DERIVING PERCEPTUAL GRADATION OF L2 ENGLISH MISPRONUNCIATIONS USING CROWDSOURCING AND THE WORKERRANK ALGORITHM

*Hao Wang and Helen Meng**

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

ABSTRACT

Pedagogically, feedback in CAPT systems can be improved by focusing on the most critical errors rather than presenting all errors to the users at the same time. This paper presents our work on the use of crowdsourcing for collection of gradations of word-level mispronunciations in non-native English speech. Quality control procedures based on the proposed WorkerRank algorithm (adapted from well-known PageRank algorithm), are performed for selecting a subset of the crowdsourced data in order to ensure reliability. Based on the selected data, we derive a set of rated word-level mispronunciations, according to a four-point gradation of no error, subtle, medium and salient errors.

Index Terms*—*CAPT, Crowdsourcing, mispronunciation gradation, WorkerRank

1. INTRODUCTION

Research in computer-assisted pronunciation training (CAPT) is gaining momentum with advancements in information and communication technologies (ICT). CAPT may provide a private, stress-free environment in which students can access virtually unlimited input, practice at their own pace and receive individualized, instantaneous feedback [1].

Giving effective feedback based on the severity of errors is of core pedagogical importance in a CAPT system. Methodologists generally advise teachers to focus attention on a few error types rather than try to address all the errors learners make [2]. One reason is that if too many mispronunciations are presented at the same time, users may get confused, be discouraged or even lose self-confidence, especially for beginner L2 learners. One criterion that can be used to select errors is perceptual relevance: listeners may accept a few "subtle" mispronunciations because those errors do not affect the intelligibility; only the "serious" errors which slow down and even hamper communication matter to the listeners' perception. Thus a system should give priority to present the "serious" errors to learners. There is a general consensus on the gradation of pronunciation errors ranging from "subtle" to "serious", although variations exist across individual listeners. Therefore, we are motivated to collect data on the severity of mispronunciations in L2 English speech in an attempt to develop an automatic means of classifying mispronunciations.

2. PREVIOUS WORK

Previous work on pronunciation quality labeling for non-native corpus presents a variety of techniques:

Witt and Young [3] collected a database consisting of 2,040 non-native utterances, and got this data scored at both sentence and word levels on a scale of 1 to 4 by trained phoneticians. These collected scores were used for measuring the performance of the goodness of pronunciation (GOP) scoring.

In Neri et al. [4], speech material of Dutch recorded from both native and nonnative speakers was evaluated by both machine and human experts on several aspects of pronunciation (e.g. segmental quality, fluency). A subgroup of speech material with low evaluation scores was selected for further investigation.

The evaluation scores in the above studies were collected from expert labelers. In recent years, crowdsourcing has become a popular technique for data collection and labeling. Crowdsourcing means outsourcing some tasks to an undefined large group of people. Amazon Mechanical Turk (AMT)[1] is one of the most well-known crowdsourcing platforms. It is a convenient mechanism for distributing human intelligence tasks (HITs) via the web to an anonymous crowd of non-expert workers who complete them in exchange for micropayments [5]. Compared with traditional data collection methods, crowdsourcing is considerably more efficient, cost-effective and diversified.

Kunath and Weinberger [6] used AMT to collect English speech accent rating from potential native English listeners, in order to construct a training data set for an automatic accent evaluation system. AMT Workers were asked to rate accentedness of the given speech (utterances read by the non-native speakers from three language groups: Arabic, Mandarin, and Russian) on the five-point Likert scale (1 as native accent, 5 as heavy accent).

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

* Corresponding author: Professor Helen Meng (hmmeng@se.cuhk.edu.hk)

Peabody [7] collected word-level judgments of pronunciation quality for each utterance in the corpus through AMT. Each utterance was assigned to 3 Workers, who were asked to provide binary judgments for each word on whether it was mispronounced (MP). The pronunciation quality of each word was classified based on the number of Workers who marked it as MP (0 MP as good, 1-2 MP as ugly, 3 MP as mispronounced).

Both of the above studies with crowdsourcing techniques considered all the collected data to be reliable. However, some of the data might be submitted by cheaters.

This work presents a new methodology to improve the reliability of crowdsourced data. A related effort in evaluating TTS systems was by Buchholz and Latorre [8]. They analyzed the issue of Workers cheating, and presented some cheater detection mechanisms, such as referring to gold standard data, to improve the test outcomes.

3. L2 ENGLISH CORPUS AND MISPRONUNCIATION GRADATION

3.1. Corpus

We use CU-CHLOE L2 English corpus which contains prompted speech collected from 100 Cantonese speakers (50 male and 50 female) and 111 Mandarin speakers (61 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

There are 86 individual prompts containing 446 unique words out of 631 total words which are designed or selected by experienced English teacher, covering representative examples of mispronunciations from Chinese-speaking learners. Each of 211 speakers read all 86 prompted texts, therefore, the corpus contain (211 × 86 =) 18,146 utterances in total.

3.2. Possible gradation of errors

We define four grades in terms of the severity of mispronunciation, 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 affects communication and understanding by the listener. Strongly advise that the deviation be rectified with high priority for improved proficiency of the learner's speech.

4. CROWDSOURCING TASK

4.1. AMT background

There are two types of AMT users – Requesters and Workers. Requesters are able to define their tasks as Human Intelligence Tasks (HITs), such as transcribing audio recordings, identifying objects in photos, etc. HITs include a task description, the task display, the format of the output, the reward to pay up on a task completion, etc. Requesters may also qualify their workforce, e.g. to require Workers to pass a qualification test, or to require a Worker to have previously completed minimum number of HITs. Then, Requesters load their HITs into the marketplace. After retrieving the results submitted by Workers, Requesters are able to review them before choosing to approve or reject. Only approved results are paid. AMT Workers can browse available HITs and choose interesting ones to complete for payments.

A key issue in crowdsourcing is that Workers perform HITs on the web without supervision. Thus cheaters may submit nonsensical results. Even if Requesters may review the results before approving them, it may sometimes be hard to verify the quality for the entire (large) volume of crowdsourced data. Requesters need mechanisms to filter Workers in terms of the reliability.

4.2. HIT design

This study collects human gradation on L2 speech of English in terms of the severity of mispronunciations with reference to native US English. Therefore, in the HIT setting, the location of the AMT Workers is required to be in the US. This aims to engage more (self-declared) native American English listeners. Each HIT (see Fig. 1) includes several L2 English speech utterances for the Worker to rate.

Workers can listen to an utterance as many times as they want before giving the corresponding rating to each word, based on the gradation criteria described in Section 3.2.

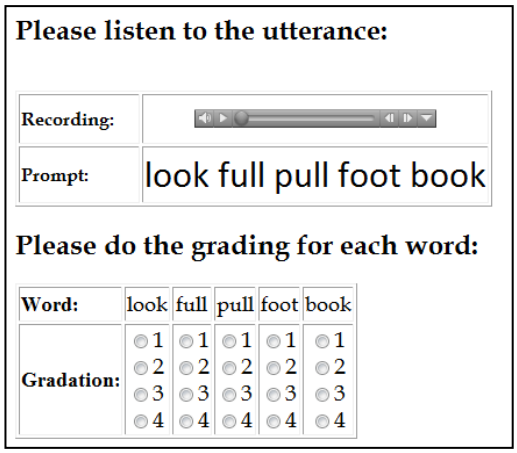

Figure 1: *An example of an utterance in an HIT.*

To constrain the length of the HITs, we split utterances of each speaker into two parts:

- The first HIT contains confusable words, phonemic sentences and the Aesop's Fable "The North Wind and the Sun" (36 utterances).
- The second HIT contains minimal pairs (50 utterances).

Each HIT is assigned to three Workers. The reward of each HIT is $0.4.

4.3. Observations from crowdsourced ratings

We published 1,266 HITs (211 speakers × 2 parts × 3 assignments) in total and collected 1,299 sets of results from 456 individual AMT Workers. 33 sets (approximately 2.5%) of the submitted results (by 26 Workers) were rejected according to some approval procedures that will be shown in the next subsection. For the approved results, there were 397,498 ratings collected. The distribution across the 4 grades (see Section 3.2) is shown in Table 2.

Table 2. *Distribution of collected results for each grade.*

Grade	Count	Percentage
1	269,307	67.75%
2	71,235	17.92%
3	33,806	8.51%
4	23,150	5.82%
TOTAL	397,498	100.00%

In fact, 399,423 ratings (211 speakers × 631 words × 3 assignments) are desired, but 1,925 words (approximately 0.5%) were missed by some AMT Workers.

4.4. Approval criteria

We present the approval criteria for screening the data. Approved data mean the Worker will be paid. But further screening follows in terms of quality assessment (see Sec. 5).

Approval conditions include:

a) Work time: The HITs completed within less than 5 minutes are rejected because the total duration of utterances contained in each HIT is approximately 5 minutes. The short work duration implies lack of care in ratings.

b) Missing inputs: While we tolerate that some AMT Workers may unintentionally skip some words, HITs containing more than 25 ratings that were left blank are rejected.

There exist some special cases that some results have a large number of identical inputs, especially for the case that more than half of the inputs are rated with grade 1 or 4 (see possible grades in Section 3.2). This kind of results seems questionable. However, they may still be possible because some of the speakers may have a perfect or a heavy non-native accent. Here we still approve such results. The approval conditions described above only filter out obviously unreasonable results. This avoids the risk of excluding genuine Workers who happen to be strict on mispronunciations, who deviate much from other Workers.

5. WORKER RELIABILITY

Verifying the quality of the crowdsourced results concerns the credibility of the mispronunciation gradation. Therefore, we devise a methodology for rating the reliability of Workers based on the approved crowdsourced data. Our aim is to identify and select reliable Workers and adopt their ratings. We assume that reliable Workers will always provide reliable ratings. Our methodology is described in this section.

5.1. Graph-based representation for Workers

We represent our problem with an undirected weighted graph $G = (W, E, K)$ where the vertex set W is the set of individual AMT Workers, the edge set E contains all connections between the Workers who have common HITs and the weight set K contains kappa values measuring how similar the ratings are between two Workers among all Worker pairs in the edge set E.

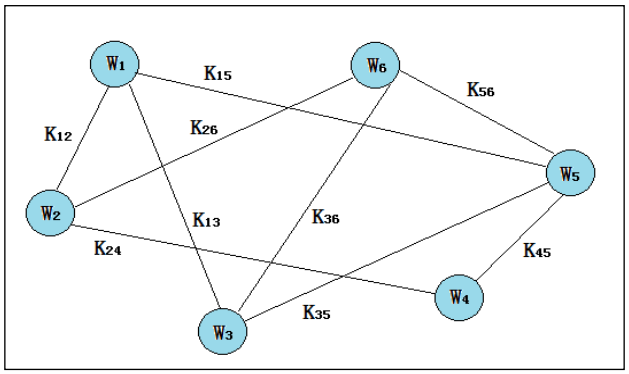

Figure 2: *A simple example of an undirected weighted graph representing AMT Workers and their relations.*

5.2. Inter-Worker agreement

We use Cohen's weighted kappa for the edge weight between Workers. It is a popular descriptive statistic for measuring the agreement between two raters on an ordinal scale. For our problem, a kappa value indicates the inter-Worker agreement between two Workers who rated a common set of HITs.

Cohen's kappa is the proportion of the total amount of agreement not explained by chance. Therefore, kappa more accurately reflects, with less ambiguity, the reliability of the data [9].

We can obtain an observed matrix from the ratings given by Workers A and B, as shown in Table 3a. n_{ij} denotes the number of words that Worker A rated with grade *i* (see possible grades in Section 3.2) and Worker B rated with grade *j*. We can also have a weight matrix (see Table 3b). With quadratic weighting scheme, the element of the weight matrix $w_{ij} = \frac{(i-j)^2}{(1-4)^2}$, indicating the degree of disagreement for each pair of ratings *i, j*.

Table 3a. *Observed Matrix.* & Table 3b. *Weight Matrix.*

(a) Observed Matrix

Worker A \ Worker B	1	2	3	4
1	n_{11}	n_{21}	n_{13}	n_{14}
2	n_{21}	n_{22}	n_{23}	n_{14}
3	n_{31}	n_{32}	n_{33}	n_{34}
4	n_{41}	n_{42}	n_{43}	n_{44}

(b) Weight Matrix

Worker A \ Worker B	1	2	3	4
1	w_{11}	w_{21}	w_{13}	w_{14}
2	w_{21}	w_{22}	w_{23}	w_{14}
3	w_{31}	w_{32}	w_{33}	w_{34}
4	w_{41}	w_{42}	w_{43}	w_{44}

Using the observed data, we obtain the proportion of the element n_{ij} of the observed matrix, $P(a)_{ij} = \frac{n_{ij}}{\sum_{i=1}^{4}\sum_{j=1}^{4} n_{ij}}$. We can also calculate the probability of Worker A rating with grade *i* and Worker B rating with grade *j* by chance, $P(e)_{ij} = \frac{\sum_j n_{ij}}{\sum_{i=1}^{4}\sum_{j=1}^{4} n_{ij}} \cdot \frac{\sum_i n_{ij}}{\sum_{i=1}^{4}\sum_{j=1}^{4} n_{ij}}$, based on the assumption of independence of ratings. Thus, we have the probability of observed weighted agreement $P(a) = 1 - \sum_i \sum_j w_{ij} P(a)_{ij}$ and the probability of expected/chance weighted agreement $P(e) = 1 - \sum_i \sum_j w_{ij} P(e)_{ij}$. We can calculate weighted kappa using the following equation [10,11]:

$$\kappa_w = \frac{P(a) - P(e)}{1 - P(e)} = 1 - \frac{\sum_i \sum_j w_{ij} P(a)_{ij}}{\sum_i \sum_j w_{ij} P(e)_{ij}} \quad (1)$$

A higher kappa value indicates a higher inter-Worker agreement. $\kappa_w \in [-1, 1]$; $\kappa_w = 1$ means perfect agreement; $\kappa_w = 0$ indicates what will be expected by chance; $\kappa_w < 0$ means agreement less than chance i.e. potential systematic disagreement.

5.3. WorkerRank

The algorithm we propose for reliable Worker selection is what we call WorkerRank, whose notion is similar to the well-known PageRank algorithm [12] 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 2:

$$\boldsymbol{W}(w_i) = \frac{1-d}{N} + d \cdot \left[\sum_{j:\{i,j\}\in E} \frac{k_{ij}}{\sum_{m:\{j,m\}\in E} k_{jm}} \boldsymbol{W}(w_j) \right], i = 1, \ldots, N \quad (2)$$

where $\boldsymbol{W}$ is the resulting WorkerRank score vector, whose *i*-th component is the score associated to Worker w_i, thus the dimension of $\boldsymbol{W}$ vector is the number of distinct Workers *N*; *d* is the damping factor which is generally assumed to be 0.85. k_{ij} is the kappa value between Worker w_i and Worker w_j. The greater is the WorkerRank score, the greater is reliability of the corresponding AMT Worker, according to its agreement with the other Workers to which it is connected [13].

The WorkerRank of each Worker depends on the WorkerRank of the Workers who rate common HITs. In computing $\boldsymbol{W}$, we give an initial $\boldsymbol{W}$ vector which is generally set uniformly to $\frac{1}{N}$ for each Worker, and repeat the calculation using the $\boldsymbol{W}$ scores calculated in the last iteration until the $\boldsymbol{W}$ values converge. We run 26 iterations to achieve convergence (i.e., the residual between two consecutive iterations changes less than 10^{-6}) and obtain a list of individual AMT Workers ranked by their WorkerRank score in descending order. We select the top 124 AMT Workers (out of 430 approved Workers) as the "reliable" set, which is the minimum set of Workers that provide ratings covering the whole corpus. According to the assumption at the beginning of Section 5, all the ratings collected from the selected 124 reliable Workers are regarded as reliable results.

5.4. Aggregated kappa

Peabody [7] proposed to use aggregated kappa (see Equation 3) for measuring agreement among a set of Workers. It computes the weighted mean of kappa values of all Worker pairs, where the weight is the number of Worker pairs for a particular number of common HITs divided by the total number of Worker pairs.

The approach is to group the words into sets for unique Worker pairs, average the kappa values computed from subsets with a common number of overlapping utterances, and then take a weighted average of all these groups. For example, consider three Worker pairs (A, B), (B, C), (A, C). Worker A annotated words 1 to 20; Worker B annotated words 6 to 15; Worker C annotated words 11 to 20. Thus, both pair (A, B) and pair (A, C) annotated 10 words which means they have an annotation overlap of 10; pair (B, C) annotated 5 words – an annotation overlap of 5.

The aggregated kappa can be computed as follows:

$$\kappa_{aggr} = \frac{1}{\sum_{s\in S}|T_s|} \sum_{s\in S} |T_s| \sum_{t\in T_s} \frac{P(a|t) - P(e|t)}{1 - P(e|t)} \quad (3)$$

where $P(a|t)$ is the probability of observed agreement; $P(e|t)$ is the probability of expected agreement; T_s is the set of Worker pairs that rated a particular number s of common words (annotation overlap of s).

We calculate both the aggregated kappa values for the whole crowdsourced dataset, as well as the reliable dataset which is obtained from top 124 AMT Workers. The values are 0.37 and 0.43 respectively. From the result we can see that the selected Workers have a higher agreement/reliability.

6. RATING OF WORD MISPRONUNCIATIONS

All speech data of the corpus are manually transcribed and the canonical pronunciations of all words can be readily obtained from electronic dictionaries (e.g., TIMIT, CMUDict, etc.). Each distinct word can have several different pronunciations which were uttered by different speakers. Based on the manual transcriptions, we group the uttered words that have the same pronunciation/transcription together. According to the reliable dataset (See Section 5.3), every uttered word in our corpus (see Section 3.1) is given ratings from at most three (at least one) reliable Workers. Therefore, each group of uttered words with the same transcription has a group of reliable ratings. We calculate the average of each group of ratings as the word mispronunciation rating for the specific pronunciation. For example, in Table 4a, the word "raid" has two pronunciations: "r ay d" which was uttered by Speakers 1 and 2, "w eh d" which was uttered by Speakers 3. In Table 4b, since Speakers 1 and 2 provided the same pronunciation "r ay d", we group the corresponding ratings "4,3" and "4,4,4" together, and calculate the average. The average value is 3.8, which is considered as the word mispronunciation rating for the pronunciation "r ay d". The average of the ratings "4,3,4" is 3.67, which is considered as the rating for the pronunciation "w eh d".

Table 4a. & 4b. *An example of word mispronunciation ratings for different transcriptions.*

(a). ratings for words

Speaker	Word	Pronunciation	Rating
1	raid	r ay d	4,3
2	raid	r ay d	4,4,4
3	raid	w eh d	4,3,4

(b). ratings for pronunciations

Word	Pronunciation	Rating	Average
raid	r ay d	4,3,4,4,4	3.8
raid	w eh d	4,3,4	3.67

Based on the procedures described above, we derive the mispronunciation ratings for all uttered words in the corpus. Some examples of words with different mispronunciation grades (see Section 3.2) are shown in Table 5.

Table 5. *Examples of uttered words with different mispronunciation grades.*

Word	Reference	Pronunciation	Rating	Grade
book	b uh k	b uh k	1	No error
adopted	ax d aa p t ix d	ax d ao p t ix d	1.31	Subtle
access	ae k s eh s	ax k s eh s	2.01	Subtle
lame	l ey m	l ae m	2.21	Subtle
wiper	w ay p axr	w ae p axr	2.89	Medium
tossed	t ao s t	t ow s t	3.01	Medium
alleged	ax l eh jh d	ax l er jh d	3.17	Medium
moan	m ow n	m aw ng	3.5	Salient
aching	ey k ix ng	ae ch ix ng	3.57	Salient
ash	ae sh	ay ch	4	Salient

7. CONCLUSIONS

A pedagogical improvement for feedback in CAPT systems is to focus on the most "serious" errors rather than to present all the errors to the learners at the same time. Our work uses crowdsourcing to collect word-level mispronunciation evaluation on L2 English speech, according to a four-point gradation of no error, subtle, medium and salient errors. In order to control the quality of crowdsourced data, we propose WorkerRank algorithm to filter Workers in terms of the reliability. Based on the data obtained from the reliable Workers, we derive a set of rated word ratings in terms of the severity of mispronunciations.

8. ACKNOWLEDGEMENT

The work is jointly supported by the research funds from the Hong Kong SAR Government's Research Grants Council (CUHK4161/08), the NSFC/RGC Joint Research Scheme (Project No. N_CUHK 414/09) and the Research Grants Council General Research Fund (Project No. 415511).

9. REFERENCES

[1] Neri, A., et al, "The pedagogy-technology interface in Computer Assisted Pronunciation Training" Computer Assisted Language Learning, 15: 441- 467, 2002.

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

[3] Witt, S and Young, S, "Language Learning Based on Non-Native Speech Recognition", Proc. EUROSPEECH1997: 633--636, Rhodes, Greece, 1997.

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

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

[6] Kunath, S. A. and Weinberger, S. H., "The wisdom of the crowd's ear: speech accent rating and annotation with Amazon

Mechanical Turk", Proc. CSLDAMT '10, Association for Computational Linguistics, June 2010.

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

[8] Bucholz, S., et al., "Crowdsourcing preference tests and how to detect cheating", Proc. Interspeech2011, Florence, 2011.

[9] Watkins, M. W. and Pacheco, M., "Interobserver Agreement in behavior research: Importance and Calculation", Journal of Behavioral Education, 10(4): 205-212, December 2000.

[10] Shoukri, M.M., Measures of interobserver agreement, 2004.

[11] Bland, M., "Measurement in Health and Disease: Assessing Agreement Using Cohen's Kappa", M.Sc. Course Material: Measurement in Health and Disease, Department of Health Sciences at University of York, 22 Jul. 2008, http://www-users.york.ac.uk/~mb55/msc/clinimet/week4/kappa.htm, 9 Jul. 2012.

[12] 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.

[13] Ienco, D., Meo, R., Botta, M., "Using PageRank in Feature Selection", In SEBD, 93-100, 2008.