Pronunciation Error Detection Method based on Error Rule Clustering using a Decision Tree

Akinori Ito, Yen-Ling Lim, Motoyuki Suzuki and Shozo Makino

Graduate School of Engineering
Tohoku University, Japan
{aito,yenling,moto,makino}@makino.ecei.tohoku.ac.jp

Abstract

We are developing a CALL system to train English pronunciation for Japanese native speakers. However, the precision of the error detection was not very high because the threshold for the detection was not optimum. To improve the detection accuracy, we propose a new method to optimize the thresholds of error detection. The proposed method makes several clusters of the pronunciation error rules, and the thresholds are determined for each cluster. An experiment was carried out to investigate the performance of the proposed method. As a result, about 90% of detection rate was obtained, which is a remarkable improvement from the conventional method.

1. Introduction

Recently, several CALL (Computer Assisted Language Learning) systems that exploit the spoken language technology have been developed. Automatic assessment of a learner's pronunciation is one of the main issue of such systems. We also have been developed a pronunciation training system for Japanese learners to learn English[1].

While there are several methodology of evaluating learner's speech, the pronunciation evaluation of our system is based on a mispronunciation detection using the multilingual phone models[2]. This method first prepares 'mispronunciation rules'[3] that are the mispronunciations a learner tends to make. When a learner utters a sentence presented by the system, the system prepares the acoustic model of the correct pronunciation of the sentence as well as that of the mispronounced sentence. Then the likelihood values of the both models are calculated using the learner's utterance, and the two likelihood values are compared each other. If the likelihood of the correct model is higher, the learner's utterance is likely to be correct. Otherwise, the utterance must be mispronounced. The error detection method based on the multilingual phone models utilizes the phone model of the target language (L2) as well as the native language of learners (L1). This method is effective because a novice learner tends to utter an L2 phoneme that is absent in the native language as an L1 phoneme.

One problem of the mispronunciation detection method is that the method tends to evaluate the utterance more strictly than human. The detection method compares the similarity between a learner's utterance and the model of the correct sentence with the similarity between the utterance and the model of the sentence with errors. However, even when the utterance is nearer in accuracy to the model with error, it does not necessarily mean that the utterance has mispronunciations because a human judges the pronunciation by referring the linguistic context. Even if the pronunciation of a phoneme is not perfect, it may not sound incorrect. To solve this problem, the strictness of detection of a mispronunciation has to be adjusted so that the system's judgment becomes similar to that by the human.

In this paper, we propose a method to solve this problem by adjusting thresholds of the error detection. This method makes several clusters of mispronunciation rules using a decision tree, and the optimum threshold is determined for each cluster.

2. Pronunciation error detection based on the mispronunciation rules

2.1. Overview of the pronunciation error detection system

First we explain the framework of the pronunciation error detection based on the multilingual phone models. In this work, the target language is English and the native language of the learners is Japanese. Figure 1 shows an overview of the pronunciation error detection. First, the system gives the learner a word or a sentence to pronounce. When the learner utters the word or sentence, the system records the speech and performs the acoustic analysis. Here the input speech is denoted by O. Next, the system prepares the model of the correct pronunciation of the presented word or sentence. The 'model' is composed of connected hidden Markov models (HMMs) each of which models a phoneme. This model of the correct pronunciation is denoted by α. Then the mispronunciation rules are applied to the correct model to generate the models $\bar{\alpha}_1, \ldots, \bar{\alpha}_K$, each of which contains one pronunciation error. An automatic labeling system calculates the likelihood values of the input speech for each pronunciation model, $L(O|\alpha), L(O|\bar{\alpha}_1), \ldots, L(O|\bar{\alpha}_K)$. Finally, $L(O|\alpha)$ and $L(O|\bar{\alpha}_i)$ are compared to determine whether i-th pronunciation error exists in the input speech.

In the multilingual phone models framework, Japanese phone HMMs are used for mispronounced part of the $\bar{\alpha}_i$. The Japanese phonemes are denoted by attaching J after the name of the phoneme in Figure 1.

2.2. Error detection by comparison of the likelihood values

The simplest method to detect the i-th pronunciation error is to compare the values of $L(O|\alpha)$ and $L(O|\bar{\alpha}_i)$ directly. It can be formulate as follows. First, let the likelihood difference $D(O|\alpha, \bar{\alpha}_i)$ be

$$D(O|\alpha, \bar{\alpha}_i) = L(O|\alpha) - L(O|\bar{\alpha}_i), \quad (1)$$

and the system detects the i-th pronunciation error when $D(O|\alpha, \bar{\alpha}_i) < 0$.

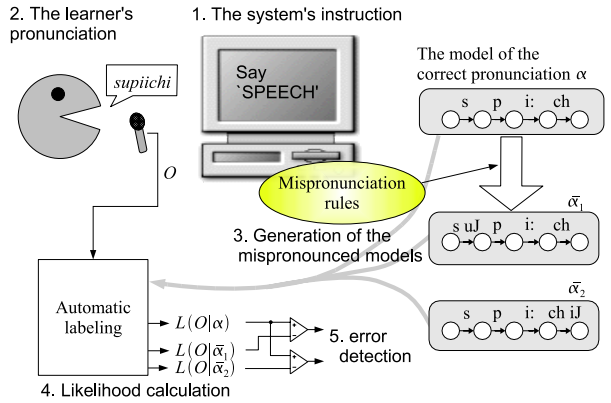

Figure 1: Overview of the pronunciation error detection system

However, this simple method does not show a good consensus with human evaluators. The reason seems to be that a human sense of badness of phoneme sounds depends on the linguistic context the phoneme is pronounced in.

Ohno et.al[4] proposed the method of finding the most suitable threshold value θ from the recognition likelihood difference of the Japanese learner and the native speakers. In their work,

$$D(O|\alpha,\bar{\alpha}_i) = L(O|\alpha) - (L(O|\bar{\alpha}_i) + \theta) \tag{2}$$

was used and θ was optimized to maximize the detection performance. However, as all the mispronunciations are given a common threshold value, it cannot be said that the most suitable determination has been conducted.

To make a pronunciation evaluation that emphasizes on the pronunciation mistakes at the respective places, the pronunciation system, which sets a different threshold value for each different type of mistakes, is proposed.

Let $R(\bar{\alpha}_i)$ be the type of mispronunciation (i.e. the rule used to generate the mispronounced model) of $\bar{\alpha}_i$. Then it is natural to use thresholds that depend on the rule. Here,

$$D(O|\alpha,\bar{\alpha}_i) = L(O|\alpha) - (L(O|\bar{\alpha}_i) + \theta(R(\bar{\alpha}_i))). \tag{3}$$

However, it is difficult to find the optimum threshold for each pronunciation error rule as a huge amount of data is required. Therefore, we cluster the types of mistakes that the learner makes, and give a common threshold for the mispronunciation rules in the cluster. common threshold point is given for each group of mispronunciations.

3. Clustering of the mispronunciation rules

3.1. Optimum threshold for a cluster

Next we explain the algorithm of clustering in the mispronunciation rules. The algorithm is based on a decision tree creation algorithm. This algorithm automatically generates a decision tree from a large amount of speech samples.

Let $O_1,\ldots,O_N$ be the speech samples, $\alpha^{(i)}$ be the model of correct pronunciation of the i-th sample, $\bar{\alpha}_1^{(i)},\ldots,\bar{\alpha}_{K_i}^{(i)}$ be the mispronounced models for i-th sample. Using human-created labels, the j-th mispronunciation candidate of the i-th sample can be determined as 0 (error) or 1 (no error). Now, we can prepare the training data $\left(D(O_i|\alpha^{(i)},\bar{\alpha}_j^{(i)}), R(\bar{\alpha}_j^{(i)}), t_j^{(i)}\right)$ for $i=1,\ldots,N$ and $j=1,\ldots,K_i$ where $t_j^{(i)}$ takes 0 or 1 depending on whether the pronunciation error exists at that position or not. These training data are denoted by $T_1,T_2,\ldots,T_M$ where $T_i=(D_i,R_i,t_i)$ hereafter.

Suppose a set of error rules (a cluster) C. Let us define the following two functions.

$$f_C(R) = \begin{cases} 1 & \text{if } R\in C \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$S(x;\theta) = \begin{cases} 1 & \text{if } x\geq\theta \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The number of samples that concern with the cluster C can be calculated as

$$N(C) = \sum_{i=1}^{M} f_C(R_i). \tag{6}$$

Now, using a threshold θ, we can calculate the following four kinds of numbers of samples.

$$N_{++}(C,\theta) = \sum_{i=1}^{M} f_C(R_i)t_iS(D_i;\theta) \tag{7}$$

$$N_{+-}(C,\theta) = \sum_{i=1}^{M} f_C(R_i)(1-t_i)S(D_i;\theta) \tag{8}$$

$$N_{-+}(C,\theta) = \sum_{i=1}^{M} f_C(R_i)t_i(1-S(D_i;\theta)) \tag{9}$$

$$N_{--}(C,\theta) = \sum_{i=1}^{M} f_C(R_i)(1-t_i)(1-S(D_i;\theta)) \tag{10}$$

$N_{++},N_{+-},N_{-+},N_{--}$ are the number of samples. The relationship of these numbers, human evaluation and system evaluation is as follows.

		human evaluation	
		error	no error
system evaluation	error	N_{--}	N_{+-}
	no error	N_{-+}	N_{++}

Now, let us think of a function $Q(C,\theta)$. This function stands for a 'quality' of the cluster and the threshold. For example, we can use the following quality function

$$Q(C,\theta) = \frac{N_{++}(C,\theta)+N_{--}(C,\theta)}{N(C)}. \tag{11}$$

This function stands for the ratio of correctly classified samples. This function is not an only quality measure for the clustering. Another quality function will be investigated in the later section.

Now, the optimum threshold with respect to the quality function Q can be determined as

$$\hat{\theta}_C = \arg\max_{\theta} Q(C,\theta). \tag{12}$$

In addition, let as define the optimum quality

$$\hat{Q}(C) = \max_{\theta} Q(C,\theta). \tag{13}$$

3.2. Determination of the optimum cluster

The clustering of the rules is performed on a top-down basis. Before performing the clustering, a set of questions about error rules is prepared. In this work, we used two kinds of questions: questions to choose one rule (*"Is R an insertion of /uJ/ after /s/?"*) or questions to choose several rules (*"Is R a replacement of a consonant?"*).

Then the clustering is performed. First, the whole rules make one cluster $C^{(0)}$. Then the cluster is divided into two clusters $C_+^{(1)}(q)$ and $C_-^{(1)}(q)$ according to a question q. Here, $C_+^{(1)}(q)$ is a set of rules that are 'yes' to the question q, and $C_-^{(1)}(q)$ is a set of the remaining rules. Then the qualities of the divided clusters are measured.

$$Q^{(1)}(q) = \frac{N(C_+^{(1)}(q))\hat{Q}(C_+^{(1)}(q)) + N(C_-^{(1)}(q))\hat{Q}(C_-^{(1)}(q))}{N(C_+^{(1)}(q)) + N(C_-^{(1)}(q))} \tag{14}$$

Then the question for the optimum division is determined as

$$\hat{q}^{(1)} = \arg\max_q Q^{(1)}(q) \tag{15}$$

and the optimum clusters are determined as $C_+^{(1)}(\hat{q}^{(1)})$ and $C_-^{(1)}(\hat{q}^{(1)})$.

The division of the cluster is performed recursively, and the division stops when the number of samples in a cluster is less than a pre-defined number or the quality of the divided clusters is not better than the quality of the parent cluster.

3.3. Quality functions

We can think of several different quality functions based on different criterion. The above-mentioned quality function,

$$Q(C,\theta) = \frac{N_{++}(C,\theta) + N_{--}(C,\theta)}{N(C)}, \tag{16}$$

minimizes the number of incorrectly classified samples. We call this function Q_1. Although this function looks good, it may cause a problem when $N_{++} \gg N_{--}$. In this case, the clustering procedure tries to minimize N_{+-}. As N_{-+} and N_{--} are relatively small, there is a tendency that they are ignored, which degrades the detection performance of the mispronunciations.

We also tried another quality function

$$Q(C,\theta) = \frac{N_{++}(C,\theta)}{N_{++}(C,\theta) + N_{+-}(C,\theta)} + \frac{N_{--}(C,\theta)}{N_{-+}(C,\theta) + N_{--}(C,\theta)}. \tag{17}$$

This function equally minimizes the misclassifications of correct pronunciations and incorrect pronunciations. We call this function Q_2. Though this function does not minimize the total number of misclassifications, Q_2 is expected to give a good result for the detection of the mispronunciations.

4. Experiment

We carried out an experiment to investigate the performance of the proposed method. The English speech data of 10 Japanese learners (5 males and 5 females) who read out 45 English sentences[5] was used for the training. A total of 9 native English speakers (4 males, 5 females) were then invited to evaluate their pronunciation. Clustering was then conducted based on the quality functions stated above. The lower limit of the samples in a cluster was set to 20. The test data was 100 sentences spoken by 10 speakers. The speakers of this data is same as that of the training data.

Table 1: Conditions of acoustic analysis

Analysis Conditions	Sampling frequency 16kHz
	16bit quantization
	25ms Hamming window
	Frame period 5ms
Feature Vector	26 dimension vector,
	$pow, MFCC(12), \Delta pow$
	$\Delta MFCC(12)$

Table 2: The mispronunciation rules used in the experiment

Kind of rule	#rules	Example
Insertion of /u/ after a consonant	14	keep /k i: p **uJ**/
Insertion of /o/ after a plosive	2	night /n ai t **oJ**/
Insertion of /i/ after a affricate	2	speech /s p i: ch **iJ**/
Replacement of consonants	7	visit /**bJ** ih z ih t/
Replacement of vowels	6	fact /f **aJ** k t/
Omission of the last /r/	1	star /s t **aJ**/
Mispronunciation from Japanized English words	1	excite /**eJ** k **i** s ai t/

Table 1 shows the conditions of the acoustic analysis, and Table 2 shows the mispronunciation rules. We prepared 40 questions: 33 questions were chosen for one rule, and the other seven was 'insertion', 'insertion of /uJ/', 'insertion of /iJ/', 'insertion of /oJ','replacement', 'replacement of a vowel' and 'replacement of a consonant'. Table 3 shows the number of mispronunciations contained in the test data.

Figure 2 and 3 show the clustering results using Q_1 and Q_2. The names of phonemes followed by 'E' denote that they are English phonemes. The numbers in the leaf nodes stand for the optimum threshold values for the clusters. The result using Q_2 makes relatively flat structure, while the result using Q_1 was finer than that using Q_2.

We measured the three indices, P_C, P_W and P_T. Let us define the following numbers:

Table 3: Distribution of mispronuciations

Insertion	Insertion of uJ	189
	Insertion of oJ	44
	Insertion of iJ	28
Replacement	Replacement of Consonants	156
	Replacement of Vowels	303
Replacement		25
Replacement due to Japanese English		114
Total		859

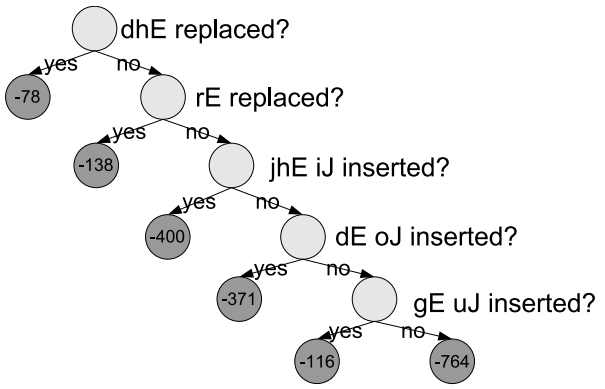

Figure 2: Clustering results based on Q_1

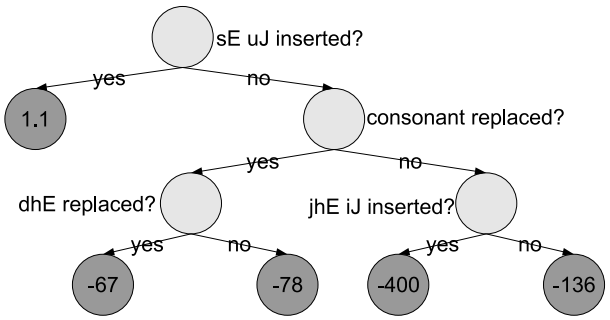

Figure 3: Clustering results based on Q_2

N_{++} Number of samples that are labeled 'correct' and judged by the system as 'correct'

N_{+-} Number of samples that are labeled 'correct' and judged by the system as 'incorrect'

N_{-+} Number of samples that are labeled 'incorrect' and judged by the system as 'correct'

N_{--} Number of samples that are labeled 'incorrect' and judged by the system as 'incorrect'

Then the indices are calculated as

$$P_C = \frac{N_{++}}{N_{++} + N_{+-}} \quad (18)$$

$$P_W = \frac{N_{--}}{N_{-+} + N_{--}} \quad (19)$$

$$P_T = \frac{N_{++} + N_{--}}{N_{++} + N_{+-} + N_{-+} + N_{--}} \quad (20)$$

$$(21)$$

P_C is a ratio of the samples judged as 'correct' among the correct pronunciation data, P_W is a ratio of the samples judged as 'incorrect' among the incorrect pronunciation data, and P_T is a ratio of correctly classified data.

Table 4 shows P_C, Table 5 shows P_W and Table 6 shows P_T obtained by all the 4 methods. From these results, it can be said that the proposed clustering method gave the remarkable improvement in the pronunciation error detection. Comparing the proposed quality functions, Q_1 was better than Q_2 not only the judgment of the correct pronunciation data (P_C) but also the detection of the mispronounced data (P_W).

Table 4: P_C results(%)

	$\theta = 0$	uniform θ	Q_1	Q_2
Insertion	62.3	77.8	100	88.8
Replacement	24.6	51.5	95.1	79.6
Deletion	31.9	42.6	98.0	74.5
Total	43.2	63.7	97.8	83.9

Table 5: P_W results (%)

	$\theta = 0$	uniform θ	Q_1	Q_2
Insertion	50.0	64.1	97.1	85.3
Replacement	32.1	57.7	84.1	71.8
Deletion	25.0	50.0	85.7	100
Total	36.7	59.2	88.2	77.5

Table 6: P_T results (%)

	$\theta = 0$	uniform θ	Q_1	Q_2
Insertion	61.7	76.0	94.9	85.0
Replacement	29.5	50.5	85.8	66.7
Deletion	38.1	43.6	87.5	66.7
Total	44.4	62.4	90.7	79.4

5. Conclusion

A pronunciation error detection method based on pronunciation error clustering is proposed. This method is based on a decision-tree based clustering algorithm. Having different threshold for each cluster, the accuracy of the pronunciation error detection was improved remarkably. We also compared the two quality functions for the clustering, and it was found that the quality function to minimize the total classification error gave the best result.

6. References

[1] M. Suzuki, H. Ogasawara, A. Ito, Y. Ohkawa and S. Makino, "Speaker adaptation method for CALL systems using bilingual speakers' utterances," Proc. ICSLP, vol. IV, pp. 2929-2932, 2004.

[2] G. Kawai, A. Ishida, and K. Hirose, "Detecting and correcting mispronunciation in non-native pronunciation learning using a speech recognizer incorporating bilingual phone models," J. ASJ, vol 57, no 9, 2001.

[3] S. Hiller, E. Rooney, J. Laver and M. Jack, "SPELL: An automated system for computer-aided pronunciation teaching," Speech Communication, vol. 13, pp. 463–473, 1993.

[4] Y. Ohno, M. Mashimo, A. Lee, H. Kawanami, H. Saruwatari and K. Shikano, "A study on pronunciation evaluation for English learner using English acoustic models," Proc. ASJ fall meeting, 1-6-1, pp. 209–210, 2002.

[5] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji and S. Makino, "Development of English speech database read by Japanese to support CALL research," Proc. 18th Int. Cong. Acoust., vol. I, pp. 557–560, 2004.