Predicting Gradation of L2 English Mispronunciations using Crowdsourced Ratings and Phonological Rules

Hao Wang, Xiaojun Qian and Helen Meng

Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR of China
{hwang,xjqian,hmmeng}@se.cuhk.edu.hk

Abstract

Pedagogically, CAPT systems can be improved by giving effective feedback based on the severity of pronunciation errors. We obtained perceptual gradation of L2 English mispronunciations through crowdsourcing, and conducted quality control utilizing the WorkerRank algorithm to refine the collected results and reach a reliable consensus on the ratings of word mispronunciations. This paper presents our work on modeling the relationship between the phonetic mispronunciations and the actual word ratings. Based on phonological rules representing phonetic mispronunciation productions, we propose two approaches to predict the gradation of word mispronunciations. Reasonable correlation and agreement are found between the human-labeled and machine-predicted gradations for both approaches, which imply that the use of phonological rules in word-level mispronunciation gradation prediction is promising.

Index Terms: CAPT, crowdsourcing, mispronunciation gradation

1. Introduction

The success of computer-assisted pronunciation training (CAPT) technology is increasing due to the fact that CAPT systems can benefit learners by offering extra learning time and material, individualized feedback and the possibility of self-paced practice in a private and stress-free environment [1]. Furthermore, we observe increasing research interest in the pedagogical effectiveness of CAPT in recent years.

A key issue in CAPT concerns the generation of corrective feedback for L2 learning. Methodologists suggest teachers focus their attention on a few error types rather than try to address all the errors [2]. This can help learners discriminate errors by priority. Another reason is that if too many mispronunciations are presented at the same time, learners may get confused, be discouraged or even lose self-confidence, especially for beginner-level learners. One criterion for selecting errors is perceptual relevance − listeners may tolerate a few "*subtle*" mispronunciations because they do not affect intelligibility greatly; but perceptually "*serious*" errors which hamper communication must be indicated and corrected promptly. Hence, a CAPT system can be pedagogically improved by providing effective feedback through prioritizing detected errors in order of their severity. We believe that while variations exist across individual listeners, there is a general consensus in the perceptual gradation of pronunciation errors ranging from *subtle* to *serious*. Therefore, we are motivated to collect data on the severity of mispronunciations in L2 English speech and attempt to develop an automatic means of predicting the gradation of mispronunciations.

We used crowdsourcing to collect perceptual gradations of word-level mispronunciations and conducted quality control using the WorkerRank algorithm [3] to filter the crowdsourced data in terms of reliability. In this paper, we propose two approaches to predicting the gradation of word mispronunciations based on crowdsourced reliable data. The rest of this paper is organized as follows: Section 2 presents some related previous work. Section 3 reviews our previous effort on the collection of perceptual gradations of word-level mispronunciations and the procedure of quality control for selecting reliable data. Section 4 introduces our proposed approaches to predicting mispronunciation gradation. Experimental results are exhibited in Section 5, together with the discussion about the results. Section 6 presents the conclusions and future work.

2. Related Work

Our work collects human perceptual ratings of L2 English speech to develop some predictive model that can mimic human ratings according to the severity of word-level mispronunciations. Related previous work includes:

Kim at al. [4] made use of an acoustic model and generated probabilistic scores for specific phone segments based on a speech recognition system developed to help American adults learn the French language. A panel of five teachers of French were asked to rate the pronunciation of selected phone segments on a scale of 1 (unintelligible) to 5 (native-like). These collected ratings were mainly used for performance evaluation, but not for training for predictive scoring (as is done in our work).

In Neri et al. [5], a subset of speech material of low overall pronunciation quality was selected for annotators to label what they considered to be the most serious phonetic errors. The annotations were used for statistical analysis and to draw up a list of suggested priority of specified phonetic errors to be addressed. The work does not perform automatic predictions for prioritizing errors.

The measures of pronunciation quality in both the above studies were collected from human expert labelers. In recent years, crowdsourcing has become a popular technique widely used for data collection and labeling. Crowdsourcing is a process of obtaining needed services, ideas or content by soliciting contributions from an undefined large group of people. Amazon Mechanical Turk (AMT) [1] is one of the best known crowdsourcing platforms. It provides a convenient mechanism

[1] https://www.mturk.com/mturk/welcome

for distributing human intelligence tasks (HITs) via the web to an anonymous crowd of non-expert workers who complete them in exchange for micropayments [6]. Compared with traditional methods for data collection and labeling, crowdsourcing is considerably more efficient, cost-effective and diversified.

Kunath and Weinberger [7] collected English speech accent ratings from native English listeners on the AMT platform. AMT Workers were asked to rate accentedness of the given non-native speech on a five-point Likert scale (ranging from ‘1’ for native accent to ‘5’ for heavy, nonnative accent). This work mentioned about a research direction in using the collected data set to train an automatic speech accent evaluation system. However it only described the data collection procedure, and did not give information about how to train an automatic system.

Peabody [8] used AMT to collect word-level judgments of pronunciation quality for each utterance in the corpus. Each utterance was assigned to three Workers, who were asked to provide binary judgments for each word on whether it was mispronounced. The pronunciation quality of each word was classified based on the number of Workers who marked it as mispronounced (0 as good, 1-2 as ugly, 3 as mispronounced). These data were further used for mispronunciation detection.

Both efforts above used crowdsourcing techniques and considered that all the collected data were reliably labeled. Our current work proposes the WorkerRank algorithm [3] to assess the quality of crowdsourced data and we only preserve data of high quality in developing a model for predicting the severity of word-level mispronunciations.

3. Crowdsourced Mispronunciation Gradations

In our previous work [3], we used the AMT crowdsourcing platform to collect perceptual gradation of word-level mispronunciations in non-native English speech. This section presents a brief description of our crowdsourcing procedure, together with new corpus-specific data.

3.1. L2 corpus

The corpus we use is the Cantonese subset of the Chinese University Chinese Learners of English (CU-CHLOE) Corpus, which contains speech recordings by 100 Cantonese speakers (50 male and 50 female) reading several types of carefully designed material, as shown in Table 1.

Table 1. *Types of prompted speech in the CU-CHLOE English corpus.*

Group	# of prompts	Example
Confusable words	10	debt doubt dubious
Phonemic sentences	20	These ships take cars across the river.
The Aesop’s Fable	6	The North Wind and the sun were…
Minimal pairs	50	look full pull foot book

The material is designed by experienced English teachers, aiming to cover representative examples of mispronunciations from Cantonese learners of English. Each of the 100 speakers reads 86 prompts that contain 436 unique words

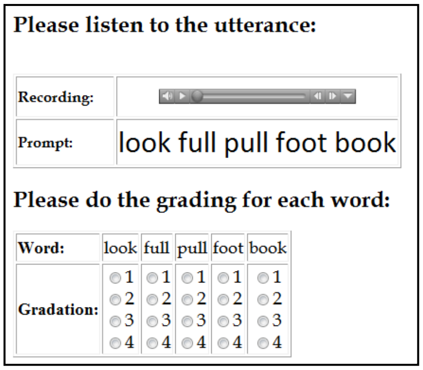
3.2. Possible gradation of errors

We defined four grades of mispronunciations in terms of the severity, as follows:

1. No mispronunciation: As good as native pronunciation.
2. Minor/Subtle: Minor deviation in word pronunciation with the native pronunciation. Can accept the deviation even if it is not rectified in the learner's speech.
3. Medium: Noticeable deviation in word pronunciation with the native pronunciation. Would prefer that the deviation be rectified for better perceived proficiency of the learner's speech.
4. Major/Salient: Very noticeable deviation in word pronunciation with the native pronunciation, to the level that it is distracting and/or affecting communication with and understanding by the listener. Strongly advise that the deviation be rectified with high priority for improved proficiency of the learner's speech.

3.3. Overview of crowdsourcing procedure

We created 200 distinct HITs (see Figure 1), each of which contains a bunch of L2 English utterances for the AMT Workers to rate according to the gradation criteria described in Section 3.2. Each distinct HIT was assigned to 3 workers.



Please listen to the utterance:

Recording:	
Prompt:	look full pull foot book

Please do the grading for each word:

Word:	look	full	pull	foot	book
Gradation:	○1 ○2 ○3 ○4	○1 ○2 ○3 ○4	○1 ○2 ○3 ○4	○1 ○2 ○3 ○4	○1 ○2 ○3 ○4

Figure 1: *An example of an utterance in an HIT.*

We ultimately obtained 600 sets of ratings (200 distinct HITs × 3 assignments) from 287 Workers.

3.4. Reliable ratings

We conducted quality control on the crowdsourced data by identifying and selecting *reliable Workers* and adopting their ratings based on an assumption that *reliable Workers* will always provide *reliable ratings*. The methodology we use for ranking the reliability of Workers is described as follows:

3.4.1. Graph-based representation for Workers

We represent the relations among Workers as an undirected weighted graph (see Figure 2) where a node is an individual Worker, a connection line between two Workers indicates that these two Workers completed common HITs and the weight for each connection is Cohen’s weighted kappa [9] value, which aims to measure the degree of agreement between two Workers on an ordinal scale.

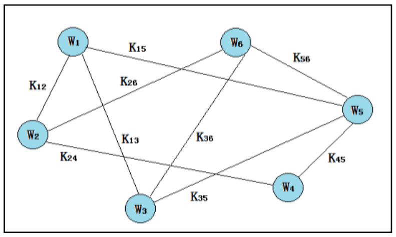


Figure 2: *A simple example of an undirected weighted graph representing AMT Workers and their relations.*

3.4.2. WorkerRank

We designed WorkerRank algorithm [3] to filter AMT Workers in terms of reliability. This algorithm is adapted from the well-known PageRank algorithm [10] that ranks web pages. We consider that a Worker is reliable if he/she gives ratings that are mostly consistent with other reliable Workers. The WorkerRank is defined in Equation 3:

$$\boldsymbol{W}(w_i) = \frac{1-d}{N} + d \cdot \left[\sum_{j:\{i,j\}\in E} \frac{k_{ij}}{\sum_{m:\{j,m\}\in E} k_{jm}} \boldsymbol{W}(w_j) \right], i = 1, \dots, N, (3)$$

where $\boldsymbol{W}$ is the resulting WorkerRank score vector, whose i-th component is the WorkerRank score associated to Worker w_i, N is the number of distinct Workers, d is the damping factor which controls the relative importance of the two involved terms, k_{ij} is the Cohen's weighted kappa value between Worker w_i and Worker w_j.

Equation 3 is a recursive expression, thus we perform iterative calculation until convergence is reached, to obtain a list of individual AMT Workers sorted by their WorkerRank according to reliability. We wish to include ratings covering the entire vocabulary of the corpus based on the most reliable Workers. Therefore, we rank all Workers in descending order according to reliability. We start by including the ratings from the top-ranking Worker, and then proceed to include the ratings from the next best Worker, and continue this procedure until all the words in the corpus are covered. We ultimately include the top 190 Workers (with the damping factor $d = 0.99$) as reliable ones, which is the minimum set of Workers that provide ratings covering all the utterances.

According to the assumption presented at the beginning of Section 3, all the ratings from reliable Workers are regarded as reliable ratings. We derived reliable ratings for the whole Cantonese subset of the corpus which has 156,709 reliable ratings. The distribution across 4 possible grades (see Section 3.2) is shown in Table 4.

Table 4. *Distribution of reliable ratings for each grade, based on Cantonese subset of CU-CHLOE corpus.*

Grade	Count	Percentage
1	109,226	69.70%
2	26,762	17.08%
3	12,109	7.73%
4	8,612	5.49%
TOTAL	156,709	100.00%

4. Predicting Mispronunciation Gradations

4.1. Baseline prediction based on table lookup for mispronunciation transcriptions

The entire corpus is phonetically transcribed by trained linguists. Based on the transcriptions, we conduct our first trial of predicting the gradations of word-level mispronunciations using the following approach: For each transcription of a word, we aggregate all the reliable ratings of the articulated words carrying the same transcription. Then, we take the average of the aggregated values and treat it as the gradation score of the corresponding transcription of that word. An example is shown in Table 5.

Table 5. *An example of how a transcription of a word mapped to crowdsourced ratings.*

Word	Transcription	Ratings from reliable Workers	Average of aggregated scores
rate	r ey t	1,1,1	1
rate	r iy t	3,3,4	3.4
rate	r iy t	4,3	
rater	r iy t	4,4	4
rater	r iy t ax	4,3,4	3.67

To predict the gradation of a given word mispronunciation, we adopt the mapped average score as its predicted rating, e.g., for an articulated word "rate" with the transcription "r iy t", we look it up in the obtained rated transcription list (See Table 5), and map it to the value 3.4 which is assigned as its predicted gradation score.

The above approach is straightforward but has an obvious limitation that the prediction can only take effect for those pronunciations (transcriptions) of words that have been observed. To solve the limitation, we attempt to use phonological rules to make the prediction have a more general coverage of mispronunciations.

4.2. Approaches based on phonological rules

The processes of phonetic mispronunciation productions are usually modeled by phonological rules. We assume that the phonological rules present in a word mispronunciation have a strong impact on the gradation of the word, so that associating each phonological rule with a certain score can help derive the gradation of word-level mispronunciations. In this section, we propose two prediction approaches by modeling the relationship between phonological rules and the crowdsourced reliable word ratings (see Section 3.4): one heuristic is to equate the word gradation with the score of the most salient phonological rule (one with the maximum score) in a word; the other one is based on linear regression – the word gradation yields from a linear combination of all the scores of rules found in a word mispronunciation

As described in [11], phonetic mispronunciation productions can be represented as context-dependent phonological rules of the form:

$$\alpha \rightarrow \beta \,/\, \sigma \_ \lambda,$$

which denotes that phone α is substituted by the phone β, when it is preceded by the phone σ and followed by the phone λ. The insertion rule can be represented by replacing α with null symbol

0 while the deletion rule is to replace β with null symbol 0. For σ and λ, they can be replaced with symbol # as a word boundary.

As mentioned previously, all speech data of the corpus are phonetically labeled by trained linguists; and the canonical pronunciations of all words can be readily obtained from electronic dictionaries (e.g., TIMIT, CMUDict, etc.). By aligning the canonical pronunciations with manual transcriptions of the corpus using phonetically-sensitive alignment [12], context-dependent phonological rules can be generated for all phonetic mispronunciations in the corpus. These derived rules are used to predict word-level mispronunciation gradation by the following two approaches.

4.2.1. *Maximum gradation score*

For each phonological rule, we aggregate the derived reliable ratings (see Section 3.4.2) of word mispronunciations that include this phonological rule; then we simply take the average of the aggregated values and treat it as the gradation score of the rule. An example is illustrated in Tables 6a and 6b.

Table 6. *An example of how a phonological rule is mapped to the crowdsourced ratings.*

(a). *word-to-rules mapping.*

Word	**Phonological rules**	**Rating**
rate	ey → iy / r _ t	3,4,4
rater	ey → iy / r _ t er → ax / t _ #	4,4

(b). *rule-to-ratings mapping.*

Phonological rule	**Rating set**	**Average**
ey → iy / r _ t	3,4,4,4,4	3.8
er → ax / t _ #	4,4	4

Using the rated phonological rules derived above, we can predict the gradation of a word mispronunciation by following the principle that the most serious error (phonological rule with the highest gradation score) dominates the gradation of the word mispronunciation. Therefore, we predict mispronunciation gradation according to the steps below:

1. get the transcription of a word mispronunciation;
2. derive a set of phonological rules of this word mispronunciation;
3. map each of the derived rules to a gradation score by referring to the rated rule list obtained previously;
4. assign the gradation score of the most serious error in a word to this word as its predicted mispronunciation gradation.

An example of the above steps is given as follows:

1. we get a mispronunciation "ae ch ih ng" of word "aching";
2. the derived phonological rules are "ey → ae / # _ k" and "k → ch / ey _ ih";
3. rule "ey → ae / # _ k" is associated with a score of 3.26, rule "k → ch / ey _ ih" is associated with a score 3.57 by referring to the rated rule list;
4. the gradation of this word mispronunciation is assigned as 3.57 which is the higher gradation score of the two rules derived previously.

4.2.2. *Linear regression*

Another approach is to model the gradation of a word mispronunciation as a linear combination of the gradation scores of the corresponding phonological rules that the word mispronunciation includes. This relationship can be expressed as:

$$G_w = \sum_r \big(G_r \cdot \delta(r)\big) + b, \tag{4}$$

where G_w is the gradation of an uttered word mispronunciation w; G_r is the gradation score of the rule r; $\delta(r)$ is an indicator function, i.e. $\delta(r) = 1$ if r occurs in w , and $\delta(r) = 0$, otherwise; b is the offset term. The summation is taking over all r in the system.

Multiple word mispronunciation gradations can be expressed in a matrix form as follows:

$$\boldsymbol{w} = \boldsymbol{A}\boldsymbol{r} + b\boldsymbol{e}, \tag{5}$$

where $\boldsymbol{w}$ is a vector containing the gradation score of each uttered word, which is calculated by averaging the crowdsourced "reliable" ratings of that word; $\boldsymbol{A}$ is a matrix with binary elements A_{ij} indicating whether the phonological rule *j* occurs in the uttered word *i*; $\boldsymbol{r}$ is a vector that contains the gradation scores of each rule; $\boldsymbol{e}$ is the all-one vector.

We run least-square linear regression analysis. A rule score vector $\boldsymbol{r}$ and an offset term b are obtained, and are used for predicting word mispronunciation gradation by Equation 4, e.g. for the mispronunciation "s ae l ax n t" of word "salient", we derive two phonological rules: "ey → ae / s _ l" and "iy → 0 / l _ ax"; the corresponding gradation scores of these two rules obtained from the previous regression analysis are 0.74 and 1.27; thus, with the trained offset term $b = 1.64$, the gradation of this mispronunciation is the summation of the above three scores, which is 3.65.

5. Experiments

5.1. Procedure

The experiments are carried out using the Cantonese subset of CU-CHLOE corpus. We split the corpus by speakers into disjoint training (25 male and 25 female) and test (25 male and 25 female) sets. 2,347 distinct context-dependent phonological rules are generated, which fully cover all phonetic mispronunciations in the training set. The gradation scores of all generated rules are trained on rated word mispronunciations using each of the two approaches as described in Section 4.2. The rules that are generated from the training set of the corpus may not cover all the mispronunciations in the test set. Thus, during prediction, we simply skip those mispronunciations that include untrained rules.

We calculate correlation and Cohen's weighted kappa between human-labeled gradations (i.e. the average of crowdsourced "reliable" ratings for each uttered word) and machine-predicted gradations by each prediction approach for the test set. To calculate kappa values, we first quantify all the word gradation scores (by rounding) to 4 integer values {1,2,3,4} which represent 4 possible grades of mispronunciations (see Section 3.2); some (less than 2% of total number of word mispronunciations) of the gradation scores obtained from linear regression approach exceed the range from 1 to 4; we quantify those gradation scores to their nearest grade values (1 or 4). For the purpose of comparison, we include the baseline approach (See Section 4.1) in the following Table.

Table 7. *Evaluation results for different prediction approaches.*

2,347 rules	**Baseline**	**Maximum score**	**Linear regression**
# of tested words	15934	15934	15934
# of predicted words	13766	14736	14736
% of predictions	86.39%	92.48%	92.48%
Correlation (*r*-value)* **95% CI**	0.627 (0.617, 0.637)	0.644 (0.635, 0.653)	0.644 (0.635, 0.653)
Kappa	0.561	0.550	0.588

5.2. Discussion

From Table 7 we see that all correlation values are above 0.6 and all kappa values exceed 0.5, which reflects a reasonable consistency between human-labeled and machine-predicted gradations. If we compare all the prediction approaches, the two approaches based on phonological rules outweighs the baseline approach in almost all evaluation measures in Table 7, and the linear regression approach has the best performance. Table 7 also illustrates that prediction based on phonological rules have a better coverage of mispronunciations than the baseline approach based on table lookup for transcriptions.

The presented approaches to predicting the gradation of word-level mispronunciations are based on the detailed phonetic transcriptions of L2 speech labeled by trained linguists. This guarantees that the phonetic mispronunciations of L2 speech are identified accurately. However, in a practical system, it is not easy to obtain accurate (manual) phonetic transcriptions of L2 utterances immediately. In that case, an acoustic model can help obtain possible transcriptions (with acoustic scores) automatically, though usually with the trade-off of lower accuracy.

Conclusions and Future Work

Giving effective feedback based on the severity of mispronunciations is of core pedagogical importance in a CAPT system. We used crowdsourcing to collect the perceptual gradation of word-level L2 English mispronunciation, and conducted quality control with the WorkerRank algorithm on the crowdsourced data to derive reliable gradations. In this paper, we propose two approaches to predict gradation of word mispronunciations using the derived reliable gradations and phonological rules. Reasonable correlation and agreement found in the experimental results shows that our proposed approaches using phonological rules for predicting word mispronunciation gradation is promising.

In future work, we will try other regression analysis to seek a better model for prediction. Directly optimizing the correlation or kappa on the training set is also an interesting direction to pursue. Besides, we also plan to use acoustic model to assist scoring the mispronunciations which cannot be covered by the phonological rules derived from the training set or whose transcriptions are not immediately available.

Acknowledgements

The work is partially supported by the grant from the Hong Kong SAR Government's Research Grants Council General Research Fund (Project No. 415511).

References

[1] Neri, A., Cucchiarini, C. and Strik, H., "ASR corrective feedback on pronunciation: Does it really work?", in Proc. of Interspeech, 1982-1985, 2006.

[2] Ellis, R, "Corrective Feedback and Teacher Development", L2 Journal, 1: 3-18, 2009.

[3] Wang, H. and Meng, H., "Deriving Perceptual Gradation of L2 English Mispronunciations using Crowdsourcing and the WorkerRank Algorithm", in Proc. of the 15th Oriental COCOSDA, Macau, China, 9-12 Decembeer 2012.

[4] Kim, Y., Franco, H. and Neumeyer, L., "Automatic pronunciation scoring of specific phone segments for language instruction", In Fifth European Conference on Speech Communication and Technology, 1997.

[5] Neri, A, Cucchiarini, C, and Strik, H, "Segmental errors in Dutch as a second language: how to establish priorities for CAPT", in Proc. of InSTIL/ICALL Symposium, 2004.

[6] McGraw, I., Glass, J., Seneff, S., "Growing a Spoken Language Interface on Amazon Mechanical Turk", in Proc. of Interspeech2011, Florence, 2011.

[7] Kunath, S. A. and Weinberger, S. H., "The wisdom of the crowd's ear: speech accent rating and annotation with Amazon Mechanical Turk", in Proc. of CSLDAMT '10, Association for Computational Linguistics, 2010.

[8] Peabody, M. A., "Methods for pronunciation assessment in computer aided language learning", [dissertation], US -- MA: Massachusetts Institute of Technology, 2011.

[9] Shoukri, M. M., "Measures of interobserver agreement", 2004.

[10] Page, L., Brin, S., Motwani, R., and Winograd, T., "The pagerank citation ranking: Bringing order to the web", Technical report, Stanford Digital Library Technologies Project, 1998.

[11] Lo, W., Zhang, S. and Meng, H., "Automatic Derivation of Phonological Rules for Mispronunciation Detection in a Computer-Assisted Pronunciation Training System," in Proc. of Interspeech, Makuhari, Japan, 26-30 September 2010.

[12] Harrison, A. M., Lo, W. K., Qian, X. J. and Meng, H., "Implementation of an Extended Recognition Network for Mispronunciation Detection and Diagnosis in Computer-Assisted Pronunciation Training," in Proc. of the 2nd ISCA Workshop on Speech and Language Technology in Education, Warrickshire, 2009.

* p-value associated with each of the three correlation values is less than 0.0001.