Regularized-MLLR Speaker Adaptation for Computer-Assisted Language Learning System

Dean Luo[1], *Yu Qiao*[1], *Nobuaki Minematsu*[1], *Yutaka Yamauchi*[2], *Keikichi Hirose*[1]

[1] The University of Tokyo, Tokyo, Japan
[2] Tokyo International University, Saitama, Japan
dean@gavo.t.u-tokyo.ac.jp

Abstract

In this paper, we propose a novel speaker adaptation technique, regularized-MLLR, for Computer Assisted Language Learning (CALL) systems. This method uses a linear combination of a group of teachers' transformation matrices to represent each target learner's transformation matrix, thus avoids the over-adaptation problem that erroneous pronunciations come to be judged as good pronunciations after conventional MLLR speaker adaptation, which uses learners' "imperfect" speech as target utterances of adaptation. Experiments of automatic scoring and error detection on public databases show that the proposed method outperforms conventional MLLR adaption in pronunciation evaluation and can avoid the problem of over adaptation.

Index Terms: Computer Assisted Language Learning (CALL), speaker adaption, pronunciation evaluation, goodness of pronunciation (GOP), maximum likelihood linear regression (MLLR)

1. Introduction

In order to deal with the mismatches between learners' speech and the acoustic models, speaker adaption has been employed for ASR (Automatic Speech Recognition) in CALL systems [1, 2]. Learners' speech often contains erroneous pronunciations. Extensive adaptations are often avoided because errors might be transformed as good pronunciations [3, 4]. In our previous study, we analyzed analytically the effects and side-effects of MLLR adaptation on pronunciation evaluation and showed that bad pronunciations can be "transformed" into good pronunciations by over-adaptation [5]. We have proposed a method that uses the average of a group of teachers' transformation matrices as constrains to the conventional MLLR transformation [6]. Although the method reduces the adverse effects of extensive adaptation and thus improves the performances, the over-adaptation problem still remains. One possible solution to the over-adaptation problem is to avoid *directly* using learners' imperfect speech data that includes erroneous pronunciation. In this paper, we formulate this idea by using a group of teachers' correctly pronounced speech data to estimate each teacher's transformation matrix, and then calculating each learner's transformation matrix as a linear combination of the teachers' matrices. We name this method as Regularized-MLLR.

In this paper, we compare the effects of conventional maximum likelihood linear regression (MLLR) speaker adaptation and the proposed Regularized-MLLR adaptation technique on pronunciation evaluation for CALL in two ways: automatic scoring and phoneme error detection. Experiments on two sets of public databases show the proposed method outperforms the conventional MLLR adaptation and can resolve the over-adaptation problem.

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 W_r for each regression class r.

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

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

where M is the HMM model set, $\hat{M}$ is the adapted model set, R is the number of the nodes of regression class tree, M_r is the number of Gaussian components that is to be tied together, $K^{(m)}$ subsumes all constants, $\hat{\mu}_{m_r}$ and $\hat{\Sigma}_{m_r}$ are the adapted mean vector and covariance matrix for the mixture component m_r respectively, and $L_{m_r}(t)$ is the occupation likelihood defined as

$$L_{m_r}(t) = p(q_{m_r}(t) \mid M, O_T) \ , \tag{2}$$

where $q_{m_r}(t)$ is the Gaussian component at time *t*, and O_T is the adaption data.

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,

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

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,

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

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

$$\xi_{m_r} = [1 \ \mu_1 \ \mu_2 \ ... \ \mu_d]^T , \qquad (5)$$

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,

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

By calculating the derivative,

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

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

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

then the linear equations become,

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

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

3.1.1. Goodness of Pronunciation

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 [7, 8]. In this study, we use HMM acoustic models trained on WSJ and TIMIT corpus with MLLR and Regularized-MLLR adaptation to calculate GOP scores defined as follows. For each acoustic segment $O^{(p)}$ of phoneme *p*, GOP($O^{(p)}$) is defined as posterior probability by the following log-likelihood ratio.

$$GOP(O^{(p)}) = \frac{1}{D_p} \log(P(p \mid O^{(p)})) \qquad (10)$$

$$= \frac{1}{D_p} \log\left(\frac{P(O^{(p)} \mid p) P(p)}{\sum_{q \in Q} P(O^{(p)} \mid q) P(q)} \right) \qquad (11)$$

$$\approx \frac{1}{D_p} \log\left(\frac{P(O^{(p)} \mid p)}{\max_{q \in Q} P(O^{(p)} \mid q)} \right), \qquad (12)$$

where $P(p \mid O^{(p)})$ is the posterior probability that the speaker uttered phoneme *p* given $O^{(p)}$, *Q* is the full set of phonemes, and D_p is the duration of segment $O^{(p)}$. The numerator of equation (12) can be calculated by scores generated during the forced Viterbi alignment, and the denominator can be approximately attained by continuous phoneme recognition with an unconstrained phone loop grammar.

We use GOP scores as measurement of automatic scoring performance.

3.1.2. Experimental Results

We use ERJ (English Read by Japanese Students) corpus [9] to measure GOP score with MLLR and Regularized-MLLR adaptation. This corpus includes proficiency labels rated by phonetic experts that a score is given for each of a subset of sentences. 42 learners (21 males and 21 females) 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 of each leaner were used as adaptation data. For Regularized-MLLR adaptation, 20 teachers' speech data were used to estimate transformation matrices that are required in equation (3) (N=20). These teachers are speakers of General American English and 60 sentence utterances of each teacher are used as adaptation data. The same amount of learners' speech data are used to determine the parameters $\{\alpha_n\}$ for calculating learners' transformation matrices as linear combinations of the teachers' matrices.

We investigate the correlations between GOP scores and human scores while increasing the number of the nodes of regression trees. Here the number 0 means without adaption,

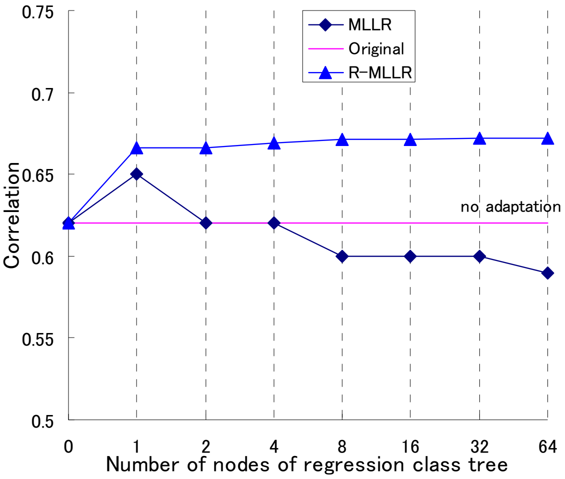

Figure 1: Correlations between GOP scores and manual scores

and 1 represents global adaption. As shown in Figure 1, in the case of MLLR, global adaptation yields the best correction of 0.65, yet while the number of nodes of regression class tree increases from 2, the performance drops, and especially when the number is larger than 4, the correlation is even worse than the original models, which indicates that over-adaptation occurs. In the case of Regularized-MLLR(R-MLLR), the correlations are higher than MLLR at each adaptation level, and even when the number of regression class increases, the performance never drops. This indicates that Regularized-MLLR adaptation can completely avoid over-adaptation and thus more reliable in terms of utilizing speaker adaptation for CALL systems than conventional MLLR.

3.2. Phoneme Error Detection

Two most popular methods of error detection are employed for our phoneme error detection experiments: one is based on pronunciation networks [1] and the other is based on GOP scores [7, 8]. The former method predicts possible error patterns and thus is able to detect specified types of errors such as phoneme-level substitution, deletion or insertion. However, the detection performance is largely depending on the size of the pronunciation networks. The latter method often uses a pre-set threshold to determine whether a phoneme is correctly pronounced or not. Although this method cannot specify the type of an error that occurs, by choosing the optimal threshold for each phoneme, much better detection performance can be obtained.

3.2.1. Database

Because the ERJ database does not contain phoneme labels with erroneous pronunciation, we use another corpus of English words spoken by Japanese students. The database [10] consists of 5950 utterances of 850 basic English words read by seven Japanese speakers. This database contains manually annotated phonemic labels that were faithfully transcribed and include erroneous phonemes. This database has been used to evaluate the performances of acoustic models for CALL [11].

We used the utterances of 4 speakers (2 males and 2 females) with many typical errors of Japanese learners. For each learner, 450 word utterances are used as adaptation data, and the remaining 400 utterances are used as test data.

3.2.2. Error Detection based on Network Grammar

The first method we use to detect pronunciation errors is using pronunciation networks that include correct pronunciations and various error patterns to predict learners' possible mispronunciations. By referring to [12], 12 major error patterns were defined and any irregular errors in the labels were added to the prediction networks. Although the error detection performance highly depends on pronunciation networks and a larger network often results in lower detection precision, when the same network is used, the relative increase or decrease of detection accuracy can be used to measure the performances of the acoustic models with MLLR and Regularized-MLLR.

3.2.3. Error Detection based on GOP Scores

We calculate the phoneme-level GOP score according to equation (12), and use phoneme-dependent thresholds, which are based on mispronunciation labels by experts, to decide if the phonemes are correct or not. We investigate the detection rate on 12 most frequently mispronounced phonemes according to the manual label to compare the performances of Regularized-MLLR and conventional MLLR. These 12 phonemes are /ih/, /er/, /ae/, /ow/, /ey/, /r/, /f/, /v/, /n/, /s/, /t/, /z/ (the phonemic descriptions are based on TIMIT database).

3.2.4. Experimental Results

We use precision and recall rates defined as below to measure the performance of acoustic models with MLLR and Regularized-MLLR.

$$\text{Precision} = \frac{N_{hit}}{N_{total}} = \frac{N_{hit}}{N_{hit} + N_{FR}} \quad , \qquad (13)$$

$$\text{Recall} \ = \frac{N_{hit}}{N_{labeled}} \quad , \qquad (14)$$

where N_{hit} represents the number of the errors that were correctly detected , N_{total} is the total number of detected errors, N_{FR} is the number of false rejections which means correct pronunciations being falsely flagged as mispronunciations, and $N_{labeled}$ is the number of all the errors that were detected by phoneticians,

Figure 2 shows the performances of error detection based on pronunciation networks with MLLR and Regularized-MLLR adaptation. In the case of MLLR, although the precision rate keeps increasing when more transforms are used for adaptation, the recall rate drops when the number of regression classes is larger than 2. This indicates that with MLLR adaptation to reduce model mismatches, the number of false rejections N_{FR} drops significantly, thus the precision rate increases. However, since the number of $N_{labeled}$ is only decided by the labels, the decrease of recall means the decrease of the number of correctly detected errors. This result shows that over adaption can cause more errors to be recognized as correct pronunciations (more false accepts), but at the same time, even with over adaptation, more false rejections can be prevented. In the case of Regularized-MLLR, it not only significantly improves recall which is easily affected by over-adaption, but also improves precision over conventional MLLR.

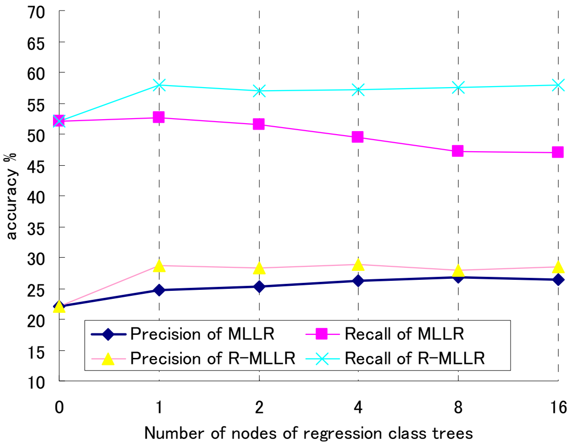

Figure 2: The performances of error detection based on pronunciation networks, comparing MLLR with Regularized-MLLR(R-MLLR)

For the error detection method based on GOP scores, the recall and precision can be adjusted by changing the values of the thresholds. According to [8], erroneously rejecting correct pronunciations would be more detrimental for learners than erroneously accepting mispronunciations. Thus we need to keep the false rejection rate at relatively low level, which means to keep the precision relatively high, and find the optimal thresholds that maximize the recall. Here, we investigate the change of recalls at precision level of 70% while increasing the number of regression classes for MLLR and Regularized-MLLR. Here, the number 0 means no adaptation, i.e. using the original acoustic models.

As shown in Figure 3, in the case of MLLR adaptation, only global adaption shows slight improvement over original models and when the number of regression classes is larger that 2, the performance drops significantly. This clearly indicates that over adaptation occurs with MLLR. In the case of Regularized-MLLR, it outperforms MLLR significantly and keeps high performance even when the number of regression classes increases. This further proves that the over-adaptation problem with conventional MLLR is resolved by our proposed Regularized-MLLR adaptation method.

4. Conclusion

This study proposed a novel speaker adaptation technique, Regularized-MLLR, for pronunciation evaluation. We compared the effects of Regularize-MLLR and conventional MLLR speaker adaption on pronunciation evaluation in terms of automatic scoring and error detection. Experiments on reliable databases show that the proposed method outperforms MLLR and can avoid the problem of over-adaptation, thus better utilize the merits of adaptation for CALL systems.

For future work, we are investigating how the number of teachers and the amount of adaptation data for Regularized-MLLR affect the evaluation performance. We are also investigating the effects of Regularized-MLLR on different phonemes.

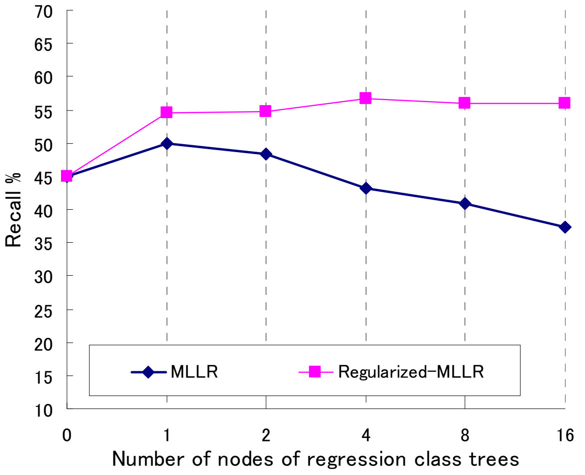

Figure 3: Recall comparison between MLLR and Regularized-MLLR at the precision level of 70%

5. References

[1] Y.Tsubota et al, “Practical Use of English Pronunciation System for Japanese Students in the CALL Classroom,” Proc. ICSLP2004, pp1689-1692 , 2004

[2] Y. Okawa et al, “A speaker adaptation method for non-native speech using learners’ native utterances for computer-assisted language learning systems”, Speech Communication, Vol 51, Issue 10, pp875-882, 2009

[3] Zhang, et al, “Generalized segment posterior probability for automatic Mandarin pronunciation evaluation,” Proc. ICASSP, pp.201-204, 2007

[4] C.Huang et al, “Improving automatic evaluation of mandarin pronunciation with speaker adaptive training (SAT) and MLLR speaker adaptation”, Proc. ISCSLP2008, pp37-40, 2008

[5] D.Luo, et al, “Quantitative analysis of the adverse effect of speaker adaptation on pronunciation evaluation”, Proc. ASJ spring meeting, 1-P-22，pp.173-176 , 2009

[6] D.Luo, et al, “Analysis and utilization of MLLR adaptation technique for learners’ pronunciation evaluation,” Proc. INTERSPEECH, pp.608-611, 2009

[7] 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

[8] Kanters et al., “The goodness of pronunciation algorithm: a detailed performance study,” Proc. SLaTE, CD-ROM, 2009

[9] Minematsu et al, ”English Speech Database Read by Japanese Learners for CALL System Development,” Proceedings of International Conference on Language Resources and Evaluation, pp896-903, 2002

[10] Tanaka et al, “Acoustic models of language-indigent phonetic code systems for speech processing,” Spring meeting of the Acoustical Society of Japan, pp191-192, 2001

[11] Y.Tsubota et al, “An English pronunciation learning system for Japanese students based on diagnosis of critical pronunciation errors”, ReCALL 16(1), pp173-188, 2004

[12] S. Kohmoto, “Applied English Phonology: Teaching of English pronunciation to the Native Japanese Speaker,” Tokyo Tanaka Press, 1965.