An Improved DNN-based Approach to Mispronunciation Detection and Diagnosis of L2 Learners’ Speech

Wenping Hu1,2,∗ Yao Qian2 Frank K. Soong2

1University of Science and Technology of China, Hefei, China
2Microsoft Research, Beijing, China

{v-wenh,yaoqian,frankkps}@microsoft.com

Abstract

We extend the Goodness of Pronunciation (GOP) algorithm from the conventional GMM-HMM to DNN-HMM and further optimize the GOP measure toward L2 language learners’ accented speech. We evaluate the performance of the new proposed approach at phone-level mispronunciation detection and diagnosis on an L2 English learners’ corpus. Experimental results show that the Equal Error Rate (EER) is improved from 32.9% to 27.6% by extending GOP from GMM-HMM to DNN-HMM and the EER can be further improved by another 1.5% to 25.5% with our optimized GOP measure. For phone mispronunciation diagnosis, by applying our optimized DNN based GOP measure, the top-1 error rate is reduced from 61.0% to 51.4%, compared with the original DNN based one, and the top-5 error rate is reduced from 8.4% to 5.2%. On a continuously read, L2 Mandarin learners’ corpus, our approaches also achieve similar improvements.

Index Terms: CALL, DNN, Goodness of Pronunciation, Mispronunciation detection and diagnosis, Non-native speech

1. Introduction

For an English-as-Second Language (ESL) learner, Computer-Aided Language Learning (CALL) can be very helpful for its ubiquitous availability and high interactivity. With the popularity of smart phones, tablets and laptop computers, etc., more language learners can use CALL for learning a new language. In an L1 independent CALL system, L2 learners can be from countries with different language dialects and accents. However, the acoustic models, used for pronunciation evaluation, are usually trained with standard native speech corpus. Therefore, it needs more refined speech technology to compensate for the performance degradation due to processing non-native speech with native acoustic models. In this paper, we propose an effective and robust pronunciation assessment for mispronunciation detection and diagnosis of L2 learners’ accented speech.

Features used for pronunciation quality evaluation or deficiency detection are usually extracted from the output of an HMM based speech recognizer. Kim et al. [1] compared three HMM based scores, e.g., log-likelihood score, log-posterior probability score and segment duration score, in pronunciation evaluation for some specific phones and found log-posterior probability scores have the highest correlation with human expert’s ratings. Besides this HMM based log-posterior probability based method, Franco et al. [2] further adopted the Log-Likelihood Ratio (LLR) between native-like and non-native models as the measure for mispronunciation detection. The results show that LLR based method has better overall performance than the posterior based method, but it needs to be trained with specific examples from the targeted non-native user population. Witt and Young [3] introduced GOP, a variation of the posterior probability, for phone level pronunciation scoring. This GOP measure is later widely used in pronunciation evaluation and mispronunciation detection. Some variations of the GOP measure are also proposed in the last decade. Zhang et al. [4] proposed a Scaling Log-Posterior Probability method for Mandarin mispronunciation detection and achieved considerable performance improvement. Wang and Lee [5] combined the GOP based method with error pattern detectors for phone mispronunciation diagnosis in a serial and parallel structure and found the serial structure can reduce the error rate and improve diagnosis feedback. To improve the scores generated by the traditional GMM-HMM based speech recognizer, some discriminative training algorithms have been applied, e.g. Maximum Mutual Information Estimation (MMIE) [6], Minimum Classification Error (MCE) [7] and Minimum Phone Error (MPE) [8]. Yan and Gong [9] introduced the discriminatively refined acoustic models by MPE for pronunciation proficiency evaluation. Qian et al. [10] investigated MWE-trained HMM models for minimizing mispronunciation detection errors in L2 English learners.

Recently, Deep Neural Network (DNN) has significantly improved the discrimination of acoustic models in speech recognition [11]. Application of using Deep Belief Nets (DBN) to mispronunciation detection and diagnosis in L2 English has been tried by Qian et al. [12], and a significant improvement on word pronunciation relative error rate was obtained on L1 (Cantonese)-dependent English learning corpus. We have used DNN trained acoustic model for English pronunciation quality scoring [13]. We find the GOP score estimated from DNN outputs correlate well with human expert’s evaluation and it yields a better conventional GOP score than that obtained from a GMM based system. In this paper, we propose an improved DNN based GOP measure to deal with L2 learners’ accented speech. The effectiveness of proposed algorithm is tested in phone mispronunciation detection and diagnosis tasks on both L2 English learners’ and Mandarin learners’ corpus.

2. Goodness of Pronunciation estimation

In the conventional GMM-HMM based system, the GOP score of phone \( p \) given the whole observations \( o \), proposed by [3], is:
\[
\text{GOP}(p) = \frac{|\log p(o|p)|}{NF(p)} \\
= \frac{|\log p(o|p)p(p)}{\sum_{q \in Q} p(o|q)p(q)} / NF(p) \tag{1}
\]

where \( Q \) is the whole phone set; \( p(p) \) is the prior of the phone \( p \), \( NF(p) \) is the number of frames occupied by phone \( p \). The numerator of Eq. (1) is calculated from forced alignment and the denominator is calculated from an output lattice, generated from automatic speech recognition with an unconstrained phone loop [3]. In practice, we use the Generalized Posterior Probability (GPP) [14] method, which relaxes unit boundary to avoid underestimating the posterior probability in a reduced search space, i.e., a lattice, in the above GOP score calculation.

### 2.1. Extend GOP to DNN-HMM based system

In this section, we extend the GOP from GMM-HMM based to DNN-HMM based system [13]. By using the maximum to approximate the summation and assuming that all phones share the same prior probability, we simplify and define GOP score as Eq.(2).

\[
\text{GOP}(p) = \frac{\log p(o)}{\sum_{q \in Q} p(o|q)p(q)} \approx \frac{\log p(o)}{\max_{q \in Q} p(o|q)p(q)} \tag{2}
\]

In DNN model training, multi-layer neural networks are trained as nonlinear basis functions to represent speech while the top layer of the network is trained discriminatively as the posterior probabilities of sub-phones (“senones”). Different from the GOP calculation in GMM-HMM based system, which uses an output lattice to approximate the denominator, we propose a frame based posterior probability method to approximate the GOP in DNN-HMM based system, since well-trained posterior probabilities can be obtained naturally.

When evaluating the pronunciation quality of segment \( o_t \), whose canonical phone model is \( p \), we obtain its most probable hidden state sequence \( s' = \{s_{t_1}, s_{t_1+1}, \ldots, s_{t_e}\} \) via forced-alignment, where \( t_s \) and \( t_e \) are the start and end frame index, respectively. Then, the likelihood score is defined as:

\[
p(o|p; t_s, t_e) = \arg \max_{s \in S} p(o,s|p; t_s, t_e) \\
= \pi_{s_{t_1}} \prod_{t = t_{s + 1}}^{t_e} A_{s_{t-1}s_t} \prod_{t = t_s}^{t_e} p(o_t|s_t) \tag{3}
\]

\[
\approx \prod_{t = t_s}^{t_e} p(o_t|s_t) \tag{4}
\]

\[
= \prod_{t = t_s}^{t_e} p(s_t|o_t)p(o_t)/p(s_t) \tag{5}
\]

where \( \pi \) is the distribution of initial states; \( A \) is the transition matrix between different states; \( p(s_t|o_t) \) is the softmax output of our DNN model, \( p(s_t) \) is obtained from the training corpus of DNN model. From Eq.(3) to Eq. (4), we ignore the transition probabilities and only keep the likelihood scores for its simplicity.

Observing that the emitting probability \( p(o_t) \) will be cancelled out in Eq. (2), we further simplify the log likelihood score as:

\[
\log p(o|p; t_s, t_e) \approx \sum_{t = t_s}^{t_e} \log p(s_t|o_t)/p(s_t) \tag{6}
\]

Compared with the proposed GOP definition in GMM-HMM systems, our DNN-based GOP estimation doesn’t need a decoding lattice and its corresponding forward-backward computations, so it is suitable for supporting fast, on-line, multi-channel applications.

### 2.2. Improved GOP toward accented speech

GOP algorithm can be defined in both GMM-HMM (Eq. 1) and DNN-HMM (Eq. 6) systems with the state (sub-phone) level segmentations, obtained by forced-alignment. In an L1-independent CALL system, the acoustic model is usually trained by native speakers’ utterances, which are uttered in standard English, while the utterances from L2 language learners tend to carry some accent. Therefore, there is a mismatch between the training native speakers’ utterances and testing non-native speakers’ utterances and the mismatch will result in some inaccuracy of state-level segmentation. In addition, the pronunciation of L2 learners sometimes is ambiguous, therefore, force allocating a frame to one single senone state at forced-alignment stage is not appropriate for the phones pronounced with heavy accent.

To robustly evaluate the pronunciation quality of non-native learners’ speech, we propose to revise the log likelihood score as:

\[
\log p(o|p; t_s, t_e) = \sum_{t = t_s}^{t_e} \log \left( \sum_{s \in \mathcal{S}} p(s|o_t) \right) \tag{7}
\]

where \( s \) is the senone label, \( \{s|s \in \mathcal{S}\} \) is the set of all senones corresponding to phone \( p \), i.e., the states belonging to those triphones (HMM models) whose correct phone is \( p \). Compared with mono phone models, not only all the triphone context of phone \( p \) but also its corresponding hidden states are considered in Eq. (7). In addition, more reliable phone segmentations can be obtained with triphone HMMs. In Eq. (7), state level path constraint is removed and only phone level segmentation results are needed.

To simplify the notation, we denote the GOP measure as in Eq. (1) of GMM-HMM systems as GMM-GOP, and two GOP measures as in Eq. (2) of DNN-HMM systems as DNN-GOP1 and DNN-GOP2, whose phone segment likelihood score is calculated by Eq. (6) and Eq. (7), respectively.

### 3. Mispronunciation detection and diagnosis

To evaluate the effectiveness of our proposed GOP algorithms, we test the performance of phone-level mispronunciation detection and mispronounced phone diagnosis on an L2 English learners’ corpus. For the second task, besides giving a binary, correct or incorrect, decision of learners’ pronunciations, our system will further predict the most probable phones spoken of the mispronunciations. The L2 learners will then receive an appropriate diagnosis of their mispronunciations.

#### 3.1. Databases

Two types of databases are used in our experiments. A speech database of native speakers (native database) is used to train.
the native acoustic model. The second one is recorded by L2 language learners (non-native database), used to evaluate the performance of different GOP approaches.

3.1.1. Native database
In this study, ‘NYNEX isolated words’ [15], a phonetically rich, isolated word, telephone speech corpus, recorded by native U.S. English speakers, is used to train the native acoustic model for phone mispronunciation detection. Each utterance contains one single isolated word. The full training set consists of 90 word lists, each list contains 75 distinctive words and each word is spoken by about 10 speakers. Neither speaker nor word is mixed across different lists. The training set contains 900 s-peakers, ~6.7k distinct words, ~20 hours data in total. Another 8 word lists, which contain 5k words, 80 speakers in total, are used to evaluate the discrimination ability of acoustic models by a speech recognition task.

3.1.2. Non-native database
To evaluate the performance of mispronunciation detection, a read English, isolated word corpus is recorded by 60 non-native English learners (all Chinese) with different level of spoken English proficiency, classified according to their TOFEL\(^1\) or IELTS\(^2\) oral scores. Each speaker records ~300 words , whose transcriptions are randomly selected from the “LDC95S27” word corpus. The “ground truth” assessments of pronunciation errors are obtained by one linguistic expert. The expert marks the phone insertion, deletion and substitution errors for each spoken word token. The number of correct and incorrect (only substitution errors are considered) tokens in the whole data sets is shown in Table 1. The mispronunciation rate, the percentage of incorrect phone tokens in all the data set, is about 13.15%.

Table 1: Phone tokens for correct and incorrect pronunciations

<table>
<thead>
<tr>
<th>Number</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Misp. rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>103,522</td>
<td>16,673</td>
<td>13.15%</td>
<td></td>
</tr>
</tbody>
</table>

3.2. Acoustic modeling
Baseline acoustic model is firstly trained as context dependent GMM-HMM models (GMM-HMM) in the Maximum Likelihood (ML) sense. All these speech data are collected and sampled in 8kHz. The acoustic features, extracted by a 25ms hamming window with a 10ms time shift, consist of 13-dim MFCC and their first and second-order time derivatives. The cepstral mean normalization is performed for each utterance. Three-states, left-to-right HMM triphone models, each state with 16 Gaussian components of diagonal covariance output distribution, are adopted. The CMU pronunciation dictionary phone set with 40 different phones is used for the acoustic model training.

Acoustic models are then enhanced by DNN training [16]. Our DNN model (DNN-HMM) is a 5 layer network, consisting of 1 input layer, 3 hidden layers (each layer with 2K units) and 1 output layer, with the same number of senones as that of GMM-HMM. The input of DNN is an augmented feature vector, which contains 5 preceding frames, the current frame and 5 succeeding frames. Each dimension is normalized to zero mean and unity variance.

We evaluate the speech recognition performance of different acoustic models on a hold-out word test set. A silence-word-silence word-net or free phone loop is adopted for word level or phone level recognition performance evaluation, respectively. The calculation of Word Error Rate (WER) is exactly the same as that of isolated word recognition, in which a word graph is built for the given vocabulary. A word with any occurred errors, including phone substitutions, deletions and insertions, are regarded as an erroneous word. Word deletions and insertions are not allowed due to the isolated, single word assumption in decoding each utterance. Compared with the baseline GMM-HMM model, the DNN-HMM model has reduced the WER from 6.95% to 3.00% and the Phone Error Rate (PER) from 38.75% to 25.43%.

3.3. Phone level mispronunciation detection
We compare those three GOP measures, i.e., GMM-GOP, DNN-GOP1, DNN-GOP2, on the non-native language learning corpus. The False Rejection Rate (FRR) and False Acceptance Rate (FAR) on different thresholds are calculated and its Receiver Operating Characteristic (ROC) curve is drawn in Figure 1. It shows that the two DNN-HMM based systems outperform GMM-HMM based system consistently. The Equal Error Rate, at the operating point where FAR equals FRR, is reduced from 32.9% to 27.0% when we replace the GOP measure from GMM-HMM to DNN-HMM system. This EER can be further optimized by another 1.5% with our revised GOP algorithm. Above observations confirm our revised GOP measure is more effective in detecting the phone-level pronunciation errors of L2 learners’ speech.

3.4. Mispronounced phone diagnosis
Besides giving a binary, correct or incorrect, decision of learner’s pronunciation, we also try to diagnose the actual pronounced phones for those incorrect pronunciations. Our system can give a short, ordered phone list for each mispronounced phone. This function can enable L2 learners to have a better understanding of their own pronunciation flaws with a summary of their personalized common error patterns. Therefore, it can help L2 learners to improve their pronunciation with a statistically meaningful mispronunciation pattern.

In the above non-native database, the linguist will write
down the actual spoken phone for some incorrect phone pronunciations when she can hear very clearly which phone is exactly pronounced and we denote these human labels as the ground truth. We use top-N error rate to evaluate system’s performance, which is defined as the fraction of test segments \( o_i^n \), where the ground truth label doesn’t appear in the top-N candidates where they are sorted in descending order of its log likelihood score \( \log p(o|p, t, t_n) \), calculated in Eq.(6) or Eq.(7), respectively. For mispronunciation diagnosis, GOP is not need to be calculated exactly as Eq. (2), since its denominator is a constant value for a given segment. But to keep the notation simplicity and consistent, we still use DNN-GOP1 and DNN-GOP2 to represent Eq. (6) and (7), respectively.

We compare the performance of two DNN-based approaches, i.e., DNN-GOP1 and DNN-GOP2, and show the top-N error rates in Fig. 2. The exact numbers of top-1 to top-5 error rates are list in Table 2. It shows that the DNN based approach is very effective and the top-5 error rate is less than 10%. Our revised GOP approach, i.e., DNN-GOP2 measure, significantly outperforms the DNN-GOP1. The top-1 error rate is reduced from 61.0% to 51.4%, or a 15.7% relative error rate reduction, and the top-5 error rate is 5.2%. We also calculate the averaged rank of ground truth label tested by those two approaches, which is 2.56 and 2.11 for DNN-GOP1 and DNN-GOP2, respectively.

Table 2: Top-N error rates for mispronounced phone diagnosis

<table>
<thead>
<tr>
<th>Top-N error</th>
<th>DNN-GOP1</th>
<th>DNN-GOP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-1 error</td>
<td>61.0%</td>
<td>51.4%</td>
</tr>
<tr>
<td>Top-2 error</td>
<td>34.5%</td>
<td>26.4%</td>
</tr>
<tr>
<td>Top-3 error</td>
<td>21.6%</td>
<td>13.8%</td>
</tr>
<tr>
<td>Top-4 error</td>
<td>14.0%</td>
<td>8.3%</td>
</tr>
<tr>
<td>Top-5 error</td>
<td>8.4%</td>
<td>5.2%</td>
</tr>
</tbody>
</table>

4. Mandarin Mispronunciation Diagnosis

To evaluate the effectiveness of our approach to mispronunciation diagnosis in other languages, we test them in a continuously read, L2 Mandarin learners’ corpus.

4.1. A brief introduction of Mandarin Chinese

Mandarin Chinese, the official common language used in China, is the most widely used tonal language in terms of its speaking population. Each Chinese character, which is a morpheme in written Chinese, is pronounced as a tonal syllable, i.e., a base syllable plus a lexical tone. All Mandarin syllables have a structure of (consonant)-vowel-(consonant), where only the vowel nucleus is an obligatory component. A mandarin syllable without tone label is referred as a base syllable and with tone label is referred as a tonal syllable. By the convention of Chinese phonology, each base syllable can be divided into two parts: initial and final. The initial (onset) includes what precedes the vowel while the final includes the vowel (nucleus) and what succeeds it (coda). Most Mandarin initials are unvoiced and the tones are carried primarily by the finals. For each vowel, there are 5 tones, i.e., 4 different tones plus a neutral tone.

4.2. Databases

A Mandarin corpus, recorded by 110 native speakers (gender balanced) with standard pronunciations, is used to train the native acoustic model for our Mandarin CALL system of about 41 hrs. The recording scripts include single tonal syllables, multisyllabic words and sentences. An extra data set of 30 speakers in about 6.5 hrs, is used to evaluate the ASR performance of the trained acoustic models.

A large scale Mandarin learning corpus, iCALL corpus [17], is used to evaluate the performance of our proposed pronunciation measure. This corpus is recorded by 300 beginning learners of Mandarin Chinese, whose mother tongues are mainly European origin, i.e., Germanic, Romance and Slavic. A randomly selected subset, about 2k utterances, are carefully labeled with its actual pronounced tonal phones by 3 native linguistic experts and this labeled set is used in the following mispronunciation phone diagnosis task.

4.3. Acoustic modeling

Similar to acoustic model training performed in English mispronunciation detection systems, we first train a context dependent GMM-HMM acoustic model and then enhance its discrimination ability by DNN training. As Mandarin Chinese is a tonal language, where F0 plays an important role to distinguish different tone labels, we embed F0 contour in the DNN model training. The pitch embedding method is the same as we used before [18]. Different from speech recognition, Tonal Syllable Error Rate (TSER) is used to evaluate the performance of different acoustic models in our language learning evaluation. On the continuously read Mandarin test set, the TSER is reduced from 54.7% to 39.9% by applying DNN discriminative training and this TSER is further reduced to 32.2% by embedding F0 contour in our DNN model.

4.4. Mispronounced phone and tone diagnosis

For a tonal phone (initial or tonal final), the mispronunciation may occur at its base-phone part or its tone part or both. Therefore, we diagnose the mispronounced phone and lexical tone independently. As introduced in section 4.1, a tonal final \( final, tone \) consists of two parts, the final part \( final \) and tone part \( tone \). Tones may be carried by the same final or different finals. In our experiments, we calculate the score of a final and tone in the following two ways:

1. Selecting the corresponding tonal final with the highest
likelihood score, which is formulated as:

\[
\log p(o|\text{final}_i) \approx \max_{\text{tone}_i} \log p(o|\text{final}_i, \text{tone}_i; t_s, t_e) \tag{8}
\]

\[
\log p(o|\text{tone}_i) \approx \max_{\text{final}_i} \log p(o|\text{final}_i, \text{tone}_i; t_s, t_e) \tag{9}
\]

where in the above equations, the log likelihood scores for each initial phone \( \log p(o|\text{initial}; t_s, t_e) \) and tonal final \( \log p(o|\text{final}, \text{tone}_i; t_s, t_e) \) are calculated as Eq. (6) or Eq. (7), which denotes DNN-GOP1 or DNN-GOP2, respectively.

2. Calculating from the frame based senone posteriors directly:

\[
\log p(o|\text{final}_i) \approx \sum_{t_s} \log \left( \sum_{t_e} \sum_{\text{tone}_i \in \{\text{tone}_i|\text{final}_i\}} p(s|o_i) \right) \tag{10}
\]

\[
\log p(o|\text{tone}_i) \approx \sum_{t_s} \log \left( \sum_{\text{final}_i} \sum_{\text{tone}_i \in \{\text{tone}_i|\text{final}_i\}} p(s|o_i) \right) \tag{11}
\]

where the log likelihood score of an initial phone is calculated as Eq. (7). We denote this approach as DNN-GOP3.

The top-N error rate is used to evaluate the performance of those three DNN based GOP measures. The results of mispronounced phone and lexical tone diagnosis are shown in tables 3 and 4, respectively. On both the mispronounced phone and tone diagnosis experiments, the DNN-GOP2 approach reduces the top-N error rates consistently in different conditions, compared with DNN-GOP1. About 10% and 4% error rate reduction is achieved for mispronounced phone and lexical tone diagnosis, respectively. These error rates can be further reduced, though slightly, by applying DNN-GOP3 approaches.

Table 3: Top-N error rates for mispronounced phone diagnosis

<table>
<thead>
<tr>
<th></th>
<th>DNN-GOP1</th>
<th>DNN-GOP2</th>
<th>DNN-GOP3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-1 error</td>
<td>61.7%</td>
<td>51.4%</td>
<td>50.6%</td>
</tr>
<tr>
<td>Top-2 error</td>
<td>39.9%</td>
<td>30.4%</td>
<td>30.4%</td>
</tr>
<tr>
<td>Top-3 error</td>
<td>31.3%</td>
<td>21.0%</td>
<td>20.5%</td>
</tr>
<tr>
<td>Top-4 error</td>
<td>26.8%</td>
<td>15.2%</td>
<td>14.8%</td>
</tr>
<tr>
<td>Top-5 error</td>
<td>22.7%</td>
<td>12.0%</td>
<td>11.6%</td>
</tr>
</tbody>
</table>

Table 4: Top-N error rates for mispronounced tone diagnosis

<table>
<thead>
<tr>
<th></th>
<th>DNN-GOP1</th>
<th>DNN-GOP2</th>
<th>DNN-GOP3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-1 error</td>
<td>41.1%</td>
<td>30.6%</td>
<td>35.6%</td>
</tr>
<tr>
<td>Top-2 error</td>
<td>19.6%</td>
<td>15.8%</td>
<td>14.8%</td>
</tr>
<tr>
<td>Top-3 error</td>
<td>9.3%</td>
<td>6.7%</td>
<td>6.1%</td>
</tr>
</tbody>
</table>

5. Conclusion

We extend the GOP evaluation from GMM-HMM to DNN-HMM and improve the pronunciation quality assessment of L2 learners’ accentuated speech in this study. We evaluate the performance of proposed GOP algorithms at phone-level mispronunciation detection and diagnosis on L2 English learning and Mandarin learning corpora. In English mispronunciation detection, the EER is reduced from 32.9% to 25.5% with our proposed DNN based GOP measure, in comparing with the conventional one in GMM-HMM system. For English mispronounced phone diagnosis, our optimized measure obtains a significantly higher accuracy than the original one and the top-5 error rate is reduced to 5.2%. The averaged rank of ground truth label is also reduced from 2.56 to 2.11. Finally, we extend the DNN based pronunciation measures to Mandarin mispronunciation diagnosis. The results show that about 10% and 4% error rates reduction is achieved for mispronounced phone and lexical tone diagnosis, respectively, with our optimized GOP measure.

6. References


