USING CLASS WEIGHTING IN INTER-CLASS MLLR

Sam-Joo Doh and Richard M. Stern

Department of Electrical and Computer Engineering and School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213, USA
{sjdoh, rms}@cs.cmu.edu

ABSTRACT

A new adaptation method called inter-class MLLR has recently been introduced. Inter-class MLLR utilizes relationships among different transformation functions to achieve more reliable estimates of MLLR parameters across multiple classes, and it produces lower word error rates (WER) than conventional MLLR in circumstances where very little speaker-specific adaptation data are available. This paper describes the application of weights to the neighboring classes to improve the effectiveness with which they are combined with the target class in inter-class MLLR. These weights are obtained from the variance of the estimation error considering the weighted least squares estimation in classical linear regression. In our experiments, the weights provided small improvements in WER for supervised adaptation but almost no improvement in unsupervised adaptation using only a small amount of adaptation data. We also discuss the effect of decreasing the number of neighboring classes as more adaptation data become available, the development of inter-class transformations from the test speaker, and the combination of inter-class MLLR with principal-component MLLR. None of the feasible variations of weighted inter-class MLLR provided significant improvements to recognition accuracy.

1. INTRODUCTION

Adaptation is a process that reduces differences between training and testing conditions, usually through the use of a small amount of adaptation data. In transformation-based adaptation such as conventional maximum-likelihood linear regression (MLLR), the parameters used in recognition (such as the means of the underlying Gaussian mixture components) are clustered into transformation classes, with all the parameters in a particular transformation class updated by the same linear transformation function. The number of the parameters characterizing the transformation function is usually much smaller than the number of recognition model parameters. While we can generally obtain useful estimates of these parameters with only a small amount of adaptation data, some of the information about individual model parameter is lost, which can impair recognition accuracy.

In conventional MLLR the linear transformation function is estimated using adaptation data from within the transformation class that it represents. Increasing the number of transformation classes enables the transformation function for each one to be modeled more specifically to the data for the class, but with reduced reliability since the estimates of the transformation functions are necessarily based on a smaller amount of adaptation data. If the estimates are not reliable, they will not be helpful for improving recognition accuracy.

It is useful to consider relationships among different parameters when only a small amount of adaptation data is available. Most previous studies use models of correlation or regression among the recognition model parameters in a Bayesian framework [*e.g.* 1, 2, 6, 7]. Because there are more than thousands of these parameters in most speech recognition systems, it may not be effective to consider the correlations among only a few parameters with a small amount of adaptation data, and it may require too much computation to consider the correlations among all the parameters.

Recently a new adaptation method called *inter-class MLLR* has been introduced [5]. Inter-class MLLR utilizes relationships among different transformation functions to achieve more reliable estimates of MLLR parameters across multiple classes. In this method, inter-class transformations given by linear regressions are used to modify the baseline mean vectors in the neighboring classes so that the neighboring classes can contribute to the estimates the MLLR parameters of the target class. The inter-class transformations are estimated from training data, and function as *a priori* information. If the inter-class transformations are identity functions, inter-class MLLR becomes the same as conventional single-class MLLR. This idea also can be applied to other types of transformation-based adaptation and general parameter estimation problems.

In inter-class MLLR, several neighboring classes are considered for each target class. In this procedure, some neighboring classes may be "closer" to the target class than other neighboring classes. In this paper we extend inter-class MLLR by applying different weights to the neighboring classes to incorporate their different contributions to the target class. We also limit the number of neighboring classes to be used as more adaptation data becomes available.

In the following sections, we first review inter-class MLLR, and then describe applying weights to the neighboring classes and limiting the number of the neighboring classes. Finally we describe our experimental results, and summarize our work.

2. INTER-CLASS MLLR

MLLR assumes that an adapted mean vector $\hat{\mu}_k$ for a Gaussian k is related to its baseline mean vector μ_k by linear regression. Consider a case in which we try to estimate the MLLR parameters $(\boldsymbol{A}_m, \boldsymbol{b}_m)$ for an MLLR class m (the "target class").

$$\hat{\mu}_k = \boldsymbol{A}_m \mu_k + \boldsymbol{b}_m, \qquad k \in \text{Class } m \tag{1}$$

In conventional MLLR, $(\boldsymbol{A}_m, \boldsymbol{b}_m)$ are estimated using the MLLR class m only, by maximizing the likelihood, or by minimizing

$$Q_{\mathrm{C}}(m) = \sum_{k \in m} \sum_t \gamma_t(k) (\boldsymbol{o}_t - \boldsymbol{A}_m \mu_k - \boldsymbol{b}_m)^{\mathrm{T}} \boldsymbol{C}_k^{-1} (\boldsymbol{o}_t - \boldsymbol{A}_m \mu_k - \boldsymbol{b}_m)$$

where $\gamma_t(k)$ is the *a posteriori* probability of being in Gaussian mixture k at time t, $\boldsymbol{o}_t$ is the input feature vector at time t (adaptation data), and $\boldsymbol{C}_k$ is the covariance matrix of Gaussian k [8].

Consider another MLLR class $n \neq m$ which has a similar relation given by $(\boldsymbol{A}_n, \boldsymbol{b}_n)$ in conventional MLLR.

$$\hat{\mu}_k = \boldsymbol{A}_n \mu_k + \boldsymbol{b}_n, \qquad k \in \text{Class } n \tag{2}$$

Inter-class MLLR assumes that the inter-class transformation which relates Class m (the target class) and Class n (neighboring classes) is given by another linear regression with $\boldsymbol{T}_{mn}$ and $\boldsymbol{d}_{mn}$, and Eq. (2) is written as follows.

$$\hat{\mu}_k = \boldsymbol{A}_m(\boldsymbol{T}_{mn}\mu_k + \boldsymbol{d}_{mn}) + \boldsymbol{b}_m, \quad k \in \text{Class } n \tag{3}$$

Defining the modified mean vector to be $\mu_k^{(mn)} \equiv \boldsymbol{T}_{mn}\mu_k + \boldsymbol{d}_{mn}$, Eq. (3) becomes

$$\hat{\mu}_k = \boldsymbol{A}_m \mu_k^{(mn)} + \boldsymbol{b}_m, \quad k \in \text{Class } n \tag{4}$$

$(\boldsymbol{A}_m, \boldsymbol{b}_m)$ are unknown parameters in Eq. (4). Therefore they can be estimated from the neighboring class n using $Q_n(m)$ which is similar to $Q_{\mathrm{C}}(m)$ except $\mu_k^{(mn)}$ and Class n.

$$Q_n(m) = \sum_{k \in n} \sum_t \gamma_t(k)(\boldsymbol{o}_t - \boldsymbol{A}_m\mu_k^{(mn)} - \boldsymbol{b}_m)^{\mathrm{T}} \boldsymbol{C}_k^{-1}(\boldsymbol{o}_t - \boldsymbol{A}_m\mu_k^{(mn)} - \boldsymbol{b}_m)$$

Now, considering Eqs. (1) and (4), $(\boldsymbol{A}_m, \boldsymbol{b}_m)$ are estimated by maximizing the overall likelihood from the target class m and neighboring class n, or by minimizing $Q_{\mathrm{I}}(m)$.

$$Q_{\mathrm{I}}(m) = Q_{\mathrm{C}}(m) + \sum_{n \in \text{ Neighbors}} Q_n(m) \tag{5}$$

where $Q_{\mathrm{C}}(m)$ is the contribution of the target class m, and $Q_n(m)$ is the contribution of the neighboring class n.

The estimates of $(\boldsymbol{A}_m, \boldsymbol{b}_m)$ from Eq. (5) will be more reliable than those from $Q(m)$ because more adaptation data are used in Eq. (5). In this procedure, $(\boldsymbol{T}_m, \boldsymbol{d}_m)$ work as *a priori* information, and are obtained from training data in advance [3].

3. APPLICATION OF WEIGHTS

In inter-class MLLR as described previously, all the neighboring classes contribute equally to updating the target class. In practice, the neighboring classes are not as important as the target class, and some neighboring classes may be "closer" to the target class than other neighboring classes. Therefore we can apply different weights to the neighboring classes to represent their different contributions to the target class. In this paper we use estimation error to measure the "closeness" of the neighboring classes to the target class.

Eq. (4) shows the relation between the modified baseline mean vector $\mu_k^{(mn)}$ and the corresponding adapted mean vector $\hat{\mu}_k$ in the neighboring class n. This equation can be interpreted as a relationship between $\mu_k^{(mn)}$ and the input feature vector $\boldsymbol{o}_{t,s}$ with estimation error $\boldsymbol{e}_{k,t,s}^{(mn)}$ [3], *i.e.*

$$\boldsymbol{o}_{t,s} = \boldsymbol{A}_{m,s}\mu_k^{(mn)} + \boldsymbol{b}_{m,s} + \boldsymbol{e}_{k,t,s}^{(mn)}, \qquad k \in \text{Class } n \tag{6}$$

where subscript s denotes the training speaker. Since the parameters $(\boldsymbol{A}_{m,s}, \boldsymbol{b}_{m,s})$ are known for each training speaker, the estimation error $\boldsymbol{e}_{k,t,s}^{(mn)}$ is easily obtained for each time frame.

$$\boldsymbol{e}_{k,t,s}^{(mn)} = \boldsymbol{o}_{t,s} - \boldsymbol{A}_{m,s}\mu_k^{(mn)} - \boldsymbol{b}_{m,s}, \qquad k \in \text{Class } n \tag{7}$$

The variance $\boldsymbol{C}_{\boldsymbol{e}_k}^{(mn)}$ of the error is obtained from all training speakers using the Baum-Welch method as the Gaussian mixtures are reestimated in regular retraining. For the target class m, the estimation error $\boldsymbol{e}_{k,t,s}$ and its variance $\boldsymbol{C}_{\boldsymbol{e}_k}$ are obtained in similar fashion.

$$\boldsymbol{o}_{t,s} = \boldsymbol{A}_{m,s}\mu_k + \boldsymbol{b}_{m,s} + \boldsymbol{e}_{k,t,s}, \qquad k \in \text{Class } m \tag{8}$$

Considering the weighted least squares estimation in classical linear regression [9], we substitute the variances $\boldsymbol{C}_{\boldsymbol{e}_k}$ and $\boldsymbol{C}_{\boldsymbol{e}_k}^{(mn)}$ for $\boldsymbol{C}_k$ in $Q_{\mathrm{C}}(m)$ and $Q_n(m)$ respectively. These variances can be considered as weights to the Gaussians. Since the inverse of the variance is used as the weight, an MLLR class with a large variance of the estimation error will get a small weight.

In our experiment we do not have enough training data to estimate accurately the variance of the error, so we estimate the average ratio of the baseline variance to the variance of the error. Consider a target class m and neighboring class n. The corresponding ratios w_m for the target class, and w_{mn} for the neighboring class n become

$$w_m = \underset{k \in \text{ Class } m}{average} \left(\frac{\sigma_{k,i}^2}{\sigma_{\boldsymbol{e}_k,i}^2} \right) \tag{9}$$

$$w_{mn} = \underset{k \in \text{ Class } n}{average} \left(\frac{\sigma_{k,i}^2}{\sigma_{\boldsymbol{e}_k^{(mn)},i}^2} \right) \tag{10}$$

where $\sigma_{k,i}^2$, $\sigma_{\boldsymbol{e}_k,i}^2$, and $\sigma_{\boldsymbol{e}_k^{(mn)},i}^2$ are the diagonal elements of $\boldsymbol{C}_k$, $\boldsymbol{C}_{\boldsymbol{e}_k}$, and $\boldsymbol{C}_{\boldsymbol{e}_k}^{(mn)}$, respectively.

The weights w_m and w_{mn} are combined as follows:

$$Q_{\mathrm{W}}(m) = w_m Q_{\mathrm{C}}(m) + \sum_{n \in \text{ Neighbors}} w_{mn} Q_n(m) \tag{11}$$

In inter-class MLLR, the number of neighboring classes to be used depends on the amount of adaptation data. The neighboring classes are ranked in order of their "closeness" to the target class. Adaptation data are selected from classes of decreasing proximity to the target class until there are sufficient data to estimate the target function. If only a very small amount of data is available, then all neighboring classes may be used. As more data becomes available, the number of neighboring classes used declines. In the limit, no neighboring classes are used and inter-class adaptation asymptotes to conventional multi-class adaptation.

The variance $\boldsymbol{C}_{\boldsymbol{e}_k}^{(mn)}$ is used to measure the closeness of neighboring class n to the target class m. For each target class, neighboring classes are sorted according to the variances of their errors. Adaptation data from the closest neighboring class are accumulated first in Eq. (11), then from the next closest neighboring class until sufficient data are used. We can consider $w_m\gamma_t(k)$ and $w_{mn}\gamma_t(k)$ in Eq. (11) as effective counts of the

adaptation data. We accumulate the counts until they exceed a preset threshold. We can control the threshold and the number of classes accumulated to get better recognition accuracy. The best threshold can be obtained through prior experimentation.

4. EXPERIMENTAL RESULTS

The methods described above are evaluated using non-native English speakers from the Spoke 3 data in the 1994 DARPA Wall Street Journal (WSJ) evaluation. The recognition test data consisted of 200 sentences, from which 10 non-native speakers read 20 sentences each. The baseline speech recognition system is the CMU SPHINX-III system which used continuous HMMs with 6000 senones, a 39-dimensional feature vector consisting of cepstra (MFCC), delta cepstra, and delta-delta cepstra, and a 5,000-word trigram language model. We used 13 phonetic-based MLLR classes which were similar to those used by Leggetter [3]. The inter-class transformation parameters T_{mn} and d_{mn} were trained from 9 speakers except the test speaker in the 10 evaluation speakers.

Table 1 summarizes the word error rates (WER) after supervised adaptation with 1 or 3 adaptation sentences for each speaker with correct transcriptions. The adaptation sentences and test sentences were different. The WER in the baseline system without adaptation was 27.3%. Inter-class MLLR without weights provided about 15% improvement in WER over conventional MLLR. Applying weights provided further relative improvements of 0.8% to 1.3% which were not statistically significant because of the small amount of data considered.

Adaptation Method	1 Adaptation Sentence	3 Adaptation Sentences
Conventional MLLR (one class)	24.1%	23.1%
Inter-class MLLR without weights	20.4% (15.4%)	19.6% (15.2%)
Inter-class MLLR with weights	20.2% (16.2%)	19.3% (16.5%)

Table 1. Word error rates after supervised adaptation. (Relative improvements over conventional MLLR are shown in parentheses.)

Table 2 shows corresponding results for unsupervised adaptation on sentences from the test set. As in Table 1, the columns in the table indicate the number of sentences used in performing the adaptation for each speaker (1,3, or 20). The MLLR parameters are estimated from blind transcriptions of the sentences by the baseline recognition system, and the test sentences are subsequently recognized again using the adapted models. The application of weights in unsupervised adaptation appeared to provide no benefit with only one adaptation sentence, and only an extremely modest benefit when 20 sentences were used for adaptation. We speculate that the limited benefit in unsupervised adaptation is a consequence of transcription errors by the baseline system. In other words, unsupervised adaptation needs more smoothing than supervised adaptation to average out the effect of incorrect transcriptions.

Adaptation Method	1 Test Sentence	3 Test Sentences	20 Test Sentences
Conventional MLLR (one class)	26.7%	25.9%	23.9%
Inter-class MLLR without weights	24.0% (10.1%)	20.9% (19.3%)	20.1% (15.9%)
Inter-class MLLR with weights	24.3% (9.0%)	20.9% (19.3%)	19.9% (16.7%)

Table 2. Word error rates after unsupervised adaptation. (Relative improvements over conventional MLLR are shown in parentheses.)

The application of weights that are smaller for the neighboring classes than for the target class enables unsupervised adaptation to focus on the target class, resulting in less smoothing. It also reduces the effective amount of input data (because the contributions of neighboring classes are multiplied by their weights). This could be a problem especially when the amount of the data is small. As we have more data, we can have enough smoothing to obtain benefits from the weights.

Fig. 1 plots word error rate as a function of the threshold after supervised adaptation. The value next to each data point is the average number of classes used to estimate the MLLR parameters of the target class. With a very small threshold, only 1 class (target class itself) is accumulated because the amount of adaptation from the target class exceeds the threshold. With a very large threshold, all the classes are accumulated. For example, with a threshold value of 1500, almost all classes (an average of 12.9 out of 13) are used in the 3 adaptation sentence case, while only 5.4 classes are used in the 10 adaptation sentence case, because there is more adaptation data from each class in the 10 adaptation sentence case (compared to the 3-sentence case). If the threshold is too small, the WER becomes high because too little adaptation data are used. We note that if 10 adaptation sentences are available, the best WER is obtained at a threshold value of 1000, which corresponds to the use of about only 3.6 transformation classes. If only 3 adaptation sentences are available, however,

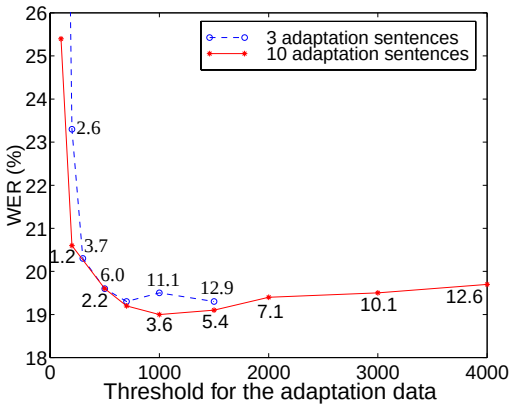

Figure 1. Word error rates as a function of the threshold for adaptation data in inter-class MLLR. The value next to each data point is the average number of classes used.

WER continues to improve until all the data are used, and all transformation classes are contributing to the final parameter estimates.

The performance of inter-class MLLR will depends on the *"quality"* of the inter-class transformations. If the inter-class transformations do not match the characteristics of the test data, they may not be helpful. Table 3 describes the WER observed when the inter-class transformations were trained from different data. The first row in Table 3 repeats the results from the case of "Inter-class MLLR without weights" in Table 1. The WER using the inter-class transformations obtained from the native speakers [5] was worse than the results from the case of the non-native speakers. We believe that this is because the native speakers have different characteristics from the test speakers who are non-native speakers.

In comparison, inter-class transformations obtained from the test speakers themselves provided very good WER. This is to be expected because the purpose of the inter-class transformations is to estimate the *a priori* characteristics of the test speaker. The test speaker is not normally available to train adaptation parameters, but in circumstances where he or she is, substantial improvements in accuracy can be obtained. One approach to improved recognition accuracy may be to prepare several sets of inter-class transformations representing different type of speakers, and selecting an appropriate set for each test speaker.

Adaptation Method	1 Adaptation Sentence	3 Adaptation Sentences
Inter-class Transformation from Non-Native Speakers	20.4%	19.6%
Inter-class Transformation from Native Speakers	22.0%	21.1%
Inter-class Transformation from Test Speaker	16.5%	16.8%

Table 3. Word error rates after supervised adaptation using inter-class transformations trained from different speakers.

We also applied principal component MLLR [3] in combination with inter-class MLLR, but we did not obtain significant additional improvement in accuracy. We believe that the principal component approach provides less benefit when applied to the inter-class MLLR than when applied to conventional MLLR because the ratio between the largest and smallest eigenvalues is much smaller for inter-class MLLR than for conventional MLLR.

5. SUMMARY

This paper describes the application of weights to neighboring classes to characterize more effectively their contributions to the target class in inter-class MLLR. The weights are proportional to the inverse of the variance of the estimation error considering the weighted least squares in classical linear regression. In our experiments, the weight provided only small improvements for supervised adaptation and virtually no improvement for unsupervised adaptation. This is because the weights makes the adaptation focus more on the target class while the unsupervised adaptation needs more smoothing to reduce the effect of incorrect transcriptions. The weights also reduce the effective amount of input data which can be a problem especially when the amount of adaptation data is small. As we have more adaptation data, we can have enough smoothing to benefit from the weights. We reduced the number of neighboring classes as more adaptation data are available, setting a threshold to control the amount of adaptation data used.

We also performed experiments using inter-class transformations obtained from different training data. Inter-class transformations obtained from training data which have similar characteristics provide better recognition accuracy. If we prepare several sets of inter-class transformations which represent different type of speakers, and select an appropriate set for the test speaker, then we can obtain a good improvement in recognition accuracy.

Finally we applied principal component MLLR in combination with inter-class MLLR, but without obtaining significant additional improvement in accuracy. The distribution of the eigenvalues changes in inter-class MLLR, and the benefits of using the principal components seems to be smaller.

ACKNOWLEGEMENTS

This research was sponsored by the Space and Naval Warfare Systems Center, San Diego, under Grant No. N66001-99-1-8905. The content of the information in this publication does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

6. REFERENCES

[1] M. Afify, Y. Gong and J.-P. Haton, "Correlation based predictive adaptation of hidden Markov models," *Proc. of Eurospeech*, pp. 2059-2062, 1997.

[2] S. M. Ahadi and P. C. Woodland, "Combined Bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, pp. 187-206, July 1997.

[3] S.-J. Doh, *Enhancements to Transformation-Based Speaker Adaptation: Principal Component and Inter-Class Maximum Likelihood Linear Regression*, Ph.D. Thesis, Carnegie Mellon University, July 2000.

[4] S.-J. Doh and R. M. Stern, "Weighted principal component MLLR for speaker adaptation," *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 1999.

[5] S.-J. Doh and R. M. Stern, "Inter-class MLLR for speaker adaptation," *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1755-1758, 2000.

[6] Q. Huo and C.-H. Lee "On-line adaptive learning of the correlated continuous density hidden Markov models for speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 4, pp. 386-397, July 1998.

[7] M. J. Lasry and R. M. Stern, "A posteriori estimation of correlated jointly Gaussian mean vectors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 4, pp. 530-535, July 1984.

[8] 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, pp.171-185, 1995.

[9] R. H. Myers, *Classical And Modern Regression With Applications*, PWS-KENT Publishing Company, Boston, 1990.