VARIANCE COMPENSATION WITHIN THE MLLR FRAMEWORK FOR ROBUST SPEECH RECOGNITION AND SPEAKER ADAPTATION

M.J.F. Gales *D. Pye* *P.C. Woodland*

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, England.

ABSTRACT

This paper investigates the use of maximum likelihood linear regression (MLLR) for both speaker and environment adaptation. MLLR transforms the mean and variance parameters of a set of HMMs. In this paper a number of different types of linear transformations of the variances are examined including full, block diagonal, and diagonal transformation matrices. Experiments on large vocabulary speaker independent data sets are described. On all the data sets examined the use of MLLR mean and variance compensation reduced the error rate compared to mean-only compensation. Furthermore, the use of a block diagonal or full transformation of the variances on the clean data task showed slight improvements over the diagonal case. However, when some environmental mismatch was present there was no difference in performance between using multiple diagonal variance transformations and a more complex single variance transform.

1. INTRODUCTION

Current state-of-the-art speaker-independent (SI) continuous-speech recognition systems are capable of achieving impressive performance in clean acoustic environments for speakers that are well represented in the training data. However, for some speakers performance can be relatively poor e.g. for non-native speakers using a system trained on speech from natives. Furthermore, the performance degrades, often dramatically, if there is a mismatch between the training and test data acoustic environments. For complex speech recognition systems a large amount of data is required to retrain the system to a particular speaker or to a new acoustic environment. Hence, it is highly desirable to be able to improve the performance of an existing system while only using a small amount of speaker-specific or environment-specific adaptation data.

A technique which is applicable to both speaker and environment adaptation is maximum likelihood linear regression (MLLR) [3, 4]. This estimates a set of linear transformations for the mean parameters of a mixture Gaussian HMM system to maximise the likelihood of the adaptation data. As these transformations can capture general relationships between the original model set and the current speaker or new acoustic environment, they can be effective in adapting all the HMM distributions with limited adaptation data. MLLR was initially developed for speaker adaptation. However, it can be used to perform environmental compensation, since it will also reduce a mismatch between the acoustic models and adaptation data due to channel or additive noise effects.

This paper investigates an extension to the basic MLLR approach for adapting the Gaussian mean parameters to also include adaptation of the variances. As for the means, the variance adaptation is performed by applying a set of transformation matrices to the variance parameters. The transformation matrices may have a full, block diagonal or diagonal structure. The estimation of the variance transformations is, as for the means, performed in a maximum likelihood (ML) framework.

Previously [8] the use of simple diagonal transformations has been investigated. Here more complex variance transformations are also examined. All of these transforms may be trained and used in recognition with little additional memory overhead, though for the more complex variance transform structures there are increased computational overheads during recognition. Experiments on large vocabulary speaker independent continuous speech recognition tasks show the utility of variance compensation especially in non-matched acoustic conditions.

2. MAXIMUM LIKELIHOOD LINEAR REGRESSION

MLLR was originally developed for speaker adaptation [3, 4] but can equally be applied to situations of environmental mismatch. A set of transformation matrices for the HMM Gaussian parameters are estimated which maximise the likelihood of the adaptation data. The set of transformations is relatively small compared to the total number of Gaussians in the system and so a number of Gaussians share the same transformation matrices. This means that the transformation parameters can be robustly estimated from only a limited amount of data, which allows all the Gaussians in the HMM set to be updated. For a small amount of data (or very robust transformation estimation) only a single global transformation is used. As more data becomes available more specific transformations can be estimated. Originally transformations were estimated only for the mean parameters but recently the approach has been extended so that the Gaussian variances can also be updated [2]. This section gives a brief overview of the basic MLLR theory for both the mean parameters and the variances.

The means and variances are adapted in two separate stages. Initially new means are found. Then, given these new means, the variances are updated. Hence, the HMMs are modified such that

$$\mathcal{L}(\mathbf{O}_T|\check{\mathcal{M}}) \geq \mathcal{L}(\mathbf{O}_T|\hat{\mathcal{M}}) \geq \mathcal{L}(\mathbf{O}_T|\mathcal{M})$$

where $\mathcal{M}$ is the original model set, the model set $\hat{\mathcal{M}}$ has just the mean parameters updated (to $\hat{\mu}_1, \ldots, \hat{\mu}_M$) and the model set $\check{\mathcal{M}}$ has both the means and the variances $\hat{\mathbf{\Sigma}}_1, \ldots, \hat{\mathbf{\Sigma}}_M$ updated, and $\mathbf{O}_T$ is the adaptation data,

$$\mathbf{O}_T = \{\mathbf{o}(1), \mathbf{o}(2), \ldots, \mathbf{o}(T)\}$$

2.1. MLLR Adaptation of the Means

The aim of MLLR is to obtain a set of transformation matrices that maximises the likelihood of the adaptation data. A transformation matrix is used to give a new estimate of the mean, where

$$\hat{\mu}_m = \hat{\mathbf{W}}_m \xi_m$$

and $\hat{\mathbf{W}}_m$ is the $n \times (n+1)$ transformation matrix (for n dimensional data) and ξ_m is the extended mean vector

$$\xi_m = \begin{bmatrix} 1 & \mu_1 & \ldots & \mu_n \end{bmatrix}^T$$

In order to ensure robust estimation of the transformation parameters, the transformation matrices are tied across a number of Gaussians, according to a regression class tree [4]. This tree contains all the Gaussians in the system, with statistics gathered at the leaves (which may each contain a number of Gaussians). The most specific transform that can be robustly estimated using the adaptation is generated for all the Gaussians in the system.

A particular transformation $\hat{\mathbf{W}}_m$ is to be tied across R Gaussians $\{m_1, \ldots, m_R\}$. For the Gaussian output probability density functions considered, $\hat{\mathbf{W}}_m$ may be found by solving

$$\sum_{\tau=1}^{T}\sum_{r=1}^{R} L_r(\tau)\mathbf{\Sigma}_r^{-1}\mathbf{o}(\tau)\xi_r^T = \sum_{\tau=1}^{T}\sum_{r=1}^{R} L_r(\tau)\mathbf{\Sigma}_r^{-1}\hat{\mathbf{W}}_m\xi_r\xi_r^T$$

where

$$L_r(\tau) = p(q_r(\tau)|\mathcal{M}, \mathbf{O}_T)$$

and $q_r(\tau)$ indicates Gaussian m_r at time τ. For the full covariance matrix case the solution is computationally very expensive [2]. However, the solution for the diagonal covariance case is computationally tractable and described in [3]. Each transformation can be a full matrix or constrained to be block diagonal or diagonal [5].

2.2. MLLR Adaptation of the Variances

The Gaussian variance vectors, or in general covariance matrices, are updated using the following transformation

$$\hat{\mathbf{\Sigma}}_m = \mathbf{B}_m^T \hat{\mathbf{H}}_m \mathbf{B}_m$$

where $\hat{\mathbf{H}}_m$ is the linear transformation to be estimated and $\mathbf{B}_m$ is the inverse of the Choleski factor of $\mathbf{\Sigma}_m^{-1}$, so that

$$\mathbf{\Sigma}_m^{-1} = \mathbf{C}_m \mathbf{C}_m^T$$

and $\mathbf{B}_m = \mathbf{C}_m^{-1}$.

In a similar fashion to the means, a variance transformation is shared over a number of Gaussians, $\{m_1, \ldots, m_R\}$. It is simple to show that the maximum likelihood estimate is given by

$$\hat{\mathbf{H}}_m = \frac{\sum_{r=1}^{R} \mathbf{C}_r^T \left[\sum_{\tau=1}^{T} L_r(\tau)(\mathbf{o}(\tau) - \hat{\mu}_r)(\mathbf{o}(\tau) - \hat{\mu}_r)^T\right] \mathbf{C}_r}{\sum_{r=1}^{R}\sum_{\tau=1}^{T} L_r(\tau)} \qquad (1)$$

where $\hat{\mu}_r$ is the previously calculated mean. It can be seen that the variance transformation matrix will be full, yielding full covariance matrices for each Gaussian. A diagonal transformation for the variances may be obtained by simply zeroing the off-diagonal terms. This is still guaranteed to increase the likelihood.

Other transformations of the variance have been proposed. A simple offset to the variance may be used [6], however, without the use of variance flooring the resultant variance is not guaranteed to be positive. Variance scaling[1] with a simple bias on the mean has also been used [5]. This is a specific case of the transformations proposed here.

If the original covariance matrices are diagonal, the use of a full covariance transform would dramatically increase the memory requirements for the model set (effectively a full covariance system would be required). In many circumstances this is impractical for large systems. Instead the transformations are stored, and the likelihoods calculated as

$$\begin{aligned} \mathcal{L}(\mathbf{o}(\tau)|\mathcal{M}_r) = K_r - \frac{1}{2}\Big[&\log(|\mathbf{\Sigma}_r|) + \log(|\hat{\mathbf{H}}_m|) \\ &+ (\mathbf{C}_r^T\mathbf{o}(\tau) - \mathbf{C}_r^T\hat{\mu}_r)^T\hat{\mathbf{H}}_m^{-1}(\mathbf{C}_r^T\mathbf{o}(\tau) - \mathbf{C}_r^T\hat{\mu}_r)\Big] \end{aligned} \qquad (2)$$

There is therefore little increase in memory requirements, though of course there is an increase in the runtime overhead of calculating the likelihoods. It is worth noting that during training there is little additional overhead in terms of both memory and computational power in estimating full variance transforms, as the additional statistics for the full variance transform may be stored at the regression class level [2].

For the experiments conducted in this paper three forms of variance transformation were used.

1. **Full**: the full matrix transforms described in equation 1 were used and the likelihood calculated using equation 2.
2. **Block Diagonal**: $\hat{\mathbf{H}}_m$ was constrained to have a block diagonal structure and hence any correlations between parameters were assumed to only occur within a block. Here three blocks consisting of the static, delta, and delta-delta parameters were used. Again, equation 2 was used to calculate the likelihood.

[1] In the terminology used here this is a diagonal transformation of the variance with just a bias (no scaling) on the mean parameters.

3. **Diagonal**: only the leading diagonal elements of $\hat{\mathbf{H}}_m$ were non-zero. With original diagonal matrices, this will yield diagonal adapted covariance matrices so that no modifications to the standard recognition system are required.

The mean transformation matrix is a function of the Gaussian variance. Thus by altering the variance, the maximum likelihood estimate of the mean transformation will also be altered. While an iterative scheme could be used it has been found that a single update of the means and variances is, in practice, sufficient.

3. RECOGNITION EXPERIMENTS

Experiments were performed on data taken from the ARPA 1994 CSRNAB Hub 1 and Spoke 5 test sets.

1. **Hub 1 (H1): Unlimited Vocabulary NAB News Baseline**. This is an unlimited vocabulary task with approximately 15 sentences per speaker. The data was recorded in a clean environment[2].
2. **Spoke 5 (S5): Microphone Independence**. This task uses data recorded with far-field microphones (only close-talking high-SNR data was available for training). It is a 5k word recognition task with ten sentences from each speaker. The average A-weighted SNR of the S5 data was 20dB.

These tasks were selected as they allow the evaluation of various variance adaptation techniques for both speaker and environmental adaptation.

It should be emphasised that the results are not comparable with the official evaluation results, as they are obtained using *static unsupervised*[3] adaptation. For the actual evaluation the results are required to be produced in a causal manner, thus only *incremental unsupervised*[4] adaptation would be allowed. The reason for using the data in a static unsupervised mode was that it allows easier comparison of results as all the systems use the same transcriptions and the models are all adapted at the same time.

3.1. System Description

The baseline system used was a gender-independent cross-word-triphone mixture-Gaussian tied-state HMM system. This was the same as the "HMM-1" model set used in the HTK 1994 ARPA evaluation system [7]. The speech was parameterised into 12 MFCCs, C_1 to C_{12}, along with normalised log-energy and the first and second differentials of these parameters yielding a 39-dimensional feature vector. The acoustic training data consisted of 36,493 sentences from the SI-284 WSJ0 and WSJ1 sets, and the LIMSI 1993 WSJ lexicon and phone set were used. For the noise corrupted task, S5, the model set parameters were initially modified using parallel model combination (PMC) [1]. For computational efficiency the PMC Log-Add approximation [1] with simple convolutional noise estimation was used to modify the means of the models. Since PMC cannot be applied to models built with the standard HTK front-end, normalised log-energy was replaced by C_0 and simple differences used instead of linear regression to generate the dynamic parameters. Furthermore, cepstral mean normalisation (CMN) was not used for the PMC models and hence it was necessary to first compensate for the different global signal levels of the WSJ0 and WSJ1 databases by applying an offset to the C_0 feature vector coefficient of the WSJ0 data such that it had the same average value as the WSJ1 database. To generate the PMC models, the standard HTK model set was initially estimated as described above. The model set was then updated using single-pass retraining [1] to be based on the PMC parameter set. An additional pass of standard Baum-Welch re-estimation was then performed.

None of the system parameters, such as the grammar scale factor or insertion penalty, were optimised for the evaluation test data. Appropriate values for these were determined from the standard clean speaker-independent system, or in the case of the PMC model sets, by optimising them on the ARPA 1994 CSRNAB S0 development data. The minimum occupancy counts for the regression classes were optimised using the H1 Development test data. All results were generated using the official adjudicated reference transcriptions and the appropriate NIST scoring software.

In all the experiments reported here, full transformation matrices for the means were used. The regression class trees used throughout this work were based on clean speech. The data transcriptions used for adaptation were generated using the standard model sets for the H1 test sets and using PMC compensated model sets for S5.

3.2. Results

Transform	Variance Transform	Error Rate (%) H1 Dev	H1 Eval	S5 Eval
—	—	9.57	9.20	10.29
Means	—	8.77	8.35	8.70
Means and Variances	Diagonal	8.54	8.05	7.21
	Block	8.42	7.94	7.30
	Full	8.20	7.91	7.30

Table 1: Error rates (%) for static unsupervised adaptation using mean and variance MLLR on 1994 ARPA H1 and S5 test sets. Results are shown for various variance transformation structures

Table 1 shows the performance of the various MLLR mean and variance compensation schemes. On all tasks the use of mean and variance compensation shows improvements over mean-only compensation. For the clean tasks, H1 Dev and H1 Eval, mean-only compensation gave 8% and 9% reductions in word error rate. Further improvements of 3% and 4% respectively were obtained with the diagonal variance transformation. Small additional improvements were obtained using the block and full variance transformation. As previously mentioned, there is an additional run-time computational overhead with this.

[2] Here the term "clean" refers to the training and test conditions being from the same microphone type with a high signal-to-noise ratio.

[3] The data is available in one block and the system adapted with no knowledge of the correct transcription.

[4] The adaptation data is made available as the system is used and the models repeatedly adapted.

On the S5 task the improvement obtained using mean-only compensation was 15%. A further reduction in the word error rate of 17% was obtained using diagonal variance compensation. However, the use of the full and block diagonal variance transforms slightly degraded the performance compared to the simple diagonal transform. On the H1 Eval data each speaker had on average about 160 seconds of data (including silence) to adapt the models, whereas for the S5 Eval data there was only 64 seconds of data and hence the full and block-diagonal variance transforms were possibly generated with insufficient data for the S5 task. For the full transform there was typically a single transformation matrix for all the speech variances, whereas for the diagonal transform there were typically three transformation matrices.

Transform	Variance Transform	Error Rate (%)		
		H1 Dev	H1 Eval	S5 Eval
—	—	9.57	9.20	10.29
Means	—	8.77	8.35	8.70
Means and Variances	Diagonal	8.61	8.15	7.80
	Full	8.20	7.91	7.30

Table 2: Error rates (%) for static unsupervised adaptation using mean MLLR and a single variance transformation

To investigate the effects of using multiple diagonal variance transformations the experiments in Table 1 were re-run using only a single variance transform and these results are given in Table 2. Comparing the results between the two tables shows that the use of multiple variance transforms for the diagonal case was important. For example, on the S5 data using multiple variance transformations gave an additional 8% reduction in word error rate. Furthermore when only a single variance transform is used it is clear that a more complex transform structure is desirable. This illustrates the trade-off that exists between the complexity of the transformation and the number of transforms that can be robustly estimated.

The improvements on the S5 data are comparable with those obtained on the 1995 ARPA H3 task [8] which included multiple passes of unsupervised adaptation using multiple variance transformations. The H3 task was an unlimited vocabulary task with data taken from a stereo database, a Sennheiser channel (H3-C0), and an unknown microphone channel (H3-P0). Here an additional 11% reduction in word error rate on the H3-P0 data was achieved using a diagonal variance transform. Initial experiments using full variance transformations on the H3-P0 data showed similar trends to the S5 results.

4. CONCLUSIONS

This paper has described mean and variance compensation using MLLR for both speaker and environmental adaptation. Three forms of variance compensation were examined, diagonal, block diagonal and full. In all cases variance compensation was found to improve recognition performance. On the clean test sets the use of a diagonal variance transformation gave an additional 3% and 4% reduction in word error rate over mean-only adaptation. These improvements were increased to 6% and 5% when block diagonal and full variance transformations were used. However, this slight improvement was gained at the additional computational overhead of non-diagonal Gaussian likelihood calculations. On the noise corrupted task greater reductions in word error rate were obtained using both mean-only and mean and variance compensation. A reduction of 17% in word error rate was obtained using mean and variance compensation compared with mean-only compensation, this was approximately the same for all variance transforms. This illustrates the greater importance of adapting the variances when there are environmental mismatches between training and testing.

5. ACKNOWLEDGEMENTS

This work is in part supported by an EPSRC grant reference GR/K25380. Mark Gales is supported by a Research Fellowship from Emmanuel College, Cambridge. Additional computing resources were provided by the ARPA CAIP computing facility.

6. REFERENCES

1. M J F Gales. *Model-Based Techniques for Noise Robust Speech Recognition*. PhD thesis, Cambridge University, 1996.
2. M J F Gales and P C Woodland. Variance compensation within the MLLR framework. Technical Report CUED/F-INFENG/TR242, Cambridge University, 1996. Available via anonymous ftp from: svr-ftp.eng.cam.ac.uk.
3. C J Leggetter and P C Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Computer Speech and Language*, 9:171–186, 1995.
4. C J Leggetter and P C Woodland. Flexible speaker adaptation using maximum likelihood linear regression. In *Proceedings ARPA Workshop on Spoken Language Systems Technology*, pages 110–115, 1995.
5. L R Neumeyer, A Sankar, and V V Digalakis. A comparative study of speaker adaptation techniques. In *Proceedings Eurospeech*, pages 1127–1130, 1995.
6. R C Rose, E M Hofstetter, and D A Reynolds. Integrated models of signal and background with application to speaker identification in noise. *IEEE Transactions SAP*, 2:245–257, 1994.
7. P C Woodland, C J Leggetter, J J Odell, V Valtchev, and S J Young. The development of the 1994 HTK large vocabulary speech recognition system. In *Proceedings ARPA Workshop on Spoken Language Systems Technology*, pages 104–109, 1995.
8. P C Woodland, M J F Gales, and D Pye. Improving environmental robustness in large vocabulary speech recognition. In *Proceedings ICASSP*, 1996.