2. The HDA transform is an ML transform for normal populations with common means and covariances in the rejected subspace [3, 6].

3. If we make the assumption that the rejected dimensions are identically distributed across all the phonetic classes, then the constrained MLLR transform becomes a speaker-dependent HDA transform.

These observations suggest that it is possible to create speaker-dependent discriminant transforms using just the apparatus of constrained MLLR in the complete feature space.

The paper is organized as follows: in section 2 we briefly revisit the constrained MLLR formulation, introduce the ex-

tension for the projection case, and show its connection to HDA. Section 3 describes the experiments and results and section 4 provides a final discussion.

2. FEATURE SPACE MLLR AND HDA

The goal of standard MLLR, as originally formulated by Leggetter *et. al* in [7], is to affinely transform the means $\{\mu_j\}$ of the diagonal Gaussian mixture components of the HMM model λ such as to maximize the likelihood of the adaptation data $\mathbf{X} = X_1 \dots X_T$, i.e. find $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$ with

$$\overline{\mu}_j = A\mu_j + b = W\xi_j \tag{1}$$

where $W = [A^T|b^T]^T$ and $\xi_j = [\mu_j^T|1]^T$, such that the auxiliary function of the EM algorithm

$$Q(\lambda, \overline{\lambda}) = \sum_{\mathbf{Q}} P(\mathbf{Q}|\mathbf{X}, \lambda) \log P(\mathbf{X}, \mathbf{Q}|\overline{\lambda}) \tag{2}$$

is maximized with respect to W. This turns out to be a linear regression problem yielding a closed form solution for W.

The MLLR formulation has been extended by Gales [4] in the following way. The author distinguishes between *constrained* MLLR where the covariances of the Gaussians have to share the same transform as the means, i.e. $\overline{\Sigma}_j = A\Sigma_j A^T$ and *unconstrained* MLLR where the covariances are transformed independently of the means. Both constrained and unconstrained MLLR are *model space* transformations in the sense that they act on the model parameters not on the features. In contrast, *feature space* MLLR transforms the observation vectors, that is $\overline{X}_t = AX_t + b$. There is a duality between constrained model space and feature space MLLR in the sense that transforming the means and the covariances is equivalent to transforming the features since the respective Gaussian likelihoods are equal:

$$|A|\mathcal{N}(AX + b|\mu; \Sigma) = \mathcal{N}(X|A^{-1}(\mu - b); A^{-1}\Sigma A^{-T}) \tag{3}$$

$|A|$ represents the determinant of the Jacobian of the transformation $X \rightarrow AX + b$ and is supposed to be positive. It is required for $\mathcal{N}(\cdot; \mu, \Sigma)$ to be a valid probability density function in the transformed space [8].

Rewriting the auxiliary function (2) for feature space MLLR yields, after some manipulations, the objective function

$$\sum_{t=1}^{T}\sum_{j=1}^{N} \gamma_t(j) \left[\log|A| - \frac{1}{2}(AX_t - \mu_j)^T \Sigma_j^{-1}(AX_t - \mu_j)\right] + \mathcal{C} \tag{4}$$

where $\mathcal{C}$ is a constant with respect to A. The bias b has been dropped since it can be taken into account by extending the matrix and the observation vectors analogous to (1). For simplicity of notation, the summation over the HMM states and over the mixture components within a state has been collapsed into a single sum over all the Gaussians in the model. $\gamma_t(j)$ represents the posterior probability of component j at time t given the complete observation sequence. The gradient of (4) with respect to A has the expression

$$TA^{-T} - \sum_{t=1}^{T}\sum_{j=1}^{N} \gamma_t(j)\Sigma_j^{-1} A X_t X_t^T - \gamma_t(j)\Sigma_j^{-1}\mu_j X_t^T \tag{5}$$

Following the terminology from [2], we define the *sufficient statistics* for feature space MLLR by:

- $K = \sum_{t=1}^{T}\sum_{j=1}^{N} \gamma_t(j)\Sigma_j^{-1}\mu_j X_t^T$ and
- $G_i = \sum_{t=1}^{T}\sum_{j=1}^{N} \frac{\gamma_t(j)}{\sigma_{ji}^2} X_t X_t^T,\ i = 1 \dots n$

where $\Sigma_j = \mathrm{diag}(\sigma_{j1}^2, \dots, \sigma_{jn}^2)$. By rewriting (5) in terms of these statistics, the rows of A can be found independently through iteratively solving a set of quadratic equations (for more details the reader is referred to [4]). Until now, we have studied the case when A is a full rank transformation operating in the complete n-dimensional space. To consider the projection to a p-dimensional subspace with $p \leq n$, the following structure will be imposed on the model parameters

$$\mu_j = \begin{bmatrix} \mu_j^{(p)} \\ \mu_0^{(n-p)} \end{bmatrix},\ \Sigma_j = \begin{bmatrix} \Sigma_j^{(p)} & 0 \\ 0 & \Sigma_0^{(n-p)} \end{bmatrix},\ 1 \leq j \leq N \tag{6}$$

meaning that, after the transformation is applied, the rejected dimensions are supposed to be identically (Gaussian) distributed across all the mixture components. This is a similar assumption to the one made in HDA [6]. Correspondingly, A can be decomposed into two parts, $A = [A^{(p)T}|A^{(n-p)T}]^T$, where $A^{(p)}$, of dimension $p \times n$, will be the useful projection and $A^{(n-p)}$, of dimension $n - p \times n$, will provide the complementary dimensions. Its role is to provide a full rank completion to $A^{(p)}$ in order to be able to make meaningful likelihood comparisons across feature spaces of equal dimension (n). Our next task at hand is to find the ML estimates for the completion parameters $\mu_0^{(n-p)}$ and $\Sigma_0^{(n-p)}$. Plugging (6) into (4) leads to the objective function

$$
\begin{aligned}
&-\frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_t(j)\left[(A^{(p)}X_t-\mu_j^{(p)})^T\Sigma_j^{(p)^{-1}}(A^{(p)}X_t-\mu_j^{(p)})\right]\\
&-\frac{1}{2}\sum_{t=1}^{T}(A^{(n-p)}X_t-\mu_0^{(n-p)})^T\Sigma_0^{(n-p)^{-1}}(A^{(n-p)}X_t-\mu_0^{(n-p)})\\
&+T\log|A|+\mathcal{C}
\end{aligned}
\tag{7}
$$

which, when differentiated with respect to $\mu_0^{(n-p)}$ and $\Sigma_0^{(n-p)}$, provides the ML solution

$$
\begin{aligned}
\mu_0^{(n-p)} &= A^{(n-p)}\mu,\\
\Sigma_0^{(n-p)} &= \mathrm{diag}(A^{(n-p)}\Sigma A^{(n-p)^T})
\end{aligned}
\tag{8}
$$

where μ and Σ represent the mean and the covariance of the adaptation data. In regular HDA [6], (8) is plugged back into (7) yielding an objective function which now depends only on A:

$$
\begin{aligned}
&-\frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_t(j)\left[(A^{(p)}X_t-\mu_j^{(p)})^T\Sigma_j^{(p)^{-1}}(A^{(p)}X_t-\mu_j^{(p)})\right]\\
&-\frac{T}{2}\sum_{t=1}^{T}\log|\,\mathrm{diag}(A^{(n-p)}\Sigma A^{(n-p)^T})|+T\log|A|+\mathcal{C}
\end{aligned}
\tag{9}
$$

However, (9) is difficult to optimize in practice due to the presence of the two determinant terms. We would like to be able to use the techniques of feature space MLLR which consist in accumulating the sufficient gradient statistics K and $\{G_i\}$ and in solving simpler, independent problems for the rows of A. In order to do this, we have to provide an explicit solution for $\mu_0^{(n-p)}$ and $\Sigma_0^{(n-p)}$. This is not a straightforward matter as both terms depend on $A^{(n-p)}$ (and $A^{(n-p)}$ depends on them). We solve this iteratively by first fixing $\mu_0^{(n-p)}$, $\Sigma_0^{(n-p)}$ and finding A which is then used to update $\mu_0^{(n-p)}$ and $\Sigma_0^{(n-p)}$ according to (8). At the beginning, A is initialized to the identity matrix and $\mu_0^{(n-p)}$ and $\Sigma_0^{(n-p)}$ to the mean and the diagonal covariance of the rejected dimensions of the adaptation data. Lastly, we derive a simplified form for the sufficient statistics of feature space MLLR:

$$
K=\begin{bmatrix}K^{(p)}\\ K^{(n-p)}\end{bmatrix}=\begin{bmatrix}\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_t(j)\Sigma_j^{(p)^{-1}}\mu_j^{(p)}X_t^T\\ T\Sigma_0^{(n-p)^{-1}}\mu_0^{(n-p)}\mu^T\end{bmatrix}
\tag{10}
$$

and

$$
G_i=\begin{cases}\sum_{t=1}^{T}\sum_{j=1}^{N}\frac{\gamma_t(j)}{\sigma_{ji}^{(p)^2}}X_tX_t^T, & i=1\ldots p\\ \frac{T}{\sigma_{0i-p}^{(n-p)^2}}(\Sigma+\mu\mu^T), & i=p+1\ldots n\end{cases}
\tag{11}
$$

3. EXPERIMENTS AND RESULTS

The speech recognition experiments were conducted on the Switchboard database. We have experimented with two systems which differ only in the number of diagonal Gaussian mixture components: 60K and 277K. Both systems were trained on 243 hours of data and have 3140 context dependent HMM states. The first system has a maximum number of 20 mixture components per state while for the second the maximum number was set to 120. They use 40-dimensional LDA+MLLT features obtained in the following way. We first compute an LDA projection from 117 dimensions to 40 dimensions. The 117 dimensional vectors are formed by splicing 9 consecutive 13-dimensional cepstral vectors for each frame in the training data. The classes for LDA are given by all the HMM states except those corresponding to the silence phones. The LDA computation required estimating one full covariance Gaussian model for each state in the original 117-dimensional space. Once the LDA matrix has been obtained, we also keep the eigenvectors corresponding to the rejected dimensions (or equivalently, the ones which correspond to the minimum eigenvalues). That is, we actually compute a full rank LDA and separate the resulting matrix into a 40×117 projection and a complement. The range of the projection is further diagonalized through a maximum likelihood linear transform leading to a composite LDA+MLLT transform. The models for MLLT are obtained by projecting the initial 117-dimensional Gaussian parameters. Next, we estimate the 60K and 277K diagonal Gaussian mixture components in the 40-dimensional LDA+MLLT space. Here the experimental setups for feature space MLLR in the subspace only and for the projection-based case differ in the following way: for the former we accumulate statistics in the LDA+MLLT space and estimate a constrained MLLR transform in that same subspace. For the latter, the statistics are accumulated in the complete LDA space (more correctly, in the LDA+MLLT space and the LDA complement). We then augment the model output distributions with a Gaussian distribution for the rejected dimensions estimated from the statistics of the adaptation data in the LDA complement. Subsequently, we compute a 117×117 feature space MLLR transform and isolate a 40×117 projection part $A^{(p)}$. The final adapted features are obtained by multiplying the spliced 117-dimensional vector with the full rank LDA matrix and then by projecting the resulting (117-dimensional) vector to 40 dimensions through $A^{(p)}$.

System	WER 60K	WER 277K
Baseline	50.0%	46.0%
MLLR	46.1%	43.3%
FMLLR	47.5%	44.3%
FMLLR-P	46.5%	42.9%
FMLLR+MLLR	–	41.4%
FMLLR-P+MLLR	–	40.7%

Table 1: Word error rates for the various adaptation schemes.

The test set consists of 40 Switchboard conversations (80 speakers) of the Eval'98 set. It contains approximately 21K words and 1.5 hours of speech with an average conversation length of 5 minutes. In table 1, we compare the performance of standard MLLR, feature space MLLR (FMLLR) and projection-based feature space MLLR (FMLLR-P) for the two different systems. For standard MLLR, we have used multiple transforms (between 3 and 5 with an average of 4) based on a regression tree obtained by bottom-up clustering the HMM states with a Gaussian likelihood metric. It can be seen that one single FMLLR-P transform outperforms both the multiple standard MLLR transforms (for the 277K system) and the constrained FMLLR transform. For the latter, the difference in performance increases with the size of the system (from 1% to 1.4%). Additionally, we have also experimented with adding MLLR on top of the FMLLR and FMLLR-P stages. These results are summarized in the last two lines of table 1. The gains from standard MLLR and FMLLR-P turned out to be additive (from 2.7% to 2.2%).

4. CONCLUSION

In this paper, we have made the connection between constrained MLLR, as a technique for speaker adaptation, and heteroscedastic discriminant analysis, as a technique for feature extraction and front-end design. We have shown that it is possible to create speaker-dependent discriminant projections by computing constrained MLLR transforms in the complete space for a particular model where the distributions of the rejected dimensions are tied across all the phonetic classes to the overall distribution of the adaptation data. Future work will be pursued along two directions: first, we will investigate speaker-adaptive training [1] in the context of this technique where a canonical model is trained on features which are transformed through speaker-dependent HDA projections. Secondly, we will allow for multiple feature space projections by making use of regression classes.

5. REFERENCES

[1] T. Anastasakos, J. McDonough, R. Schwartz, J. Makhoul. A compact model for speaker-adaptive training. *Proceedings of ICSLP'96*, Philadelphia, 1996.

[2] M. Bacchiani. Using maximum likelihood linear regression for segment clustering and speaker identification. *Proceedings of ICSLP 2000*, Beijing, 2000.

[3] N. A. Campbell. Canonical variate analysis - a general model formulation. *Australian Journal of Statistics*, 26(1):86–96, 1984.

[4] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. Technical Report CUED/F-INFENG, Cambridge University Engineering Department, 1997.

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

[6] N. Kumar and A. G. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communcation*, 26:283–297, 1998.

[7] C. J. Leggetter and P. C. Woodland. Speaker adaptation of HMMs using linear regression. Technical Report CUED/F-INFENG, Cambridge University Engineering Department, 1994.

[8] A. Papoulis. Probabiliy, random variables and stochastic processes. *WCB/McGraw-Hill*, 1991.

[9] G. Saon, M. Padmanabhan, R. Gopinath and S. Chen. Maximum likelihood discriminant feature spaces. *Proceedings of ICASSP 2000*, Istanbul, 2000.

[10] G. Saon and M. Padmanabhan. Minimum Bayes error feature selection. *Proceedings of ICSLP 2000*, Beijing, 2000.