INTER-CLASS MLLR FOR SPEAKER ADAPTATION

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

This paper examines the use of interdependencies of parameter classes in transformation-based speaker adaptation algorithms such as maximum likelihood linear regression (MLLR). In transformation-based adaptation, increasing the number of transformation classes can provide more detailed information for adaptation, but at the expense of greater estimation error with small amounts of data. In this paper we introduce a new procedure, inter-class MLLR, which utilizes relationship between different classes to achieve both detailed and reliable transformation-based adaptation using limited data. In this method, the inter-class relation is given by a linear regression which is estimated from training data. In experiments using non-native English speakers from the Spoke 3 data in the 1994 DARPA Wall Street Journal evaluation, inter-class MLLR provided a relative reduction in word error rates of 11.3% compared to conventional MLLR.

1. INTRODUCTION

Many adaptation methods have been developed to reduce the mismatch between training and test conditions in a speech recognition system. If we can modify the model parameters of a speech recognition system so that they more closely resemble the test conditions, we can obtain better recognition accuracy.[1] If a large amount of data representing the test conditions is available, improved performance can be achieved by simply retraining the model parameters. We usually do not have enough data for complete retraining, and we would like to achieve good recognition accuracy even with only a small amount of adaptation data. One way of obtaining improved adaptation performance with limited adaptation data is by constraining the model parameters to be estimated, and thereby reducing their number. In this paper we present a new method which utilizes relationship between different classes in transformation-based adaptation as exemplified by maximum likelihood linear regression (MLLR) [11].

In transformation-based adaptation, we assume that a group of model parameters are transformed by the same function. This reduces the number of parameters to be estimated because the number of the parameters in the transformation function is usually much smaller than the original number of model parameters. While we generally obtain better estimates of these parameters with a smaller amount of adaptation data (which tends to improve recognition accuracy), some of the information of the individual model parameter is lost (which can impair recognition accuracy).

1. In this paper we focus on the transformation of the recognition model parameters, especially Gaussian means, even though adaptation can also be obtained by transformation of input feature vectors.

Conventional transformation-based adaptation typically proceeds as follows. The model parameters are clustered into sub-classes. A form for the transformation function is chosen (usually a simple linear function). The parameters of the transformation function are then estimated for each class by using the adaptation data associated with the class. Finally, the model parameters in each class are modified by applying the estimated class-dependent transformation function to them. Each class is modified independently and does not affect any other.

As an example, suppose that the Gaussian means underlying a set of mixture densities are as shown by the dots in Fig. 1. These means can be classified either as a single class (Method I) or as two classes (Method II). For Method I we assume that all the Gaussian means (μ_i) are transformed to adapted means ($\hat{\mu}_i$) by a single function $f(\cdot)$:

$$\hat{\mu}_i = f(\mu_i), \qquad i \in \text{all Gaussians} \tag{1}$$

For Method II we develop two functions $f_1(\cdot)$ and $f_2(\cdot)$, one for each class:

$$\hat{\mu}_i = f_1(\mu_i), \qquad i \in \text{Class 1} \tag{2}$$

$$\hat{\mu}_i = f_2(\mu_i), \qquad i \in \text{Class 2} \tag{3}$$

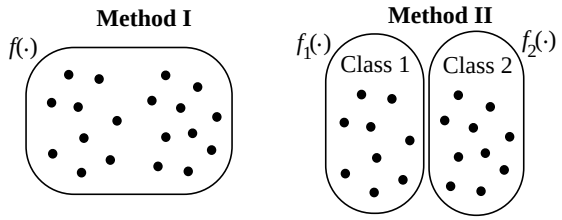

Figure 1. Classification of Gaussian features as either a single class (Method I) or as two classes (Method II)

If only a small amount of speaker-specific data are present, single-class adaptation (Method I) may be more effective than multi-class adaptation (Method II). With multi-class adaptation, we have a smaller amount of data available per class, so we may not obtain reliable estimates of the functions $f_1(\cdot)$ and $f_2(\cdot)$. In single-class adaptation, all data are used to estimate the function $f(\cdot)$ so that function can be estimated more reliably. Multi-class adaptation may be more effective than single-class adaptation, however, if enough data are available to estimate the transformation function reliably for each class because the parameters of each class are modified in a fashion that is appropriate for the individual class. Hence, the optimum number of classes will depend on the amount of available adaptation data.

Gales [7] used regression class trees to find optimal regression classes in conventional MLLR. He developed a binary regression class tree and dynamically selected the classes so that either each class had sufficient data or the overall likelihood of the adaptation data was maximized. He also described a method in which MLLR was applied to the root node and the sub-nodes of the regression class tree iteratively. Chien *et al.* [5] described a hierarchical adaptation procedure that used a tree structure to control the transformation sharing and to obtain reliable transformation parameters.

As we have noted above, new parameters are estimated independently for each class in conventional multi-class regression. Nevertheless, under many circumstances, estimation accuracy (and hence recognition accuracy) can be improved if information is shared across parameter classes. Correlation between different model parameters, for example, has been used by several research groups to improve estimation of parameter values with sparse data. Most previous work using parameter correlation has been done within a Bayesian framework (*e.g.* [1,2,4,9,10,13]). Within the MLLR framework, Bocchieri *et al.* [3] used correlation between the shift terms (*i.e.* the b term in $\hat{\mu}_i = A\mu_i + b$) to refine further the transformation function. Correlation information was limited to the shift terms in their work, and the relation between the multiplication term (*i.e.* the A term in $\hat{\mu}_i = A\mu_i + b$) was not considered.

In other work, principal component analysis (PCA) has been used to reduce parameter dimensionality. For example, Kuhn *et al.* [12] used "eigenvoices" to capture *a priori* knowledge about speakers' characteristics. Hu [8] applied PCA to describe the correlation between phoneme classes for speaker normalization. These efforts are not directly related to MLLR. Doh and Stern [6] also used PCA, but not as prior knowledge, in developing the weighted principal component MLLR method that reduces the estimation error in MLLR.

In the following sections, we will describe a new method which utilizes the inter-class relationship as prior knowledge to obtain better estimates of the transformation function in MLLR, including the multiplication term as well as the shift term.

2. EXPLOITING INTER-CLASS RELATIONSHIPS

Let us consider some of the issues involving inter-class relationships in terms of the approach depicted schematically as "Method II" in Fig. 1. If there is a known relation between two classes which is given by an inter-class function $g_{12}(\cdot)$ then we can rewrite the transformation of Class 2 in Eq. (3) using the function $f_1(\cdot)$ of Class 1 and a modified Gaussian mean $\mu_i^{(12)}$.

$$\hat{\mu}_i = f_2(\mu_i) = f_1(g_{12}(\mu_i)) = f_1(\mu_i^{(12)}), \quad i \in \text{class 2} \qquad (4)$$

Since $f_1(\cdot)$ is now the only unknown function for both classes in Eqs. (2) and (4), we can use all the data from Classes 1 and 2 to estimate $f_1(\cdot)$. This approach can also be applied to the estimation of the function $f_2(\cdot)$ of Class 2. In this case, we keep the original function $f_2(\cdot)$ in Eq. (3) as it is and substitute Eq. (2) into Eq. (5) using the inter-class function $g_{21}(\cdot)$. Then we can use all the data with Eqs. (3) and (5) to estimate $f_2(\cdot)$ for Class 2.

$$\hat{\mu}_i = f_1(\mu_i) = f_2(g_{21}(\mu_i)) = f_2(\mu_i^{(21)}), \quad i \in \text{class 1} \qquad (5)$$

The functions $f_1(\cdot)$ and $f_2(\cdot)$ estimated by this method will be more reliable than those estimated by conventional multi-class methods. They also provide more effective adaptation to each class than the class-independent function $f(\cdot)$ used in Method I because while the adaptation is primarily based on observations within a given class, it benefits from relevant information from other classes as well.

This method can be extended to any form of transformation-based adaptation. First, multiple transformation classes are defined using clustering or *ad hoc* techniques, and the inter-class functions are estimated using training data. For each target class, the inter-class functions are used to modify the models in other *neighboring* classes. The incoming adaptation data from all classes are then combined to estimate the transformation functions for the target class. This process is repeated for all other target classes.

In this process, the number of neighbor classes to be used depends on the amount of adaptation data. The neighbor 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. Choosing the identity function as the inter-class function is equivalent to conventional single-class adaptation.

The advantages of the inter-class method are illustrated by a series of simulations with artificial data whose results are summarized in Fig. 2. We assume that there are 5 classes and that each class has 5 Gaussians. We have adaptation samples from new Gaussians whose mean values are related to original mean values by linear regression (LR) with known inter-class functions. New Gaussian mean values were estimated using linear regression for the conventional single-class and multi-class adaptation procedures, as well as for inter-class adaptation. We calculated the mean square

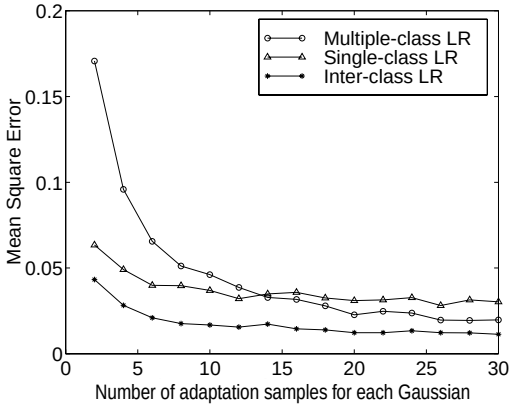

Figure 2. Mean-square errors from simulated estimates of Gaussian means using single-class, multi-class, and inter-class MLLR.

estimation error (MSE) as a function of the number of available samples. As can be seen in Fig. 2, the inter-class LR always provides lower MSE than the other methods in this simulation. We note that while single-class LR provides better MSE than multi-class LR when the number of samples is small, multi-class LR surpasses it as the number of samples increases

3. INTER-CLASS MLLR

Conventional MLLR [11] assumes that a new mean values $\hat{\mu}_i$ is related to its baseline mean value μ_i by linear regression. For an arbitrary regression class m, the relation is given by the multiplication matrix A_m and shift vector b_m.

$$\hat{\mu}_i = A_m \mu_i + b_m, \quad i \in \text{class } m \tag{6}$$

Another regression class $n \neq m$ has a similar relation given by A_n and b_n.

$$\hat{\mu}_i = A_n \mu_i + b_n, \quad i \in \text{class } n \tag{7}$$

In this section we describe inter-class MLLR, in which we assume the inter-class relation from class n (neighbor classes) to class m (target class) is given by linear regression with T_{mn} and d_{mn}, so Eq. (7) can be written as follows.

$$\hat{\mu}_i = A_m(T_{mn}\mu_i + d_{mn}) + b_m, \quad i \in \text{class } n \tag{8}$$

We estimate T_{mn} and d_{mn} from training data. We assume that there are many training speakers with sufficient data to estimate reliable speaker-dependent regression parameters $A_{s,m}$ and $b_{s,m}$ for each speaker s and regression class m. First we estimate $A_{s,m}$ and $b_{s,m}$ from training data using conventional MLLR. We then use these parameters to estimate the inter-class regression parameters T_{mn} and d_{mn}.

For an adapted mean $\hat{\mu}_i^{(s)}$ for speaker s, Eq. (8) becomes

$$\hat{\mu}_i^{(s)} = A_{s,m}(T_{mn}\mu_i + d_{mn}) + b_{s,m}, \quad i \in \text{class } n \tag{9}$$

Letting $\hat{\mu}_i^{(s,m)} = A_{s,m}^{-1}(\hat{\mu}_i^{(s)} - b_{s,m})$, we obtain the linear regression between μ_i and $\hat{\mu}_i^{(s,m)}$.

$$\hat{\mu}_i^{(s,m)} = T_{mn}\mu_i + d_{mn}, \quad i \in \text{class } n \tag{10}$$

When $\hat{\mu}_i^{(s)}$ is the adapted mean for the training data $x^{(s)}$ from speaker s, $\hat{\mu}_i^{(s,m)}$ will be the adapted mean for the modified data $x^{(s,m)} = A_{s,m}^{-1}(x^{(s)} - b_{s,m})$. To estimate T_{mn} and d_{mn} which represent the inter-class relation for different speakers, we apply conventional MLLR technique on Eq. (10) by using $x^{(s,m)}$ from all training speakers together. We estimate T_{mn} and d_{mn} for each neighbor class $n \neq m$ and repeat the same process for other target classes. In this method, we assume that the baseline variances are unchanged.

Once we have T_{mn} and d_{mn}, we can use them in adapting to a new test speaker. In this case, A_m and b_m are unknown parameters in Eq. (8). For $\mu_i^{(mn)} = T_{mn}\mu_i + d_{mn}$, Eq. (8) becomes

$$\hat{\mu}_i = A_m \mu_i^{(mn)} + b_m, \quad i \in \text{class } n \tag{11}$$

This produces a linear regression between the modified baseline mean $\mu_i^{(mn)}$ and adapted mean $\hat{\mu}_i$ given by the A_m and b_m of the regression class m. To estimate A_m and b_m, we apply conventional MLLR technique using Eqs. (6) and (11). We can use either all the neighbor classes or the top N neighbor classes as described in the previous section.

4. EXPERIMENTS

We evaluated inter-class MLLR using non-native English speakers from the Spoke 3 data in the 1994 DARPA Wall Street Journal (WSJ) evaluation. The recognition test data consist of 200 sentences from which 10 non-native speakers read 20 sentences each. We selected 5 adaptation sentences for each speaker which were different from the test sentences, using the correct transcriptions in supervised adaptation fashion. We used SPHINX-III as the baseline speech recognition system. SPHINX-III uses continuous HMMs with 6000 senones, a 39-dimensional feature vector consisting of MFCC cepstra, delta cepstra, and delta-delta cepstra, and a 5,000-word trigram language model. We choose 13 phonetic-based regression classes which are similar to those used by Leggetter [11], and 25 training speakers from WSJ training data to estimate the inter-class regression parameters T_{mn} and d_{mn}. Each training speaker has about 1000 sentences. We used weighted principal component MLLR [6] to obtain more reliable estimates of the speaker-dependent regression parameters $A_{s,m}$ and $b_{s,m}$ in Eq. (9) to estimate the inter-class regression parameters T_{mn} and d_{mn}. We compare results obtained using three types of inter-class regressions for which T_{mn} is assumed to be a full matrix (which allows the parameters to be rotated, scaled, and shifted), diagonal matrix (which allows them to be scaled and shifted only), or the identity matrix (which allows them to be shifted only). The shift vector d_{mn} is used in all three cases. We used all the neighbor classes to estimate each target class in inter-class MLLR. Silence and noise phones are not adapted in these experiments.

Table 1 summarizes the word error rates (WER) obtained using the three types of inter-class MLLR, along with conventional single-class MLLR. Inter-class MLLR with a full matrix reduces the

Adaptation Method	WER
Baseline (unadapted)	27.3%
Conventional MLLR (one class)	23.7% (13.1%)
Inter-class MLLR (shift only)	22.0% (19.4%)
Inter-class MLLR (diag + shift)	21.0% (23.0%)
Inter-class MLLR (full + shift)	21.0% (23.0%)

Table 1. Word error rates for the Spoke 3 in the 1994 DARPA WSJ evaluation after only 5 adaptation sentences. (Relative improvement over the baseline is shown in parenthesis).

WER by 11.3% relative to conventional MLLR, which is greater than the reduction in WER observed using inter-class MLLR with the identity matrix (shift only). We were surprised that the WER obtained with the full matrix was no better than the WER obtained using the diagonal matrix. We speculate that this may be a consequence of greater errors in estimating the larger number of parameters in the full matrix. The improvement in WER from adaptation is less than that reported in the 1994 DARPA evaluation primarily because we use only five sentences of adaptation data.

It is useful to consider the extent to which the benefits obtained using inter-class MLLR as described above approach the greatest possible benefit from adaptation. In a side experiment, we performed conventional MLLR using supervised adaptation from 120 sentences and 13 regression classes, achieving an error rate of 17.3%. This WER is probably about the best that can be expected using this type of adaptation. In a second side experiment we performed inter-class MLLR using only 5 adaptation sentences but obtaining estimates of T_{mn} and d_{mn} in supervised fashion using the same 120 sentences of data from each test speaker. The WER obtained using this form of adaptation was 17.9%, which is close to the best case, suggesting that the transformation parameters T_{mn} and d_{mn} do indeed capture virtually all of the relevant adaptation information contained in the 120 adaptation sentences. When we applied inter-class MLLR using 5 sentences with T_{mn} and d_{mn} estimated from training speakers, the WER was 21.0% as shown in Table 1. It is worse than the previous cases as it is expected.

Finally, we also considered the performance of inter-class MLLR using native English speakers in the Spoke 0 data from the 1994 DARPA WSJ evaluation. In this case inter-class MLLR did not provide better WER than conventional MLLR. We observed that the improvement in WER from single-class to multi-class implementations of conventional MLLR was correspondingly small. This suggests to us that conventional single-class MLLR captures most of the benefit to be expected from MLLR for speaker adaptation for native English speakers.

In previous related work, we have described weighted principal component MLLR [6] which provided greater improvement in WER than inter-class MLLR for the same test data. We are currently evaluating the extent to which combining principal-component and inter-class MLLR is worthwhile.

5. SUMMARY

In this paper, we introduced a new method called inter-class MLLR which utilizes relationship between different classes to achieve both the detail and reliability in transformation-based adaptation, and we applied this method to the MLLR framework. We described how to estimate inter-class regression parameters from training data and apply them for adaptation. In our experiments using non-native speakers from the Spoke 3 data in the 1994 DARPA Wall Street Journal evaluation, inter-class MLLR provided a relative reduction in word error rates of 11.3% compared to conventional MLLR.

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. Sam-Joo Doh is partially supported by Korea Telecom Overseas Education Program.

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] E. Bocchieri *et al*., "Corrrelation modeling of MLLR transform biases for rapid HMM adaptation to new speakers," *Proc. of ICASSP*, pp. 2343-2346, 1996.

[4] S. S. Chen and P. DeSouza, "Speaker adaptation by correlation (ABC)," *Proc. of Eurospeech*, pp. 2111-2114, 1997.

[5] J.-T. Chien, J.-C. Junqua, and P. Gelin, "Extraction of reliable transformation parameters for unsupervised speaker adaptation," *Proc. of Eurospeech*, pp. 207-210, 1999.

[6] S.-J. Doh and R. M. Stern, "Weighted principal component MLLR for speaker adaptation," *to appear at the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Cololado, December 1999.

[7] M. J. F. Gales, "The generation and use of regression class trees for MLLR adaptation," *Technical Report*, CUED/F-INFENG/TR 263, Cambridge University Engineering Department, August 1996.

[8] Z. Hu, *Understanding and adapting to speaker variability using correlation-based principal analysis*, Ph.D. Thesis, Oregon Graduate Institute of Science and Technology, 1999.

[9] 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.

[10] 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.

[11] C. J. Leggetter, *Improved Acoustic Modeling for HMMs Using Linear Trandformations*, Ph.D. Thesis, Cambridge Unversity, 1995.

[12] R. Kuhn *et al*., "Fast speaker adaptation using *a priori* knowledge," *Proc. of ICASSP*, pp. 1587-1590, 1999.

[13] S. Takahashi and S. Sagayama, "Tied-Structure HMM Based on Parameter Correlation for Efficient Model Training," *Proc. of ICASSP*, pp. 467-470, 1996.