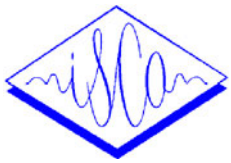

Second Language Studies - Acquisition, Learning, Education and Technology
Tokyo, Japan
September 22-24, 2010

Pronunciation Proficiency Estimation Based on Multilayer Regression Analysis Using Speaker-independent Structural Features

Masayuki Suzuki[1], *Yu Qiao*[2], *Nobuaki Minematsu*[1], *Keikichi Hirose*[1]

[1]The University of Tokyo, Japan, [2]Shenzhen Institutes of Advanced Technology, China

suzuki@gavo.t.u-tokyo.ac.jp

Abstract

Teachers can assess the pronunciations of students independently of extra-linguistic features such as age and gender observed in the students' utterances. This capacity is, however, difficult to realize on machines because linguistic differences and extra-linguistic differences change acoustic features commonly. Therefore, the performance of automatic pronunciation assessment is inevitably affected by the extra-linguistic features. Recently, we proposed acoustic features that are independent of extra-linguistic factors, called *structural features* and realized a technique for pronunciation proficiency estimation that is extremely robust to these factors. In this paper, we extend this technique with multilayer regression analysis, where supervised learning is done at each layer by using teachers' scores of that layer. Experiments of estimating the proficiency show that higher correlations between teachers and machines are obtained compared to our previous structure-based assessment.

Index Terms: structural features, multilayer regression analysis, proficiency estimation, CALL

1. Introduction

It is obvious that direct and acoustic comparison between a teacher's utterance and a student's imitative utterance leads not to a goodness score of pronunciation but to that of impersonation. For example, DTW-based distance between the two utterances quantifies how well that student can impersonate his/her teacher. Therefore, when DTW is applied to pronunciation assessment, the teacher's utterance has to be acoustically modified so that it can match with the voice quality of the student [1]. In other words, for each student, a teacher who sounds like him/her is needed. It is the case with HMMs, often adapted to new students [2, 3]. Although pronunciation training is pedagogically different from impersonation training, technically speaking, a pronunciation assessment system can be built based on the impersonation assessment techniques combined with acoustic adaptation. This strategy surely leads to a technical solution but we doubt whether it is a good pedagogical solution.

In language learning and acquisition, students and infants do not try to produce utterances acoustically matched with the utterances of teachers and parents, respectively. Humans can ignore the extra-lingusitic features very easily and teachers also ignore these features when assessing the pronunciations of students. As far as we know, however, the conventional speech technologies can hardly remove only the extra-linguistic features and this is why engineers have tried to solve the pronunciation assessment problem based on acoustic comparison.

To build a human teacher-like and probably pedagogically-sound pronunciation assessment system, in [4], we proposed new speech features that are independent of the extra-linguistic factors and highly accordant to a linguistic theory of "relational invariance" [5]. The resulting system also has some technical merits. It shows the extremely robust performance against students' age and gender with no explicit adaptation [6] and it successfully classifies the students exclusively based on their pronunciation differences with age and gender ignored [7].

The new features are relational and contrastive features, which are obtained by removing absolute features completely. Speaker differences are often modeled as space mapping in many studies of voice conversion. This indicates that speaker-indepedent features are defined as transform-invariant features. We proved that f-divergence between two distributions (events) is invariant with any kind of continuous and invertible transform and that completely invariant features, if any, have to be f-div. [8]. Then, if a student's pronunciation is represented only by f-div., that representation is speaker-independent. Although this speaker-independence is very beneficial, the contrastive representation also has a drawback, i.e., high dimensionality. If a student's pronunciation contains M acoustic events, then, ${}_MC_2 = \frac{M(M-1)}{2}$ contrastive features are calculated, some of which may be irrelevant to estimate the pronunciation proficiency of that student.

In our previous studies [6, 7], we adequately selected a part of the contrastive features to increase the robustness and the correlation between teacher scores and machine scores simultaneously. In this paper, we generalize this approach by regression analysis. If automatic estimation of the overall pronunciation proficiency scores for individual students is the only research target, what we have to do is to calculate regression coefficients of the ${}_MC_2$ parameters to predict the overall score. Or, if only one wants to reduce the parameter dimension, PCA/LDA-based compression can be used. In developing a good CALL system, however, in addition to the overall score, diagnostic instructions should be generated adequately. For example, instructions on which vowels or consonants should be corrected at first may be helpful to students. To reduce the parameter dimension and realize the mechanism for instruction generation simultaneously, however, direct regression or PCA/LDA-based compression is not a good choice. This is because direct regression can give us no diagnostic instruction and the compressed feature representation is often difficult to interpret. In this paper, we introduce step-wise regression analysis instead, called as multilayer regression analysis, where pronunciation analysis with different resolutions is performed on different layers. While the final regression predicts the overall score, the intermediate regression predicts the proficiency scores for individual phonemes.

2. Pronunciation structure

2.1. Theory of invariant pronunciation structure

Two speakers have different vocal tract lengths and shapes. Speaker difference is often modeled mathematically as a lin-

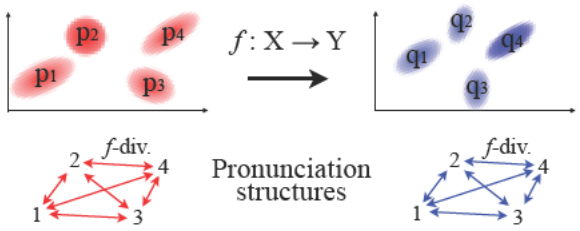

Figure 1: Transform-invariant pronunciation structures

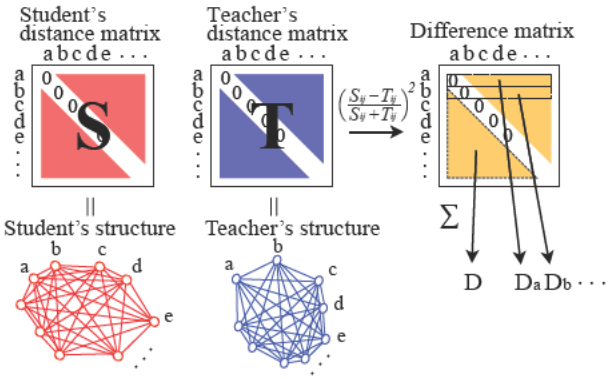

Figure 2: Structure-based pronunciation assessment

ear or non-linear transformation of cepstrum coefficients. Then, transform-invariant features, if any, can be robust features.

Consider feature space $\boldsymbol{X}$ and pattern $\boldsymbol{P}$ in $\boldsymbol{X}$. Suppose that $\boldsymbol{P}$ can be decomposed into M events $\{p_i\}_{i=1}^{M}$. Each event is described as distribution $p_i(\boldsymbol{x})$ in the space. Assume that there is an invertible transformation $f : \boldsymbol{X} \rightarrow \boldsymbol{Y}$ which transforms $\boldsymbol{X}$ into new space $\boldsymbol{Y}$. In this way, pattern $\boldsymbol{P}$ in $\boldsymbol{X}$ is mapped to pattern $\boldsymbol{Q}$ in $\boldsymbol{Y}$ and event p_i is transformed to event q_i. Here, what we want is invariant metrics in both the spaces.

As described in Section 1, f-div. is invariant with any kind of invertible and differentiable transform. Figure 1 shows two invariant pronunciation structures composed of only f-divs. f-div. is a family of divergence measures defined as

$$f_{div}\left(p_1, p_2\right) = \oint p_2(\boldsymbol{x}) g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right) d\boldsymbol{x}, \tag{1}$$

where $g : (0, \infty) \rightarrow R$ is a real convex function and $g(1)=0$. Many well known distances and divergences can be seen as special examples of f-div. For example, when $\sqrt{t}$ is used for $g(t)$, $-\ln(f_{div})$ becomes the Bhattacharyya distance (BD),

$$BD(p_1, p_2) = -\ln \oint \sqrt{p_1(\boldsymbol{x}) p_2(\boldsymbol{x})} d\boldsymbol{x}. \tag{2}$$

We use $\sqrt{BD}$ to form the pronunciation structures in this paper.

2.2. Structure-based pronunciation assessment

Figure 2 shows a diagram of our previous structure-based pronunciation assessment. A student's structure S and a teacher's structure T are extracted from their respective utterances. A structure is represented as a distance matrix and the structural difference between two structures is calculated as

$$D(S,T) = \sqrt{\frac{1}{M} \sum_{i<j} \left(\frac{S_{ij} - T_{ij}}{S_{ij} + T_{ij}}\right)^2}, \tag{3}$$

Figure 3: Two-layered regression analysis

where S and T are two distance matrices whose elements, S_{ij} and T_{ij}, are calculated as $\sqrt{BD}$ [6]. M is the number of distributions, which typically indicate phonemes. From these two distance matrices, we derive a difference matrix whose element D_{ij} is $((S_{ij} - T_{ij})/(S_{ij} + T_{ij}))^2$, shown in Figure 2. In [6], through structural comparison between each student in a Japanese-English database and a specific teacher, the pronunciation proficiency of that student was automatically estimated. The obtained score was compared to the proficiency scores provided by the database and a high correlation was found. In [9], D is decomposed into a phoneme-specific score D_a defined as

$$D_a(S,T) = \sqrt{\frac{1}{M} \sum_{i \neq a} \left(\frac{S_{ai} - T_{ai}}{S_{ai} + T_{ai}}\right)^2}. \tag{4}$$

D_a is used to generate diagnostic instructions for phoneme a.

3. Multilayer regression analysis

3.1. Two-layered regression analyis

Generally speaking, a pronunciation structure has high dimensionality. When the number of distributions of a structure is M, the number of parameters is ${}_M C_2$. The high dimensionality not only increases the computational cost but also degrades the performance. In our previous studies [6, 7], a part of S_{ij} and T_{ij} were selectively used to calculate $D(S,T)$. In this paper, we generalize this approach by integrating two-layered regression analysis with the structure-based assessment.

Figure 3 shows a diagram of two-layered regression analysis. The first layer regression analysis is performed using each row vector of the difference matrix as independent variable and teachers' score for each phoneme as dependent variable. The estimated weight vector $\boldsymbol{w_i}$ gives us the information on which contrast to phoneme i is more important to evaluate phoneme i. The results of the regression are estimated proficiency scores for the phonemes. Then, the second layer regression analysis is carried out using these scores as independent variables to predict the teachers' rating of overall student proficiency. The estimated weight vector $\boldsymbol{w_{all}}$ shows on which phonemes more focus should be put. This two-layered regression analysis reduces dimensionality like PCA or LDA, but unlike these, it can estimate a score for each phoneme.

3.2. Three-layered regression analysis

We can obtain multiple difference matrices by using multiple teachers. These matrices surely have more information than a single difference matrix, but the dimensionality of n matrices is naturally higher than that of a single matrix.

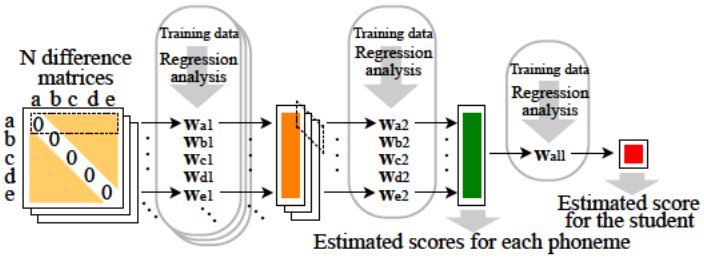

Figure 4: Three-layered regression analysis

Table 1: Inter-rater correlations for each vowel

ɑː	æ	ʌ	ɔː	ɛ	ɚː	ɪ	iː	ʊ	uː	avg
0.69	0.66	0.61	0.43	0.19	0.83	0.63	0.66	0.46	0.46	0.56

Table 2: Intra-rater correlations for each vowel

ɑː	æ	ʌ	ɔː	ɛ	ɚː	ɪ	iː	ʊ	uː	avg
0.79	0.85	0.37	0.61	0.33	0.82	0.73	0.84	0.59	0.49	0.64

Table 3: Standard deviation for each vowel

ɑː	æ	ʌ	ɔː	ɛ	ɚː	ɪ	iː	ʊ	uː
0.87	0.95	0.70	0.64	0.36	1.13	0.93	0.84	0.58	0.57

For multiple difference matrices, we extend the two-layered regression analysis to three-layered analysis. Figure 4 shows its diagram. The first layer regression and the third layer regression is almost the same as the first layer regression and the second layer regression in the case of two-layered regression analysis, respectively. At the second layer regression in Figure 4, the results of the first layer regression of each phoneme are used as independent variables. The estimated weight vector $\boldsymbol{w_{i2}}$ tells us which difference matrix is more important.

4. Experiments

4.1. Speech materials used in the experiments

We collected word utterances of 36 male Japanese students of Tokyo Sugamo junior high school and those of 5 male Japanese teachers of English working at that school. The average age over the students is 13.5 and that over the teachers is 45.2. Recording was done not in a sound-proof room but in a normal classroom with headset microphones. Each speaker was asked to read ten monosyllabic words only once, which correspond to ten American English monophthongs excluding schwa: pot ([ɑː]), bat ([æ]), but ([ʌ]), bought ([ɔː]), bet ([ɛ]), bird ([ɚː]), bit ([ɪ]), beat ([iː]), put ([ʊ]), and boot ([uː]). In the experiments, we focus on the vowel structures for the individual speakers. These utterances were digitized at 16 kHz sampling rate and 16 bit accuracy. From them, we calculated 10-dimensional mel-cepstral features with 25 ms window length and 1 ms frame shift. The vowel portion of each word utterance was detected by forced alignment using speaker-independent HMMs. Then, we estimated a Gaussian distribution (event) for each vowel segment and a vowel structure (distance matrix) for each speaker.

4.2. Teachers' rating of the segmented vowels

We asked three phoneticians to assess the correctness of the vowel quality of each segmented vowel on a web page. As reference, we prepared "correct" vowel productions, which were provided by one of the phoneticians. Two assessment tasks were carried out. One is rating the vowel quality of each segmented vowel by hearing its corresponding "correct" vowel and a score was assigned to each vowel and 10 scores were given in total to each speaker. The other is rating the overall proficiency of producing American English vowels and a single score was given to each speaker. After preliminary discussions with the phoneticians, four-level rating (1 to 4) was adopted in either task, where 4 means phonetically almost the same as the "correct" vowel and 1 means the worst. One phonetician did these tasks twice and the others did them once. Thus, we obtained four scores for each vowel and each speaker. In the regression analysis, using the four scores, the averaged vowel-based scores and the averaged speaker-based scores were used as prediction targets.

4.3. Analysis of inter-rater and intra-rater agreement

Before the machine assessment experiments, we calculated correlations of the scores between different phoneticians and those within the phonetician who did the assessment tasks twice. Table1 shows the average of the correlations of the vowel-based scores between two different phoneticians and Table2 shows the correlations of the vowel-based scores between two assessment experiments done by a phonetician. In [ɛ] and [uː], agreement is low (cor.$<$0.5) in both the tables. This is because standard deviations of these vowels are small among speakers (see Table3) since Japanese vowels can be substituted for them. For other vowels whose deviations are less than 0.8 ([ʌ], [ɔː], [ʊ]), their correlations are less than 0.5 in either Table1 or 2. To estimate the proficiency of Japanese students speaking English, their pronunciations of the other vowels of [ɑː], [æ], [ɚː], [ɪ], and [iː] are expected to reflect their proficiency level better, but an adequate vowel subset will be dependent on the mother tongue.

4.4. Multilayer regression with structural features

In the two-layered regression analysis shown in Figure 3, the vowel structure of the reference phonetician was used as teacher matrix T of Figure 2. In the first regression and the second regression of Figure 3, the vowel-based scores and the speaker-based scores were used as prediction targets, respectively. To calculate the regression coefficients, we used ridge regression with sign constraint because larger differences in the different matrix are logically expected to lower the proficiency scores.

In the three-layered regression analysis of Figure 4, multiple teacher matrices can be used. Further, even from a single teacher, multiple matrices can be obtained by using different features. In the experiments, we added the vowel structures of other 6 teachers. Further, the vowel structures created by using low-pass filtered speech data were also added. This is because, in [10], the upper bands of the spectrum of vowels were shown to carry a large portion of speaker identity, which is irrelevant to pronunciation assessment. Thus [7 teachers] $\times$ [2 features] $=$ 14 matrices were used for the three-layered regression analysis. In the experiments, 3.0 kHz was used as cut-off frequency.

Using the 41 speakers (36 male young students and 5 male teachers), leave-one-out cross-validation experiments were carried out and the correlation between human scores and machine scores was calculated. For comparison, our previous structure-based assessment [9] was conducted using the same data. Here, Equation (3) was used to predict the overall proficiency score. We should note that our previous method is unsupervised learning, where human labels are not referred to, and that our new method is supervised learning with the labels. The three-layered regression with low-pass filtered speech and that without it were

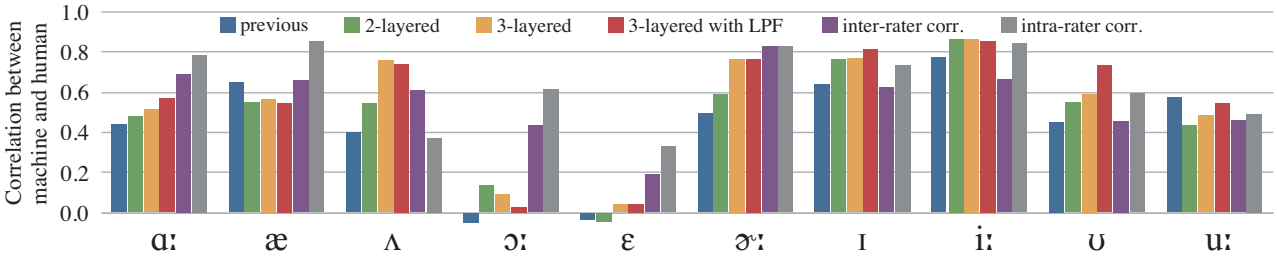

Figure 5: Correlations of the vowel-based scores between machine and human

Table 4: Correlations of the overall pronunciation proficiency

Previous	2-layered	3-layered	3-layered with LPF
0.76	0.79	0.87	0.88

also compared in terms of their estimation performances.

4.5. Results and discussion

Table4 shows the performance of predicting the overall proficiency using four methods: our previous structure-based assessment [9], two-layered regression analysis with a single teacher matrix, three-layered regression with seven teacher matrices, three-layered regression with 14 teacher matrices (seven teachers and two features). The prediction performance is improved by regression analysis and the use of multiple teacher matrices is very effective. Matrices with low-pass filtered speech data improves the performance only slightly. This implies that the structural features are independent of extra-linguistic features.

Figure 5 shows the correlations for each vowel between human scores and machine scores, which are calculated using the four methods. In the figure, the inter-rater and intra-rater correlations are also plotted as reference. In the five vowels of higher standard deviations in Table3, [ɑː], [æ], [ɚː], [ɪ], and [iː], the correlations based on the multilayer regression are higher than those of our previous structure-based assessment except for the case of [æ]. In the other five vowels, for [ʌ] and [ʊ], our proposed method outperforms our previous one. For [ɔː] and [ɛ], all the four methods give us only a very low performance. Since [ɛ] shows the smallest standard deviation in Table3, this result is acceptable. As for [ɔː], however, the structural features may not be sufficient to evaluate it and need further analysis.

The proposed framework captures only the relational aspects of the pronunciation and ignores the absolute aspects. In other words, the framework does not know the spectral shape of the individual vowels. This strategy is originated from a hypothesis that students learn and imitate not absolute features of individual sounds but their systemic features in teachers' pronunciation [5]. The imitation of absolute features results in impersonation. For unvoiced consonant sounds, however, since they are less dependent on speakers, the imitation of these sounds has to be like impersonation. With three-layered regression analysis, absolute features or scores can be easily integrated and, in [11], GOP (Goodness of Pronunciation) scores [12] are used as new and absolute scores. In [6], [7], and [11], comparison is done between the GOP and the structure for both vowels and consonants. Interested readers should refer to them.

5. Conclusions

In this paper, we extended our previously proposed technique for automatic estimation of the pronunciation proficiency. This technique is based on speaker-independent and contrastive features of speech and models a student's pronunciation as structure (distance matrix). To avoid a high dimensionality problem, we proposed multilayer regression analysis where phoneme-level and speaker-level regressions were implemented. Experiments showed that the proposed method outperforms our previous method in terms of the performance of estimating both speaker-level and phoneme-level proficiency.

For future work, we are planning to integrate structural features and absolute features to enable a system to capture different acoustic features according to different phonemes. [11] shows some results of our initial attempt for that integration and also shows the extremely high robustness of our proposed framework against age and gender. Besides, we are planning to apply multilayer regression analysis on other tasks, such as pronunciation error detection, dialect analysis, and so on.

6. References

[1] H. Hamada, S. Miki, and R. Nakatsu, "Automatic evaluation of English pronunciation based on speech recognition techniques," *IEICE Trans. Inf. & Syst.*, E76-D, 3, 352–359, 1993.

[2] Y. Ohkawa, M. Suzuki, H. Ogasawara, A. Ito, and S. Makino, "A speaker adaptation method for non-native speech using learners' native utterances for computer-assisted language learning systems," *Speech Communication*, 51, 875–882, 2009.

[3] M. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech," *Proc. SLaTE*, CD-ROM, 2007.

[4] N. Minematsu, "Pronunciation assessment based upon the phonological distortions observed in language learners' utterances," *Proc. INTERSPEECH*, 1669–1672, 2004.

[5] R. Jakobson and L.R. Waugh, *The sound shape of language,* Mouton De Gruyter, 1987.

[6] M. Suzuki, D. Luo, N. Minematsu, and K. Hirose, "Improved structure-based automatic estimation of pronunciation proficiency," *Proc. SLaTE*, CD-ROM, 2009.

[7] M. Suzuki, N. Minematsu, D. Luo, and K. Hirose, "Sub-structure-based estimation of pronunciation proficiency and classification of learners," *Proc. ASRU*, 574–579, 2009.

[8] Y. Qiao and N. Minematsu, "A study on invariance of f-divergence and its application to speech recognition," *IEEE Transactions on Signal Processing*, 58, 2010 (to appear).

[9] N. Minematsu, K. Kamata, S. Asakawa, T. Makino, T. Nishimura, and K. Hirose, "Structural assessment of language learners' pronunciation," *Proc. INTERSPEECH*, 210–213, 2007.

[10] T. Kitamura and M. Akagi, "Speaker individualities in speech spectral envelopes," *Journal of the Acoustic Society of Japan*, 16, 5, 283–289, 1995.

[11] M. Suzuki, Y. Qiao, N. Minematsu, and K. Hirose, "Integration of multilayer regression analysis with structure-based pronunciation assessment," *Proc. INTERSPEECH*, 2010 (submitted)

[12] S.M. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, 30, 95–108, 2000.