Chinese language pronunciation proficiency estimation based on structural features*

Tongmu Zhao, Masayuki Suzuki, Chengshuo Wang, Nobuaki Minematsu, Keikichi Hirose
(The University of Tokyo)

1 Introduction

Assessment of pronunciation aims at evaluating the linguistic and para-linguistic aspects of speech [1]. The acoustic features in speech are, however, affected and changed easily by the non-linguistic aspect of speech such as age and gender. This fact makes the assessment task harder. The current framework of acoustic modeling of speech basically models speech sounds with their non-linguistic features included. To solve the problem, model adaptation or feature normalization is often done but it does not always give a good solution [2].

Recently, a novel model, speech structure, was proposed, which works efficiently to discard the non-linguistic features and only keeps the linguistic and para-linguistic information [3]. Besides, this structure model has been applied to speech recognition [4], speech synthesis [5], pronunciation assessment [6], and dialect-based speaker clustering [7].

In the structure-based pronunciation assessment, however, the excessively high dimension of a pronunciation structure often degrades the performance. To reduce the dimension, selection of the edges in a pronunciation structure was examined [8] and multilayer regression analysis was applied to give weights to the individual edges of a pronunciation structure [2].

This paper introduces the structure-based pronunciation assessment to Japanese learners of Chinese for the first time. Here, prior phonetic knowledge of the difference between Chinese and Japanese is used for edge selection. Size normalization and frequency effects are also considered. Results show that higher performance is obtained than the original structure-based assessment.

2 Data preparation

Based on the phone coverage and difficulty, 3 levels of materials were selected. The high level contains 8 paragraphs from HSK test. HSK is the People's Republic of China's only standardized test of Modern Standard Chinese language proficiency for non-native speakers. The middle level materials contain 8 paragraphs from a middle level textbook [9]. The low level materials contain 60 sentences and 2 paragraphs, all of which are from a beginning level textbook [10]. Table 1 shows the used materials in different levels, and the number of sentences (S) and words (W).

Table 1 Materials in different levels

Level	Materials	S	W
High	8 paragraphs from HSK	69	2042
Middle	8 paragraphs from a middle level textbook [9]	61	1771
Low	60 sentences and 2 paragraphs from a beginning level textbook [10]	78	1139

Students were grouped into 3 levels. Advanced students read the high and middle level materials. Middle students read only the middle materials. Beginning students read the low materials and paragraphs 2&8 in the middle. Table 2 shows the number of students in different levels for each gender. We're still continuing the recording and this paper describes the experimental results obtained so far using the data of Table 2.

Table 2 Students in different levels

	Reading Materials	M	F
Advanced	High and Middle	2	1
Middle	Middle	3	1
Beginning	Low and Middle 2&8	2	1
Native	High and Middle	1	3

In recording, transcripts of all the materials were visually presented to students. Here, tone

*構造表象に基づく中国語発音評価，趙 童牧，鈴木 雅之，王 程碩，峯松 信明, 広瀬 啓吉(東京大学)

information was discarded. Figure 1 shows an example of transcription.

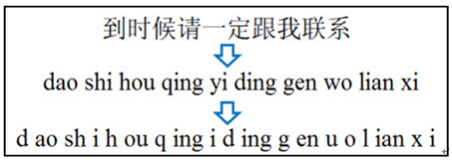

Fig. 1 An example of transcription.

Finally, human assessment was done to be used as reference to train an assessment machine. In this research, 3 Chinese native speakers listened to the utterances, and rated every Japanese student based on his/her pronunciation.

3 Structure assessment

The process of obtaining a pronunciation structure was described in detail in [2]. When two sets of utterances of the same sentence set, one is from a teacher and the other is from a student, are represented as two phoneme-based structures (distance matrices), the structure difference between them is calculated by (1) and used to assess that student.

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

In (1), S and T are two distance matrices, whose elements are calculated as root of Bhattacharyya distance (BD). BD is one kind of f-divergence.

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

M is the total number of distributions, which can be phonemes or states. In (1), all the edges in a matrix are utilized to get an assessment score. Figure 2 illustrates the entire process of the structure-based pronunciation assessment.

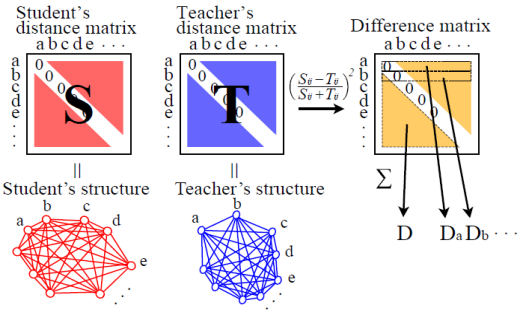

Fig. 2 Process of structure-based assessment

4 Experiments and results

4.1 Structure-based analysis

Through the conditions described in Table 3, speaker-dependent monophone HMMs were trained from all the Middle level materials. After training the HMMs of each speaker, his/her phoneme-based pronunciation structure was calculated, in which a distance between two phonemes was obtained as the average of three BDes between the corresponding three states.

Table 3 Conditions for acoustic analysis

Sampling	16bit/16kHz
Windows	25ms length and 10ms shift
Parameters	MFCC (13dim.)
HMMs	speaker-dependent
Topology	5 states and 3 distributions
Monophones	58 in total

The human scores used in all the experiments were obtained by averaging assessment scores of 3 native speakers. Except for the experiment aimed at comparing results of using different teachers, all the experiments used the average structure over all the 4 native speakers as teacher structure. Evaluation of the proposed methods was done by investigating the correlation between human and machine assessments.

4.2 Edge selection based on phonetic differences between Chinese and Japanese

In Chinese Pinyin, initials contain stops, nasals, and so on. Finals contain single vowels, double vowels and the combination of vowel and others.

Fig. 3 All the phonemes of Chinese

Compared with Japanese, some phonemes are found only in Chinese (red circles in Figure 3). These phonemes may cause special difficulty to Japanese learners of Chinese. So, these kinds of

phoneme pairs (edges) were selected for automatic assessment.

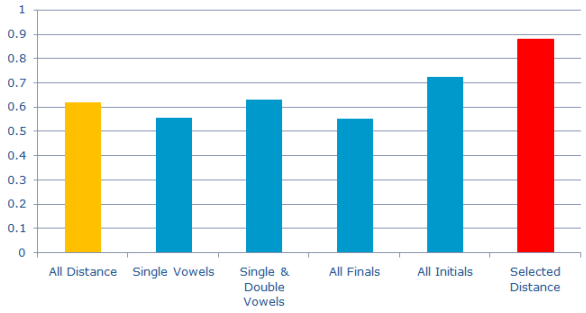

Fig. 4 Results of experiments

4.3 Results and Discussion

Figure 4 shows the correlations between human and machine assessments. In machine assessment, several conditions were tested. "All Distance" means that all the 58 phonemes were used. "Single Vowels" only used single vowels. "Single & Double Vowels" contained single and double vowels. "Finals" and "Initials" used all finals and all initials, respectively. "Selected phone pairs" contained f-h, l-r, j-q, x-s, c-s, v-i, ou-o, ie-ve, van-ian, er-e, un-vn, iii-i, ii-i, iii-ii.

Table 4 Number of phone in different groups

Phonetic Grouping	All Distance	Single Vowels	Single & Double Vowels	Finals	Initials	Selected phone pairs
Number of Phones	58	6	16	37	23	14

Table 4 shows the number of phonemes in the six conditions. From Figure 4 and Table 4, we can see that the correlations were not totally influenced by the number of phonemes used. Although the Selected phone pairs only contain 14 phones, it works the best among all the methods examined. It experimentally proves that the phones existing only in Chinese are essential to the assessment of foreign speakers.

4.4 Size normalization

The size of the pronunciation structure of a student tends to be larger when he/she speaks louder with larger articulation efforts. This variation should be canceled for assessment because it is not related to pronunciation proficiency. For this aim, size normalization is applied. Figure 5 shows the correlation with/without size normalization. It's clear that through size normalization, the correlation scores get improved.

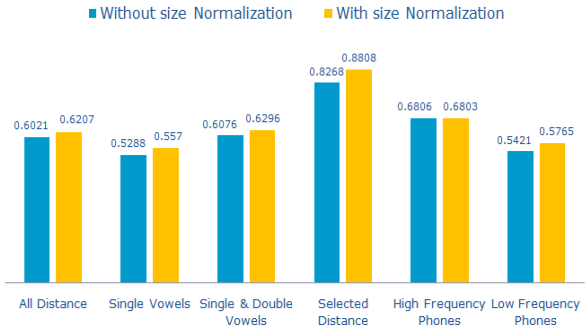

Fig. 5 Results with and without normalization

4.5 Phonemes with different frequency

The number of occurrences heavily depends on kinds of phonemes. The minimum frequency was found to be 3 and the maximum was 172 in the training set. It is expected that the HMMs with a smaller number of training data will be unreliable to be used for assessment.

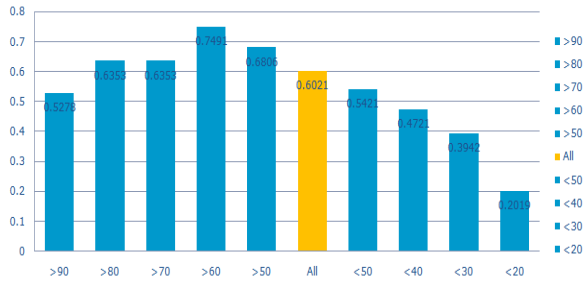

Fig. 6 Human-machine correlations as a function of phoneme frequency

Figure 6 shows the results of assessment using the phonemes with different frequencies. For example, ">90" means that the phonemes which were found more than 90 times were used for assessment. "<50" means that the phonemes which appeared less than 50 times were used. We can see the correlations in the left side (high frequency) are better than that in the right side (low frequency). Each bin of Figure 6 has a different number of phonemes. Then, we carried out other experiments using the same number of phonemes with different frequencies. Figure 7 shows the results of using a fixed number of phonemes in two cases, the most frequent phonemes and the

least frequent ones. Here, “9 phonemes” means that the most frequent 9 phonemes. Clearly shown in the figure, phonemes with higher frequencies work much better than those with lower frequencies (almost twice in the correlation).

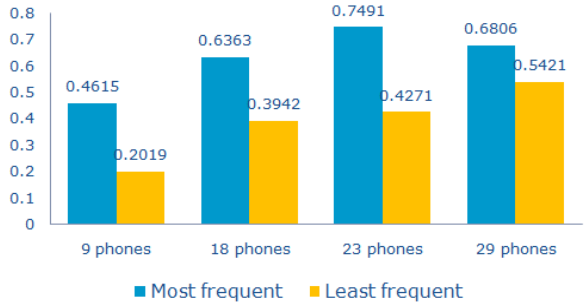

Fig. 7 Results of using phonemes with different frequencies

4.6 Structures of different teachers

In this experiment, we recorded 4 Chinese native speakers (1 male and 3 female speakers) as teacher data. So, we conducted experiments using the averaged teacher structure and four teacher structures each corresponding to each Chinese.

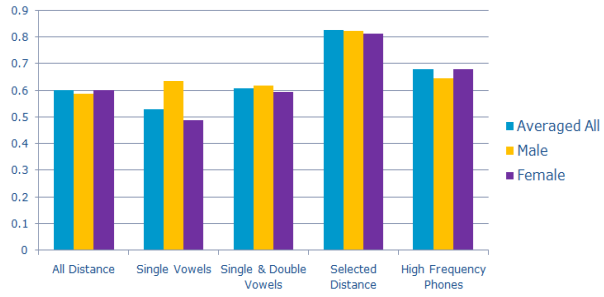

Fig. 8 Assessment with different teachers

The result shows that there is no big difference among 3 kinds of settings. Generally speaking, the result proves that the structure model works efficiently to discard speaker difference and only keeps phonetic information.

5 Conclusion

This paper discussed utilizing structure models to assess pronunciation quality of non-native speech at the phone level. From the results, it was proved that not all edges in a structure matrix are with the equal importance for assessment. Based on the comparison between Chinese and Japanese in their characteristics of phonetic level, selected edges achieved the best correlation with human scores. Size normalization was shown to be useful to improve assessment performance. Furthermore, when doing assessment with different settings of teacher structures, the result proves that the structure model works efficiently to discard speaker difference and only keeps phonetic information.

Acknowledgements

Recording is a very time-consuming task. Here, I need to show special thanks to the third author, who recorded data of all the non-native speakers.

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] M. Suzuki et.al., “Integration of Multilayer Regression Analysis with Structure-based Pronunciation Assessment,” Proc. INTERSPEECH, pp.586–589, 2010.

[3] N. Minematsu, “Mathematical evidence of the acoustic universal structure in speech,” ICASSP, pp. 889-892, 2005

[4] Y. Qiao, “f-divergence is a generalized invariant measure between distributions,” INTERSPEECH, 1349-1352, 2008

[5] S. Saito et al., “Structure to speech conversion speech generation based on infant-like vocal imitation,” INTERSPEECH, 1837–1840, 2008

[6] N. Minematsu et al., “Speech Structure and Its Application to Robust Speech Processing,” New Generation Computing, 28, 299-319, 2010

[7] N. Minematsu, “Training of pronunciation as learning of the sound system embedded in the target language,” in Proc. Int. Symposium on Phonetic Frontiers, 2008

[8] M. Suzuki et al., “Sub-structure-based estimation of pronunciation proficiency and classification of learners,” Proc. ASRU, pp.574-577, 2009

[9] 基礎漢語，駒場中国語教育研究会，2010

[10] 中国語講読教材「行人」，東京大学教養学部中国語部会，2008