Electrical & Computer ENGINEERING

School of Computer Science

# 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
Pittsburgh, PA 15213, USA
{sjdoh, rms}@cs.cmu.edu, http://www.cs.cmu.edu/~robust/

## ABSTRACT

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 relationships among different classes to achieve more reliable estimates of the transformation parameters across multiple classes using limited adaptation 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 15.3% compared to conventional MLLR with a small amount of adaptation data.

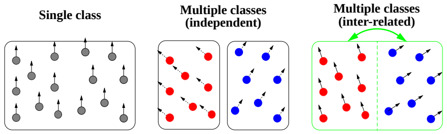
## TRANSFORMATION-BASED ADAPTATION

◆ **Conventional transformation-based adaptation:**

- More effective than Bayesian approaches for small amounts of adaptation data
- Each **transformation class** is considered **independently**
- With limited adaptation data we cannot simultaneously achieve both
  - highly reliable estimates of the transformation parameters ***A*** and ***b***
  - estimates of the Gaussian mean vectors that can be different for each transformation class

◆ **Goal of this work:**

- Utilize **relationships among transformation classes** to achieve more reliable estimates of the transformation parameters across multiple classes

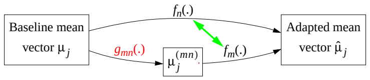

Single class
Multiple classes (independent)
Multiple classes (inter-related)

All data available for reliable estimation of transformation function, but all Gaussians must move in the same direction

Each class can move in a different direction, but less data are available within each adaptation class

More reliable estimates are possible by using information from all classes

◆ **Inter-class transformation $g_{mn}(.)$** relates transformations $f_m(.)$ and $f_n(.)$


Baseline mean vector $\mu_j$
$f_n(.)$
$g_{mn}(.)$
$\mu_j^{(mn)}$
$f_m(.)$
Adapted mean vector $\hat{\mu}_j$

**Carnegie Mellon**

## RELATED WORK

◆ **Some previous work using Correlations/Regressions:**

- Lasry & Stern (1984): Extended MAP (EMAP)
- Huo & Lee (1998): Pair-wise correlation
- Cox (1995): Prediction using correlation-based model
- Ahadi & Woodland (1997): Regression-based model prediction
- Bocchieri (1999): Refinement of shift parameters in MLLR using correlation models

◆ **Comments:**

- **Correlations/Regression**s among model parameters (Gaussian mean vectors) have been used mostly in a **Bayesian framework.**
- **Bayesian formulation** describes shift but not rotation of mean vectors
- Because there are thousands of Gaussians in speech recognition systems,
  - Consideration of the values of **only a few neighboring Gaussians** may not have much effect on Bayesian estimates of means for a new speaker
  - Consideration of **larger numbers of neighboring Gaussians** may require too much computation.

## INTER-CLASS MLLR

◆ **Comparison of the estimates of $(A_m, b_m)$ for a target class $m$**

**(1) Conventional MLLR**

- Assumption: $\hat{\mu}_i = A_m \mu_i + b_m, \quad i \in$ target class $m$
- Estimate $(A_m, b_m)$ by minimizing $Q_c$ (from EM algorithm)

$$Q_c = \sum_t \sum_{i \in \text{class } m} \gamma_t(i)(o_t - A_m\mu_i - b_m)^T C_i^{-1}(o_t - A_m\mu_i - b_m)$$

where $o_t$ is the input feature vector at time $t$ (adaptation data)
$\gamma_t(i)$ is the *a posteriori* probability of being Gaussian $i$ at time $t$
$C_i$ is the covariance matrix of Gaussian $i$

- Data are considered only from Class $m$, and small amounts of data may not provide reliable estimates of $(A_m, b_m)$.

**(2) Inter-class MLLR**

- Origianl assumption: $\hat{\mu}_j = A_n \mu_j + b_n, \quad j \in$ neighboring class $n$
- New assumption: $\hat{\mu}_j = A_m \mu_j^{(mn)} + b_m, \quad j \in$ neighboring class $n$
- Introduce inter-class transformation: $\mu_j^{(mn)} = T_{mn}\mu_j + d_{mn} = g_{mn}(\mu_j)$
- Estimate $(A_m, b_m)$ by minimizing $Q_I$

$$Q_I = \sum_t \sum_{i \in \text{class } m} \gamma_t(i)(o_t - A_m\mu_i - b_m)^T C_i^{-1}(o_t - A_m\mu_i - b_m)$$
$$+ \sum_t \sum_{n \in \text{neighbors}} \sum_{j \in \text{class } n} \gamma_t(i)(o_t - A_m\mu_j^{(mn)} - b_m)^T C_j^{-1}(o_t - A_m\mu_j^{(mn)} - b_m)$$

- Data are considered from neighboring classes as well as from Class $m$, and more reliable estimates of $(A_m, b_m)$ can be obtained, preserving the details of Class $m$.

◆ **Estimate $(T_{mn}, d_{mn})$ from training data**

- From a training corpus we estimate speaker-specific values of $(A_{m,s}, b_{m,s})$ for each training speaker $s$ using conventional MLLR according to

$$\hat{\mu}_{i,s} = A_{m,s}\mu_i + b_{m,s}, \qquad i \in \text{target class } m$$

- Since $\hat{\mu}_{j,s} = A_{m,s}(T_{mn}\mu_j + d_{mn}) + b_{m,s}, \quad j \in$ neighboring class $n$,

we get $\hat{\mu}_{j,s}^{(mn)} = T_{mn}\mu_j + d_{mn}$, where $\hat{\mu}_{j,s}^{(mn)} \equiv A_{m,s}^{-1}(\hat{\mu}_{j,s} - b_{m,s})$

- $(T_{mn}, d_{mn})$ can be estimated using **conventional MLLR**
  - Minimize $Q_{mn}$ with all training speaker's data

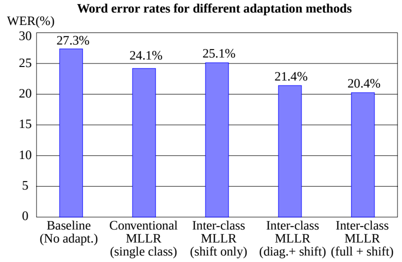
$$Q_{mn} = \sum_s \sum_t \sum_{j \in \text{class } n} \gamma_{t,s}(i)(o_{t,s}^{(mn)} - T_{mn}\mu_j - d_{mn})^T C_j^{-1}(o_{t,s}^{(mn)} - T_{mn}\mu_j - d_{mn})$$

where $o_{t,s}^{(mn)} = A_{m,s}^{-1}(o_{t,s} - b_{m,s})$

## EXPERIMENTS

- **Speech recognition system:** SPHINX-III (continuous HMMs)
- **Test data:** Spoke 3 of the DARPA 1994 Wall Street Journal evaluation, 10 Non-native speakers x 20 test sentences
- Small amount of adaptation data in supervised mode: 1 sentence (5-6 sec.)
- 13 phonetic-based classes for inter-class MLLR

**Word error rates for different adaptation methods**



- Full matrix ***A*** and shift vector ***b*** are used in all adaptation methods.
- Different (***T***, ***d***) are used in inter-class MLLR, as described in the figure above.

## SUMMARY

- **Conventional transformation-based adaptation**
  - assumes that each class is **independent**
  - trades off **reliable estimation of the transformation function** with the ability for **Gaussian mean vectors to adapt differently** from class to class
- We developed a new adaptation algorithm: **Inter-class MLLR.**
  - It utilizes **relationships among different classes** to achieve both **detailed and reliable** transformation-based adaptation using limited data.
- In our experiments, **inter-class MLLR** provides **15.3% relative improvement** over conventional MLLR.

## ACKNOWLEGEMENT

The authors thank the Speech Group members for their contributions. 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.

## REFERENCES

[1] 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.

[2] E. Bocchieri *et al.*, "Corrrelation modeling of MLLR transform biases for rapid HMM adaptation to new speakers," *Proc. of ICASSP*, pp. 2343-2346, 1999.

[3] S. Cox, "Predictive speaker adaptation in speech recognition," *Computer Speech and Language*, vol. 9, pp.1-17, 1995.

[4] 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.

[5] 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.

[6] 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.

**Robust Speech Group**