AUTOMATIC MISPRONUNCIATION DETECTION FOR MANDARIN

Feng Zhang [1, 2], Chao Huang [2], Frank K. Soong [2], Min Chu [2], Renhua Wang [1]*

[1] iFlytek Speech Lab, University of Science and Technology of China, Hefei, China
[2] Microsoft Research Asia, Beijing, China
zhangf@ustc.edu, {chaoh, frankkps, minchu}@microsoft.com, rhw@ustc.edu.cn

ABSTRACT

This paper presents the methods to improve the performance of mispronunciation detection at syllable level for Mandarin from two aspects: proposing scaled log-posterior probability (SLPP) and weighted phone SLPP to get the better measure of pronunciation quality; introducing speaker normalization of speaker adaptive training (SAT) and speaker adaptation of selective maximum likelihood linear regression (SMLLR) to get a better statistical model. Experiments based on a database, consisting of 8000 syllables pronounced by 40 speakers with varied pronunciation proficiency, confirm the promising effectiveness of these strategies by reducing FAR from 41.1% to 31.4% at 90% FRR and 36.0% to 16.3% at 95% FRR.

Index Terms— Automatic mispronunciation detection (AMD), log-posterior probability, speaker adaptive training (SAT), selective maximum likelihood linear regression (SMLLR)

1. INTRODUCTION

Computer Assisted Language Learning (CALL) has received a considerable attention in recent years. In a CALL system, it is very useful to offer a real feedback on pronunciation quality of the speaker. Many investigations have been reported in this area [1, 2]. Furthermore, providing detailed feedbacks on mispronunciation problem is also very important to help correct or improve pronunciation in addition to giving a general proficiency score, especially in interactive language learning environment. This paper focuses to improve the performance of automatic mispronunciation detection (AMD) for Mandarin syllable.

In Mandarin, each syllable normally consists of three parts: an initial (mainly consisting of the consonant), a final (mainly consisting of the vowels) and a tone while the tone is usually reflected in the final part. Therefore, any pronunciation problem on either part is classified as the mispronunciation of a syllable.

Some methods have been proposed to detect mispronunciation. Franco [3] use posterior probability score based on Hidden Markov Model (HMM) and log-likelihood ratio score based on Gaussian mixture model for pronunciation error detection. Ito [4] adopts multi-thresholds based on decision tree to detect pronunciation error. In this paper, scaled log-posterior probability is introduced to measure the goodness of pronunciation (GOP). Considering the consistent structure of Mandarin syllable, improved syllable GOP based on weighted phones-SLPP is proposed.

Besides investigating proper measures, improving modeling strategies can also improve the performance. In this paper, the strategies based on speaker normalization and speaker adaptation schemes originally proposed for automatic speech recognition (ASR) are presented to build the referenced model as standard as possible for detecting mispronunciation. Taking account of the difference tasks of using these strategies for ASR and AMD, speaker adaptation based on selective maximum likelihood linear regression (SMLLR) is proposed for the special purpose of AMD.

This paper is organized as follows: Section 2 presents improving the performance of mispronunciation detection from two aspects: improved GOP measures and improved modeling strategies. Section 3 introduces our database. In Section 4, in addition to the experiment results are compared, the detailed analysis and discussions are present. The conclusions are drawn in Section 5.

2. IMPROVED MISPRONUNCIATION DETECTION

In this section, methods from two aspects are investigated to improve AMD. In terms of measure of pronunciation quality, scaled log-posterior probability score and weighted phone SLPP at syllable level are proposed; in terms of model, speaker adaptive training (SAT) and SMLLR speaker adaptation are investigated to get a better model. The detailed description is introduced as follows.

2.1. Improved GOP measures

2.1.1. Scaled Log-posterior probability

To assess GOP score, log-posterior probability (LPP) has been reported as a good parameter since it is less affected by the changes in the spectral match due to particular speaker characteristics or acoustic channel variations and more focused on the phonetic quality.

In an HMM-based speech recognizer, given an isolated phone acoustic observation sequence: O and its corresponding transcription: q_i, its LPP can be written as:

$$P(q_i \mid O) = \log\left(\frac{\sum_{k=1}^{K} p(O \mid L_{k,i})\, p(q_i)}{\sum_{l=1}^{L} p(O \mid L_l)\, p(q_i)}\right) \quad (1)$$

where L is the number of all the path in the lattice from Viterbi decoding, K is the number of the path including phone p_i, $p(O|L_l)$ is the likelihood of the *l-th* path in the lattice, $p(O|L_{k,i})$ is the likelihood of the *k-th* path which includes the phone p_i, $p(q_i)$ is the prior probability of the phone q_i.

In practical implementation, if the acoustic model probabilities are not scaled appropriately, the sums in the denominator of Formula (1) are dominated by only a few hypotheses because of

*Join the work as an intern at Microsoft Research Asia

the very large dynamic ranges of the acoustic scores. Therefore, the acoustic probabilities have to be scaled in order to obtain the useful results as shown in Formula (2). With the proper scaling factor α, the resultant values of posterior probability can be more meaningful between 0 and 1 instead of either nearly 0 or 1 that calculated without rescaling. The induced LPP after introducing scaling factor is called scaled log-PP (SLPP).

$$P(q_i \mid O) = \log\left(\frac{\sum_{k=1}^{K} p^{\alpha}(O \mid L_{k,i})\, p(q_i)}{\sum_{l=1}^{L} p^{\alpha}(O \mid L_l)\, p(q_i)}\right) \quad (2)$$

2.1.2. Improved GOP based on weighted phones-SLPP

To detect mispronunciation for syllable in Mandarin, SLPP of phone is calculated at first. Then, GOP score of syllable can also be calculated in two ways: the averaged or weighted SLPP scores of the phones. Since each syllable consists of two phones, called an initial and a tonal final (combined final with tone), the GOP score of syllable can be calculated as follows:

$$P(s_k \mid O) = w_i \times P(q_i \mid O_i) + w_f \times P(q_f \mid O_f) \quad (3)$$

Where w_i and w_f are the weights of the initial q_i and final phone q_f, of the syllable s_k respectively. In the average way, w_i / w_f =1. In the weighted way, the value of w_i / w_f can be tuned from a development set based on their relate contributions.

2.2. Improve modeling strategies

To improve the performance of AMD, searching for better measures or algorithms is one of the key techniques. Furthermore, improving the model also plays an important role. In this section, two strategies are adopted here to general a better model. In training, speaker adaptive training (SAT) based on constrained maximum likelihood linear regression (CMLLR) is used to reduce the variation by the characteristics of the speakers of the training data [5, 6]. In testing, speaker adaptation based on maximum likelihood linear regression (MLLR) is used to reduce the mismatch between the training and testing data. Because of the difference task purpose between ASR and AMD, SMLLR are especially proposed in this paper. The flowchart to improve modeling strategies is shown in Fig.1. The detailed description is introduced as follows.

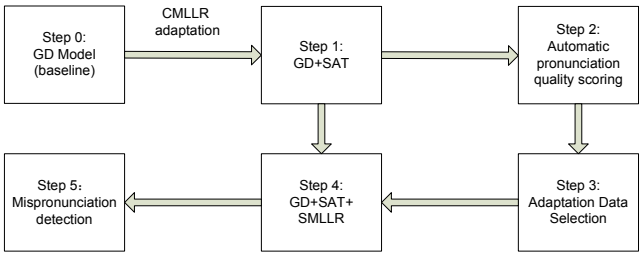

Fig.1. Flowchart of schemes to improve modeling

2.2.1. Speaker normalization using SAT

To estimate the parameters of the hidden Markov Model, the training data from a large number of speakers is usually exploited. Therefore, the parameters are affected by inter-speaker acoustic variability induced by the different characteristics of the speaker in the training data. It is not only one of the major causes of error in ASR, but also greatly affects the performance of mispronunciation detection. SAT is proved to be a useful approach of speaker normalization to reduce the overlap of speaker independent model caused by variation among the speakers of the training data. With linear transforms estimated by maximum likelihood formulation, this approach aims at separating the two processes, one being the speaker specific variation and the other the phonetically relevant variation of the speech signal. There are two main forms of SAT. One is the unconstrained case (MLLR) [5], the mean and variance transformations are unrelated to each other. The other is the constrained case (CMLLR) [6] where the transformation *A'* on the means μ and variances Σ is required to have the same form, other than the bias *b'*. Thus, the general form of CMLLR is

$$\hat{\mu} = A'\mu - b' \quad (4)$$

$$\hat{\Sigma} = A'\Sigma A'^{T} \quad (5)$$

In our system, CMLLR is adopted for its comparatively simple implementation [7].

2.2.2. Speaker adaptation using SMLLR

SAT can obtain a more neutral model with less speaker variability in training. Given the testing features from target that includes both speaker-specific and phonetic-specific variations, adaptation is needed to reduce such mismatch further. MLLR is proved to be a useful method to compensate for the mismatch between the target model and the testing data in ASR. For mispronunciation detection, it's also helpful. However, there is an important difference between using it for ASR and mispronunciation detection. The purpose of ASR is to increase the recognition accuracy. Therefore, speaker adaptation in ASR is to reduce the mismatch as far as possible, no matter such mismatch is caused by the characteristics or the mispronunciation of the speaker. The purpose of mispronunciation detection is to judge the pronunciation correct or not instead of increasing the recognition accuracy. Speaker adaptation should be used carefully to reduce the mismatch only induced by the characteristics, not the mispronunciation of the speaker. Two strategies are proposed here to approach this intention. The first one is to use a global transformation matrix during MLLR adaptation where a global transformation matrix can carry more about characteristics of speaker but less phonetic-specific variations of the speaker. Nevertheless, such transformation matrix is still unavoidably affected by the mispronunciations of a speaker. As an alternative, MLLR adaptation based on well-selected data is proposed to avoid such issues. After GOP score for each syllable of each speaker is calculated, the syllables with high GOP scores can be considered as "good" ones without mispronunciation from the speaker. MLLR adaptation based on the data consisting of these pre-selected syllables is called selective MLLR (SMLLR). It can deeply reduce the effect of mispronunciation during the adaptation.

The flowchart of improving modeling strategies including SAT and SMLLR has been shown in Fig.1. In Step 1, a compact model is generated from SAT strategy based on GD model. In Step2, the GOP scores for each syllable are calculated with the Formula (2). In Step 3, the syllables with high GOP scores are selected as the adaptation data. MLLR with these selected adaptation data is used in Step 4. In Step 5, mispronunciations are detected with the GOP scores calculated with GD+SAT+SMLLR and proper thresholds.

3. DATABASE

Our database is carefully designed in order to be consistent with Putonghua Shuiping Ceshi (PSC), which is a national test to evaluate the proficiency of spoken Mandarin.

There are totally 140 native speakers (70 males and 70 females). 100 speakers (50 males and 50 females) with standard pronunciations are chosen from them to train the gold standard model. The rest 40 speakers whose pronunciation qualities varied from very bad with strong accent to standard are reserved as the testing set.

Each speaker pronounces two full sets (Set A and Set B) and each set consists of 4 parts: 100 single syllablic word utterances (Part1), 49 multi-syllablic word utterances consisting of 100 single syllables (Part2), a reading paragraph (Part3) and a spontaneous talking (Part 4).

Part1 and Part2 from 100 gold standard speakers are set as the training data to generate the gender dependent mono-phone models. Mispronunciation detection experiments in the paper are carried out based on Part1 of the rest 40 speakers.

To get the mispronunciation references, three expert raters with national certificate are invited to evaluate the whole set and made a tag for any pronunciation with errors or defectives. Those pronunciations tagged with errors or defectives at least by one rater are taken as the mispronunciations references used for machine detection. There are totally 1746 mispronunciations in 8000 testing syllables from 40 speakers, 2 Part1 per speaker and 100 single syllables per Part1.

4. EXPERIMENT

There are two error types for any detection tasks. In our mispronunciation detection task, any pronunciations with errors or defectives are the targets we try to identify. Therefore, we define the following two measures, called false rejection rate (FRR) and false acceptance rate (FAR). With different threshold, a hypothesis syllables can be accepted as mispronunciation or not. Any mispronunciation detected as correct is classified as missing or false rejection and any correct pronunciation detected as incorrect one is treated as false acceptance. To fully reflect the changing performance FAR/FRR with different thresholds, Detection-Error Tradeoff (DET) curve is used in following experiments. Before discussing the detailed effects of each strategy, we will briefly review the baseline modeling method in the experiment.

4.1. MSD-HMM

As we know, Mandarin is a tonal language. Tones are more difficult to be pronounced correctly because they are much easily influenced by the dialect of the speaker. Studies have indicated that F0 related features can greatly improve tone recognition accuracy, but how to deal with the no observation of F0 in the unvoiced region is always a big problem.

Multi-space distribution (MSD) approach, first proposed by Tokuda [8] for speech synthesis, can deal with the discontinuity of F0 elegantly and achieve good performance in tonal language speech recognition and tone mispronunciation detection [9]. In our experiment, mono-phone MSD-HMM consisting of 184 tonal phones is adapted. The acoustic feature vector contains 39-dimension spectral features and 5-dimension F0 related features.

4.2. Effect of SLPP

Because of the very large dynamic ranges of the acoustic scores, the acoustic probabilities have to be scaled in order to obtain the useful results as shown in Formula (2). The optimal α is tuned as 1/80.0 experimentally. The comparison DET curves based on GOPs with or without scaling factor, called LPP and SLPP respectively, are shown in Fig.2. It is clear that SLPP makes better performance especially for low FAR regions and α is fixed as 1/80 during the resting experiments.

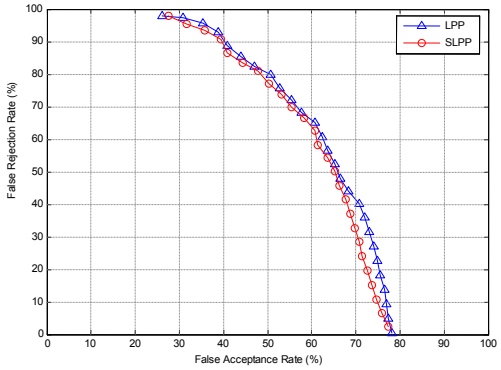

Fig.2. DET curves with LPP and SLPP

4.3. Effect of improved GOP based on weighted phones-SLPP

In Mandarin, an initial phone and a final phone consisted into a valid syllable are very different and their contributions to final GOP calculation at syllable level should be different too. DET based on different GOP calculation by average and weighting are compared in Fig.2. The optimal ratio w_f / w_i in formula (3) is tuned as 3.0 experimentally. GOP based on weighted phones-SLPP improves the performance slightly better, especially for high FAR regions. It may attribute that, compared with initial phone, final phone is more stable and longer and its calculation is more stable. In addition, all tone information is also taken by final phone.

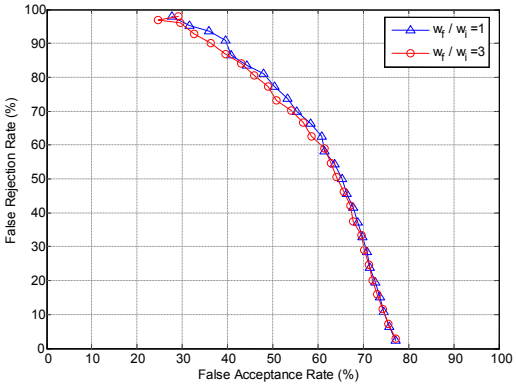

Fig.3. DET curves with GOP of syllable based on averaged and weighted phones SLPP

4.4. Effect of SAT

As an efficient speaker normalization technique, SAT based on CMLLR can obtain a more standard model with less speaker-specific variation. Its performance on AMD can be shown in Fig. 3. It is observed that SAT improves performance more at high FAR and middle FAR regions.

4.5. Effect of SMLLR

SAT is applied in training to obtain more compact model. In testing, MLLR speaker adaptation can generate speaker-specific transformations to eliminate the mismatch between the compact model and the adaptation data. A global transformation is adopted for MLLR in experiment since it focuses more on characteristics of the speaker while ignoring the pronunciation variations. To reduce the effect of mispronunciation possibly existing in adaptation data, SMLLR is proposed. The detailed procedure is shown in Fig.1.

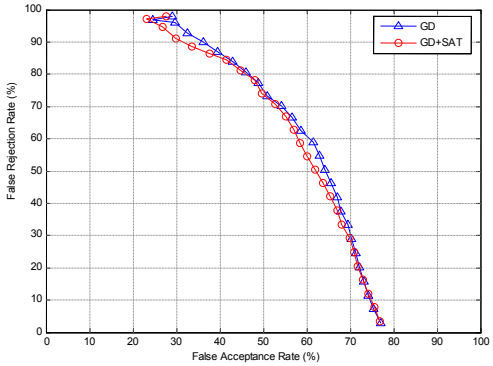

Fig.4. DET curves with SAT

Syllables with higher GOP scores calculated in Section 4.4 are selected as the adaptation data. In our experiment, 40% syllables of each speaker are used to do speaker adaptation although it can be dynamically adjusted based on proficiency level evaluated for the speaker. The result of SMLLR is shown is Fig 5. For high FAR regions, SMLLR can reduce FAR greatly, e.g. for 95% FRR, FAR reduces from 25.6% to 16.3%. It is what we expect in real CALL that providing correct feedback instead of misleading one is more critical.

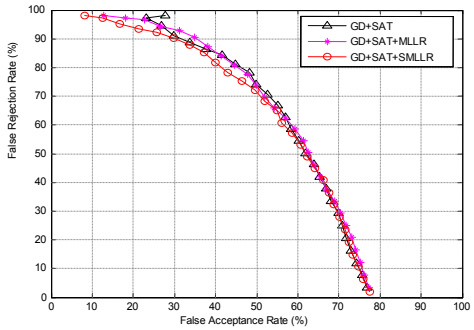

Fig.5. DET curves with MLLR and SMLLR

4.6. Results summary and discussions

The summarized results for above experiments are given in Table 1. We care more about FAR and therefore performance of FAR conditioned on the same FRR are clearly compared among all the improved schemes.

Table.1. Performance of all schemes in experiment

FRR (%)	FAR (%)					
	Improved GOP measures (w_i: w_f)			Improved modeling strategies (based on GOP of SLPP, w_i: w_f=1:3)		
	LPP (1:1)	SLPP (1:1)	SLPP (1:3)	SAT	SAT +MLLR	SAT +SMLLR
50.0	65.7	65.3	64.2	61.8	62.6	61.7
60.0	62.5	61.3	61.0	58.0	58.6	56.6
70.0	56.9	55.3	54.2	52.8	51.6	51.3
80.0	50.6	47.8	46.5	45.7	45.7	42.4
90.0	41.1	39.6	35.4	31.4	35.7	31.4
95.0	36.0	31.5	30.4	25.6	25.6	16.3

The proposed schemes can improve the detection performance step-by-step. However, we must admit that there are still great gap on performance between pronunciation evaluation task and mispronunciation detection task, where the former has nearly approached the expert scoring. Detection is a more challenged task in that we have to make a decision based on very limited resource, e.g. observation of a syllable while APE can be done with lots of syllables of a speaker. However, APE score can be used later to normalize the GOP score of each syllable as a prior in AMD.

A more detailed performance of AMD on different groups with varied proficiency level is shown in Fig. 6. As we observe, the higher proficiency level, the worse performance of AMD. AMD on worst group on pronunciation achieved the best performance far beyond the average. It may be explained that more errors factually come from the worst group and are easily identified by the experts.

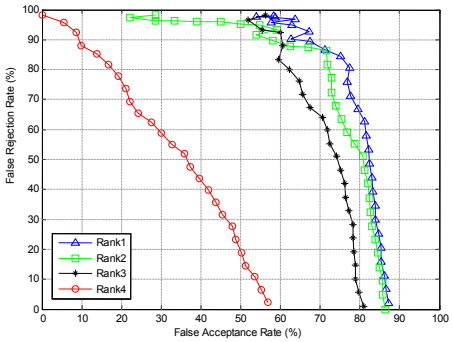

Fig.6. DET curve for different proficiency groups (10 speakers per group, pronunciation proficiency incrementally reduce from Rank1 to Rank 4, speakers in Rank4 have the worst proficiency level)

5. CONCLUSION

In this paper, several approaches from two aspects are proposed to improve the performance of mispronunciation detection: improved GOPs measure for each syllable and improved modeling strategies based on SAT and SMLLR. Experiment based on a database collected internally show they are very promising by reducing FAR from 41.1% to 31.4% at 90% FRR and 36.0% to 16.3% at 95% FRR. In addition, details analysis and discussions are given for current results and the main gap between automatic evaluation task and automatic mispronunciation detection task.

6. REFERENCES

[1] Zheng, J., Huang, C., Chu, M., Soong, F. K., Ye, W., "Generalized Segment Posterior Probability for Automatic Mandarin Pronunciation Evaluation", in *Proc. ICASSP*, pp.201-204, Hawaii, USA, 2007.

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

[3] Franco, H., Neumeyer, L., Kim, Y., Ronen, O., Bratt, H., "Automatic Detection of phone-level mispronunciation for language learning", in *Proc. Eurospeech*, Vol. 2, pp. 851-854, 1999.

[4] Ito, A., Lim, Y., Suzuki, M., Makino, S., "Pronunciation Error Detection Method based on Error Rule Clustering using a Decision Tree", in *Proc. EuroSpeech*, pp. 173-176, 2005.

[5] Anastasakos, T., McDonough, J., Schwartz, R. & Makhoul, J. "A compact model for speaker-adaptive training", in *Proc ICSLP*, Philadelphia, pp. 1137-1140, 1996.

[6] Giuliani, D., Gerosa, M., Brugnara, F., "Improved automatic speech recognition through speaker normalization", *computer speech and language*, 20, pp.107-123, 2006.

[7] Gales, M.J.F, "Maximum likelihood linear transformations for HMM-based speech recognition", *Computer Speech and Language*, 12, pp. 75-98, 1998.

[8] 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.

[9] Zhang, L., Huang, C., Chu, M., Soong, F. K. "Automatic detection of tone mispronunciation in Mandarin Chinese", in *Proc. ISCSLP*, LNAI 4272, pp. 590-601, Springer, 2006.