# IMPROVEMENTS IN LINEAR TRANSFORM BASED SPEAKER ADAPTATION

*L.F. Uebel & P.C. Woodland*

Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK.
Email: {lfu20,pcw}@eng.cam.ac.uk

**ABSTRACT**

This paper presents three forms of linear transform based speaker adaptation that can give better performance than standard maximum likelihood linear regression (MLLR) adaptation. For unsupervised adaptation, a lattice-based technique is introduced which is compared to MLLR using confidence scores. For supervised adaptation, estimation of the adaptation matrices using the maximum mutual information criterion is discussed which leads to the MMILR approach. Recognition experiments show that lattice MLLR can reduce word error rates on a Switchboard task by 1.4% absolute. For recognition of non-native speech from the Wall Street Journal database, a reduction in word error rate of 10-16% relative was obtained using MMILR compared to standard MLLR.

## 1. INTRODUCTION

Maximum likelihood linear regression (MLLR) [6, 7, 2] is a widely-used technique for speaker adaptation. It can be successful with fairly limited amounts of training data and can operate in all adaptation modes including supervised and unsupervised. MLLR estimates linear transformation matrices for HMM Gaussian means and variances to maximise the likelihood of the adaptation data, and the relatively small number of parameters estimated gives the technique its robustness and data efficiency. This paper aims to improve on the estimation of transform matrix parameters in two directions: unsupervised adaptation via the use of confidence measures and lattices; and discriminative training techniques for transform estimation in supervised adaptation.

Unsupervised adaptation uses recognition hypotheses to provide the adaptation supervision. The effect of errors in adaptation supervision can mean that either fewer transform parameters can be estimated from a certain amount of adaptation data, or that performance of unsupervised adaptation is degraded significantly with respect to supervised adaptation. This is particularly important for high error rate tasks such as Switchboard. The potential beneficial effect of word correctness confidence scores is discussed. This information is included either explicitly or implicitly by using a lattice-based estimation process. The techniques are considered in the context of transcription-mode MLLR in which the (unsupervised) adaptation data is also the test data. In this case, iterative MLLR adaptation [13] can be used to interleave adaptation and decoding.

It is well-known that maximum likelihood estimation (MLE) relies on model correctness assumptions and hence other parameter estimation criteria which have a closer relationship to minimising the training data word error rate, such as maximum mutual information estimation (MMIE), can improve performance. Furthermore, it has recently been shown that MMIE training techniques can significantly improve performance for large vocabulary recognition over the best MLE systems [14]. Hence, in this paper, we also address the issue of estimating the linear transform parameters by MMIE which we denote MMILR. This is applied to the task of supervised adaptation to non-native speech.

A common thread for all the adaptation techniques discussed is the use of word lattices. These are either used to explicitly extract word-level confidence scores; to represent an utterance in lattice-based MLLR or as a compact representation of confusable utterances for use in MMILR. The lattices consist of nodes representing points in time and corresponding to the ends of particular words. These are joined by arcs that record the language model probability of a particular word transition and, if necessary, the acoustic score. For some purposes we also use model-marked lattices, in which the HMM model segmentation points are explicitly encoded for each lattice arc.

The rest of the paper is organised as follows. First confidence score based MLLR is described; then lattice-based MLLR, which implicitly uses a confidence score measure, is discussed. These techniques are evaluated using iterative transcription mode adaptation using Switchboard data. The use of MMILR adaptation is then described and evaluated in the context of recognition of non-native speech from the Wall Street Journal corpus.

## 2. CONFIDENCE SCORE BASED MLLR

The use of confidence scores in unsupervised MLLR-based speaker adaptation has previously been investigated in a number of papers (e.g. [15, 1, 11]). The general approach is to compute a confidence score for each word of an automatically generated transcription and then, during adaptation, only use data which has a high confidence score to accumulate statistics for transform generation. This method may be particularly useful in situations where the automatically generated transcription has a high word error rate, such as for the Switchboard corpus. In this paper the incorporation of a confidence score in MLLR adaptation is used as a point of comparison for the lattice-based technique discussed in Section 3.

### 2.1. Confidence Score Calculation

To calculate the confidence score, a version of the approach presented in [3] was used which can compute the confidence score for any particular word sequence from a lattice.

First, the forward-backward algorithm is used to calculate a lattice arc posterior probability $P(l|\mathcal{O})$ for each arc in the lattice

$$P(l|\mathcal{O}) = \frac{\sum_{q \in Q_l} p_{acc}(\mathcal{O}|q)^{\frac{1}{\gamma}} P_{lm}(w) P_{pr}(q|w)}{p(\mathcal{O})} \qquad (1)$$

where $\gamma$ is the language model scale factor, $q$ is a path through the lattice corresponding to the word sequence $w$, $Q_l$ is the set of paths passing through arc $l$, $p_{acc}(\mathcal{O}|q)$ is the acoustic likelihood, $P_{lm}(w)$ is the language model probability, $P_{pr}(q|w)$ is the pronunciation probability and $p(\mathcal{O})$ is the overall likelihood of all paths through the lattice. Note that in this process the acoustic model likelihood and language model probabilities are combined by scaling down the acoustic scores rather than scaling the language model probabilities as is commonly done in decoding. While this scaling process makes no difference when finding only the best path, when probabilities are added the way scaling is performed is very important. This acoustic scaling process leads to a much broader posterior distribution of arc probabilities and is essential when computing confidence scores.

The arc posteriors are used to calculate time-dependent word posteriors for each time frame in the utterance. For a given time the arc posteriors of all arcs spanning this time which correspond to the same word are summed. The final word posterior probability of a word, with particular start and end times, is calculated as the geometric mean of the corresponding time-dependent posteriors in this interval and this value is used as a confidence score.

## 3. LATTICE-BASED MLLR

One problem with confidence score based MLLR is that a reasonable amount of adaptation data may need to be discarded which limits the accuracy of the estimated transformation matrices. As an alternative, a method was developed to directly use a lattice representation of each utterance which is traversed to provide the statistics needed for MLLR adaptation. This, in principle, means that no data needs is discarded but rather is included so that each frames gives a weighted contribution to the statistics gathered for several HMM states. A similar lattice-based MLLR adaptation method to that presented here, which appears to have been developed contemporaneously with the current work, was recently reported in [9].

Standard MLLR uses a forward-backward pass through just a single HMM model sequence when computing the posterior probability of each Gaussian at each frame and accumulating the necessary statistics for MLLR. The idea behind lattice based MLLR is that the forward-backward pass is performed through the recognition lattice of alternatives paths. Therefore the posterior probability of a particular state at a particular time will include weighted contributions from all relevant word instances that were in the lattice at that time.

The implementation used here employs model-marked lattices which give the HMM boundary information for each arc of a word lattice. This boundary information is used to compact the word level lattice to a model-level structure while still retaining the associated language model information. A full forward-backward pass through the lattice using the current model set is then performed with pruning performed using the times associated with model boundaries with an additional margin of typically 50ms. This process is the same as that used for MMIE training in [14]. The forward-backward pass computes the posterior probability of being in each Gaussian of each HMM state for every lattice arc at each time. This is equivalent to computing the product of an arc posterior probability (from a forward-backward pass at the lattice node level) with a Gaussian posterior probability given the arc.

During the forward-backward pass it is necessary to combine the likelihoods from an HMM-based acoustic model and the language model. For similar reasons to those discussed in Section 2, this is again done by scaling the acoustic model log likelihoods by the inverse of of the normal language model scale factor. This is important in the context of lattice-based MLLR since it greatly broadens the posterior distribution of Gaussians at each time.

In order not to take into account very unlikely Gaussians when gathering MLLR statistics, a threshold on the Gaussian posteriors can be set. It should be noted that unlike the use of a confidence threshold in in Section 2, the lattice MLLR posterior threshold still (in general) retains the contribution of all data frames to the MLLR transformations.

Finally it should be noted that both lattice based MLLR and confidence based MLLR solely alter the way that the posterior probability of Gaussian occupation during the forward-backward pass is computed. Therefore the techniques can be applied to estimate either unconstrained MLLR or constrained MLLR transforms [4]. In this paper, results for only unconstrained MLLR adaptation are presented in which the Gaussian mean and variance transforms are calculated separately.

## 4. MMILR

Maximum mutual information linear regression (MMILR) estimates the parameters of the linear transformation matrices to optimise the MMI criterion for the adaptation data.

The MMIE objective function can be computed over $R$ adaptation observation sequences $\{\mathcal{O}_1, \ldots, \mathcal{O}_r, \ldots, \mathcal{O}_R\}$ with the corresponding word level transcription $w_r$ by

$$\mathcal{F}_{\mathrm{MMIE}}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda(\mathcal{O}_r|\mathcal{M}_{w_r})P(w_r)}{\sum_{\hat{w}} p_\lambda(\mathcal{O}_r|\mathcal{M}_{\hat{w}})P(\hat{w})} \qquad (2)$$

where, $\mathcal{M}_w$ is the composite model corresponding to the word sequence $w$ and $P(w)$ is the probability of the corresponding sequence given the language model. The numerator term in (2), $p_\lambda(\mathcal{O}_r|\mathcal{M}_{w_r})$, is the MLE objective function. The denominator can be represented by the likelihood of each utterance given the full recognition model that encodes all possible word sequences.

However, computation of the denominator for a large vocabulary task is still very computationally expensive and an approximation using word lattices to compactly encode the most likely word sequences can be used to make large vocabulary MMIE estimation feasible. Further details of lattice-based MMIE training can be found in [14].

Assume for simplicity of notation that there is a single adaptation observation sequence of length $T$, and that a particular mean transformation matrix $W_m$ is shared by $P$ Gaussians $\{m_1 \ldots m_P\}$ with means $\mu_{m_p}$ and diagonal covariance matrix $\Sigma_{m_p}$. It can be shown that the MMIE objective function is optimised with respect to to a mean transformation matrix $W_m$ by solving the following equation

$$\sum_{t=1}^{T}\sum_{p=1}^{P}(L_{m_p}^{\mathrm{num}}(t) - L_{m_p}^{\mathrm{den}}(t))\Sigma_{m_p}^{-1}o(t)\xi'_{m_p} = \sum_{t=1}^{T}\sum_{p=1}^{P}(L_{m_p}^{\mathrm{num}}(t) - L_{m_p}^{\mathrm{den}}(t))\Sigma_{m_p}^{-1}W_m\xi_{m_p}\xi'_{m_p} \qquad (3)$$

where $\xi_{m_p}$ is the extended mean vector Gaussian component $m$, $o(t)$ is the observation at time $t$, and $L_{m_p}(t)$ is the posterior probability of occupying Gaussian $m$ at time $t$. The use of $num$ and

$den$ refer to the use of either the numerator (correct word sequence) or the denominator (word lattice approximation to recognition model) of (2) when computing the $L_{m_p}(t)$ values.

Equation (3) can be viewed as simply the standard MLLR formulation with the normal Gaussian occupation probabilities replaced by $(L^{\text{num}}_{m_p}(t) - L^{\text{den}}_{m_p}(t))$, and hence can be solved in the same way as MLLR [6].

The implementation of lattice-based MMIE was used to compute the statistics required for MMILR transform estimation. The forward-backward pass again uses acoustic model likelihood scaling to broaden the posterior distribution and also uses unigram language model scores for the same reasons of improved generalisation as discussed in [14].

## 5. SWITCHBOARD EXPERIMENTS

The confidence score and lattice-based MLLR techniques were evaluated using the Switchboard-1 corpus. Each speech frame is represented by a 39 dimensional feature vector with 13 MF-PLP (including $c_0$) cepstral parameters with their first and second differentials. The basic HMM set uses decision tree clustered cross-word triphones with 2945 speech states and 12 Gaussians per state and is trained on the 18 hour Minitrain corpus as defined by BBN. Recognition uses a trigram language model trained on 2 million words of Switchboard transcriptions trigrams, a 24k word vocabulary and a pronunciation dictionary based on the 1993 LIMSI pronunciation dictionary. The data used for testing was from the Minitrain Test set. For recognition lattices generated by a non-adapted system were rescored. Further details of the system setup can be found in [10].

The test set was warped using a bilinear transform [8]. Block diagonal mean and diagonal variance unsupervised MLLR transforms were generated using interleaved decoding and adaptation passes (iterative MLLR). Adaptation setups that either used a single global transform for speech (global) or a 128 leaf regression class (tree) were used. In both cases a separate silence transform was estimated. When the regression class tree was used, an average of 10 speech transforms were calculated from the available data.

Table 1 presents the word error rates (WER) for standard unsupervised MLLR (Standard), MLLR using confidence scores (Confidence) and lattice MLLR (Lattice) for six adaptation/decoding iterations. The baseline unadapted system has a WER of 40.73%.

| It. | Standard | | Confidence | | Lattice | |
|---|---|---|---|---|---|---|
| | global | tree | global | tree | global | tree |
| 1 | 38.33 | 38.36 | 37.97 | 37.89 | 38.35 | 37.94 |
| 2 | 38.21 | 38.27 | 37.88 | 37.82 | 38.22 | 37.30 |
| 3 | 38.30 | 38.21 | 37.83 | 37.85 | 37.94 | 37.00 |
| 4 | 38.24 | 38.18 | 37.83 | 37.86 | 37.93 | 36.88 |
| 5 | 38.16 | 38.18 | 37.86 | 37.86 | 37.80 | 36.61 |
| 6 | 38.11 | 38.18 | 37.83 | 37.86 | 37.86 | 36.75 |

**Table 1**. % WER for standard unconstrained MLLR, confidence score MLLR and lattice MLLR on the Switchboard Minitrain test set.

It should be noted that for standard MLLR the use of the regression class tree and more transformations is of no help (performance is slightly poorer) since the transformation estimation lacks robustness in the face of transcription errors. The use of confidence scores yields a 0.32% absolute redcution in WER using a regression class tree. In this case also, the global transform and the tree give similar performance. The threshold used in confidence-based MLLR was 20%.

For lattice MLLR, a global transform gives similar results to a global transform with confidence-based MLLR and hence slightly better than standard MLLR. However when the regression class tree is used and several iterations of adaptation/decoding are used, improved results are obtained. For instance, after 6 iterations of adaptation and decoding a reduction in WER of 1.43% absolute is obtained using lattice MLLR relative to the standard case.

## 6. WSJ/NAB S3 EXPERIMENTS

This section describes experiments used to evaluate the effectiveness of MMILR for the case of supervised adaptation to non-native speakers of an HMM system trained on natives. In a case such as this, there is a severe mismatch between the original HMMs and the adaptation data, and effective transform parameter estimation is important.

The speaker independent system used gender independent decision tree clustered triphone HMMs with 6399 speech states and 12 component Gaussian mixture output distributions. It was trained using the SI-284 WSJ0+1 data set. The speech is represented by 39 dimensional feature vectors with 13 MF-PLP coefficients and their first and second differentials with cepstral mean normalisation applied to each utterance. This setup is an MF-PLP version of the HMM-1 model set described in [12].

The MMILR technique was tested on the 1994 North American Business News (NAB) Spoke 3 (S3) task. There are 40 sentences of adaptation data for each speaker. For MMILR, the adaptation sentences were recognised using the standard Lincoln Labs 20k bigram grammar (modified to include the words missing in the adaptation data) and word lattices generated. The actual denominator lattices used unigram scores from this grammar during MMILR transform estimation. The test data for the task is limited to a 5k word vocabulary and the standard Lincoln Labs 5k trigram language model was used. Results are reported using both the 1994 S3 development and evaluation sets.

The baseline word error rates for the system using the native speaker models are 21.42% for the development test data and 17.61% for the evaluation data. Standard supervised mean and variance MLLR adaptation was used and test-set lattices created from these adapted models, which gave error rates of 13.70% and 11.68% respectively. It should be noted that the word error rates do not use the official NIST tools/mappings to compute word error rate which results in an increase in WER values.

| It. | Std | MMILR | | | | |
|---|---|---|---|---|---|---|
| | U1 | M1 | M2 | M5 | M6 | M7 |
| 1 | 13.70 | 12.55 | 12.24 | 11.38 | 11.16 | 11.00 |
| 2 | 12.77 | 12.46 | 12.19 | 11.12 | 11.12 | 10.90 |
| 3 | 12.75 | 12.46 | 12.27 | 11.36 | 11.16 | 11.00 |
| 4 | 12.55 | 12.12 | 12.29 | 11.43 | 11.19 | 11.04 |
| 5 | 12.70 | 12.12 | 12.17 | 11.38 | 11.16 | 11.02 |
| 6 | 12.60 | 12.10 | 12.19 | 11.48 | 11.16 | 11.12 |

**Table 2**. % WER for the 1994 NAB Spoke 3 development test corpus. U1 is result of standard MLLR and columns M1 to M7 use MMILR with adaptation/test lattices generated column by column. It. denotes the iteration of adaptation.

| It. | Std | MMILR | | | | |
|---|---|---|---|---|---|---|
| | U1 | M1 | M2 | M5 | M6 | M7 |
| 1 | 11.68 | 11.01 | 10.56 | 10.56 | 10.12 | 10.06 |
| 2 | 11.77 | 10.98 | 10.68 | 10.37 | 10.12 | 10.15 |
| 3 | 11.43 | 11.07 | 10.62 | 10.26 | 10.17 | 9.95 |
| 4 | 11.12 | 10.84 | 10.65 | 10.26 | 10.03 | 10.12 |
| 5 | 11.35 | 10.70 | 10.56 | 10.20 | 10.17 | 10.15 |
| 6 | 11.01 | 10.40 | 10.54 | 10.17 | 10.23 | 10.17 |

**Table 3**. % WER for the 1994 NAB Spoke 3 evaluation test corpus. U1 is the result of standard MLLR and columns M1 to M7 use MMILR with adaptation/test lattices generated column by column. It. denotes the iteration of adaptation.

The effect of iterating MLLR i.e. performing multiple iterations of MLLR estimation with the same supervised transcription is shown in the U1 column in Tables 2 and 3. Further improvements were obtained with this approach. The initial transforms from the first line of U1 were used to generate test-set lattices for further iterations of MLLR. A significant improvement in performance was obtained by this iterative process especially for the development corpus.

The results of using MMILR are also given in Tables 2 and 3. Using the initial standard MLLR lattices, MMILR transforms were estimated. The results in column M1 were generated first and all use the adaptation and test data lattices generated by models U1. The first iteration of MMILR transform estimation gives an especially large improvement relative to standard MLLR (1.15% absolute for the development data), although the size of the improvement is reduced slightly with more iterations. The first iteration of MMILR was used to generate new adaptation and test lattices for the next column of (M2), and this process is repeated until the lattices for M7 were generated. Note that at each stage, six iterations of MMILR were performed with the same adaptation and test lattices. The missing columns (M3 and M4) show the same trends as the other columns; which give a continued improvement with more iterations.

The final results presented in M7 show relative improvements of 19% and 13% for development and evaluation data sets respectively over a single iteration of standard MLLR and 12% and 8% respectively using iterative MLLR. The process of re-generating adaptation/test lattices is especially important for column M2 since this is the first case where MMILR transforms are used. By the time this process has been iterated to column M7 the performance has converged.

## 7. CONCLUSIONS

This paper has described three techniques for estimating the parameters of linear transforms for speaker adaptation. It was shown that for unsupervised adaptation, lattice-based MLLR significantly outperforms confidence score MLLR and allows the robust unsupervised estimation of more adaptation transforms for high error rate data such as Switchboard. The lattice-based adaptation procedure could also be applied to other techniques, such as MAP [5], if unsupervised adaptation is being performed.

MMILR replaces the maximum likelihood estimation from standard MLLR with the maximum mutual information objective function. It was shown that significant improvements in performance could be achieved using this method for the task of supervised adaption of a native-speaker system to non-natives.

## 8. ACKNOWLEDGEMENTS

Luis Felipe Uebel is supported by a scholarship from the CNPq (Brazilian Council of Research). The work is in part supported by a grant from GCHQ.

## 9. REFERENCES

[1] T. Anastasakos & S.V. Balakrishnan (1998). The Use of Confidence Measures in Unsupervised Adaptation of Speech Recognizers. *Proc. ICSLP'98*, pp. 2303–2306, Sydney.

[2] M.J.F. Gales & P.C. Woodland (1996). Mean and Variance Adaptation within the MLLR Framework. *Computer Speech and Language*, vol. 10, pp. 249–264.

[3] G. Evermann & P.C. Woodland (2000). Large Vocabulary Decoding and Confidence Estimation using Word Posterior Probability. *Proc. ICASSP'00*, pp. 1655–1659, Istanbul.

[4] M.J.F. Gales (1998). Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition. *Computer Speech and Language*, vol. 12, pp. 75–98.

[5] J.L. Gauvain & C.H. Lee (1994). Maximum A-Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Trans. SAP*, Vol. 2, pp. 291–298.

[6] C.J. Leggetter & P.C. Woodland (1995). Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs. *Computer Speech and Language*, Vol. 9, pp. 171–185.

[7] C.J. Leggetter & P.C. Woodland (1995). Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression. *Proc. Eurospeech'95*, pp. 1155-1158, Madrid.

[8] J.W. McDonough, G. Zavaliagkos & H. Gish (1996). An Approach to Speaker Adaptation based on Analytic Functions. *Proc. ICASSP'96*, pp. 721–724, Atlanta.

[9] M. Padmanabhan, G. Saon & G. Zweig (2000). Lattice-Based Unsupervised MLLR for Speaker Adaptation. *Proc. ISCA ITRW ASR2000*, pp. 128–131, Paris.

[10] L.F. Uebel & P.C. Woodland (1999). An Investigation into Vocal Tract Length Normalisation. *Proc. Eurospeech'99*, pp. 2527–2530, Budapest.

[11] F. Wallhoff, D. Willett & G. Rigoll (2000). Frame-Discriminative and Confidence-Driven Adaptation for LVCSR. *Proc. ICASSP 2000*, pp. 1835–1838, Istanbul.

[12] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev & S.J. Young (1995). The 1994 HTK Large Vocabulary Speech Recognition System. *Proc. ICASSP'95*, pp. 73-76, Detroit.

[13] P.C. Woodland, D. Pye & M.J.F. Gales (1996). Iterative Unsupervised Adaptation Using Maximum Likelihood Linear Regression. *Proc. ICSLP'96*, pp. 1133-1136, Philadelphia.

[14] P.C. Woodland & D. Povey (2000). Large Scale Discriminative Training for Speech Recognition. *Proc. ISCA ITRW ASR2000*, pp. 7–16, Paris.

[15] T. Zeppenfeld, M. Finke, K. Ries, M. Westphal & A. Waibel (1997). Recognition of Conversational Telephone Speech using The JANUS Speech Engine. *Proc. ICASSP'97*, pp. 1815–1818, Munich.