

MLLR method for Environmental Adaptation in a Continuous Farsi Speech Recognition

K. Hosseinzadeh¹, H. Sameti, A. Fazel
Department of Computer Engineering, Sharif University of Technology
hosseinzadeh@ce.sharif.edu

Abstract. In this paper, MLLR¹ adaptation of continuous density HMM² is investigated in a Farsi speaker independent large vocabulary continuous speech recognition system in attempt to improve recognition rate in real world situations. In the MLLR framework, we have experienced the use of Gaussian mean transformations in global adaptation and regression tree based adaptation. Besides full and block-diagonal transformations of Gaussian means, transformation of Gaussian variances is examined. We have used MLLR technique in batch-supervised fashion since it is more beneficial in situations of severe mismatch. Our experiments on 4 different tasks, show that by using this technique the system recognition performance in a new environment can be significantly improved.

1 Introduction

When there is an acoustic mismatch between train and test conditions, serious reduction of performance occurs, even if the model has been carefully trained. One way to compensate differences in acoustic environments is to adapt model parameters to the new environment. MLLR technique has proved to be a powerful tool for this purpose. This method was first used for speaker adaptation effectively [1, 2]. Environmental adaptation requires some samples of speech from several speakers recorded in the new environment. In supervised mode as used in this paper, transcription of adaptation data is also needed. Since this technique can be very robust to transcription errors [3], unsupervised mode in which system generates transcription could be effective as well.

One of the main advantages of MLLR is that the amount of adaptation data is relatively small compared to for example MAP³ method [4]. This is due to the use of an identical transformation across a number of Gaussians which causes parameters with and without adaptation data to be transformed. Gaussians could be grouped into broad phone classes or regression tree for dynamic tying of the transformations can be used [2]. When only a small amount of data is available, Gaussian parameters may be transformed by only one matrix which is referred to as global adaptation.

The structure of the paper is as follows: MLLR technique is introduced in section 2. In section 3 transform sharing is discussed. Section 4 describes some implementation issues. In section 5 we describe the baseline system and results of our experiments and section 6 gives a brief conclusion of this paper.

¹ Maximum Likelihood Linear Regression

² Hidden Markov Model

³ Maximum A Posteriori

2 MLLR Overview

This section briefly reviews the MLLR technique which is covered in greater details in [4].

2.1 Maximum Likelihood Estimation (MLE)

MLLR computes the parameters of linear transforms to better match the characteristics of adaptation data, using an Expectation Maximization approach and by optimizing the standard auxiliary function [5],

$$Q(\lambda, \hat{\lambda}) = K - \frac{1}{2} L(O_T | \lambda) \sum_{m=1}^M \sum_{\tau=1}^T \gamma_m(\tau) \left[K_m + \log(|\hat{\Sigma}_m|) + (o(\tau) - \hat{\mu}_m)^T \hat{\Sigma}_m^{-1} (o(\tau) - \hat{\mu}_m) \right] \quad (1)$$

Where λ is model parameters, O_T is the adaptation data. $\gamma_m(\tau)$ is the occupation likelihood of Gaussian m at time τ obtained by the forward-backward procedure.

2.2 Mean Transform

Lets define μ_m , as an n -dimensional mean vector for Gaussian mixture m then the adapted mean is found by the following linear regression [5],

$$\hat{\mu}_m = W_m \xi_m \quad (2)$$

$$\xi_m = \begin{bmatrix} 1 \\ \mu_m \end{bmatrix} \quad (3)$$

Where W_m is the $n \times (n+1)$ transformation matrix and ξ_m is the extended mean vector.

By formulating and maximizing the standard auxiliary function in Eq. (1) with respect to a particular transformation, W_m , that is tied across R Gaussians, $\{m_1, \dots, m_R\}$, and by having diagonal covariance matrix, W_m can be calculated row by row using

$$w_i^T = G_{(i)}^{-1} z_i^T \quad (4)$$

Where w_i and z_i are the i^{th} rows of W_m and Z respectively. $G_{(i)}$ is an $(n+1) \times (n+1)$ matrix defined as

$$G_{(i)} = \sum_{r=1}^R \left(\sigma_{m_r}^{-1}(i, i) \xi_{m_r} \xi_{m_r}^T \sum_{\tau=1}^T \gamma_{m_r}(\tau) \right) \quad (5)$$

$$Z = \sum_{\tau=1}^T \sum_{r=1}^R \gamma_{m_r}(\tau) \Sigma_{m_r}^{-1} o(\tau) \xi_{m_r}^T \quad (6)$$

2.3 Variance Transform

The unconstrained optimization for variance transform [4], first the mean transformation is found, given the current variance and second the variance transform is found given the current mean. The Gaussian covariance for a system using diagonal covariance matrix is modified using,

$$\hat{\Sigma}_m = B_m^T H_m B_m \quad (7)$$

where H_m is the linear transformation to be estimated and B_m is the inverse of the Choleski factor of Σ_m^{-1} .

After rewriting the auxiliary function and sharing the transform over R Gaussians, $\{m_1, \dots, m_R\}$, the transform matrix, H_m , is estimated from,

$$H_m = \frac{\sum_{r=1}^R \left\{ C_{m_r}^T \left[\sum_{\tau=1}^T \gamma_{m_r}(\tau) (\alpha(\tau) - \hat{\mu}_{m_r}) (\alpha(\tau) - \hat{\mu}_{m_r})^T \right] C_{m_r} \right\}}{\sum_{r=1}^R \sum_{\tau=1}^T \gamma_{m_r}(\tau)} \quad (8)$$

Since diagonal covariance matrix is used, the off-diagonal terms should be set to zero.

2.4 Statistics Required

The statistics required to be gathered during forward-backward may be stored at either the Gaussian level or at the base class level [4]. We examined both methods and realized that the second method is computationally more expensive. Thus by defining $s_{m_r}^{(1)}$, $s_{m_r}^{(2)}$ and $s_{m_r}^{(3)}$ for Gaussian m_r as

$$s_{m_r}^{(1)} = \sum_{\tau=1}^T \gamma_{m_r}(\tau) \quad (9)$$

$$s_{m_r}^{(2)} = \sum_{\tau=1}^T \gamma_{m_r}(\tau) \alpha(\tau) \quad (10)$$

$$s_{m_r}^{(3)} = \sum_{\tau=1}^T \gamma_{m_r}(\tau) \alpha(\tau) \alpha(\tau)^T \quad (11)$$

Then the formulae for adapted mean and variance can simply be rewritten to use Eqs. (9)-(11).

3 Transform Sharing

The tying of each transformation over a cluster of Gaussian components makes it possible to adapt components for which there were no observations at all. This transform sharing can allow all the Gaussians in a system to be updated with only a relatively small amount of adaptation data and the adaptation process is dynamically refined when more data becomes available. In early MLLR formulations, the clusters are defined based on broad phonetic classes (knowledge-based) that leads to an early saturation as the amount of adaptation data increases. MLLR better works when the optimal number of transforms is known, given the amount of adaptation data. This optimized adaptation is achieved by clustering all the Gaussians of the models in the form of a tree, where the root contains all the Gaussians. The regression class tree is constructed so as to cluster Gaussians that are close in acoustic space, so that similar Gaussians can be transformed in similar way [7]. We construct the tree with a centroid splitting algorithm which uses Euclidean distance measure. The terminal nodes of the tree are called base classes. The tree is grown until number of base classes reaches a predefined number. The node occupancy for a node with R mixtures is

$$NodeOcc = \sum_{r=1}^R s_{m_r}^{(1)} \quad (12)$$

When transformations are to be computed, a top-down traversal is done to construct transforms at nodes with occupation count above a threshold. The nodes that share a single transform called regression classes. If there is not sufficient data, a global adaptation can be applied [9]. In this case, all the Gaussians are modified with a single transform matrix. We have used global adaptation as a first pass of adaptation to provide better models for the next local transforms. In our implementation, it is also possible to form a separate regression class for silence models (silence, noise, closures, ...) besides regression tree [9]. Due to this sharing these models may be better adapted if sufficient amount of data for these models is available.

4 Implementation Issues

In this section we discuss some useful notes that should be cared in the implementation.

4.1 Numerical Accuracy

To compute mean transforms, it is necessary to inverse $G_{(i)}$, which may be ill-conditioned for some transforms. By using singular value decomposition, the problem can be handled [5]. However using Gauss-Jordan method and replacing dividing by zero with a tiny, as we did, works quite well.

4.2 Computation Reduction

Since there is no need to compute log-likelihoods of all frames to all mixtures in the forward-backward, the computation can be efficiently reduced.

The mean transformation matrix can be stored as a block diagonal matrix [6]. This representation makes the assumption that there is no correlation between the statics, first derivatives and second derivatives of the feature vector.

It makes the adaptation process faster and reduces the adaptation data required by reducing the parameters to be learnt.

To gather statistics faster, we should properly define minimum $\gamma_{m_r}(\tau)$ that must contribute in Eqs. (9)-(11). Experiments show that the value 10^{-6} is suitable.

5 Performance Evaluation

In this section, the baseline system is first summarized then the results of MLLR on the amount of adaptation data and effect of variance transform, block diagonal transforms, global adaptation and silence class are presented.

5.1 Experimental Setup

To evaluate the performance of the MLLR, a small database is recorded and transcribed. The database consists of 4 tasks. Each task includes two subsets identified as adaptation subset and test subset which are uttered by 7 speakers including 5 males and 7 females. On average, each utterance length is 3.5 seconds. The sentences are selected from FARSDAT [9].

The adaptation subset has 175 sentences that are 10 sentences common to all plus 15 different sentences for each one. The test subset contains 140 sentences uttered equally among speakers. The properties of each task are shown in Table 1.

Table 1. Tasks characteristics in our experiments

	Environment	Microphone	SNR(dB)
Task A	Office	Condenser	12
Task B	Office	Dynamic	30
Task C	Exhibition	Condenser	9
Task D	Volvo noise	Condenser	7

We have used Sharif Farsi SI-LVCSR system [8] in our experiments. In the baseline system, 44 phonemes were modeled using FARSDAT database. MFCC was used as a feature extraction. Each model contains 6 states and each state has 44 Gaussian mixtures. The FARSDAT database consists of 6080 Farsi utterances, uttered by 304 speakers. The sentences are formed by using over 1000 Farsi words. The average signal to noise ratio is 31dB.

5.2 Results

In this section we present some experiments carried out to evaluate the performance of the MLLR technique. Our baseline for comparison is the word error rates (WERs) of 4 test subsets on clean model mentioned in section 5.1, with respect to a 1000-word lexicon. In all of our tests, minimum mixture in a tree node is set to 100.

5.2.1 Initial Experiments

For initial tests, we examined MLLR for just mean transformations. MLLR used the whole utterances of each adaptation subset with 87 base classes and minimum occupation of 1024. WER reductions obtained are shown in Table 2.

Table 2. Initial tests on mean adaptation

Environment	Relative reduction [%]
Task A	47.8
Task B	43.9
Task C	12.6
Task D	24

It shows considerable reduction in WER even though we didn't work on tuning of the parameters. Task C comprises babble noise which is difficult to overcome so MLLR on task C has the least WER reduction among others.

5.2.2 Amount of Adaptation Data

The number of adaptation utterances is varied from 5 to 175 for task A. Results are shown in Figure 1. It demonstrates fast adaptation with a little data but saturation in reduction of WER as more data are used.

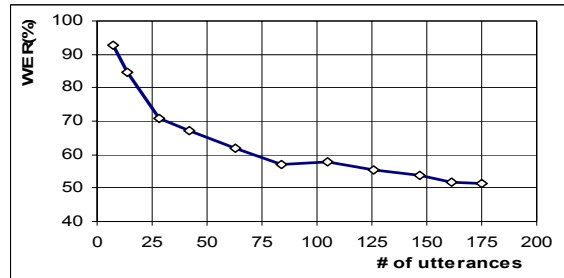


Figure 1. Result of varying the amount of adaptation data on task A

5.2.2 Various Experiments

Some experiments have been done to find the effect of using variance transform, block diagonal transform, global adaptation and silence class as described in section 3.2. Number of base classes and minimum occupation were 50 and 1024 respectively. 126 utterances are used for adaptation. The reductions in WER are shown in Table 3.

Table 3. Various tests results(% WER reduction)

Environment	Mean	Silence	3B-Block	Variance	Global
Task A	46.2	47.8	41.2	45.8	50.8
Task B	44.7	43.2	32	44.5	50.3
Task C	17.8	18	17.7	20	38
Task D	34.6	37.5	30.1	38	55

For high-noise tasks (tasks C and D) grouping the silences separately has better results perhaps because there were enough data for them. Since for low-noise tasks we have better alignments that is why using full transformation matrix has better effect on these tasks. Using variance and mean transformation together has a slightly positive effect on tasks C and D, possibly because of more severe changes in models distributions. We also used global adaptation as a pre-adaptation as described in section 3.2. It has better effect on tasks C and D, because global-adapted model improves the very poor frame/state alignments of the clean model.

6 Conclusion

Number of base classes depends strongly on the number of utterances available for accumulation of adaptation statistics. MLLR is dependent on the frame/state alignment thus when models can not provide good alignment, performance can be improved using multiple iterations of MLLR.

One of the drawbacks of variance transformation is that it does not simultaneously optimize the mean and the variance transformations [4] because the ML estimate of the mean transformation matrix is a function of the current estimate of the covariance matrix. To solve this problem an iterative scheme can be used to alternately optimize the mean and then the covariance transformations. The other way is to use constrained transformations [5].

References

1. C.J. Leggetter & P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs," *Computer Speech and Language*, Vol. 9, pp.171-186, 1995.
2. C.J. Leggetter & P.C. Woodland, "Flexible Speaker Adaptation using Maximum Likelihood Linear Regression," *Proc. ARPA Spoken Language Technology Workshop*, pp.104-109, Morgan Kaufmann, 1995.
3. P.C. Woodland, D. Pye & M.J.F Gales, "Iterative Unsupervised Adaptation Using Maximum Likelihood Linear Regression," *Proc. ICSLP*, pp.1133-1136, Philadelphia, 1996.
4. M.J.F. Gales & P.C. Woodland, "Mean and Variance Adaptation within the MLLR Framework," *Computer Speech and Language*, Vol. 10, pp.249-264, 1996.
5. M.J.F. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition," *Computer Speech & Language*, Vol. 12, pp.75-98, 1998
6. J.E. Hamaker, "MLLR: A Speaker Adaptation Technique For LVCSR," ISIP course lecture, Mississippi State University, 1999.
7. P.C. Woodland, "Speaker adaptation: techniques and challenges," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp.85-90, 1999.
8. B. Babaali & H. Sameti, "The Sharif Speaker Independent Large Vocabulary Speech Recognition System," *The 2nd Workshop on IT & Its Disciplines*, Kish Island, Iran, 2004.
9. M. Bijankhan et al., "FARSDAT–The Speech Database of Farsi Spoken Language," *Proc. The Fifth Australian Int. Conf. on Speech Science and Tech.*, Vol. 2, perth, 1994.