GENERALIZED SEGMENT POSTERIOR PROBABILITY FOR AUTOMATIC MANDARIN PRONUNCIATION EVALUATION

Jing Zheng [1,2], Chao Huang [1], Min Chu [1], Frank K. Soong [1], Wei-ping Ye [2]*

[1] Microsoft Research Asia, Beijing, China

[2] School of Information Science and Technology, Beijing Normal University, China

{chaoh, minchu, frankkps}@microsoft.com, hattiezheng@gmail.com, bnuywp@yahoo.com

ABSTRACT

In this paper, we investigate the automatic pronunciation evaluation method for native Mandarin. Multi-space distribution (MSD) Hidden Markov Model (HMM) is adopted to train the gold standard model. Machine scores derived from the generalized segment posterior probability on both syllables and phone level are proposed and investigated to measure the Goodness of Pronunciation (GOP). They are evaluated on the database collected internally and shown better performance than other well-known methods. In addition, detailed analyses of human scoring such as inter/intra-rater on utterance/speaker level are also given.

Index Terms— pronunciation evaluation, posterior probability, goodness of pronunciation (GOP)

1. INTRODUCTION

Computer assisted language learning (CALL) has become more and more popular with the integration of the ASR technology. In a CALL system, it is critical to get the real feedback on pronunciation quality of the speaker. In other word, pronunciation evaluation plays a key role here.

Most of the studies focused on the second language learning and evaluation. It also should be concerned on the native speakers' pronunciation quality. The proficiency test of Mandarin which is called Putonghua Shuiping Ceshi (PSC) is an official standard to evaluate a person's Mandarin level. A certificate of PSC is required for many vocations in China, such as teacher, announcer and official. It is a time-consuming and costly work to organize experts to score speakers' pronunciation. Therefore, several attempts have been made to do the evaluation automatically.

All kinds of scores based on statistic modeling, e.g. Hidden Markov Models (HMM) are proposed as the Goodness of Pronunciation (GOP) scores which show comparatively correlation with human scores [1][2]. Among them, the log-likelihood score and recognition accuracy are the most direct way derived from the recognizer to measure the pronunciation quality, as we will review and study in the paper. Rate of speech (ROS), reflecting the fluency of the speaker, is pointed out by some researchers [1] [7] as a good indicator for non-native speakers. However, for native speaker fluency is not an issue and ROS is not investigated.

Posterior probability is a more elegant measure since it is less affected by spectral changes due to the particular speaker characteristics or channel variations and more focused on the pronunciations quality and therefore is thought as one of the most promising indicator. Based on it, we propose a refined GOP based on the generalized segment posterior probability (GSPP) and apply it into both syllable level and phone level to do the evaluation.

This paper is organized as follows: several main GOP measures are present in Section 2. The database configuration is described in Section 3. The evaluation results and analysis are given in Section 4. The paper concludes with a discussion of the results and the comments on future work.

2. AUTOMATIC SCORING METHODS

Before discussing the proposed generalized segment posterior probability based GOP, we will review measures in Section 2.1 and 2.2 that have been used by other researchers and they will be also investigated on our database.

2.1. Log-likelihood based scoring

In this method, using HMM, the log-likelihoods of spectral observations extracted from short-time windows of speech are used as scores. For each utterance the phone segment is obtained along with the corresponding log-likelihood. The "global average log-likelihood" score G, is defined as:

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

[*] Join the work as an intern at Microsoft Research Asia

where l_i is the log-likelihood corresponding to the *i*-th phone and d_i is the duration in terms of frames.

To compensate the different duration, the following "local average log-likelihood" score can be used:

$$L = \frac{1}{N}\sum_{i=1}^{N}\frac{l_i}{d_i} \quad (2)$$

There is no normalization against speaker variability for these two measures. As we will see in Section 4.2.1, they exhibit low correlations to experts' ratings.

2.2. Recognition accuracy as a measure

Decoded with a canonical model trained based on gold standard speakers, recognition performance in terms of word, syllable, and even phone accuracy of a speaker can also be a measure of the utterances' pronunciation.

2.3. Generalized segment posterior probability scoring

In speech recognition, posterior probability, as a kind of confidence measure, tries to estimate the probability of a recognized entity given all the acoustic observations. It should be a good measure to evaluate the pronunciation quality. The frame based posterior probability was proposed by Neumeyer *et al* [1]. However, the state-of-the-art tie-states tri-phone model can not be used to calculate the score in his formula. As a natural extension of the GWPP [3], we propose generalized segment posterior probability (GSPP) and apply them into both syllable level and phone level. Several GOPs are derived correspondingly for the evaluation.

The generalized segment posterior probability is computed by summing the posterior probabilities of all string hypotheses in the search space bearing the focused segment w, starting at time s and ending at time t and the exponential weight of the acoustic and language models are labeled as α and β , given as [3]:

$$p([w;s,t]\,|\,x_1^T) = \sum_{\substack{\forall M,\ [w;\ s,\ t]_1^M \\ \exists n,\ 1\le n\le M \\ w=w_n \\ [s,t]\cap[s_n,t_n]\neq\phi}} \frac{\prod_{m=1}^{M} p^{\alpha}(x_{s_m}^{t_m}\,|\,w_m)\cdot p^{\beta}(w_m\,|\,w_1^{m-1})}{p(x_1^T)} \quad (3)$$

where a segment hypothesis is defined by the corresponding triple, $[w;s,t]$; x_s^t is the sequence of acoustic observations; M, the no. of segments in a string hypothesis; $p(x_1^T)$, the probability of the acoustic observations; T , the length of the complete acoustic observations.

Start time s and end time t can be obtained from the forced alignment using the given reference. Introducing of α and β in Formula (3) can efficiently prevent the segment posterior probability from dominated by just a few top string hypotheses with high likelihoods in the practical implementation.

In practice, recognized lattice is used to calculate the denominator instead of using all the string possibilities. Therefore, compared with GWPP in recognition task, we made following modifications for pronunciation evaluation,

1. Transcribed references should be added into the lattice as a path given it does not exist in the recognized lattice;
2. GSPPs are computed based on both syllable and phone level instead of word level.

Several GOPs can be derived based on GSPP such as syllable level and phone level. On phone level, GOPs can be computed in two ways. One is to calculate the average of the phone segments in each syllable (including both initial and final), the other is to weight the scores of each segment, e.g. initial and final, by their duration respectively, as shown in Formula (4)

$$G = G_i\left(\frac{D_i}{D_i + D_f}\right) + G_f\left(\frac{D_f}{D_i + D_f}\right) \quad (4)$$

where G_i and G_f indicate the phone level GSPP scores of initial and final; D_i and D_f are the duration of initial and final.

The posterior probability score for a whole utterance G_u is defined as the average of the individual posterior probability scores over all the segments in an utterance:

$$G_u = \frac{1}{N}\sum_{j=1}^{N} G_j \quad (5)$$

We expect the posterior probability score could be less affected by the variety of speakers' spectral characteristics since the same changes will affect both numerator and denominator in Formula (3), making the score more consistent and more focused on the acoustic pronunciation quality. It is also verified by the experiment in Section 4.2.2 that the weighted phone scores can achieve the best performance.

The optimal parameter for the acoustic and language model weights $(r = \alpha/\beta)$ is obtained from a development set.

3. DATABASE

In order to keep consistent with official PSC, we carefully design and collect the database as the setup of the PSC. It can be described as follows: There are 140 speakers (70 males and 70 females), in which 100 are gold standard speakers with conical pronunciations and they are used to train the gold standard model to obtain all kinds of machine scores. And rest 40 speakers with different pronunciation proficiency varied from strong accent to standard pronunciation are reserved for both subjective and automatic evaluations.

Each speaker pronounces two full sets and each set consists of 4 parts and they are 100 single syllable word utterances (P1), 49 multi-syllable word utterances (P2), a reading paragraph with approximately 400 syllables (P3) and a spontaneous talking in 3 minutes (P4).

For the subjective evaluation, three experts with national certification are invited to score each part of each speaker and finally total scores per speaker are summed. For P1 and P2, they score them utterance by utterance with three grades: good, defective and bad. All results reported in the experiment section are based on P2 except that explicitly mentioned.

4. EXPERIMENTS

As a tonal language, tone pronunciation plays a very important role in Mandarin evaluation where most of the accents in Chinese are due to the tone mispronunciation that differentiates them from Mandarin.

Multi-space distribution (MSD) approach, first proposed by Tokuda [4] for speech synthesis, has also achieved good performance in the tonal language speech recognition [5] and tone mispronunciation detection [6]. The MSD assumes that the observation space can be made of multiple subspaces with different priors and different distribution forms (discrete or continuous probability density function) can be adopted for each subspace. It can deal with the discontinuity of F0 elegantly. MSD-HMM is used to model the observation space consisting of F0 related features and spectral features.

4.1. Human scoring

The human scores are the reference against which the performance of automatic scoring systems should be evaluated and calibrated. Therefore it is important to assess the consistency of these scores between raters and within each rater, called inter-rater and intra-rater correlation respectively. All the human results reported here are based on P2 although the similar results can be found on P1.

4.1.1. Inter-rater correlation

The inter-rater correlation is calculated as the correlation of a rater and the mean of all other raters excluding the current one, which is also called open correlation. It is obvious that correlation on speaker level is much higher that that on utterance level and shown in Table 1. And the details of the correlation improving with the amount of utterances evaluated can be found in Figure 1.

Table 1: Utterance and speaker level inter-rater correlations.

Inter-rater correlations	Rater ID			Avg.
	1	2	3	
Utterance	0.559	0.491	0.532	0.527
Speaker	0.879	0.806	0.851	0.845

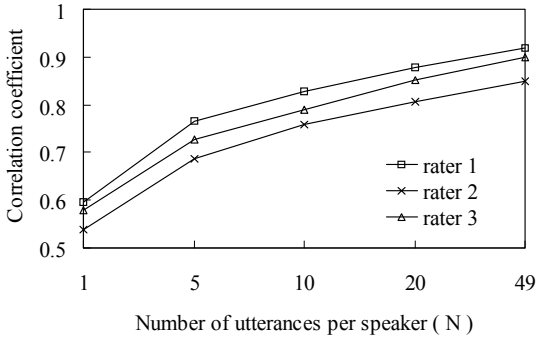

Fig.1 Inter-rater correlation based on different number of utterances per speaker

4.1.2. Intra-rater correlation

The intra-rater correlation describes the consistency of a specify rater. The correlation still increases with the number of utterances. However, we also observe that consistency varied greatly from raters, it is 0.96 for the best rater and 0.75 for the worst one, although they all have national level qualifications. It is inevitable for subjective evaluation given even by experts.

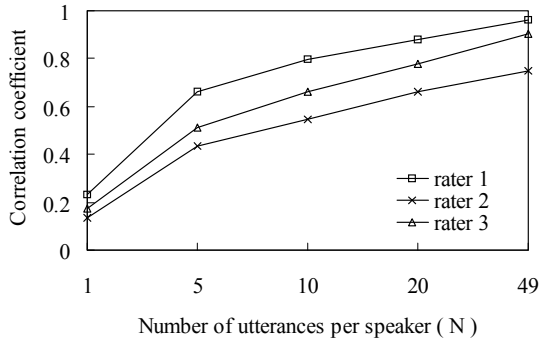

Fig.2 Intra-rater correlation based on different number of utterances per speaker

4.2. Automatic scoring

For automatic evaluation, a phone set consisting of 184 mono-phones is designed where both the neutral tone and Érhuà [†] phenomena are specially considered for PSC evaluation requirement. A gold standard model with the tie-states tri-phone based on MSD-HMM is therefore trained.

The automatic scores are calculated based on the above gold standard model. We also generate the phonetic time alignment for all data using the Viterbi decoding. With the confidence error rate as a reference, the optimal parameters for the acoustic and language model weights $(r = \alpha / \beta)$ are

[†] Érhuà refers to the r-coloring or addition of the "ér" sound (transcribed in IPA as /ɚ/) to syllables in spoken Mandarin Chinese

tuned on a development set and they are 23 in syllable level and 14 in phone level respectively.

Table 2: Utterance and speaker level correlation between human and machine scores.

Machine Scores	Utterance level	Speaker level
Global log-likelihood	0.078	0.218
Local log-likelihood	0.072	0.222
Phone recognition accuracy	0.185	0.518
GSPP (Syllable level)	0.169	0.537

4.2.1. Comparisons of scoring measures

We investigate all kinds of GOP measures described in Section 2 experimentally at both utterance and speaker level. The silence part is discarded in calculating GOP. The comparison results are shown in Table 2.

As we observed, both log-likelihood based scores obtain very low correlations, either at utterance level or speaker level. Phone accuracy achieves much better performance and increases a lot with the number of utterance as posterior probability score does while later provides the best results among all these methods.

4.2.2. Generalized posterior probability based scoring

In Section 4.2.1, the GSPP based on syllable level achieves the best performance among all the measures. As we know, phone has better granularity of pronunciation than syllables. Furthermore, we extend GSPP from syllable level to phone level as described in Section 2.3. Three kind of GOP based on posterior probabilities scores are compared as shown in Table 3. They are comparable in utterance level. In speaker level, weighted posterior score based on phone segments achieves better correlation than that of syllable level, e.g. 0.610 vs. 0.537.

Table 3: Utterance and speaker level correlation between human and GSPP scores.

GOPs based on GSPP	Utterance level	Speaker level
Syllable	0.169	0.537
Average phone	0.194	0.588
Weighted phone	0.194	0.610

The performance of weighted phone segment posterior probability by their duration with the number of utterances can be seen in Figure 3.

There is still a gap even between the most promising GOP measure, weighted phone posterior score and human scoring. One possibility is that the utterance currently used for computing machine score only consists of single word with 2-3 syllables. Compared with the continuous utterance with dozens of words, it is not enough to evaluate the speaker's pronunciation quality reliably. And we are working to testify the proposed GOPs on the continuous and spontaneous speech database, e.g. P3 and P4.

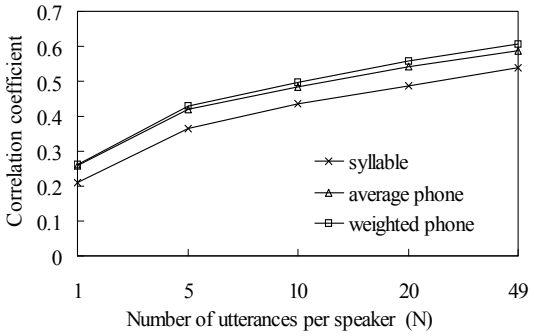

Fig.3 Machine-Rater correlation based on different number of utterances per speaker

5. CONCLUSIONS

In this paper, we investigate automatic pronunciation evaluation methods for native Mandarin and propose several measures for the Goodness of Pronunciation (GOP) based on the generalized segment posterior probability. Evaluation on the database collected internally shows the weighted phone posterior score achieves the best result and the GOP currently calculated from utterances with 2-3 syllable is still a promising indicator of the pronunciation quality. In addition, detailed analyses of human scoring like inter/intra-rater on utterance/speaker are also provided for the reference and calibration for automatic scoring.

6. REFERENCES

[1] Franco, H., Neumeyer, L., Y, Kim and O. Ronen, O., "Automatic pronunciation scoring for language instruction," In proceedings of ICASSP, pp. 1471–1474, Munich, Germany, April 1997.

[2] Witt, S., M., "Use of Speech recognition in Computer assisted Language Learning", PhD Thesis, the University of Cambridge, Nov.1999.

[3] Soong, F. K., Lo W. K. and Nakamura, S., "Generalized Word Posterior Probability (GWPP) for Measuring Reliability of Recognized Words", Proc. SWIM, 2004.

[4] Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T., Multi-space Probability Distribution HMM", IEICE Trans. Inf. & Syst., E85-D(3): pp. 455-464, 2002.

[5] Wang, H., Qian, Y., Soong, F. K., Zhou, J.L., Han, J., "A multi-space distribution (MSD) approach to speech recognition of tonal languages," Proc. ICSLP2006.

[6] Zhang, L., Huang, C., Chu, M., Soong, F. K. "Automatic detection of tone mispronunciation in Mandarin Chinese," Proc. ISCSLP2006, LNAI 4272, pp. 590-601, Springer.

[7] Cucchiarini, C., Strik, H., Boves, L., "Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms," Speech Communication, 30 (2-3), pp. 109-119, 2000