Mispronunciation Detection for Mandarin Chinese

Chao Huang 1, Feng Zhang 1,2*, Frank K. Soong 1, Min Chu 1

1 Microsoft Research Asia, Beijing, China
2 iFlytek Speech Lab, University of Science and Technology of China, Hefei, China

{chaoh, frankkps, minchu}@microsoft.com, zhangf@ustc.edu

ABSTRACT

In this paper, we propose several reliable weighting factors based on the speaker’s proficiency level, which can be used to normalize the scaled log-posterior probability (SLPP) to further improve mispronunciation detection at syllable level for Mandarin Chinese. Experiments based on a database consisting of 8000 syllables, pronounced by 40 speakers with varied pronunciation proficiency, shows the very promising effectiveness of these normalization schemes by reducing FAR from 44.4% to 35.1% on average and greatly improving automatic mispronunciation detection (AMD) performance greatly. In addition, we have attempted to investigate and analyze underlying behavior of such normalization factors. Some modifications, extensions and possible applications of such factors in real usage cases are also discussed.

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 real time feedback to the speakers on the quality of their pronunciation. Many investigations have been reported in this area [2] [3]. Furthermore, providing detailed feedback on mispronunciation in addition to giving a general proficiency score is very important in helping to correct or improve pronunciation especially in interactive language learning environments. This paper focuses on improving the performance of automatic mispronunciation detection (AMD) for Mandarin syllables by applying efficient normalizations to log-posterior probability (LPP).

There are roughly 2000 tonal syllables in Mandarin, and each syllable normally consists of three parts: an initial (mainly consisting of a consonant), a final (mainly consisting of a vowel) and a tone which is usually reflected in the final part. Any pronunciation problem on either part is classified as a mispronunciation of the syllable. In addition, there is only one correct pronunciation for any syllable in Mandarin.

Various methods have been proposed to detect mispronunciation. Franco et al. [4] use posterior probability scores based on Hidden Markov Models (HMM) and log-likelihood ratio score based on Gaussian mixture models for pronunciation error detection. Ito et al. [5] adopt multiple thresholds based on a decision tree to detect pronunciation error. In our previous work [1], scaled log-posterior probability is introduced as a way to measure the goodness of pronunciation (GOP). Taking account of the consistent structure of Mandarin syllables, we also proposed weighted phone SLPP as a way to improve syllable GOP.

Additionally, we also investigate how both speaker normalization based on speaker adaptive training (SAT) and speaker adaptation based on selective MLLR can be used to implemented by building the referenced model as standard as possible.

However, we still find it is challenging to detect speaker’s mispronunciation. We notice that there is a great performance gap between the automatic pronunciation evaluation (APE) task [3] and the mispronunciation detection task in that the former has nearly approached the expert scoring [1]. A main reason is that we have to make a decision based on a single data point for AMD, e.g. an observation of a single syllable, while APE can be applied to many syllables from the speaker. This inspires us try to detect pronunciation mismatch in the context of the speaker’s proficiency level.

This paper is organized as follows: Section 2 presents improvements in AMD performance resulting from different kinds of normalization schema. Section 3 introduces the database used in experiment. In Section 4, the experimental setups and results are described along with the detailed analysis and discussion. Our conclusions are presented in Section 5.

2. IMPROVED MISPRONUNCIATION DETECTION

In this section, after a brief introduction to SLPP that we proposed in [1], we propose several kinds of normalization schema on SLPP to further improve detection performance.

2.1. Scaled Log-posterior probability (SLPP)

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

In practical implementations, if the acoustic model probabilities are not scaled appropriately, LPPs are usually dominated by only a few hypotheses because of the very large dynamic ranges of the acoustic scores. Therefore, the acoustic probabilities have to be scaled as shown in Formula (1) in order to obtain meaningful results as. With the proper scaling factor $\alpha$, the resultant values of posterior probability are more meaningfully spread between 0 and 1 instead of the un-scaled values that are either nearly 0 or 1 that calculated without rescaling. The LPP after introducing scaling factor is called scaled log-posterior probability (SLPP).

$$p(q_i|O) = \log \left( \frac{\sum_{j=1}^{n} p^\alpha(O|q_j)p(q_j)}{\sum_{j}^{n} p^\alpha(O|q_j)p(q_j)} \right)$$  (1)
Calculations of SLPP can be extended to phone level. 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 in two ways: as either averaged or weighted SLPP scores of the phones, as shown in Formula (2)

\[ P_{sx} | O) = w_i \times P(q_i | O) + w_f \times P(q_f | O) \]  

Where \( w_i \) and \( w_f \) are the weights of the initial \( q_i \) and final phone \( q_f \) of the syllable \( s \) respectively. When using the weighted approach, the value of \( w_i / w_f \) can be tuned from a development set based on their relative contributions. Details can be found in [1].

### 2.2. Improve SLPP by normalization

According to the Formula (2), we can easily deduce the SLPP for the referenced syllable and the decoded syllable as described in P1 and P2, shown by Formula (3) and Formula (4) respectively. P1 is mainly used to judge if the current referenced pronunciation is correct or not based on a pre-defined threshold. P2 is the SLPP of the most competing syllable in term of the acoustic similarity between the current pronunciation and the model; it also represents the score of factual pronunciation determined by the model. In other words, P1 and P2 represent the Goodness of Pronunciation (GOP) of the prompted or referenced syllable and its most competing syllable. P3 and P4 are the average of P1 and P2 over many syllables respectively. P3 stands for the overall pronunciation proficiency score of a speaker. P4 is a newly proposed measure here and it can be explained as the pure acoustic distance of a speaker to the golden model without considering pronunciation at syllables, P4 can be explained as the pure acoustic distance of a similar to the golden model. As the average of P2 on many syllables respectively. P3 stands for the overall pronunciation proficiency score of a speaker. P4 is a newly proposed measure here and it can be explained as the pure acoustic distance of a speaker to the golden model without considering pronunciation at syllables, P4 can be explained as the pure acoustic distance of a similar to the golden model.

\[ P_{sy} | O) = \frac{1}{N_s} \sum_{i=1}^{N_s} P(q_i | O) \]  

\[ P_{sy} | O) = \frac{1}{N_s} \sum_{i=1}^{N_s} P(q_i | O) \]  

\[ P_{Spk} | O) = \frac{1}{N_s} \sum_{i=1}^{N_s} P(q_i | O) \]  

\[ P_{Spk} | O) = \frac{1}{N_s} \sum_{i=1}^{N_s} P(q_i | O) \]  

\[ (1/P3) = (P4/P3-1) \]

With above weighting factors, we can obtain following several normalized SLPPs:

\[ P_{sy} | O) = P_{sy} | O) \times \frac{1}{P_{Spk} | O) \]  

\[ P_{sy} | O) = P_{sy} | O) \times (P_{Spk} | O) - P_{Spk} | O) \]  

\[ P_{sy} | O) = P_{sy} | O) \times \frac{1}{P_{Spk} | O) \]  

\[ P_{sy} | O) = P_{sy} | O) \times \frac{1}{P_{Spk} | O) - 1) \]

### 3. DATABASE

Our database is carefully designed in order to be consistent with Putonghua Shuiing 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 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-syllable word utterances (Part1), 49 multi-syllabic word utterances consisting of 100 single syllables (Part2), a reading paragraph (Part3) and a spontaneous talking (Part4).

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 AMD 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).

\[ FRR = \frac{all\ mispronunciations\ that\ detected\ as\ correct\ ones}{all\ the\ mispronunciations} \]

\[ FAR = \frac{all\ correct\ ones\ that\ detected\ as\ mispronunciations}{all\ the\ detected\ mispronunciations} \]

To fully reflect the changing performance FAR/FRR with different thresholds, Detection-Error Tradeoff (DET) curve is used in following experiments. Before reporting the full results of the different normalization factors, we will briefly review the baseline modeling method in the experiment.
4.1. MSD-HMM

Mandarin is a tonal language. As we know, learning tones are more difficult for foreigner or native speakers with accents. About 29% mispronunciations in our database are due to tonal issue.

Multi-space distribution Hidden Markov Model (MSD-HMM) first proposed by Tokuda [6] for speech synthesis, can deal with the discontinuity of F0 elegantly and achieve good performance in both speech recognition and tone mispronunciation detection [7]. In our experiment, mono-phone MSD-HMM consisting of 184 tonal phones is adopted. The acoustic feature vector contains 39-dimension spectral features and 5-dimension F0 related features.

4.2. Normalized SLPP with speaker’s proficiency level: 1/P3

Based on the normalization scheme proposed in Section 2, we first try to normalize the SLPP with the speaker’s proficiency score, or 1/P3. Fig.1 show the DET curve based on original SLPP and normalized SLPP with 1/P3. We can observe with 1/P3 as the normalization factor, the mispronunciation detection performance has been consistently improved. The SLPP values of all 8000 syllables change with or without normalization are illustrated in Fig.2. Without impacting the detection results, all the values have been normalized to [-1, 0] for better comparison. It somehow show that values SLPP of right pronunciation (shown in red) have been improved relatively compared with the SLPP of mispronunciation (shown in blue). The impact of factor 1/P3 can be explained as follows: a syllable with the same SLPP value will be judged as mispronunciation more likely if it is from speaker with worse proficiency level.

4.3. Normalized SLPP with averaged distances: P4-P3

We tried the second normalization factor: P4-P3. P2 is larger than P1 and then P4 is larger than P3, therefore P4-P3 is non-negative. It can be understood as the averaged distance between speaker’s pronunciation and the golden acoustic model and can be used as a prior to measure the distance of a referenced syllable to the golden model. DET curve in Fig.3 shows the corresponding effects.

4.4. Combined normalization: P4/P3-1

As normalization factors, 1/P3 and (P4-P3) both improve the detection performance a lot. The question is that if they are complimentary and can be combined together to improve the performance. The answer is YES and the results are shown in Fig.4. Corresponding SLPP values change with the normalization are illustrated in Fig.5.

4.5. Analysis of normalization factors
Three kinds of normalization factors are illustrated in Fig. 6. The x-axis is syllable index. Since such factors keep unchanged for any syllable of one set (100 syllables) from one speaker, the factor values show below are factually speaker- and set-dependent. From the Fig.6, we can observe that the first two factors (1/P3 and P4-P3) have much larger ranges while range for (P4/P3-1) is much smaller. We can explain them and their effects on SLPP (P1) as follows. Firstly all three factors are the average from many syllable of one speaker. Therefore, their calculations are much more reliable and can be used as a prior to guide the detection for mispronunciation of each single syllable. However, they also show difference when used as the normalization factors.

- 1/P3 is like a weighting factor to encourage to detect more mispronunciations from speakers with worse proficiency;
- P4-P3, by introducing another score P4, it measures absolutely the averaged distance (or distortion) of pronunciation of a speaker to the golden model.
- P4/P3-1 measures relatively the averaged distance (or distortion) of pronunciation of a speaker to the golden model compared with the absolute one from P4-P3. Therefore, it shows much better performance than either of the factors: 1/P3 or P4-P3.

$$\text{FAR} = \frac{1}{\sum_{i=1}^{N} \text{SLPP}(i)}$$

Formulas (5) and (6). In such a way, form of factors should be using averaged SLPP directly instead of using sections are still superficial and need to be investigated further.

Fact, the attempts to getting underlying interpretations in above cases are often not as intuitive. Using their product is natural while using (P4-P3) is more intuitive. Using their product is natural but under many situations it has much larger ranges when compared with (P4/P3-1) which shows much better performance than either of the factors: 1/P3 or P4-P3.

Additionally, we attempt to investigate and analyze underlying behavior of such normalization factors and explain why they work. Some modifications, extensions and possible applications in real case such as are also discussed. More work is still to be done, such as understanding in-depth the underlying meaning of such factors and their generalization ability to other databases and conditions.

### Table 1. Results summary with different normalization factors

<table>
<thead>
<tr>
<th>FRR (%)</th>
<th>baseline</th>
<th>1/P3</th>
<th>(P4-P3)</th>
<th>(P4-P3)/P3</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>63.0</td>
<td>58.1</td>
<td>62.0</td>
<td>57.3</td>
</tr>
<tr>
<td>60</td>
<td>57.2</td>
<td>52.5</td>
<td>53.1</td>
<td>50.5</td>
</tr>
<tr>
<td>70</td>
<td>52.1</td>
<td>46.4</td>
<td>47.5</td>
<td>41.5</td>
</tr>
<tr>
<td>80</td>
<td>43.3</td>
<td>39.3</td>
<td>37.6</td>
<td>32.4</td>
</tr>
<tr>
<td>90</td>
<td>29.1</td>
<td>20.1</td>
<td>20.1</td>
<td>16.7</td>
</tr>
<tr>
<td>95</td>
<td>21.6</td>
<td>19.4</td>
<td>16.3</td>
<td>12.1</td>
</tr>
<tr>
<td>Aver.</td>
<td>44.4</td>
<td>39.3</td>
<td>39.4</td>
<td>35.1</td>
</tr>
</tbody>
</table>

### 5. CONCLUSION

In this paper, we propose several reliable weighting factors from speaker’s proficiency level, called 1/P3, P4-P3 and P4/P3-1 respectively, to normalize the SLPP and improve the AMD performance greatly by reducing relatively FAR 21% on average. Additionally, we attempt to investigate and analyze underlying behavior of such normalization factors and explain why they work. Some modifications, extensions and possible applications in real case such as are also discussed. More work is still to be done, such as understanding in-depth the underlying meaning of such factors and their generalization ability to other databases and conditions.

**ACKNOWLEDGEMENT**

The authors gratefully acknowledge Darren Edge’s help which improves significantly the readability of the paper.

### 6. REFERENCES


