CALLにおける発音評価のための制約付きMLLR話者適応手法*

○羅徳安（東大），喬宇（中国科学院），峯松信明，広瀬啓吉（東大）

1 Abstract

This study focuses on speak adaptation techniques for Computer-Assisted Language Learning (CALL). We first investigate the effects and problems of Maximum Likelihood Linear Regression (MLLR) speaker adaptation when used in pronunciation evaluation. Automatic scoring and error detection experiments are conducted on two publicly available databases of Japanese learners' English pronunciation. As we expected, over-adaptation causes misjudge of pronunciation accuracy. Following the analysis, we propose a novel method, Regularized Maximum Likelihood Regression (Regularized-MLLR) adaptation, to solve the problem of adverse effects of MLLR adaption. This method uses a group of teachers' data to regularize learners' transformation matrices so that erroneous pronunciations will not be transformed as correct ones. We implement this idea in two ways: one is using the average of the teachers' transformation matrices as a constraint to MLLR, and the other is using linear combinations of the teachers' matrices to represent learners' transformations. Experimental results show that the proposed methods can better utilize MLLR adaptation and avoid over-adaptation.

2 Regularized-MLLR Adaptation

2.1 Definition of Regularized-MLLR

In order to regularize MLLR transformation so that the erroneous pronunciation will not be "transformed" to good pronunciation, we use the transformation matrices calculated through a group of teachers' speech data with conventional MLLR and use their linear combination to derive each specific learner's transformation matrix. Since a learner's transformation matrix is not estimated directly from his/her data, the resulting matrix is expected not to over-transform that learner's data.

The standard auxiliary function for MLLR is defined as below to estimate the transform for each regression class r.

$$Q(M, \hat{M}) = \frac{1}{2}\sum_{r=1}^{R}\sum_{m_r=1}^{M_r}\sum_{t=1}^{T} L_{m_r}(t) \times$$

$$[K^{(m)} + \log|\hat{\Sigma}_{m_r}| + (o(t)-\hat{\mu}_{m_r})^T \hat{\Sigma}_{m_r}^{-1}(o(t)-\hat{\mu}_{m_r})] \quad (1)$$

Here we obtain a set of transforms estimated from a group of teachers who are native speakers of General American English. Teachers' transforms are used to represent the transforms of ideal students and their combination is applied for others to avoid bad pronunciations being transformed into good pronunciations.

Let $\{W_r^{C_1}, ..., W_r^{C_N}\}$ denote a set of transformation matrices estimated from a group of N teachers, and we assume that each learner's transformation matrix W_r must be written as a linear combination of the teachers' transformation matrices,

$$W_r = \sum_n \alpha_n W_r^{C_n} \quad (3)$$

By calculating the optimal parameters $(\alpha_1, \alpha_2, ..., \alpha_N)$, we can obtain the learner's transformation matrix.

We assume diagonal covariance matrices and the adaptation is only applied to the mean vector for each Gaussian component,

$$\hat{\mu}_{m_r} = W_r \xi_{m_r} \quad (4)$$

where ξ_{m_r} is the extended mean vector for the Gaussian component m_r,

$$\xi_{m_r} = [1\ \mu_1\ \mu_2\ \ldots\ \mu_d]^T \quad (5)$$

where d is the dimensionality of the data. Thus the parameters $(\alpha_1, \alpha_2, ..., \alpha_N)$ can be estimated using the following objective function,

*Regularized-MLLR for pronunciation evaluation in CALL, by Dean Luo (University of Tokyo), Yu Qiao (Chinese Academy of Science), Minematsu Nobuaki and Keikichi Hirose (University of Tokyo).

$$\max_{\{\alpha_n\}} g(\alpha_1, \alpha_2, ..., \alpha_N) = \sum_{m_r=1}^{M_r} \sum_{t=1}^{T} L_{m_r}(t)(o(t) - \sum_n \alpha_n W_r^{C_n} \xi_{m_r})^T \Sigma_{m_r}^{-1} \times (o(t) - \sum_n \alpha_n W_r^{C_n} \xi_{m_r}) \quad (6)$$

By calculating the derivative,

$$\frac{\partial g}{\partial \alpha_n} = -2 \sum_{m_r=1}^{M_r} \sum_{t=1}^{T} L_{m_r}(t) \Sigma_{m_r}^{-1} (o(t) - \sum_n \alpha_n W_r^{C_n} \xi_{m_r}) \times (W_r^{C_n} \xi_{m_r})^T = 0 \quad (7)$$

and changing $n = 1,2,...,N$, we have N linear equations on $\{\alpha_n\}$. For simplicity, if we set

$$\xi'_{m_r,n} = W_r^{C_n} \xi_{m_r} \quad (8)$$

then the linear equations become,

$$\sum_{m_r=1}^{M_r} \sum_{t=1}^{T} L_{m_r}(t) \Sigma_{m_r}^{-1} (o(t) - \sum_n \alpha_n \xi'_{m_r,n}) \xi'^T_{m_r,n} = 0 \quad (9)$$

By solving these linear equations, we obtain the optimal $\{\alpha_n\}$. Then we can use equation (3) to calculate the target learner's transformation matrix.

3 Experiments

We compared the effects of MLLR and Regularized-MLLR adaptations on pronunciation evaluation based on HMM acoustic models in two ways: automatic scoring and error detection.

3.1 Automatic Scoring

The confidence-based pronunciation assessment, which is defined as the Goodness of Pronunciation (GOP), is often used for assessing speakers' articulation and shows good results [1].

We use ERJ corpus to measure GOP score with MLLR and Regularized-MLLR adaptation. 42 learners with higher agreement among raters and a variety of proficiency were selected. The average phoneme GOP score over 30 sentences read by each learner is calculated as automatic score for that learner. 60 sentence utterances were used as adaptation data. For Regularized-MLLR adaptation, 20 teachers' speech data were used to estimate transformation matrices. The result is shown in Fig.1.

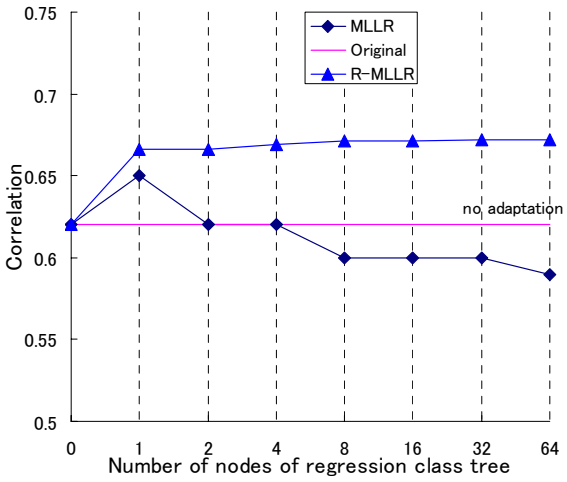

Fig. 1 Correlations between GOP scores and manual scores

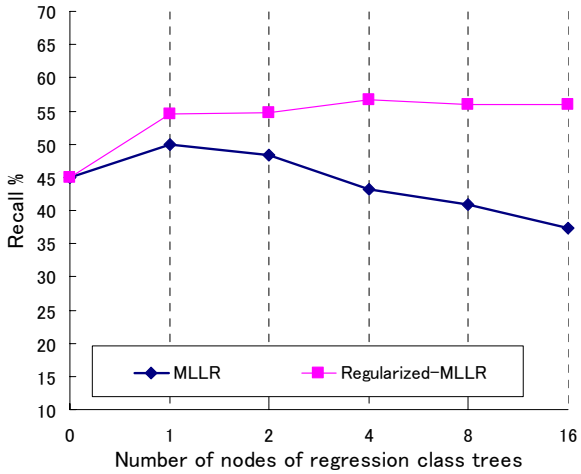

Fig. 2 Recall comparison between MLLR and Regularized-MLLR at the precision level of 70%

3.2 Error detection

We used the utterances of 4 speakers (2 males and 2 females) with many typical errors of Japanese learners from Basic English Words Read by Japanese corpus or error detection based on GOP threshold. The result is shown in Fig.2.

4 Conclusion

R-MLLR not only out-performed MLLR but also prevent over-adaptation problem.

Reference

[1] S.M. Witt and S.J. Young, "Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning," Speech Communications, 30 (2–3): pp.95-108, 2000