AUTOMATIC CHINESE PRONUNCIATION ERROR DETECTION USING SVM TRAINED WITH STRUCTURAL FEATURES

Tongmu Zhao[1], Akemi Hoshino[2], Masayuki Suzuki[1], Nobuaki Minematsu[1], Keikichi Hirose[1]
[1]University of Tokyo, [2]Toyama National College of Technology

ABSTRACT

Pronunciation errors are often made by learners of a foreign language. To build a Computer-Assisted Language Learning (CALL) system to support them, automatic error detection is essential. In this study, Japanese learners of Chinese are focused on. We investigated in automatic detection of their typical and frequent phoneme production errors. For this aim, four databases are newly created and we propose a detection method using Support Vector Machine (SVM) with structural features. The proposed method is compared to two baseline methods of Goodness Of Pronunciation (GOP) and Likelihood Ratio (LR) under the task of phoneme error detection. Experiments show that the proposed method performs much better than both of the two baseline methods. For example, the false rejection rate is reduced by as much as 82%. However, the results also indicate some drawbacks of using SVM with structural features. In this paper, we discuss merits and demerits of the proposed method and in what kind of real applications it works effectively.

Index Terms— Pronunciation error detection, Chinese, SVM, structural feature, GOP, LR, robustness

1. INTRODUCTION

Pronunciation errors are often made by learners of a foreign language. Especially when the target language contains some phonemes that are not found in learners' native language, learners tend to replace these phonemes with ones existing in their native language. Automatic detection of these errors is an essential and requisite technique in CALL systems [1]. For this task, GOP is extensively studied [2][3] and often used with phoneme-dependent thresholds and, when a difficult phoneme is often replaced with a specific competitive phoneme, LR between the two phonemes is useful [4]. These methods, however, have the well-known mismatch problem, where training speakers of pronunciation models, i.e., teachers, and testing speakers, i.e., students, are mismatched, e.g., adults and kids, the performance readily decreases. To solve this, pronunciation models may be adapted to students, but the adapted models tend to falsely accept wrong pronunciations simply because the models are adapted to students [5].

Recently, a novel structural model of pronunciation was proposed [6], which works effectively to remove the non-linguistic aspects of speech from speech acoustics and keep the linguistic aspects well at the same time. Since the non-linguistic change of speech features is often modeled as feature transformation, the novel model is based on completely transform-invariant features, which is f-divergence [7]. This model has been already applied to overall pronunciation scoring in CALL [8], large vocabulary continuous speech recognition [9], speech synthesis [10], and dialect-based speaker clustering [11]. In these studies, remarkable robustness of our invariant model to speaker differences was experimentally shown. This paper reports our first trial to apply our invariant model to phoneme error detection.

2. MATERIALS

In our study, 5 databases are used for different purposes: Chinese Read by Natives (CRN), Chinese Read by Japanese (CRJ), Chinese Read by Natives with Errors (CRN-E), Chinese generated by a Text-To-Speech (TTS) converter [12] and NICT Chinese database [13].

We created the first three databases by asking speakers to read given sentences. At first, considering phoneme coverage and level of reading content difficulty, 2 paragraphs (17 sentences) were selected from a Chinese textbook [14] as reading material. In the CRN database, 4 Chinese speakers (2 females and 2 males) were asked to read the material, which will be used as teachers' data. In the CRJ database, 7 Japanese learners (3 females and 4 males) read the material. As for Chinese spoken by Japanese, through good discussion with Chinese teachers, 8 phonemes were defined that are the most problematic and difficult phonemes for Japanese learners to pronounce correctly. In this paper, we call these 8 phonemes as target phonemes. Further, for each target phoneme, its competitive one is selected by teachers, i.e., the one which is often substituted by Japanese learners for the target phoneme. If we follow the description in the previous section of how learners substitute phonemes, teachers should select Japanese phonemes as competitive ones. However, for designing the CRN-E database below, we asked teachers to select competitive phonemes out of the Chinese phoneme set. Table 1 shows the 8 phoneme pairs. When a Japanese wants to pronounce /sh/, he may pronounce /x/ instead.

Table 1 Eight target phonemes and their competitive ones

Targets	zh	ch	Sh	v	er	ing	eng	ang
Competitive	j	q	x	u	a	in	en	an

Generally speaking, techniques for pronunciation scoring and error detection should be built using real learners' data that have many pronunciation errors. However, preparation of a non-native speech database with phone-level annotation is a very laborious task for teachers and phoneticians. This often blocks efficient technical development of new methods for error detection. To solve this, in [3], a database including phoneme production errors was prepared by using native utterances. Through changing phoneme-based transcripts of native utterances based on production error characteristics of learners, production errors were artificially simulated in the database. [3] also shows technical validity and effectiveness of this "artificial" preparation. In this study, we prepared speech samples with phoneme errors in a similar way, i.e., modifying the transcripts of the NICT Chinese database. Further, we made another version of artificial data, which was created by asking native speakers to read sentences with intentional errors based on Table 1. In the reading sheet of CRN, 48% of the instances of the 8 target phonemes are replaced by their competitive (confusing) phonemes. 9 native speakers (5 females and 4 males) were asked to read this material. Each speaker read 3 times per sentence. Their speech samples formed a database called CRN-E. For more efficient collection of utterances including phoneme errors, we tentatively tested the use of a commercial Chinese text-to-speech synthesizer [12]. By using it, we can easily obtain utterances of the same original reading sheet with mispronunciations added at different positions in the sheet. This database will be called henceforth as TTS. This tedious recording is difficult to ask human speakers.

The above databases will be used for different purposes in the following way. The NICT Chinese database contains utterances of 200 native speakers from 4 big cities. For our study, Beijing speakers (15 females and 15 males) were used for training native phoneme HMMs (monophones) and the same material was used also to determine GOP thresholds through modification of the transcripts. Out of the 17 sentences in CRJ, 5 sentences (35 utterances) were selected and acoustic realizations of the 8 target phonemes were checked by a Chinese phonetician (the second author). In this paper, both the CRN-E database and the CRJ database are used in testing GOP, LR, and SVM with structural features. As for the TTS database, since we have a male synthesizer and a female one, we ask them to read 10 transcripts with mispronunciations at different positions and different error rates. The resulting TTS database as well as CRN-E is used for SVM model training. The summary of the databases is shown in Table 2. "#M/F" represents the number of male and female speakers. "#U" represents the number of utterances. Usage of each database is also shown in the table.

Table 2 Summary of the 5 databases

database	#M/F	#U	Usage
CRN	2/2	80	Teachers' structural model
CRJ	3/4	35	Training and/or testing samples for the three models of GOP, LR, and SVM with structural features.
CRN-E	5/4	459	
NICT	10/10	5000	Native HMM training
	5/5	2500	GOP thresholds estimation
TTS	1/1	340	SVM model training

3. METHODS

3.1. GOP with thresholds

The GOP score is a well-known pronunciation measurement. It calculates posterior probability of phoneme x given its acoustic observation O, which is approximated by equation (1). Here, Q is the inventory of phonemes. A student's utterance is subjected to both forced alignment and phoneme-loop speech recognition [2].

$$\mathrm{GOP}(x,O) = \mathrm{P}(x|O) \approx \log\left(\frac{\mathrm{P}(O|x)}{\max_{y\in Q}\mathrm{P}(O|y)}\right) \qquad (1)$$

By using correctly pronounced data and incorrect data, distribution of the GOP scores of correct pronunciation and those of errors can be obtained. By observing the two distributions, GOP thresholds for error detection can be obtained. If GOP(x|O) ≥ α, segment O is judged as correct and otherwise not, where α is a threshold often determined dependently on target phonemes [2][3]. Estimation procedures of the thresholds will be explained in section 4.1.

3.2. Likelihood Ratio

As was done in data collection, we supposed that the phoneme-level substitution pattern found in Japanese Chinese is stable and that each target phoneme has its own competitive one. Hereafter, we use x as intended phoneme and y as substituted phoneme. In preparing the CRN-E database, we used the information of y although GOP does not exploit this information. When phoneme confusion is stable, Likelihood Ratio (LR) is useful [4], which uses the information of y as well. An LR score of phoneme x is calculated by taking the absolute difference of the log probability calculated through forced alignment as x and the log probability of forced alignment as y.

$$\mathrm{LR}(x,y,O) = \log\left(\frac{\mathrm{P}(O|x)}{\mathrm{P}(O|y)}\right) \qquad (2)$$

The LR is based on binary classification, determining whether O is more like x or y. Actually, in the case of GOP, if phoneme loop recognition claims that segment O is y, then the LR is basically the same as GOP. In LR error detection,

if the LR score is higher than 0, segment O is judged as correct and otherwise, not.

3.3. Structural features

The GOP and LR scores use only the pronunciation features in the segment of O (absolute features), while structural features are contrastive (relative) features between the segment O and other segments. The process of constructing a speech structure from an input utterance is shown in Fig. 1. An utterance is represented by a sequence of feature vectors. Then, it is converted into a sequence of distributions. This conversion process can be viewed as the training process of an HMM from an utterance. Distance between every distribution pair is calculated as the root of the Bhattacharyya distance. A full set of distances, i.e., distance matrix, is used to represent this utterance [6]. Note that this representation does not keep any information of the spectral shape of the segment O but keeps only how different O is to other segments. In other words, the GOP and LR methods are based on phonetic features of sound substances but the speech structure method is based on phonological features of sound contrasts [15].

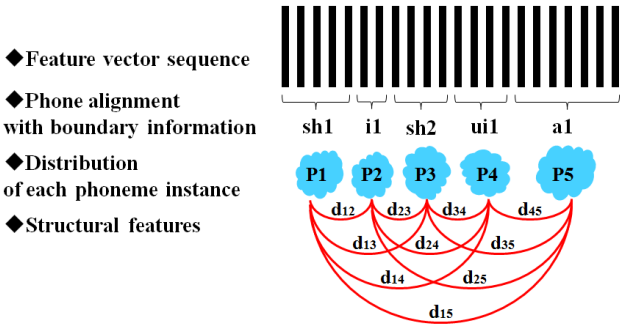

Fig. 1 Extraction of structural features

Suppose that a teacher and a student read the same sentences and both the utterances are converted into two distance matrices, $\{T_{ij}\}$ and $\{S_{ij}\}$. In [8], the structural deviation related to phoneme i is calculated by (3), which quantifies the magnitude of structural difference as for phoneme i between the teacher and the student. Fig. 2 schematically shows the process of calculating DEV(S,T,i), where $\{D_{ij}\}$ is a difference matrix between $\{S_{ij}\}$ and $\{T_{ij}\}$.

$$\mathrm{DEV}(S,T,i) = \sum_{j=1}^{M} D_{ij} = \sum_{j=1}^{M} \left\| \frac{S_{ij}-T_{ij}}{S_{ij}+T_{ij}} \right\| \qquad (3)$$

In (3), M is the number of distributions, which is the number of phoneme instances of the input utterance. Explanation of how to convert an utterance to a distribution sequence will be explained in detail in section 4.3.

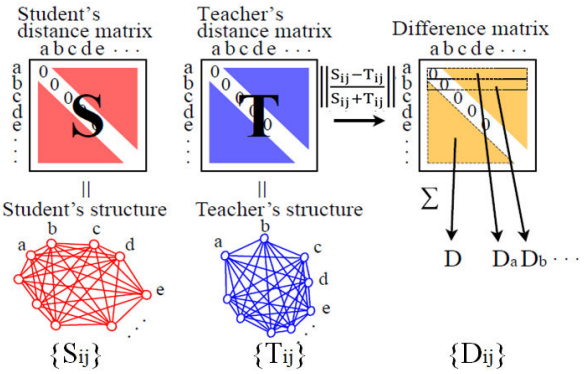

Fig. 2 Structural difference between a student and a teacher

3.4. Support vector machine

Structural features are expected to tell us which phoneme instance is likely to be pronounced incorrectly based on its relations to other phoneme instances in the utterance. One problem is that equation (3) claims that all the elements in a difference matrix contribute with the same importance, but this claim will not be good. Especially when multiple phoneme instances are incorrectly pronounced in a sentence, the distance from phoneme i in $\{S_{ij}\}$ to one of these erroneous phoneme instances will impede the detection performance. One possible solution is to introduce weights and use a regression model. For example, when correct phonemes are labeled as 0 and incorrect phonemes are labeled as 1, these scores can be predicted by the following regression.

$$\mathrm{DEV}_{\mathrm{weight}}(S,T,i) = \sum_{j=1}^{M} w(j) * \left\| \frac{S_{ij}-T_{ij}}{S_{ij}+T_{ij}} \right\| \qquad (4)$$

Based on consideration of this binary classification, we introduce the Support Vector Machine. Let x_i represent a structural difference vector of phoneme i ($\{D_{ij}\}_{j=1,2\ldots M}$), and y_i represent a 1/0 label of x_i, indicating whether phoneme i is correctly pronounced, shown in Fig.3.

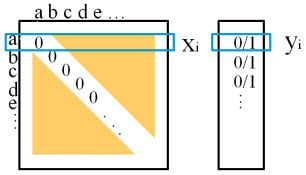

Fig. 3 Adoption of structural features in SVM

Given a training set of instance-label pairs of (x_i,y_i), the SVM is obtained by solving the following problem [16]: x_i is mapped into a hyperplane by function $\emptyset$. b is the bias term of the hyperplane. C(>0) is the penalty parameter of the error term ε_i. W is the weight vector of $\emptyset(x_i)$.

$$\min_{w,b,\varepsilon} \quad \frac{1}{2}W^T W + C\sum_{i=1}^{M}\varepsilon_i$$
$$\text{subject to} \quad y_i(W^T\emptyset(x_i) + b) \geq 1 - \varepsilon_i$$
$$\varepsilon_i \geq 0$$

Here, the linear kernel and the radial basis function kernel (RBF) are considered in the model training process. Generally speaking, the result of the linear kernel is similar to that of linear regression. The RBF kernel can handle the case when the relation between class labels and instances is nonlinear, and it has fewer parameters than other kernels so that it can reduce computational difficulty [16].

$$\text{Linear: } K(x_i, x_j) = x_i^T x_j.$$
$$\text{RBF: } K(x_i, x_j) = \exp\left(-\gamma\left\|x_i - x_j\right\|^2\right), \gamma > 0.$$

3.5. Performance measures

Error detection can produce four types of outcomes [3]: 1) correct acceptance (CA), i.e., the number of correct pronunciations that are judged as correct, 2) correct rejection (CR), the number of mispronunciations that are judged as incorrect, 3) false acceptance (FA), i.e., the number of mispronunciations that are judged as correct and 4) false rejection (FR), i.e., the number of correct pronunciations that are judged as incorrect. Using these four outcomes, False Acceptance Rate (FAR), False Rejection Rate (FRR), and Average Error Rate (AER) are calculated [17] for comparison among GOP, LR, and SVM. FAR = FA / (CR + FA), FRR = FR / (CA + FR), and AER = (FAR + FRR)/2.

4. EXPERIMENTS AND RESULTS

4.1. GOP-based error detection

In the NICT database, artificial pronunciation errors are created by changing the transcript as in [3]. Some instances of the phonemes in the second row of Table 1 are replaced by their target phonemes in the first row. We simulated that the speaker intended to pronounce a target phoneme but actually pronounced its competitive one. Using these data, the GOP scores of correct pronunciations and those of mispronunciations were calculated separately.

In Fig.4, the GOP distribution of /sh/ (correct pronunciation) is drawn in blue, while the GOP distribution of incorrect /sh/ (real pronunciation is /x/) is drawn in red. We set the threshold so as to minimize the classification error. The thresholds of all the target phonemes were obtained from their corresponding distributions, shown in Table 3.

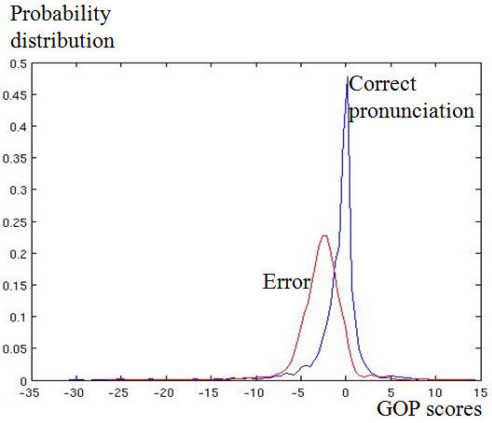

Fig. 4 Probability distribution of /sh/ GOP scores

Table 3 GOP thresholds of the 8 target phonemes

Phonemes	zh	ch	sh	v	er	ing	eng	ang
Thresholds	1.2	2	1.3	5	2.7	5	2.5	2.6

Finally, phoneme error detection is done in the following way. First, the GOP scores of all the individual phonemes in test data (CRN-E) are calculated. Then, if the phoneme is one of the eight target phonemes, its GOP score is compared with its threshold. Table 4 shows our results and the results of another study just as reference although these scores should not be compared directly due to differences of experimental conditions. From the table, our AER is worse than that in [17]. The reason is that, although we have a better result for FRR, we have a much worse result for FA. FAR and FRR have a trade-off relation and two FARs should be compared under the same score of FRR.

Table 4 GOP-based error detection results

	CRN-E	[17]
Language	Mandarin	Mandarin
FAR	0.75	0.42
FRR	0.12	0.24
AER	0.43	0.33

4.2. LR-based error detection

Results of LR-based error detection in the CRN-E database are shown in Table 5. AER of the LR-based error detection improved a lot because FAR is reduced greatly. The LR scores show its high capacity in detecting errors but FRR increases compared with the GOP-based error detection.

Table 5 LR-based error detection results

	GOP	LR
FAR	0.75	0.26
FRR	0.12	0.24
AER	0.43	0.25

LR can be used for fair comparison with SVM using structural features when the CRN-E database in used. This is because both the models are trained based on competitive (confusing) phoneme pairs.

4.3. SVM with structural features

When using SVM with structural features for error detection, firstly, structural features should be extracted. Compared with structural matrix calculation for overall pronunciation scoring [8], there are two different steps. The first one is that data used to extract a distance matrix in this study is only one utterance. The second one is that each phoneme instance should be treated separately although, in [8], the instances of a phoneme are used together to estimate a distribution of that phonemic category. In [8], all the data of a student were used to estimate an N×N distance matrix, where N is the number of the kinds of phonemes. In this study for error detection, however, an M×M distance matrix has to be estimated for an utterance, where M is the number of phoneme instances observed in the utterance.

An input utterance is converted into its distance matrix in the following way. Forced alignment is firstly done using the HMMs trained with the NICT database. Then, using the boundary information, Viterbi training is done to train an HMM only for that utterance. Each utterance of each student and that of each teacher is converted to its HMM and its distance matrix. Here in a distance matrix, element S_{ij} or T_{ij} is a phoneme-to-phoneme distance, defined as the averaged distance among three state-to-state distances calculated as the root of the Bhattacharyya distance.

As for SVM, LIBSVM [18] is used. The CRN-E database is divided into two parts: training and testing. For each sentence, the teachers' matrix of that sentence is obtained as the average matrix among the four teachers. Then, equation (3) is used to calculate the structural deviation of each phoneme instance in each of the students' utterances. Data scaling was done to improve the accuracy. When using the RBF kernel, we used cross-validation and grid search to find the best parameters C and γ, explained in section 3.3.

Then, a leave-one-out cross-validation experiment was done. For a sentence, there are 27 utterances spoken by 9 speakers (pseudo students). We set one speaker as testing speaker and the other speakers as training speakers of SVM. By changing speaker assignment, we used all the speakers as testing speakers. Table 6 shows the results, which are the average performance over the 9 experiments using the linear kernel. The performance of the RBM kernel is very close. We can see that the proposed SVM with structural features works better than the baseline LR-based method. Especially, FRR is decreased by 81.5%. Generally speaking, when the training data size is small, the obtained model tends to be dependent on the extra-linguistic factors found in the training data. Considering a very high performance of SVM using a small number of training speakers, this problem seems to be solved well by using structural features.

Table 6 Comparison of error detection using LR and SVM with structural features

	LR	SVM + structural features	Relative comparison
FAR	0.26	0.21	-20.8%
FRR	0.24	0.04	-81.5%
AER	0.25	0.13	-49.8%

We ran another test to evaluate the robustness of the structure-based SVM experimentally. Here, cross-gender experiments were done. Table 7 shows the results of the two cases where training speakers for SVM were only 3 males and 3 females, respectively. The RBF kernel was used. The testing speakers were of the opposite gender to the training speakers. 3 measures show similar scores between the two cases and these scores are very close to the results of Table 6. This indicates very high robustness of our proposed method.

Table 7 Results of cross-gender tests using CRN-E

Training speakers	3 males	3 females
Optimal parameters	$C=2^{-7}$, $\gamma=2^{22}$	$C=2^{-7}$, $\gamma=2^{22}$
FAR	0.25	0.27
FRR	0.05	0.05
AER	0.15	0.16

4.4. Results using the CRJ database

Error detection experiments using 3 methods were done using the CRJ database. Unlike section 4.3, CRJ cannot be divided into training and testing parts because of the size of the database. Then, we used CRN-E or TTS data to train SVM. Further, as well as the feature vectors used in section 4.3, a structural vector and a GOP score were combined to make a new vector in order to take advantage of both absolute and relative features. Results are shown in Table 8.

Comparing the performance of LR with that of GOP, we find that AER improves a little, but not as much as its improvement when using CRN-E both in training and testing. One possible reason may be that Japanese speakers' pronunciations do not always follow our expectation. More various patterns of substitution may be found.

AER of SVM trained from TTS data is slightly better than that of SVM trained from CRN-E data. This is probably because TTS data contain utterances of more various occurrence patterns of phoneme errors, i.e., different positions and different rates of errors. In the table, it is shown that feature combination certainly improves the performance but its effect is rather minor.

Looking at tables 6, 7, and 8, however, the most remarkable finding about the performance of SVM is that SVM with structural features is very accurate in the condition of artificially prepared utterances of CRN-E and also

very robust to gender differences. However, it is very weak at mismatch between error production patterns between training and testing. This is very natural because structural features are relational features between a segment of interest and other surrounding segments in the utterance. If assumptions on the surrounding segments are invalid, the effect of our proposed method will decrease. These results lead us to consider merits and demerits of our proposed method. If we want to solve the above problem, a sufficiently large training corpus of non-native utterances with correct phone-based annotation is needed. With this kind of database, SVM training will learn which segments in the surrounding contexts are more reliable and will estimate weight vector W in an adequate way. For example, [19] develops a non-native English corpus with IPA annotation, where a fixed paragraph is read by over 1,500 speakers all over the world. The number of speakers is still increasing today. The aim of this project is to analyze and cluster world types of English on an individual basis. We consider that our structural approach can be directly applied to this aim because a structural SVM can be trained for each phoneme in the fixed paragraph using phone-based IPA annotations. We've already started testing our structural model using this corpus.

Table 8 Results of experiments using CRJ database

	GOP	LR	SVM with structural features		SVM with structures and GOP
Training			CRN-E	TTS	TTS
FAR	0.76	0.36	0.67	0.54	0.48
FRR	0.14	0.46	0.26	0.33	0.36
AER	0.45	0.41	0.47	0.44	0.42

5. CONCLUSION

In this paper, the most problematic 8 phonemes for Japanese learners of Chinese were defined and automatic error detection for these phonemes was investigated. For experimental investigation, we designed four new databases: CRN, CRJ, CRN-E and TTS databases. Three methods of error detection were tested using the four databases and the NICT Chinese database. Our proposed SVM with structural features worked much better than both of the GOP and the LR in the CRN-E database. Moreover, structural features turned out to be robust against gender differences. However, the superiority of SVM with structural features over the GOP or the LR can be said to depend on a more complicated training process and need some additional utterances. The GOP and the LR only require native HMMs for error detection. In the SVM with structural features, however, in addition to the native HMMs, teachers' utterances of the target sentences are always needed for structural comparison and learners' incorrect utterances are also needed for SVM training.

Besides, by using the CRJ database, some drawbacks of our proposal were made clear. The SVM with structural features is very robust against acoustic mismatch but still weak at proficiency level mismatch between training and testing. Since this mismatch is due to lack of labels for non-native utterances, this problem can be solved by using a sufficiently large non-native speech corpus with labels. Even under this condition, we showed a concrete example of possible and practical application of our proposed method.

6. REFERENCES

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, 51, 832-844, 2009

[2] S. M. Witt *et al*., "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communications*, 30, 95-108, 2000

[3] S. Kanters *et al*., "The goodness of pronunciation algorithm: a detailed performance study," *Proc. SLaTE*, CD-ROM, 2009

[4] H. Franco, *et al*., "Combination of machine scores for automatic grading of pronunciation quality," *Speech Communications*, 30, 121-130, 2000

[5] D. Luo et al., "Regularized maximum likelihood linear regression adaptation for computer-assisted language learning systems," *IEICE Trans. Inf. & Syst*., E94-D, 2, 308-316, 2011

[6] N. Minematsu *et al*., "Speech structure and its application to robust speech processing," *Journal of New Generation Computing*, 28, 3, 299-319, 2010

[7] Y. Qiao and et al., "A study on invariance of f-divergence and its application to speech recognition," *IEEE Trans. on Signal Processing*, 58, 7, 3884-3890, 2010

[8] M. Suzuki *et al*., "Integration of multilayer regression with structure-based pronunciation assessment," *Proc. INTERSPEECH*, 586-589, 2010

[9] M. Suzuki *et al*., "Discriminative reranking for LVCSR leveraging invariant structure," *Proc. INTERSPEEC*H, 2012 (to appear)

[10] S. Saito *et al*., "Structure to speech conversion speech generation based on infant-like vocal imitation," *Proc. INTERSPEECH*, 1837–1840, 2008

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

[12] HOYA Chinese TTS: http://voicetext.jp/

[13] NICT Chinese database: http://alagin.jp/

[14] *Chinese reading materials: Xingren*, University of Tokyo Faculty of Arts Committee, 2008

[15] R. Jakobson *et al*., *The sound shape of language,* Mouton de Gruyter, 1987

[16] C. W. Hsu et al., *A practical guide to support vector classification*. Tech. rep., Department of Computer Science, National Taiwan University, 2003

[17] Y.B. Wang, "Improved Approaches of Modeling and Detecting Error Patterns with Empirical Analysis for Computer-Aided Pronunciation Training," *Proc. ICASSP*, 5049-5052, 2012

[18] C. Chang et al., LIBSVM: a library for support vector machines, 2001.

[19] W. Steven, *Speech accent archive*, George Mason University, 2012 (http://accent.gmu.edu)