INTERSPEECH
2009 BRIGHTON

Analysis and Utilization of MLLR Speaker Adaptation Technique for Learners' Pronunciation Evaluation

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 investigate the effects and problems of MLLR speaker adaptation when applied to 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 these experiments, two novel methods, Forced-aligned GOP scoring and Regularized-MLLR adaptation, are proposed to solve the adverse effects of MLLR adaption. Experimental results show that the proposed methods can better utilize MLLR adaptation and avoid over-adaptation.

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

1. Introduction

One of the largest challenges in CALL system development is to deal with the acoustic mismatches between learners' speech and the acoustic models. In ASR, speaker adaptation techniques have been proved effective in reducing the model mismatches. However, instead of recognizing the intended words by the speaker, the purposes of CALL are to evaluate and detect mispronunciations in learners' speech. When conventional adaptation techniques are directly applied to the acoustic models used in CALL, the incorrect pronunciation might be recognized as correct due to over-adaption. Although there are some studies using global adaption for CALL system to avoid over-adaptation [1, 2], to the best of the authors' knowledge, no quantitative analysis has been reported to investigate the adverse effects of speaker adaptation.

This study investigates the effects of conventional maximum likelihood linear regression (MLLR) speaker adaptation on pronunciation evaluation for CALL in two ways: automatic scoring and phoneme error detection. Based on the analysis results, we provide solutions to the over-adaption problem. Experimental results show the high validity of the proposed methods.

2. Pronunciation evaluation with MLLR

2.1. Automatic scoring

2.1.1. Goodness of Pronunciation

The confidence-based pronunciation assessment, which is defined as the Goodness of Pronunciation (GOP), is often used for assessing learners' articulation and shows good results [3,4]. In this study, we use HMM acoustic models trained on WSJ and TIMIT corpus 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.

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

$$= \frac{1}{D_p}\log\left(\frac{P(O^{(p)} \mid p)P(p)}{\sum_{q\in Q} P(O^{(p)} \mid q)P(q)}\right) \tag{2}$$

$$\approx \frac{1}{D_p}\log\left(\frac{P(O^{(p)} \mid p)}{\max_{q\in Q} P(O^{(p)} \mid q)}\right), \tag{3}$$

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 3 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.

Since the boundaries of phoneme p yielded from forced alignment do not necessarily coincide with the boundaries of phoneme q resulted from continuous phoneme recognition, the frame average log likelihoods of the same speech segment are often used in traditional GOP calculation [3].

2.1.2. Experimental results

We use ERJ (English Read by Japanese Students) corpus [5] to measure GOP scores with MLLR adaptation. This corpus contains proficiency labels rated by phonetic experts. 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 are calculated as automatic score for the learner. 60 sentence utterances of each leaner were used as adaptation data.

We investigate the correlations between GOP scores and human scores while increasing the number of the nodes of regression class tree. Here the number 0 means without adaptation, and 1 represents global adaption. As shown in Figure 1, global adaptation yielded the best correlation of 0.65, yet while the number of nodes of regression class tree increases from 2, the performance drops. When the number is larger

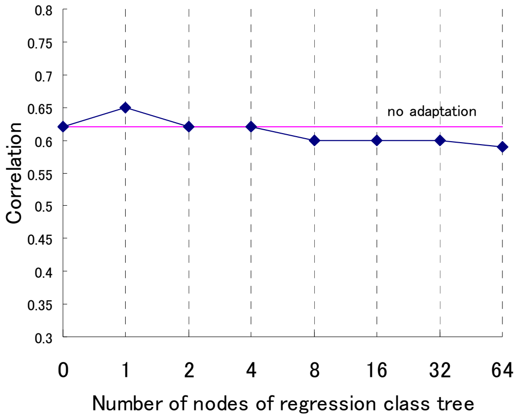

Figure 1: Correlations between GOP scores and manual scores as the number of classes in MLLR increases

than 4, the correlation is even worse than the original model.

2.2. Phoneme error detection

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 [6] consists of 5950 utterances of 850 basic English words read by seven Japanese learners. 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 [7].

We used the utterances of 4 speakers (2 males and 2 females) with many typical errors of Japanese learners.

2.2.1. Error detection network grammar

One of the major methods to detect pronunciation errors is using pronunciation networks that include correct pronunciation and various error patterns to predict learners' possible mispronunciations. By referring to [8], 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 always used, the relative increase or decrease of detection accuracy can be used to measure the performance of the acoustic models with MLLR.

2.2.2. Experimental results

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

$$\text{Precision} = \frac{N_{hit}}{N_{total}} = \frac{N_{hit}}{N_{hit} + N_{FA}} \qquad (4)$$

$$\text{Recall} = \frac{N_{hit}}{N_{labeled}} \qquad , \qquad (5)$$

where N_{hit} represents the number of the errors that were correctly detected , N_{total} is the total number of detected errors, N_{FA} is the number of false alarms and $N_{labeled}$ is the number of all the errors that were detected by phoneticians, and F-measure defined as below is also calculated to combine the two measures.

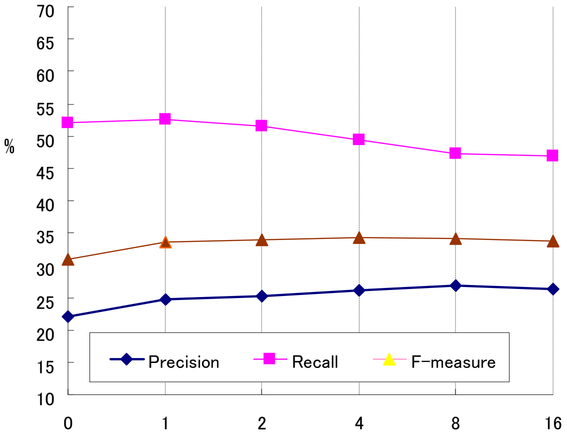

Figure 2: Error detection performances

$$\text{F} - \text{measure} = \frac{2\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \qquad (6)$$

Figure 2 shows the performances of error detection with adaption. Although the precision rates keep increasing when more transforms were used for adaptation, the recall rate drops when the number of nodes is larger than 2. This indicates that with adaptation to reduce model mismatches, the number of false alarms N_{FA} drops significantly, thus the precision rate increases. However, since the number of $N_{labeled}$ is only decided by the label, 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 pronunciation. In the following two sections, to solve the problem of over-adaptation, two novel methods are proposed and evaluated for automatic scoring and error detection.

3. Forced-aligned GOP scores

As mentioned in section 2.1.1, conventional GOP calculation refers to the results of both forced alignment and continuous phoneme recognition. This causes a problem as depicted in (a) of Figure 3, that 3 phonemes from detected by continuous phoneme recognition might correspond to one forced aligned phoneme p. In this case, GOP score for p is calculated using the log likelihood of p and average log likelihood of q1, q2 and q3 within the segment of p.

As an alternative way of calculating GOP score, we can first obtain the phoneme boundaries for phoneme p based on the result of forced alignment, and then calculated the posterior probability of that segment using equation (3) directly. We call this method Forced-aligned GOP (F-GOP). This method refers only to the boundaries of forced alignment and actually separates the calculation of GOP score into two processes, one is detecting the phoneme boundaries and the other is calculating the posterior probability for that segment. We can use different models for the two processes. Because the ERJ database does not contain labels with phoneme boundary information, we couldn't examine directly the improvement of boundary detection by using adaptation, but we can still compare the result of conventional GOP and F-GOP with MLLR adaptation. We tested different combinations of acoustic models for detecting phoneme boundaries and calculating posterior probabilities.

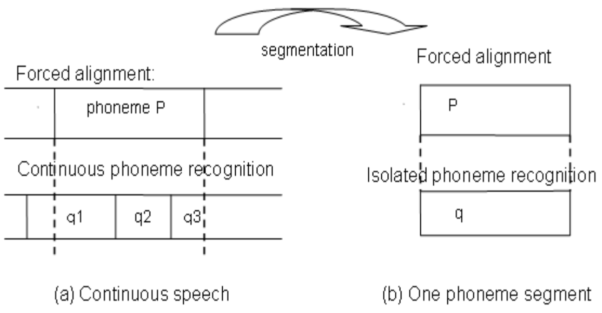

Figure 3: Forced-aligned GOP method

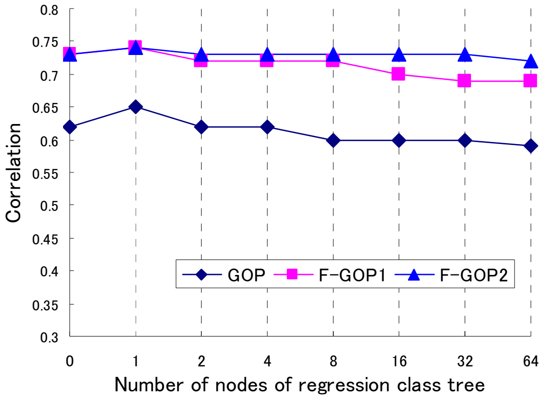

Figure 4: Correlations between human scores and Forced-aligned GOP, comparing with conventional GOP.

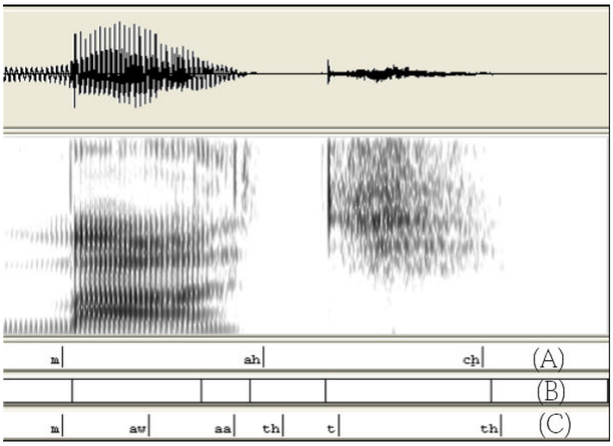

Figure 5: Phoneme segmentation results, A) forced alignment, B) unsupervised bottom-up clustering, C) continuous phoneme recognition

Figure 3 shows the results of three conditions: F-GOP1, which used the same models for both phoneme boundary detection and posterior probability calculation, F-GOP2, which used the adapted models to detect phoneme forced alignment boundaries, and the original models to calculate posterior probabilities, comparing with conventional GOP scores.

As shown in Figure 4, two kinds of F-GOP outperformed the conventional GOP. We consider this is because F-GOP did not refer to the results of continuous phoneme recognition which is often unreliable. Figure 5 shows an example of phoneme segmentation results of A) forced alignment, B) unsupervised bottom-up clustering and C) continuous phoneme recognition. In this example, the result of continuous phoneme recognition is even worse than segmentation based on unsupervised clustering [9], which uses no prior knowledge at all.

F-GOP2 shows better performance than F-GOP1, especially when the number of the nodes of regression class tree is larger than 2. The only difference between F-GOP1 and F-GOP2 is that while F-GOP1 used the adapted models to calculate posterior probabilities, F-GOP2 used the original models to evaluate the same phoneme segment. This indicates that with more transforms used for adaption, the “judgment” of the acoustic model becomes worse. By utilizing only the better phoneme alignment results based on the adapted models, F-GOP can better benefit from speaker adaptation.

4. Regularized-MLLR Adaptation

The results of automatic scoring and error detection experiments clearly show the adverse and good effects of MLLR adaptation on pronunciation evaluation. If we can solve the problem of “bad judgment” of adapted models, we might be able to achieve both better recall and precision. Regularized-MLLR is one possible solution to this problem.

4.1. Definition of Regularized-MLLR

In order to regularize MLLR transformation so that the erroneous pronunciation will not be “transformed” to good pronunciation, we add constraints to conventional MLLR.

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

$$Q(M,\hat{M}) = \frac{1}{2}\sum_{r=1}^{R}\sum_{m_r=1}^{M_r}\sum_{t=1}^{T} L_{m_r}(t)\times \left[K^{(m)} + \log\left|\hat{\Sigma}_{m_r}\right| + (o(t)-\hat{\mu}_{m_r})^T\hat{\Sigma}_{m_r}^{-1}(o(t)-\hat{\mu}_{m_r})\right] \tag{7}$$

where M is the HMM model set, $\hat{M}$ is the adapted model set, and 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, and $L_{m_r}(t)$ is the occupation likelihood that O_T is given from Gaussian component $q_{m_r}(t)$,

$$L_{m_r}(t) = p(q_{m_r}(t)\,|\,M,O_T)\ . \tag{8}$$

Here we obtained a set of transforms estimated from a group of teachers who are native speakers of General English. Teachers’ transforms are used to constrain the transforms for the learners to avoid bad pronunciation being transformed into good pronunciation.

Let $\{W_r^{C_1},\ldots,W_r^{C_N}\}$ denote a set of transformation matrices estimated from a group of N teachers, and $W_r^C = (1/N)\sum_n W_r^{C_n}$ represents the mean of these matrices. The objective function for Regularized-MLLR is defined as

$$\max_{W_r}\{Q(M,\hat{M}) - \lambda\sum_{r=1}^{R}\left\|W_r - W_r^C\right\|_F^2\}\ , \tag{9}$$

where λ is a parameter depending on the acoustic characteristics of the speaker. In conventional MLLR, W_r is estimated by maximizing $Q(M,\hat{M})$. In the proposed method, however, over-adaptation is avoided by the 2-nd term of Equation (9). This term functions as penalty of changing the model parameters so radically.

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

$$\hat{\mu}_{m_r} = W_r \xi_{m_r} \quad , \tag{10}$$

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

$$\xi_{m_r} = [1 \;\; \mu_1 \;\; \mu_2 \;\cdots\; \mu_d\,]^T , \tag{11}$$

where *d* is the dimensionality of the data.
Considering the row decomposition $W_r = [w_{r,1}; w_{r,2}; ...; w_{r,d}]$, Equation (9) can be decomposed into minimization of a set of cost functions related to $w_{r,j}$,

$$f(w_{r,j}) = K_j + w_{r,j} H_r^j w_{r,j}^T - 2 w_{r,j} N_r^{(j)T} , \tag{12}$$

Where K_j is a constant and,

$$H_i^j = \sum_{m_r=1}^{M_r} \frac{1}{\sigma^2_{m_{r,j}}} \xi_{m_r} \xi_{m_r}^T \sum_{t=1}^{T} L_{m_r}(t) - \lambda I \tag{13}$$

$$N_r^j = \sum_{m_r=1}^{M_r} \sum_{t=1}^{T} L_{m_r}(t) \frac{1}{\sigma^2_{m_{r,j}}} o_j(t) \xi_{m_r}^T - \lambda w_{r,j}^C \tag{14}$$

The optimal $w_{r,j}$ is given by solving

$$\frac{\partial f(w_{r,j})}{\partial w_{r,j}} = 0 \quad , \tag{15}$$

which yields,

$$w_{r,j} = N_r^j (H_r^j)^{-1} \tag{16}$$

4.2. Experimental results

We used 10 native teachers' utterances of General English from the ERJ corpus to calculate the mean of transformation matrices, W_r^C, to regularize transforms for Japanese learners. The parameter λ was experimentally estimated for each of the 4 learners.

As shown in Figure 6, Regularized-MLLR improved the performance of recall rate and kept relatively high precision that were achieved by adaptation with more nodes of regression class tree. This indicates that by setting the right parameter for each speaker, Regularized-MLLR can not only benefit from a lower number of false alarms due to mismatches, but also high ability to detect erroneous pronunciation is obtained.

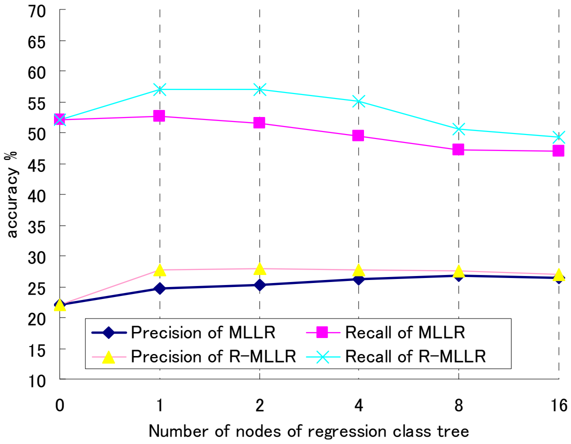

Figure 6: Performances of acoustic models with Regularized-MLLR(R-MLLR) and MLLR

5. Conclusion

This study analyzes the effects of MLLR speaker adaption on pronunciation evaluation based on experiments of automatic scoring and error detection on reliable databases. Forced-aligned GOP and Regularized-MLLR have been proposed for automatic scoring and error detection. Experimental results show that the proposed methods can better utilize the merits of adaptation and reduce side effects of over-adaptation than conventional methods for CALL systems.

For future work, we are working on automatic estimation of parameters for different learners and combine the two proposed methods together. We are also planning on improving the Regularized-MLLR to better constrain transformation matrices for adaptation.

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

[3] 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

[4] L.Neumeyer et al., "Automatic scoring of pronunciation quality," *Speech Communications*, 30(2-3): pp.83-93, 2000

[5] 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

[6] 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

[7] 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

[8] S. Kohmoto, "*Applied English Phonology: Teaching of English pronunciation to the Native Japanese Speaker*," Tokyo Tanaka Press, 1965.

[9] Y. Qiao, N. Shimomura, N. Minematsu, "Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons," *Proc. ICASSP*, pp.3989-3992, 2008