Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Performance improvement in automatic evaluation system of English pronunciation by using various normalization methods

Masaru Kusumi, Masaharu Kato, Tetsuo Kosaka and Itaru Matsunaga

Graduate school of science and engineering, Yamagata University, 4-3-16 Jonan Yonezawa-city Yamagata, Japan

PACS: 43.72 Ne

ABSTRACT

We investigate the performance improvement in an automatic evaluation system of English pronunciation uttered by Japanese learners. In this system, Japanese and English acoustic models are used to detect mispronunciation of a phoneme level. We use hidden Markov models (HMMs) as acoustic models. English and Japanese HMMs are trained by using speech data uttered by native English and Japanese speakers, respectively. Mispronunciation is detected by comparing output likelihoods of the two models. In order to improve the performance of this system, we investigate the following points: (1) Reduction in an acoustic mismatch. Because of the use of speaker-independent acoustic models, a mismatch in speaker characteristics arises between an input speech and acoustic models. In addition, the mismatch between recording environments must be considered. Therefore, we attempt to reduce the acoustic mismatch by using cepstral mean normalization (CMN) and histogram equalization (HEQ) methods. (2) Analyses of the effectiveness of pronunciation error rules. In order to detect the pronunciation errors in a phonetic level, the system uses pronunciation error rules. We compare some error rules to clarify which rules are effective in evaluating pronunciation. In order to evaluate the proposed methods, we investigated the correlation between an objective evaluation value returned by the system and the subjective evaluation value given by English experts. We used the English Read by Japanese (ERJ) speech corpus as evaluation data. In this corpus, each utterance was given a score on the basis of a five-grade evaluation made by the experts. We use the score as the subjective evaluation value. The experimental results showed that the combination of CMN and HEQ was most effective. From the results of comparison of error rules, four error rules were found to be particularly effective: vowel insertion at the end of a word, vowel substitution, vowel insertion between consonants, and consonant substitution.

INTRODUCTION

We develop an automatic evaluation system of English pronunciation uttered by Japanese learners. Until now, various researches have been conducted on automatic pronunciation evaluation. Kawai et al. have proposed the method of detecting a pronunciation error by using a speech recognition technique in which acoustic models for two languages are used [1]: One model is for non-native speakers; the other, for native speakers. In a pronunciation evaluation system of English uttered by Japanese, the former is the acoustic model trained by Japanese speakers and the latter is the model trained by American speakers. In this system, a mismatch arises between an input and an acoustic model because of speaker characteristics, the acoustic difference between recording environments, and so on. The use of speaker adaptation can be considered to solve this problem. However, if speaker adaptation is carried out by using inaccurate English pronunciation, the English acoustic model will express inaccurate pronunciation, and will hence adversely affect pronunciation evaluation. In order to overcome this disadvantage, a speaker adaptation method employing two language acoustic models that use bilingual speech data has been proposed by Ogasawara et al. [2]. However, although the speech of a bilingual speaker who can pronounce both languages correctly is useful for speaker adaptation, it is difficult to obtain it.

This research examines normalization methods that do not adversely affect pronunciation evaluation. As normalization methods, cepstral mean normalization (CMN) and histogram equalization (HEQ) are employed. These methods are widely used in speech recognition. HEQ is a technique often used in image processing. Recently, it has been applied to speech processing as a feature normalization technique [3] [4], and an improvement has been achieved in the performance of speech recognition under noisy conditions. In this study, we attempt to improve the performance by reducing the difference of distribution between various acoustic models by using the above normalization methods. Moreover, by comparing CMN and HEQ we investigate which normalization is appropriate for pronunciation evaluation.

Pronunciation errors are frequently detected in conventional systems, even in the case of native speakers. In order to avoid this problem, a weighting method is also employed. In the weighting method, errors can be reduced by adding low weights to Japanese phonemes. In this case, although excessive weighting may have a bad influence on the system, we demonstrate that the use of above-mentioned normalization methods can reduce the influence. Moreover, in this evaluation system, pronunciation error rules are used to detect the errors in a phoneme level. In this study, error rules of eight categories are used. In the evaluation experiments, we compare which error rules are effective for the system.

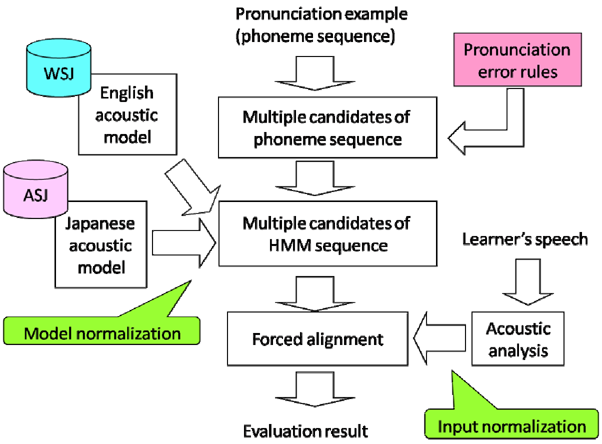

Figure 1. Block diagram of automatic pronunciation evaluation system

AUTOMATIC PRONUNCIATION EVALUATION METHOD

Overview

This section describes the method and system of automatic English pronunciation evaluation. An overview of the system is given as follows. First, the system displays an English sentence to be pronounced, and a learner speaks the sentence. Next, the system shows an evaluation score and points out pronunciation errors in a phoneme level. In order to build the system, a phonetic alignment of the utterances made by the learners is required. The alignment is performed by using an HMM-based Viterbi algorithm in which both English and Japanese models are used. In order to detect errors in a phoneme level, pronunciation error rules are used in the forced alignment procedure.

A block diagram of the system is shown in Fig. 1, briefly describing the procedures of the system. First, the system displays an English sentence to be pronounced. When the learner utters a sentence, the system performs acoustic analysis in the acoustic analysis module where the utterance is analysed to obtain feature vectors. Next, the displayed sentence is automatically translated into a phoneme sequence. This sequence contains phonemes for both correct and incorrect pronunciations. The incorrect phoneme sequence is generated from mispronunciation rules. The rules represent mispronunciations that non-native learners can make. Next, the phoneme sequence is converted to an HMM sequence. The HMM sequence is used for a process called "forced alignment." This process is used to find the best assignment of feature vectors, which are derived from the speech analysis module, with HMM states by using the Viterbi algorithm. In the alignment procedure, the single best state path is determined by selecting either the Japanese or English HMMs. By following the above procedures, the English utterance made by a Japanese learner can be automatically evaluated at a phoneme level. In this study, we focus on the following topics. We make an effort to reduce the acoustic mismatch between the Japanese and English models by using various normalization methods and to reduce error detection rates by using the weighting method. In addition, we study the effectiveness of pronunciation error rules in detail. For the evaluation system, we provide eight categories of error rules. We compare these categories in order to clarity which of them are effective for automatic evaluation.

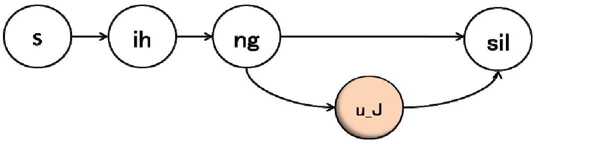

Figure 2. Error in vowel insertion at the end of the word "sing"

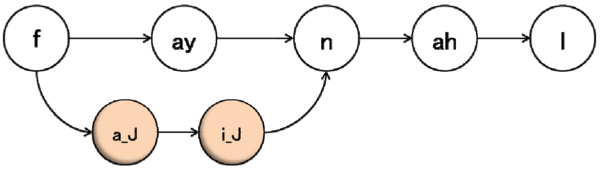

Figure 3. Diphthong substitution error in the word "final"

Error rules

In the proposed system, pronunciation error rules are used to detect mispronunciation by Japanese learners.

These rules are categorized into eight groups. In the case of insertion errors, a italic type shows an insertion of a Japanese phoneme in the following descriptions. In the case of substitution errors, the italic type shows a replacement of a Japanese phoneme with an English one. In the case of omission errors, () indicates an English consonant that may be omitted.

- Vowel insertion (at the end of a word)
  The rule in which a Japanese vowel is inserted after an English consonant at the end of a word.
  Example: sing (s ih ng *u*)
- Vowel substitution
  The rule in which an English vowel is replaced with a Japanese vowel.
  Example: the (dh *ah* → dh *a*)
- Vowel insertion (between consonants)
  The rule in which a Japanese vowel is inserted between English consonants.
  Example: study (s *u* t ah d iy)
- Consonant substitution
  The rule in which an English consonant is replaced with a Japanese consonant.
  Example: child (ch ay *l* d → ch ay *r* d)
- Consonant omission (from the end of a word)
  The rule in which the English consonant /r/ after a vowel drops out at the end of a word.
  Example: far (f aa (*r*))
- Vowel substitution (diphthong)
  The rules in which the English diphthong /ay/, /aw/, or /oy/ is replaced with a Japanese vowel.
  Example: final (f *ay* n ah l → f *a i* n ah l)
- Consonant insertion (loanword)
  The rule in which a Japanese phoneme is inserted in loanwords borrowed by the 19th century.
  Example: extra (eh k *i*|*u* s t r ah)
- Consonant omission (from the beginning of a word)
  The rule in which /w/ or /y/ is omitted from the beginning of a word.
  Example: would ((*w*) uh d)

The examples of the error rules for vowel insertion at the end of a word and diphthong substitution are illustrated in Fig. 2 and Fig. 3, respectively. In these figures, ○ indicates a phoneme model. A phoneme symbol added with "_J" at the end represents a Japanese phoneme. In this system, when a Japanese phoneme is detected, it is judged as a pronunciation error.

NORMALIZATION METHODS WITH LIKELIHOOD WEIGHTING

The above-mentioned evaluation system faces a problem that many pronunciation errors are detected even if an evaluation speaker is a native speaker. Regarding this problem, the system can reduce the errors by adding a low weight to the output likelihood of a Japanese phoneme. However, the likelihood difference between the Japanese and English models may decrease due to the weighting, and the performance of pronunciation evaluation may deteriorate. In order to solve this problem, we propose a combination of a normalization method and the weighting method.

Normalization methods

In order to reduce the mismatch between acoustic models, the difference in speaker characteristics or recording environments needs to be adapted without adapting the difference of acoustic characteristics between different languages. However, it is difficult to extract only the difference in speaker characteristics or recording environments. Therefore, it is assumed that cestrum distributions of each speaker's speech are similar over different languages. In fact, it was observed that the difference in recording environments or speaker characteristics was larger than the difference in languages. Therefore, we investigated the relation between the Japanese and English acoustic models by measuring the Bhattacharyya distances (B distances) between them. The B distances before normalization are listed in the upper part of Table 1. In the table, WSJ indicates the English acoustic model trained by using the Wall Street Journal (WSJ) database. ASJ indicates the Japanese acoustic model trained by the Japanese corpus called the Acoustical Society of Japan Japanese Newspaper Article Sentences (ASJ-JNAS). J-E denotes the English acoustic model based on speeches uttered by Japanese students. This model was trained by the English Read by Japanese Students (ERJ) corpus [8]. E-E expresses the English acoustic model based on speeches uttered by Americans. This model was trained by ERJ. From the results without normalization, the E-E model is closer to the Japanese model (ASJ) than the American English model (WSJ) is. The same can be said for the J-E model. Accordingly, it turns out that the distances between the models are strongly affected by the difference between the databases, rather than that between the languages. In order to visualize the distances between the models, the models are plotted by the COSMOS method [6]. In this method, the distribution of the acoustic models is plotted in a two-dimensional diagram by means of multidimensional liner measurement. The B distance between two models was used to calculate the similarity of the probability distributions of the models. The results without normalization are shown in Fig. 4. Each point of a scatter plot represents a speaker. From the results of Fig. 4, it is observed that the distributions of WSJ and ASJ are greatly separated, but the distributions of J-E and E-E are close to the distribution of ASJ despite the fact that they are English models.

In the HEQ method, a transform function is calculated directly from the histograms of both training and test data, and the method can compensate for nonlinear effects. The transform function *HEQ()* is given by,

$$o_t^{'} = HEQ(o_t) = C_T^{-1}(C_E(o_t)), \quad (1)$$

where C_E and C_T denote the CDFs estimated form the test data and training data, respectively. The distributions of acoustic models obtained by applying CMN and HEQ are shown in Fig. 5. The distances between the acoustic models after normalization are shown in the lower part of Table 1.

Table 1. Bhattacharyya distances between acoustic models

Before normalization	WSJ	ASJ	J-E	E-E
WSJ	0.0	24.21	12.11	25.80
ASJ		0.0	2.86	3.21
J-E			0.0	5.60
E-E				0.0
After normalization	**WSJ**	**ASJ**	**J-E**	**E-E**
WSJ	0.0	1.70	1.66	1.33
ASJ		0.0	0.29	0.19
J-E			0.0	0.34
E-E				0.0

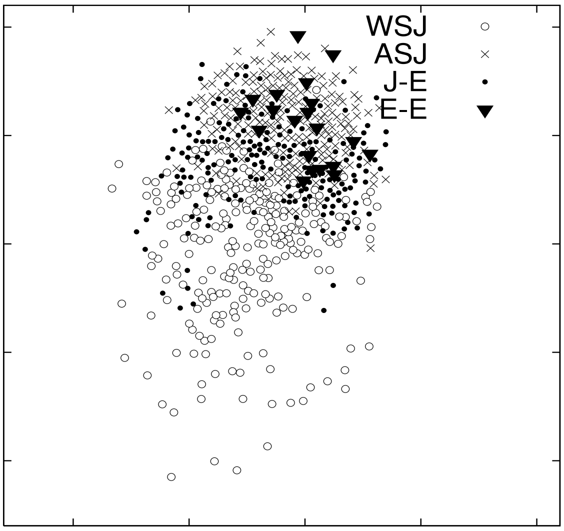

Figure 4. Distribution of acoustic models trained by various databases without normalization

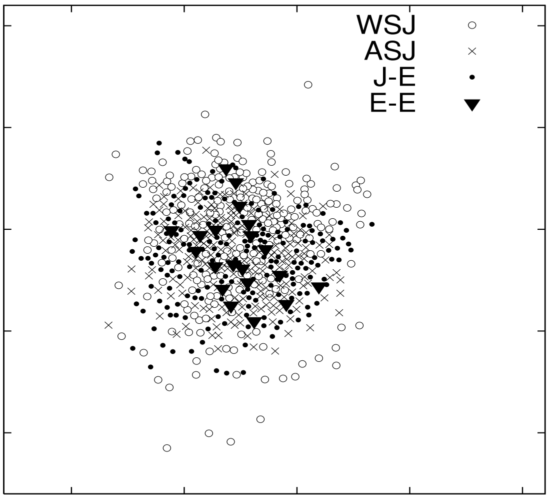

Figure 5. Distribution of acoustic models trained by various databases normalized with HEQ and CMN

The figure indicates that both the distributions of ASJ and ERJ overlap with WSJ. Since the difference in recording environments could be reduced by CMN and HEQ, each data set was considered to have a distribution similar to WSJ. From the above results, we found that the mismatch between acoustic models could be substantially reduced by the normalization methods. However, the influence of the normalization methods on the difference between the languages is unclear. In order to clarify the influence, we study the effect of CMN and HEQ on automatic pronunciation evaluation.

Weighting method

In our previous study, we faced a problem that many pronunciation errors were detected even in the case of a native speaker. Therefore, in this study, a weighting method is used to reduce such errors. Weighting of the output likelihood is performed as follows. $b_i(o_t)$ denotes the output probability in state S_i. Weighting is carried out by calculating $\lambda b_i(o_t)$, where λ represents a weight for a Japanese phoneme and is set to be less than 1.0. However, the evaluation performance may degrade because the likelihood difference between the English and Japanese models decreases on assigning a weight. We investigate whether the degradation can be suppressed by combining the weighting method with the normalization methods.

EXPERIMENTAL CONDITIONS

Training and evaluation data

In the speech analysis module, a speech signal is digitized at a sampling frequency of 16 kHz with a quantization size of 16 bits. The length of the analysis frame is 32 ms, and the frame period is set to be 8 ms. A 13-dimensional feature (12-dimensional MFCC and log power) is derived from the digitized samples for each frame. Further, the delta and the delta-delta features are calculated from the MFCC feature and the log power. Then, the total number of dimensions is 39. For training of English models, 69,094 sentences uttered by 238 American speakers (119 males and 119 females) in the WSJ corpus are used. For training of Japanese models, 31,511 sentences uttered by 204 Japanese speakers (102 males and 102 females) in the ASJ-JNAS corpus are used. For evaluation data, we used 1,900 English sentences uttered by 190 Japanese speakers (95 males and 95 females), 3,215 English sentences uttered by 8 English teachers (4 males and 4s female), and 4,827 English sentences uttered by 12 general Americans (4 males and 8 females). Two types of acoustic models, English and Japanese, are used in the system. Each monophone HMM consists of three states and 16 mixture components per state. The phonemes are listed in Table 2. The English phonemes were determined by referring a CMU phoneme dictionary [7].

Evaluation of system performance

On the basis of five-grade evaluation made by an English teacher, a score is given to each learner's utterance. The score is considered as a subjective evaluation value. The performance evaluation of the system is conducted by comparing the subjective evaluation with pronunciation error rates detected by the system. In general, performance evaluation should be conducted by using subjective evaluation of phoneme errors. However, it is difficult to conduct subjective evaluation at a phoneme level. Hence, subjective evaluation is conducted at a sentence level. Since evaluation values are assigned by four English teachers to each sentence, the average of those values is calculated and used as a sentence evaluation value. In order to investigate the accuracy of the subjective evaluation values, the correlation of values between teachers is calculated. Ten sentences uttered by each of the 190 Japanese speakers are evaluated and assigned subjective evaluation values. Table 3 lists the correlations between each particular combination of English teachers. R1–R4 are the IDs of the four English teachers. The results show that a high average correlation of 0.797 was obtained. This implies that the subjective values assigned by the English teachers are reliable.

Table 2. Phoneme lists

34 Japanese phonemes	a i u e o aa ii uu ee oo ei ou w y xy r h f z j s sh ch ts p t k b d g m n N cl
39 English phonemes	aa ae ah ao ih iy uh uw ey eh er aw ay ow oy ch l m n ng b d dh f g hh p r s sh jh k t th v w y z zh

Table 3. Correlation of subjective evaluation values between English teachers

Teacher combination	Correlation
R1/R2	0.783
R1/R3	0.845
R1/R4	0.766
R2/R3	0.796
R2/R4	0.841
R3/R4	0.750
Average	0.797

RESULTS AND DISCUSSIONS

In order to evaluate the accuracy of the proposed automatic pronunciation evaluation system, the error detection rates given by the system and the subjective evaluation values assigned by the English teachers were compared. For the comparison, average values of the error rates and those of the evaluation values are calculated for each set of 50 sentences. The system performance can be evaluated from the correlation between the subjective evaluation values and the error detection rates. In the experiments, CMN and HEQ were used as normalization methods. CMN was applied to every sentence in the training data and evaluation data. For applying HEQ, histograms were derived from each of the databases (ERJ, ASJ-JNAS, and WSJ). The evaluation data in ERJ and the training data in ASJ-JNAS were normalized to bring them closer to the training data in WSJ.

Table 4 shows the correlation between the error rates determined by the system and the subjective evaluation values assigned by the English teachers in various normalization methods. Since the number of phonemes corresponding to "consonant insertion (loanword)" is insufficient (only 28 phonemes in 1900 sentences), correlation coefficients are not computed. From the results, both CMN and HEQ were found to be effective. In particular, the combination of CMN and HEQ (CMN + HEQ) achieved the best performance. On the other hand, the weighting method causes a lower performance with or without normalization. However, performance degradation can be suppressed by using the normalization methods. As for "vowel insertion (at the end of a word)," "vowel insertion (between consonants)," and "consonant substitution," the system performance is high and the correlation coefficients are –0.861, –0.721, and –0.822, respectively. These coefficients are close to the correlation value of the subjective evaluation (0.797) performed by the teachers shown in Table 3. Thus, we can conclude that these error rules are notably effective in evaluating pronunciation. Conventional systems have a problem that their error detection rate is high even for a native speaker. The weighting method is used in order to solve this problem. The results of the weighting method, obtained by using CMN + HEQ, are listed in Table 5. The numbers shown in the column "Japanese speaker" or "American speaker" are the error rates for Japanese phonemes, and it is found that the rates are high even if the speaker is an American. Relative difference shows the percentage of the difference in the error rates between Japanese and American speakers. It can be said that the system performance is high if the difference is large. From the results, it turns out that error rates decline in the case of both

Table 4. Correlation between error rates returned by the system and the subjective evaluation values assigned by English teachers in various normalization methods

Normalization	w/o normalization		CMN		HEQ		CMN+HEQ	
Weighting	no	yes	no	yes	no	yes	no	yes
Vowel insertion (at the end of a word)	–0.791	–0.553	–0.779	–0.557	–0.812	–0.819	–0.802	–0.861
Vowel substitution	–0.058	0.134	–0.448	–0.281	–0.239	–0.059	–0.375	–0.335
Vowel insertion (between consonants)	–0.667	–0.363	–0.646	–0.468	–0.713	–0.651	–0.804	–0.721
Consonant substitution	–0.787	–0.713	–0.846	–0.702	–0.830	–0.791	–0.857	–0.822
Consonant omission (at the end of a word)	–0.462	–0.299	–0.579	–0.305	–0.539	–0.253	–0.595	–0.348
Vowel substitution (diphthong)	0.322	0.552	0.241	0.274	0.150	0.198	0.174	0.209
Consonant omission (at the beginning of a word)	–0.206	–0.114	0.047	–0.029	–0.150	–0.148	–0.291	–0.167

Table 5. Error detection rates obtained by using CMN + HEQ and relative difference rates between Japanese and English speakers (%)

	With weighting			Without weighting		
	Japanese speaker	American speaker	Relative difference	Japanese speaker	American speaker	Relative difference
Vowel insertion (at the end of a word)	37.37	12.22	67.31	15.30	3.16	79.37
Vowel substitution	53.50	22.86	57.27	15.06	3.44	77.16
Vowel insertion (between consonants)	38.56	12.86	66.66	14.20	2.94	79.32
Consonant substitution	64.63	15.41	76.15	35.30	3.55	89.93
Consonant omission (at the end of a word)	63.56	43.57	31.45	26.53	18.17	31.53
Vowel substitution (diphthong)	47.49	34.62	27.10	8.99	5.50	38.84
Consonant insertion (loanword)	60.71	26.74	55.96	28.57	9.63	66.31
Consonant omission (at the beginning of a word)	24.23	19.08	21.25	7.97	5.77	27.63

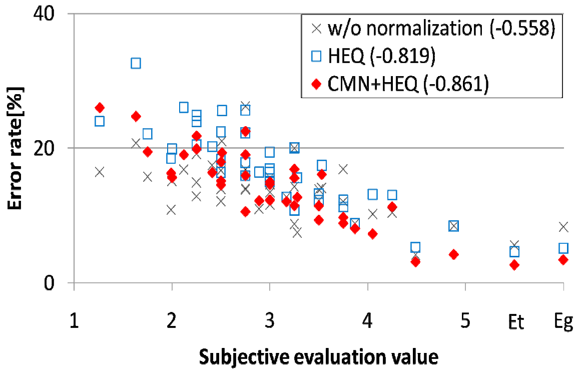

Figure 6. Relation between subjective evaluation value and error detection rate for vowel insertion (at the end of a word)

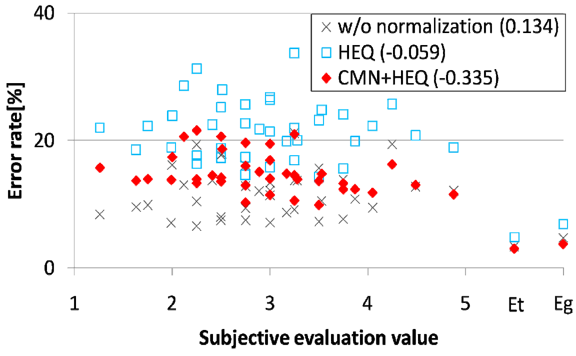

Figure 7. Relation between subjective evaluation value and error detection rate for vowel substitution

Japanese and American speakers on using the weighting method. Furthermore, it is found that the relative differences become large. As described above, it can be concluded that the weighting method is effective. In “vowel insertion (at the end of a word),” “vowel substitution,” “vowel insertion (between consonants),” and “consonant substitution,” the relative difference is large and the error detection rates of American speakers are low (less than 5%). Thus, we can say that these rules are notably effective. On the other hand, the error detection rates of “consonant omission (from the end of a word)” for English speakers are high. Therefore, this rule is considered to be unsuitable for pronunciation evaluation. From the results given in Table 4, it is observed that the weighting method causes performance degradation. However, by considering the effect of a decline in the error detection rate, it can be concluded that the combination of CMN + HEQ and the weighting method is the most effective approach.

The relations between the subjective evaluation values and the error detection rates are shown in Fig. 6 and Fig. 7. Fig. 6 shows the results of “vowel insertion (at the end of a word)”, and Fig. 7 shows those of “vowel substitution.” In these figures, the results of “without normalization” (×), HEQ (□), and CMN + HEQ (◆) are shown. Furthermore, the weighting values for Japanese phonemes are indicated. If there is a high negative correlation between the two axes, the system performance is high. For reference, the error rates of 8 English teachers (Et) and 12 general Americans (Eg) are also indicated. From the results of Fig. 6, a significant improvement can be achieved by using CMN + HEQ. In addition, the error rates of the English teachers and the general Americans can be reduced by the normalization methods. On the other hand, regarding “vowel substitution,” the system performance is very low without normalization or with HEQ. Although the correlation increases on using CMN + HEQ, the absolute value of the correlation is still low (–0.335). However, the difference of error rate between Japanese students and native English speakers is not small. Thus, although this rule cannot be used for automatic pronunciation evaluation of beginners or learners at the intermediate level, there is a possibility that it can be used for upper-grade learners.

CONCLUSIONS

In this study, we proposed a new pronunciation evaluation system using normalization methods. In order to reduce the mismatch of speaker characteristics or the acoustic difference between recording environments, CMN and HEQ were used. In addition, a weighting method was applied to the output likelihood of Japanese phoneme in order to avoid the problem that pronunciation errors were detected even for a native speaker. From the comparison of normalization methods, both CMN and HEQ were found to be effective, and the combination of CMN and HEQ exhibited the best performance. In addition, the weighting method was effective in reducing the error detection rate. By using the rule “vowel insertion (at the end of a word),” “vowel insertion (between consonants),” or “consonant substitution,” better performance could be obtained than that achieved using other error rules. Finally, we conclude that the best performance can be achieved by using CMN + HEQ with the weighting method. We plan to conduct detailed analysis of error rules for making a further improvement in pronunciation evaluation.

REFERENCES

1 G. Kawai and K. Hirose, “A Method for Measuring the Intelligibility and Nonnativeness of Phone Quality in Foreign Language Pronunciation Training”, Proc. of ICSLP98, vol. 5, pp.782-785 (1998).

2 M. Suzuki, H. Ogasawara, A. Ito, Y. Ohkawa and S. Makino, “Speaker Adaptation Method for CALL System Using Bilingual Speakers' Utterances”, Proc. of ICSLP2004, vol. 4, pp.2929-2932 (2004)

3 A.Torre and J.C.Segura, “Non-linear transformations of the feature space for robust speech recognition”, Proc. of ICASSP 2002, pp.401-404 (2002).

4 Y. Obuchi, “Delta cepstral mean normalization for robust speech recognition”, Advanced Research Laboratory, Proc of ICA2004, pp.2587-2590 (2004).

5 I. Matsunaga, M. Katoh and T. Kosaka, “Improvement of an automatic pronunciation evaluation of English by using a histogram equalization”, The ASJ spring meeting, 1-R-11, pp.405-408 (2009) (in Japanese).

6 M. Shozakai and G. Nagino, “Two-dimensional Visualization of Acoustic Space by Multidimensional Scaling”, Proc. of ICSLP2004, vol. 1, pp.717-720 (2004).

7 The CMU Pronouncing Dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict?in=C+M+U+Dictionary

8 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. of ICA2004, pp.557-560 (2004)

9 O. Viikki and K. Laurila, “Cepstral domain segmental feature vector normalization for noise robust speech recognition.” Speech Communication, vol. 25, pp.133-147 (1998).