Automatic Mandarin Pronunciation Scoring for Native Learners with Dialect Accent

Si Wei, Qing-Sheng Liu, Yu Hu, Ren-Hua Wang

Department of Speech and Hearing
iFlytek Speech Lab, University of Science and Technology of China, Hefei
{tonyw, ustc_qsliu}@ustc.edu, yuhu@iflytek.com, rhw@ustc.edu.cn

Abstract

This paper studies pronunciation scoring algorithm in CALL system aiming at teaching native Chinese learn standard Mandarin. Most of the pronunciation scoring algorithms focus on non-native environment, which may not be suitable for native speakers. We bring up a new algorithm based on traditional posterior log-likelihood algorithm by weighting the initial part of Mandarin syllables, where final-initial's duration ratio is introduced to control the weight. Experiments show that the proposed algorithm is much more effective than traditional posterior log-likelihood algorithm in the Mandarin learning system. The correlation with human score achieves an increase of 11%.

Index Terms: CALL, duration ratio, Mandarin pronunciation scoring

1. Introduction

Computer Assisted Language Learning (CALL) systems can provide many potential benefits for both the language learner and teacher [1, 2, 3]. They can point out errors made by students and give corrective advice without teacher's instruction. To be efficient, CALL system should evaluate the pronunciation and judge whether a segment is error. The aim of this work is to obtain an effective method to evaluate the pronunciation of native students under the framework of hidden Markov model (HMM) speech recognition.

Existing work on automatic pronunciation scoring mainly focuses on L2 language learning such as the system (VILTS) developed by SRI [4, 5] and the system developed by S.M.Witt and S.J. Youg [1,6].

This work is part of an effort aiming at developing computer assisted system for Chinese to learn standard Mandarin. That is to evaluate the pronunciation quality of dialect Chinese. As pointed by other studies [7], the quality of pronunciation consists of many factors such as Segmental Quality (SQ), Speech Rate (SR), Fluency (FL) and Overall Pronunciation (OP). Speech Rate and Fluency is not a problem for native speakers. The main problems of our learners are the influence of dialect. As introduced in the papers [4, 5], the basic algorithm for pronunciation scoring is HMM-based phone log-posterior probability [1, 4, 5, 6]. This algorithm utilizes posterior log-likelihood based 'Goodness of Pronunciation' (GOP) [6] measure to generate machine score.

When using GOP algorithm as evaluation method in our system, we find the result is not as good as published in the papers. A new method is brought up to improve the performance. Our evaluation system uses HMM-based continuous speech recognition system (built via HTK) [8] to generate phonetic segmentations. Based on the segmentations and the log-likelihood produced by the speech recognition system, machine score based on the GOP algorithm is obtained. Based on this score, we bring up a new algorithm to weight the initial part of syllables where the final-initial duration ratio is used to control the weight. At the same time, we use an adjustable factor to control the initial weight, which can be tuned on the developing set. Effectiveness of the new algorithm is evaluated via the machine score's correlation with human score on a dialect Mandarin Database. Results indicate the new algorithm is better than the GOP algorithm.

The rest of this paper is organized as follows. Section 2 presents the basic pronunciation evaluation algorithm. In section 3, the characteristic of Mandarin initial and final is investigated. In section 4, the new algorithm is brought up. Section 5 describes experiments based on the new algorithm. Then conclusions are shown in section 6.

2. Computer Assisted Pronunciation Evaluation Algorithm

Various pronunciation scoring algorithms were developed to evaluate the pronunciation qualities. A brief view of the existing algorithm is necessary for further study. Two kinds of algorithm are investigated below: Hidden Markov Models (HMM) log-likelihood based score and segment duration based score [1, 4, 5, 6, 7]. The details about these algorithms are described below.

2.1. HMM log-likelihood based scores

One of these algorithms [4] uses HMM log-likelihood as machine score. The underlying assumption is the log-likelihood of speech data is a good measure of

pronunciation quality. For each sentence, the phone segmentation is obtained along with the corresponding log-likelihood of each segment. However, the log-likelihood depends on the length of the sentence. In order to remove the effect of the sentence length, an average log-likelihood score is defined as follows:

$$G = \frac{\sum_{i=1}^{N} l_i}{\sum_{i=1}^{N} d_i} \quad (1)$$

Where l_i denotes the log-likelihood of ith phone and N_i denotes the duration of ith phone.

This log-likelihood score correlates badly with human score (about 0.33 at sentence level) [4]. Word posterior log-likelihood score is introduced to improve the correlation [2, 4, 5, 6]. Word posterior log-likelihood score is obtained by the following process. Under the assumption that word posterior log-likelihood score represent the pronunciation quality, the quality of phone p is defined to be the duration normalized posterior probability $P(p|O^{(p)})$, where p is the orthographic transcription and $O^{(p)}$ is the spoken data. Using Bayes decision theory, we can get (2).

$$\begin{aligned} G &= (\sum_{i=1}^{N} |\log(P(p_i \mid O^{(p_i)}))| / NF(p_i)) / N \\ &= (\sum_{i=1}^{N} |\log(\frac{P(O^{(p_i)} \mid p_i)P(p_i)}{\sum_{q \in Q^{(p_i)}} P(O^{(p_i)} \mid q)P(q)})| / NF(p_i)) / N \\ &\approx (\sum_{i=1}^{N} |\log(\frac{P(O^{(p_i)} \mid p_i)P(p_i)}{\max_{q \in Q^{(p_i)}} P(O^{(p_i)} \mid q)P(q)})| / NF(p_i)) / N \end{aligned} \quad (2)$$

where $Q^{(p_i)}$ is the phone set which phone p will be misread as. $NF(p_i)$ denotes the number of frames in the segment $O(p_i)$. In order to compute word posterior log-likelihood score it is assumed that the orthographic transcription p is known to determine the likelihood $p(O^{(p)} \mid p)$ of the acoustic segment $O^{(p)}$ corresponding to each phone p.

The correlation between human and machine rises from 0.33 to 0.58 at sentence level and from 0.50 to 0.88 at speaker level [4,5] after using this posterior probability (we call this GOP introduced by [6]) as machine score.

2.2. Duration score

As described in existing systems, duration score is used to evaluate the quality of pronunciation. [4,5,7]. The procedure to compute the phone-based duration score is as follows: first, from the Viterbi alignment we measure the duration in frame for the i-th segment, then its value is normalized to compensate for rate of speech. To obtain the corresponding phone segment duration score, the log-probability of the normalized duration is computed using a discrete distribution of duration for corresponding phone. The discrete duration distributions have been previously trained from alignments generated from the standard Mandarin training database. The corresponding sentence duration score is defined as the average of the phone segment scores over the sentence.

Duration score is efficient for non-native pronunciation evaluation especially for text-independent task. For native speakers, there is no problem to speak fluently. Duration score is not as valuable as for non-native speakers. We should investigate the duration characteristic of Mandarin Chinese and decide how to bring duration's affection into the evaluation method.

3. Characteristic of Mandarin Initial and Final

Mandarin is quite different from European languages. It is a tonal and monosyllabic language with about 1200 tonal syllables. If disregarding the lexical tones, there are 410 basic toneless syllables. The structure of Mandarin syllables consists of initial plus final or final alone. There are 21 initials and 37 finals in Mandarin. Initials and finals are the smallest natural pronunciation units in Mandarin. The main differences between Chinese dialect and standard Mandarin are the pronunciation of initials. At the same time, initials are much shorter and more changeable than finals. Initial's errors are main problem for most of the native speakers. An experiment is carried out to prove this. The experiment uses an ASR system to recognize the accented speech database in a restricted way and acquire the recognition error rate of phonemes. The recognition model is trained from the national Chinese recognition database using HTK.

The result is shown in figure 1 and 2, the phonemes whose error rate is below 1% are neglected. From figure 1 and 2, initial's error rates are much higher than finals' except a few exceptions. That is to say initials are more important than finals in the pronunciation evaluation system. Another investigation on Mandarin pronunciation is the initial-final's duration ratio, which is $\frac{Dur_{final}}{Dur_{initial}}$. An experiment is done to investigate this. The result is shown in figure 3. From figure 3, we find that syllables with low or high initial-final's duration ratio are more difficult to recognize than syllables with medium duration ratio.

These two experiments are the foundation of the new algorithm. We find that attention should be paid to initials and the final-initial's duration rate.

4. New Evaluation Algorithm Based on Chinese Characteristic

As described in preceding sections, GOP is the most efficient algorithm in text-dependent environment.

Section 3 indicates that initials tend to be more misread than finals. That means initials should have more weight than finals. At the same time, the larger the final-initial's duration ratio, the more frequently misreading happens. The original GOP algorithm for Chinese is as (3).

$$G_{sent} = (\sum_{i=1}^{N} G_i)/N$$
$$G_i = G^i_{initial} + G^i_{final} \quad (3)$$

Where N is the phone num in the sentence, G_i is the GOP score of i-th syllable. $G^i_{initial}$ is the GOP score of initial of the i-th syllable and G^i_{final} is the GOP score of final.

Based on the conclusions of section 3, the new algorithm improves the weight of initials according to the final-initial's duration ratio as (4).

$$G_{sent} = (\sum_{i=1}^{N} G_i)/N$$
$$G_i = \begin{cases} G^i_{initial} \times (1 + \frac{Dur^i_{final}}{Dur^i_{initial}} \times COFF) + G^i_{final} & if \quad Dur^i_{final} > Dur^i_{initial} \\ G^i_{initial} \times (1 + \frac{Dur^i_{initial}}{Dur^i_{final}} \times COFF) + G^i_{final} & if \quad Dur^i_{final} < Dur^i_{initial} \end{cases} \quad (4)$$

Where Dur^i_{final} is duration of i-th syllable's final and $Dur^i_{initial}$ is duration of i-th syllable's initial. $COFF$ is an adjustable factor which can be tuned on the developing set.

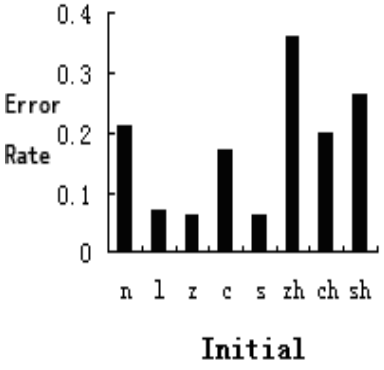

Figure 1 *Average error rate of ChongQing accented Mandarin initials.*

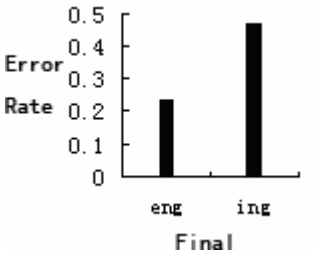

Figure 2 *Average error rate of ChongQing accented Mandarin finals.*

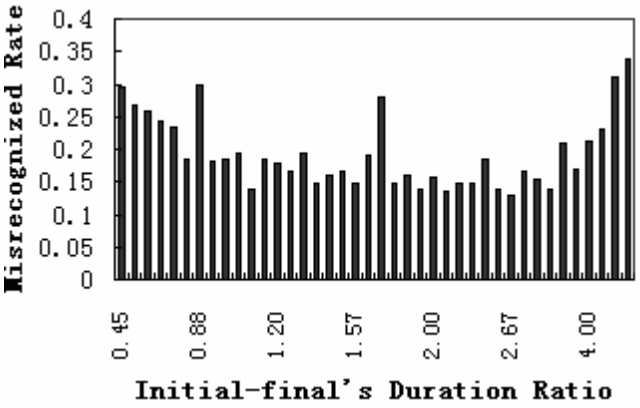

Figure 3 *Misrecognized syllables' rate with different initial-final's duration ratio*

5. Experiments Based on the New Algorithm

In preceding section, new algorithm is described in detail. In this section, experiments based on the new algorithm are carried out to prove the effectiveness of the new algorithm.

5.1. Database and human evaluation

The Database used in this paper is a speech database recorded for ChongQing accented speech recognition. There are 22 males and 23 females in the database and each of them reads 110 phonetic balanced sentences. The spoken intelligence of the speakers varies from very high intelligent to quiet accented. All the speech was recorded in quiet offices using close talking headed-microphone. Five human raters are selected for the human evaluation process. 20 sentences per speaker are taken out and split into two sets. Each set contains 12 sentences and there are four same sentences between the two sets. Raters evaluate each set at different time. They give each speaker two speaker-level scores for two sets and 24 sentence-level scores for 24 sentences [5]. Human evaluation's correlations are shown in table 1.

Corr. Type	Level	Rater ID					Avg.
		1	2	3	4	5	
Inter	Sent	0.73	0.73	0.62	0.68	0.70	0.69
Inter	Spkr	0.82	0.82	0.76	0.78	0.80	0.80
Intra	Sent	0.88	0.83	0.53	0.75	0.71	0.74
Intra	Spkr	0.92	0.88	0.64	0.85	0.78	0.81

Table 1: *Human evaluation's sentence-level and speaker-level correlations. Inter correlations are correlation between two raters. Intra correlations are correlation between two scores of one rater at different time. "Sent"*

means the correlations at sentence-level. "Spkr" means the correlations at speaker-level.

These human evaluation's correlations are the up-boundary for scoring algorithm. The correlation between human score and machine score indicates the performance of scoring algorithm.

5.2. Experiments based on the new algorithm

Experiment is carried out on the database described in section 5.1. The new algorithm is evaluated by computing the correlation between machine score and human score at sentence and speaker level.

Figure 4 and 5 give the correlations at sentence and speaker level on the developing set. Changing with the *COFF* factor in (4), the correlation goes up first and then falls. The peak is the best point to evaluate the pronunciation. Figure 4 and 5 shows the changing process.

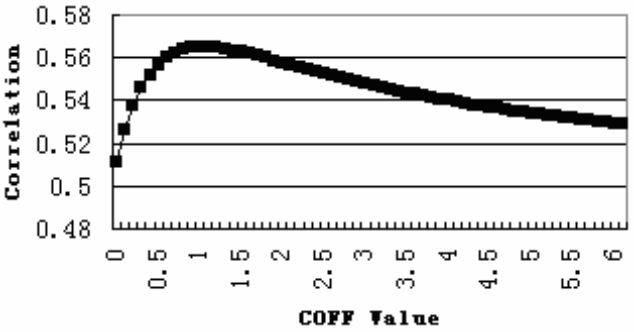

Figure 4 *The sentence-level correlation curve when COFF factor changes from 0.0 to 6.0.*

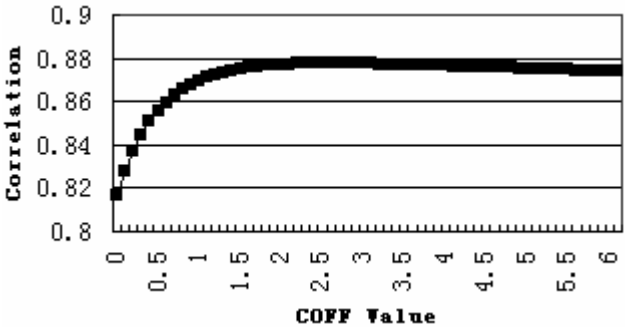

Figure 5 *The speaker-level correlation curve when COFF factor changes from 0.0 to 6.0.*

When the COFF factor changes, the sentence-level correlation goes up and then falls down. When COFF equals 1.0, the correlation is 0.565, which is much higher than the original GOP algorithm (The new algorithm is exactly same as GOP algorithm when COFF equals 0.0). The speaker-level correlation goes from 0.816 to 0.870.

Table 2 gives the correlation on testing set using the best COFF got on the developing set (COFF=1.0).

Algorithm	Correlation	
	sentence	speaker
GOP	0.490	0.790
NEW	0.542	0.855

Table 2 *GOP and new algorithm's correlation with human on testing set*

Table 2 shows the new algorithm with tuned COFF factor gets much better correlation with human score than GOP algorithm. This proves the new algorithm is efficient.

6. Conclusions

A new algorithm for pronunciation scoring by enhancing the weight of initials according to the initial-final's duration ratio is presented. Compared with the GOP algorithm, the performance of the new algorithm is much better. At the sentence-level, the new algorithm gets 11% higher correlation than the posterior algorithm, which is proved to be most efficient in text-dependent environment. At the speaker-level, it also shows better performance. But compared with human correlations, the sentence-level correlation is still low. This is the direction of future work.

7. References

[1] Silke Maren Witt, "Use of Speech Recognition in Computer-assisted Language Learning", PhD's thesis, November 1999.

[2] S.M.Witt and S.Young, "Language Learning based on Non-native Speech Recognition", Proc. EUROSPEECH, Rhodes, Greece, pp.633--636, 1997.

[3] C.H.JO, "Studies on Computer-Assisted Pronunciation Learning System for Non-native Learners based on Speech Recognition Techniques", PhD's thesis, 1999.

[4] H.Franco, L.Neumeyer, Y.Kim and O.Ronen, "Automatic Pronunciation Scoring for Language Instruction.", Proc. ICASSP, pp. 1471-1474, 1997.

[5] L.Neumeyer, H.Franco, M.Weintraub and P.Price, "Automatic Text-independent Pronunciation Scoring of Foreign Language Student Speech", Proc. ICSLP 96, Philadelphia, pp.1457-1460, 1996.

[6] S.M. Witt, S.J.Young, "Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning", Speech Communication, pp.95-108, 2000.

[7] C.Cucchiarini, H.Strik, L.Boves, "Automatic Evaluation Of Dutch Pronunciation by Using Speech Recognition Technology", IEEE workshop ASRU, pp. 622–629, 1997.

[8] S.Young, D.Kershaw, J.Odell, D.Ollason, V.Valthev, "The HTK Book (for HTK Version 3.0)", Microsoft Corporation, July 2000.