Integration of Multilayer Regression Analysis with Structure-based Pronunciation Assessment

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

Automatic pronunciation assessment has several difficulties. Adequacy in controlling the vocal organs is often estimated from the spectral envelopes of input utterances but the envelope patterns are also affected by other factors such as speaker identity. Recently, a new method of speech representation was proposed where these non-linguistic variations are effectively removed through modeling only the contrastive aspects of speech features. This speech representation is called speech structure. However, the often excessively high dimensionality of the speech structure can degrade the performance of structure-based pronunciation assessment. To deal with this problem, we integrate multilayer regression analysis with the structure-based assessment. The results show higher correlation between human and machine scores and also show much higher robustness to speaker differences compared to widely used GOP-based analysis.

Index Terms: CALL, speech structure, regression, GOP

1. Introduction

Automatic pronunciation assessment is a task used to evaluate only the linguistic aspect of utterances. However, speech features inevitably include acoustic variations caused by non-linguistic factors such as the speaker, communication channel and noise. The same pronunciation can lead to different acoustic observations due to different speakers and different environments. To deal with these variations, modern pronunciation assessment approaches mainly make use of statistical methods to model the distributions of the acoustic features [1]. These methods can achieve relatively high performance when there is a good match between training and testing conditions. But their performance always degrades significantly when these conditions are mismatched. In Automatic Speech Recognition (ASR), speaker adaptation techniques have proved effective at reducing mismatches. However, if the acoustic models used in pronunciation assessment are adapted to learners, incorrect pronunciations might be recognized as correct due to over-adaptation [2].

To solve the mismatch problem, the third author of this paper proposed a new speech representation, called speech structure, which aims at removing the non-linguistic factors in speech features [3]. In contrast to classical speech models, speech structures make use of f-divergence to model only the contrastive aspects of speech and discard the absolute features completely. This novel approach has been applied to pronunciation assessment [4, 5], ASR [6], dialect-based speaker classification [7], and speech synthesis [8].

However, the often excessively high dimensionality in speech structures can degrade the performance of structure-based pronunciation assessment. In order to solve this problem, in this paper, we integrate multilayer regression analysis with the structure-based assessment. Through this integration, we can reduce the dimensionality and thereby make it possible to estimate pronunciation proficiency more accurately. We also propose an appropriate combination of structure-based multilayer regression analysis and the widely used Goodness Of Pronunciation (GOP) based analysis [1]. The results show that the proposed methods achieve higher performance than our previous structure-based method and the GOP-based method.

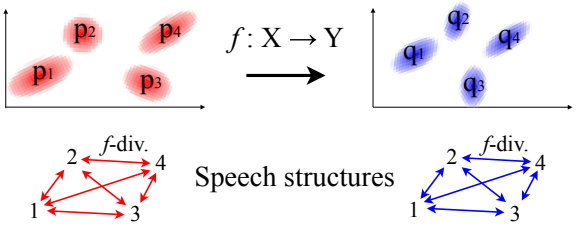

Figure 1: Transform-invariant speech structures

2. Speech structure

2.1. Theory of invariant speech structure

Two speakers have different vocal tract lengths and shapes. In studies of voice conversion, speaker difference is often modeled mathematically as a linear or non-linear transformation of the cepstrum. Especially, vocal tract length difference can be modeled well by monotonic frequency warping in the spectral domain, which can be converted into a linear transformation in the cepstrum domain [9]. These facts indicate that some transform-invariant features can be robust features.

Consider a feature space $\boldsymbol{X}$ and a pattern $\boldsymbol{P}$ in $\boldsymbol{X}$. Suppose $\boldsymbol{P}$ can be decomposed into M events $\{p_i\}_{i=1}^{M}$. Each event is described as a distribution $p_i(\boldsymbol{x})$ in the feature space. Assume there is an invertible transformation $f : \boldsymbol{X} \rightarrow \boldsymbol{Y}$ (linear or non-linear) which transforms $\boldsymbol{X}$ into a new feature 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. Thus what we want is invariant metrics in both space $\boldsymbol{X}$ and space $\boldsymbol{Y}$.

The second author of this paper proved that f-divergence between two distributions are invariant with any kind of invertible and differentiable transform [10]. Fig. 1 shows two invariant speech structures made by only f-divergences in both spaces. f-divergence is a family of divergence measures defined as

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

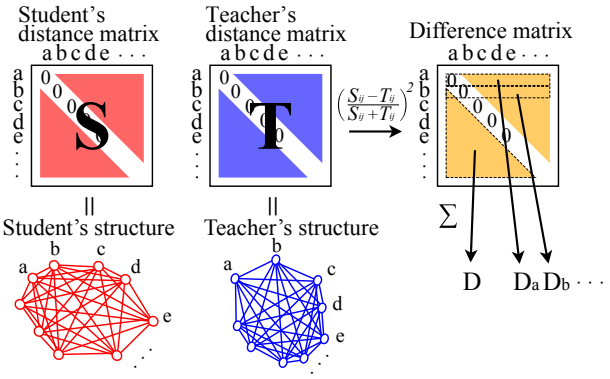

Figure 2: Structure-based pronunciation assessment

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

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

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

2.2. Structure-based pronunciation assessment

Fig. 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

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

where S and T are two distance matrices whose elements are calculated as $\sqrt{BD}$ [5]. 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 Fig. 2 In [5], through structural comparison between each student in a Japanese-English database and a specific teacher, a proficiency score 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 [4], D is decomposed into a phoneme-specific score D_a,

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

D_a was used to generate diagnostic instructions for phoneme a.

3. Multilayer regression analysis

3.1. Two-layered regression analyis

Generally speaking, a speech structure has high dimensionality. Let M denote the number of distributions of it. Then the number of parameters is $M(M-1)/2$. The high dimensionality

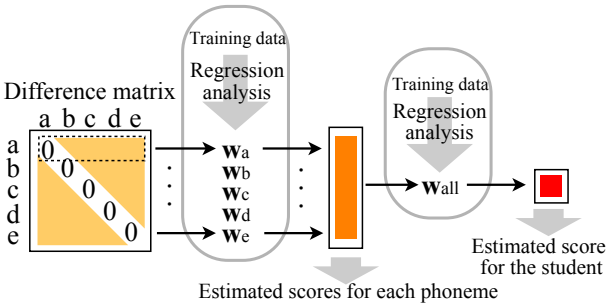

Figure 3: Two-layered regression analysis

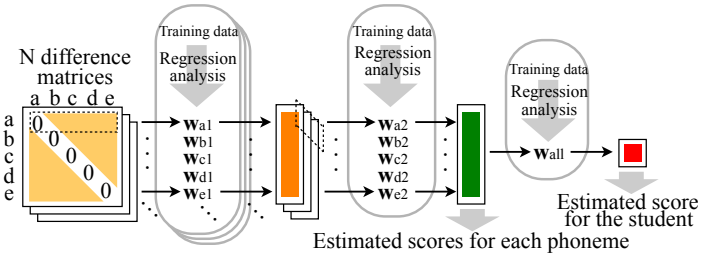

Figure 4: Three-layered regression analysis

not only increases the computational cost but also degrades the performance. In structure-based ASR studies, PCA and LDA were examined, and dimension reduction proved effective at improving the performance. To build a pronunciation learning system, however, diagnostic instructions often have to be generated automatically. Considering this function, dimensionality reduction using PCA or LDA is not appropriate for the system because the reduced parameters are difficult to analyze.

To deal with this problem, we integrate two-layered regression analysis with structure-based pronunciation assessment. Fig. 3 shows a diagram of two-layered regression analysis. The first layer regression analysis is done using each row vector of the difference matrix as independent variable and teacher's 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 done using these scores as independent variables and teacher's overall proficiency as a dependent variable. The estimated weight vector $\boldsymbol{w_{all}}$ shows on which phonemes more focus should be put. The results of the regression can be used as a proficiency score for the student. This two-layered regression analysis reduces dimensionality like PCA or LDA, but unlike these, it can estimate a score for each phoneme at intermediate stages. Those scores can be used for diagnostic instructions although instruction generation is not dealt with in this paper.

3.2. Three-layered regression analysis

We can obtain more than one difference matrix using more than one teacher. Multiple difference matrices have more information than a single difference matrix, but the dimensionality of n difference matrices is higher than that of a single difference matrix.

We extend two-layered regression analysis to three-layered regression analysis for n difference matrices. Fig. 4 shows a

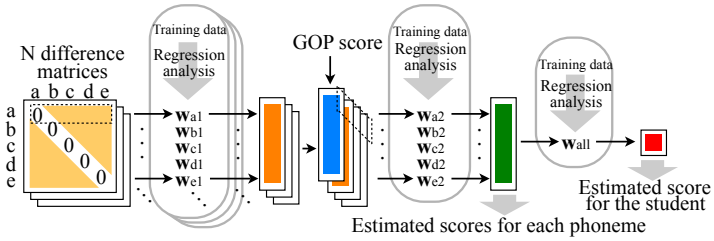

Figure 5: Three-layered regression analysis with GOP scores

diagram of three-layered regression analysis. The first layer regression and the third layer regression is almost the same as the first layer regression and the second layer regression for two-layered regression analysis, respectively. At the second layer regression in Fig. 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.

3.3. Multilayer regression analysis combined with GOP

The speech structure uses the contrastive aspect of speech and discards the absolute features. In contrast, GOP mainly focuses on the absolute aspects of speech. A GOP score of phoneme p_i is posterior probability of the phonemes given input utterances approximately calculated as follows.

$$GOP(O, p_i) \approx \log \left\{ \frac{P(o^{p_i}|p_i)}{\max_{q \in Q} P(o^{p_i}|q)} \right\}, \quad (5)$$

where o^{p_i} is the speech segment obtained for p_i through forced alignment. Q is the inventory of phonemes.

The speech structure and the GOP capture different aspects of speech, so a combination of them might be useful. For example, speech structures are useful for vowels because the acoustic features of vowels are strongly influenced by speaker difference. In contrast, GOP scores are expected to be effective for unvoiced consonants because the features of unvoiced consonants are much less influenced[11].

We propose a method to appropriately combine them. Fig. 5 shows a diagram of three-layered regression analysis combined with GOP scores. The GOP score is combined with the results of the first layer regression. At the second layer regression, the n results of the first layer regression and the GOP score of each phoneme are used as independent variables. The estimated weight vector $\boldsymbol{w_{i2}}$ reflects the importance of each of the n matrices and the GOP score to evaluate phoneme i.

4. Experiments

4.1. The database used in the experiments

The English Read by Japanese (ERJ) database was used in our experiments, which contains 8 sets of read sentences [12]. Each set is composed of about 60 sentences, read by 25 university students, among whom about half are male. Those sentences are a part of the sentences used in the TIMIT database. Proficiency scores are also provided for all the $8 \times 25 = 200$ students, which were rated by five native teachers of American English with good knowledge of phonetics and Japanese English. In the database, the utterances of the same sentences read by 20 native speakers of General American are also included. 18 of them read only half of the sentences and the remaining two (M08, F12) read all the sentences.

Table 1: *Conditions for the acoustic analysis*

sampling	16bit / 16kHz
windows	25ms length and 10ms shift
parameters	MFCC (12dim.)
HMMs	speaker-dependent, context-independent, and 1-mixture monophones with diagonal matrix
topology	5 states and 3 distributions per HMM
monophones	aa, ae, ah, ao, aw, ax, axr, ay, b, ch, d, dh, eh, er, ey, f, g, hh, ih, iy, jh, j, k, l, m, n, ng, ow, oy, p, r, s, sh, t, th, uh, u, w, v, w, y, z, zh, sil

4.2. Structure-based analysis and GOP-based analysis

Table 1 shows the acoustic analysis conditions. From the database, 200 sets of speaker-dependent monophone HMMs were trained. From the two teachers who read all the sentences, 8 sets of HMMs were trained, each corresponding to a sentence set in the database. From the HMMs of a speaker, a speech structure was calculated. A distance between phonemes was obtained as the average over three $\sqrt{BD}$ values between the corresponding states. Eventually, 216 distance matrices were formed in total. In two-layered regression analysis, only M08 was used as a common reference teacher for all the 200 students. In three-layered regression analysis, M08 and F12 were used. In addition, another structure using different features was prepared. Low-pass filtered speech data were used to calculate the structure. This is because [13] showed that the upper bands of the spectrum of vowels carry a large portion of speaker identity, which is irrelevant to pronunciation assessment. Thus $2 \times 2 = 4$ difference matrices were used for three-layered regression analysis. Using the students' 8×25 distance matrices, we did 8-fold cross-validation. We used ridge regression to estimate the weight vectors. In Fig. 3, Fig. 4 and Fig. 5, phoneme-specific scores were used as dependent variables. In this paper, however, since phoneme-level scores were not provided in the database, we used speaker-level scores commonly for any layer. Using the obtained optimal weights, each student's structure of the open set was compared to the teacher's structure of that set. Then, the correlation between human and machine scores was calculated.

To calculate the GOP score, we prepared speaker-independent and 4-mixture monophone HMMs trained with all the utterances of the 20 native speakers in the database. Using 60 sentences from each student, we adapted the HMMs with Maximum Likelihood Linear Regression (MLLR).

We examine three proposed methods: two-layered regression, three-layered regression, and three-layered regression with GOP. For comparison, a sub-structure-based method [5] and two GOP-based methods are tested, i.e., GOP with and without MLLR adaptation. Both in the GOP-based methods, regression analysis is also performed.

4.3. Results of pronunciation assessment experiments

Fig. 6 shows that averages and standard variations of correlation coefficients between the human and machine scores. The average correlation over all the teacher pairs is also plotted as reference. As one can see, the multilayer regression method achieves a higher correlation than our previous sub-structure-based method [5], and three-layered regression with GOP scores achieves the highest correlation (0.75) which is almost equal to the average correlation over teacher pairs (0.77).

Next, we examine the robustness of the speech structure

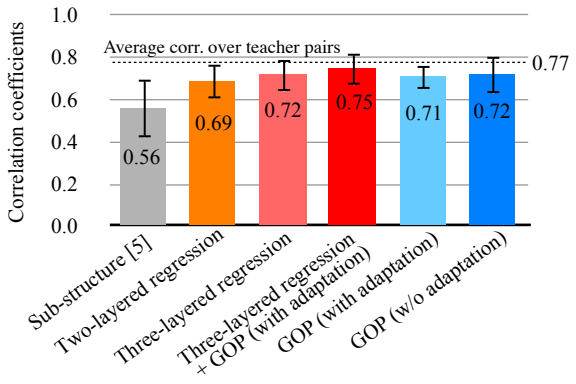

Figure 6: Averages and standard deviations of correlation coefficients between human and machine scores

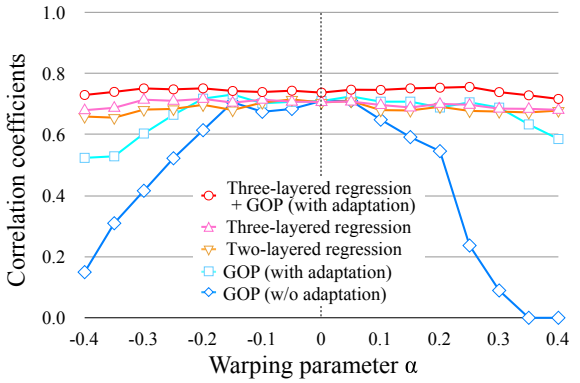

Figure 7: Averages of correlation coefficients between human and machine scores with warped utterances

with respect to the variation in vocal tract length (VTL). Differences in VTL are a major cause of non-linguistic variations, and this difference can be modeled by warping the frequency axis of the power spectrum. Let ω denote angular frequency of a base speaker and $\hat{\omega}$ angular frequency of another (warped) speaker ($0 < \omega, \hat{\omega} \leq \pi$). One popular warping function has the following form,

$$e^{j\hat{\omega}} = \frac{e^{j\omega} - \alpha}{1 - e^{j\omega}\alpha}, \tag{6}$$

where α represents a warping parameter ($-1 < \alpha < 1$). With negative/positive values of α, the VTL is lengthened/shortened. $\alpha = -0.4/+0.4$ approximately doubles/halves the VTL. As it is difficult to collect speech samples with large VTL variations in practice, we artificially generate utterances with various VTLs by applying the above warping function on each utterance in the database using the STRAIGHT morphing technique [14].

The results are shown in Fig. 7. As one can see, three-layered regression with GOP scores obtains the highest correlation for every α. When $|\alpha|$ is large, the correlations of GOP without adaptation drop significantly. The correlations of GOP with adaptation are higher than those of GOP without adaptation, but drop slightly with larger $|\alpha|$. On the other hand, the structure-based methods show high and constant correlations even when $|\alpha|$ is large, and that without adaptation. Especially, the two-layered regression uses only a single teacher's structure for all the cases of α. This indicates that the speech structure is much more robust to changes in VTL.

5. Conclusions

This paper integrated multilayer regression analysis with the structure-based pronunciation assessment technique and proposed an appropriate combination of the structure-based method and the GOP-based method. The experimental results showed that our proposed methods achieved a high correlation coefficient (0.75) on the ERJ database, which is higher than the results of our previous structure-based method and the GOP-based method. The results also showed much higher robustness of the proposed method to changes in VTL compared with the GOP-based method.

6. References

[1] S. M. Witt *et.al.*, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, Vol. 30, pp.95–108, 2000.

[2] D. Luo *et.al.*, "Analysis and utilization of MLLR speaker adaptation technique for learners' pronunciation evaluation," *Proc. INTERSPEECH*, pp.608–611, 2009.

[3] N. Minematsu, "Yet another acoustic representation of speech sounds," *Proc. ICASSP*, pp.585–588, 2004.

[4] N. Minematsu, *et.al.*, "Structural assessment of language learners' pronunciation," *Proc. INTERSPEECH*, pp.210–213, 2007.

[5] M. Suzuki *et.al.*, "Sub-structure-based estimation of pronunciation proficiency and classification of learners," *Proc. ASRU*, pp.574–579, 2009.

[6] Y. Qiao *et.al.*, "A study on invariance of f-divergence and its application to speech recognition," *IEEE Trans. on Signal Processing*, Vol 58, 2010 (to appear).

[7] X. Ma, *et.al.*, "Structural analysis of dialects, sub-dialects, and sub-sub-dialects of Chinese," *Proc. INTERSPEECH*, pp.2219–2222, 2009.

[8] D. Saito, *et.al.*, "Optimal event search using a structural cost function –improvement of structure to speech conversion–," *Proc. INTERSPEECH*, pp.2047–2050, 2009.

[9] M. Pitz *et al.*, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Speech and Audio Processing*, Vol.13, No.5, pp.930–944, 2005.

[10] Y. Qiao, *et.al.* "f-Divergence is a generalized invariant measure between distributions," *Proc. INTERSPEECH*, pp.1349–1452, 2008.

[11] R. H. Heinz, "A model of the regularities underlying speaker variation: evidence from hybrid synthesis," *Proc. INTERSPEECH*, pp.1249–1252, 2006.

[12] N. Minematsu, *et.al.*, "English speech database read by japanese learners for CALL system development," *Proc. Int. Conf. Language Resources and Evaluation*, pp.896–903, 2002.

[13] T. Kitamura, *et.al.* "Speaker individualities in speech spectral envelopes," *Journal of the Acoustic Society of Japan*, Vol.16, No.5, 1995.

[14] H. Kawahara, "STRAIGHT, exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustic Science and Technology*, Vol. 27, No. 6, 2006.