PHONOLOGICAL MODELING OF MISPRONUNCIATION GRADATIONS IN L2 ENGLISH SPEECH OF L1 CHINESE LEARNERS

Hao Wang, Xiaojun Qian and Helen Meng

Human-Computer Communications Laboratory,
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China

ABSTRACT

Generation of corrective feedback carries significant pedagogical importance in the design of computer-aided pronunciation training systems. Such feedback generation should take into account the severity of detected mispronunciations, in order to prioritize different kinds of corrections to be conveyed to the learner. However, mispronunciation gradation is highly dependent on the phonetic context and acoustic context of the word pronunciation, as well as human perception. We have defined several categories of mispronunciation gradation, ranging from subtle to salient, and collected crowdsourced ratings from a large number of listeners. This work aims to capture the phonetic context of word mispronunciation by phonological rules, which are then augmented with statistical scoring to quantitatively model mispronunciation gradations. The model can thus be used to generate gradation ratings of word mispronunciations, especially those that are previously unseen in the training set. We will report the results of automatic gradation classification, as well as its correlation(s) with human perception.

Index Terms— CAPT, Mispronunciation gradation, Crowdsourcing

1. INTRODUCTION

Pedagogical effectiveness of corrective feedback generation is one of the key concerns in the design of computer-assisted pronunciation training (CAPT) systems. Researchers in pedagogy suggest that attention should be focused on a few salient mispronunciation error types rather than on all errors at the same time [1]. This can avoid discouraging the learners, as well as help learners prioritize their training strategies for progressive improvements. Such a learning model calls for the discrimination between "perceptually salient" versus "perceptually subtle" mispronunciations. Errors that are deemed perceptually salient tends to hamper communication and should receive high priority in pronunciation training. However state-of-the-art mispronunciation detection techniques based on automatic speech recognition primarily identifies phonetic errors and does not include a gradation in terms of perceptual severity. This situation motives our current investigation in developing a methodology for objective characterization of the severity of L2 English mispronunciations. Our approach leverages crowdsourcing in obtaining perceptual evaluation from a large number of listeners for word-level mispronunciations automatically detected from an L2 English corpus. We have devised a method to filter for "reliable" assessment of perceptual gradations from the crowdsourced data. Furthermore, we developed a linear model that can transform mispronunciation gradations based on the word unit into gradations based on phonological rules. Such rules can then be applied to previously unseen word mispronunciations in order to compute a predicted word-level gradation automatically.

Previous work in gradation of mispronunciations has relied primarily on expert knowledge. We leverage on the "wisdom of crowds" [2] and collected ratings from a large number of diverse listeners, with an aim to obtain an aggregated gradation that can closely approximate an expert's opinion. Crowdsourcing refers to outsourcing tasks to a large group of anonymous people and has recently become a popular technique for data collection, labeling and evaluation. Web-based crowdsourcing platforms are available, with a popular site being the Amazon Mechanical Turk (AMT). In AMT, Requesters publish human intelligence tasks (HITs) to the marketplace, while Workers browse and choose to work on some of the published HITs and receive micropayments upon task completion. This study collects word-level human perceptual gradations of word-level L2 English mispronunciations on AMT platform. In order to control the quality of the crowdsourced data, we proposed the "WorkerRank" [3] algorithm to filter for the gradation ratings from "more reliable" human raters. We have also designed phonological rules to model L2 English mispronunciations using a data-driven approach [4]. This is enhanced with a linear regression approach that transforms word-level mispronunciation gradations into rule-level gradations. Our previous work is based on manual phonetic transcriptions. The current work extends our research with the incorporation of acoustic models for automatic categorization of mispronunciation gradations.

The rest of this paper is organized as follows: Section 2 presents an overview of related work. Section 3 presents a brief review of our WorkerRank algorithm that is developed

to filter for "reliable" word-level mispronunciation gradations in crowdsourced data. Section 4 describes our approach for deriving phonological rules that capture the mispronunciation gradations and the rules can be used to generalize to other previously unseen word-level mispronunciations. Gradations can be predicted based on perfect and imperfect phonetic transcriptions. The former is based on manual transcriptions and the latter is dependent on the acoustic models used. We performed a comparative investigation between the two and present the results in Section 5. Finally the last section presents our summary and conclusions.

2. RELATED WORK

Crowdsourcing has previously been used in data evaluation. Related work was done by Kunath and Weinberger [5], who collected ratings for non-native English speech accent from native English listeners on AMT platform. These ratings were based on a five-point Likert scale (ranging from '1' for native accent to '5' for heavy, nonnative accent). It was stated that the objective was to create a training data set for an automatic speech accent evaluation system.

Peabody [6] also used AMT for collecting word-level judgments of pronunciation quality. Each utterance in the corpus was assigned to three AMT Workers, who were asked to provide binary judgment for each word on whether it was mispronounced. These collected data were mainly used for mispronunciation detection.

Both efforts conducted speech data assessment by crowdsourcing. However, procedures to ensure quality control of the crowdsourced data and how the data was used in predictive scoring were not provided.

3. CROWDSOURCED PERCEPTUAL GRADATIONS

This section presents a brief review of our previous effort [3] on the collection of perceptual gradation of word-level mispronunciations in L2 English speech and the procedure to filter for a "reliable" subset of the crowdsourced data.

3.1. L2 English corpus

In this study, we use the Cantonese subset of Chinese University CHinese Learners Of English (CU-CHLOE) corpus, which contains speech recordings by 100 native Cantonese speakers (50 male and 50 female) reading confusable words, phonemic sentences, minimal pairs and the Aesop's Fable "The North Wind and the Sun".

3.2. Categories for the gradation of mispronunciations

We defined the following four grades of mispronunciations:

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. Moderate: 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. Crowdsourcing task and reliable data selection

We published 200 distinct HITs, each of which contains several L2 English utterances for the AMT Workers to listen to and give word-level ratings according to the gradation criteria described in Section 3.2. Each distinct HIT was assigned to 3 workers.

The veracity of the collected data is one of the most critical issues in crowdsourcing since it has direct effect on the predictive efficacy of the trained system. When a gold standard is not well defined as in our study that involves human perception, unsupervised approaches using interworker agreement can be considered in some cases to be sufficiently accurate for measuring the quality of individual performance [7]. In our previous work [3], we devised a graph-based ranking algorithm – the WorkerRank algorithm to categorize workers in terms of their reliability. We perform data quality control by adopting only the crowdsourced data contributed by Workers who are deemed "reliable" in terms of their WorkerRank scores.

Our approach for data filtering involves first representing the Workers as nodes in an undirected weighted graph, with an edge forming between nodes if the two Workers completed common HITs. The weight for each edge is the Cohen's weighted kappa [8] value measuring the degree of inter-Worker agreement. An example is given in Figure 1.

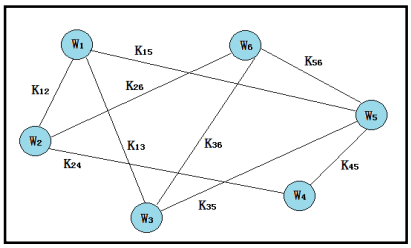

Figure 1: *An example of an undirected weighted graph representing AMT Workers and their relations.*

We developed the WorkerRank algorithm [3] which is inspired by the well-known PageRank algorithm [9] that ranks webpages. The calculated WorkerRank score (see Equation 1) reflects the reliability of a Worker based on his/her rating consistency compared with other Workers.

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

where $\vec{\boldsymbol{W}}$ is the 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 (we set $d = 0.99$ in this study), k_{ij} is the weighted kappa value between Worker w_i and Worker w_j.

Equation 1 is a recursive expression. We perform iterative calculations until convergence to obtain the resulting WorkerRank score vector. To select "reliable" Workers, we rank all of them in descending order according to their WorkerRank scores. Then we adopt the ratings from the top-ranking Worker, followed by the ratings from the next best Worker, and continue with this procedure until all the words in the corpus are covered. The selected Workers were deemed "reliable"; hence we adopt their ratings as "reliable" ratings. Our previous work [3] showed that the "reliable" subset of Workers indeed had a higher level of inter-Worker agreement ($\kappa_r = 0.47$) compared with the entire set ($\kappa_e = 0.39$).

4. PHONOLOGICAL MODELING OF MISPRONUNCIATION GRADATIONS

In our study we consider that mispronunciation gradation is highly dependent on the phonetic context of the word pronunciation.

4.1. Phonological rules

A phonetic mispronunciation can be modeled as a context-dependent phonological rule which is represented using the following expression [10]:

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

meaning phone α, which is between phone σ and phone λ, is substituted by phone β; α or β can be replaced with null symbol 0 to represent the insertion and the deletion rule, respectively; σ and λ can be replaced with symbol # to denote a word boundary.

For any word mispronunciation, we can generate corresponding phonological rules by aligning its canonical pronunciation (obtained from electronic dictionaries such as CMUDict) with the given transcription using phonetically-sensitive alignment [11].

4.2. Linear model of relationship between word-based and rule-based graduation scores

We modeled the gradation score of a word mispronunciation as a linear combination of the gradation scores of all the phonological rules occurring in that word mispronunciation [4]. The relationship can be expressed as Equation 2:

$$G_m = \sum_p \left(G_p \cdot \delta(r) \right) + b, \qquad (2)$$

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

Hence we may obtain the gradation scores of multiple word mispronunciations based on the following matrix form.

$$\vec{\boldsymbol{w}} = \boldsymbol{A}\vec{\boldsymbol{r}} + b\vec{\boldsymbol{e}}, \qquad (3)$$

where $\vec{\boldsymbol{w}} = \left[G_{m_1}, G_{m_2}, \cdots\right]^{\mathbf{T}}$ is a vector that contains the gradation score of each uttered word, $\boldsymbol{A}$ is a matrix with binary elements A_{ij} indicating whether phonological rule j occurs in uttered word i, $\vec{\boldsymbol{r}} = \left[G_{p_1}, G_{p_2}, \cdots\right]^{\mathbf{T}}$ is a vector containing the gradation scores of all the rules in the system; $\vec{\boldsymbol{e}}$ is the all-one vector.

Based on the above model, when we have the gradation scores and the transcriptions of a set of word mispronunciations, we can conveniently apply the above model to predict the gradation scores of word mispronunciations in testing data.

5. EXPERIMENTS

The experiments are carried out using the Cantonese subset of CU-CHLOE corpus (see Section 3.1). All speech data of the corpus are phonetically labeled by trained linguists. 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 phonological rules are generated, which fully cover all phonetic mispronunciations in the training set. Based on the "reliable" ratings (see Section 3.3), we derive the gradation score of each word mispronunciation in the corpus by averaging all its corresponding "reliable" ratings. Then, we are ready to predict mispronunciation gradation.

5.1. Prediction based on manual phonetic transcriptions

Manual transcriptions are used both in training and testing. According to Equation 3, we use training data set to run least-square linear regression analysis to obtain the gradation scores of all 2,347 phonological rules (generated from the training set) and the offset term b in Equation 3. The following example illustrates that to some extent, the obtained gradation scores reflect the severity of mispronunciations captured in the phonological rules: the mispronunciation "b l ao d" for the word "block" has two phonological rules "aa → ao / l _ k" with the score of -0.136 and "k → d / aa _ #" with the score of 2.511; the latter rule represents a more "serious" error perceptually and it corresponds to a much higher gradation score than the former rule. For testing, since we have manual transcriptions of the test data set, we can easily obtain the phonological rules, and then, we can get the gradation score of a word mispronunciation by adding up all the trained gradation scores of the phonological rules occurring in that word and the offset term.

However, this prediction procedure has an obvious limitation: in practical systems, manual transcriptions are usually not available immediately. Hence we are motivated to incorporate acoustic models into gradation scoring.

5.2. Prediction based on automatically recognized transcriptions

We create a dictionary for the speech recognizer by building extended recognition network (ERN) for each of the word in the corpus. An ERN is extended from a standard recognition network (SRN) that is built by using the canonical pronunciation of a word. By applying given phonological rules which are represented as finite state transducers [11,12] to an SRN, we can derive the corresponding ERN. Traversing all the paths present in an ERN generate all possible mispronunciations of a word according to the given phonological rules. Figure 2 shows an example of an ERN.

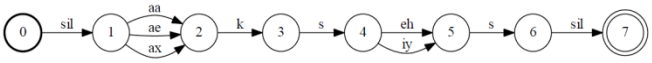

Figure 2. *Extended recognition network of the word "access" by applying the following 3 phonological rules: ae → aa / # _ k, ae → ax / # _ k, and eh → iy / s _ s.*

We find that many of the 2,347 generated rules (from the training data set) have rare (3 or fewer) occurrences; these rules are probably due to misreading or guessed pronunciations for words unfamiliar to the speakers. Therefore, in order to reduce the number of possible noisy pronunciations, we prune those rules with rare (3 or fewer) occurrences. After pruning, 765 rules are remained, which are used to build an ERN for each distinct word in the corpus using AT&T Finite State Machine Library [13] and create a dictionary that covers a number of possible word mispronunciations.

Then, we use a standard Maximum-likelihood HMM model trained on L1 English speech (from the TIMIT corpus) with 39 dimensional PLP features [14] to generate recognized transcriptions, and apply them both in training and testing based on the linear regression model (see Section 4.2).

5.3. Evaluation and discussion

We calculate correlation and Cohen's weighted kappa [3,8] between human-labeled gradations (i.e. the average of the crowdsourced "reliable" ratings for each uttered word) and machine-predicted gradations for each type of above predictions. To calculate weighted kappa values, we first quantify all the word gradation scores (by rounding) to 4 integer values {1,2,3,4} representing 4 possible grades of errors (see Section 3.2); for those words whose gradation scores exceed the range from 1 to 4; we quantify those gradation scores to their nearest grade values (1 or 4). The evaluation results are given in Table 1.

Table 1. *Evaluation results of predictions based on the manual transcription (MT) and the acoustic model (AM) used.*

	MT	***AM***
Correlation	0.64	0.53
Weighted kappa	0.59	0.45

From Table 1, we find that the prediction based on manual transcriptions achieves better performance in terms of the two evaluation measures, since the manual transcriptions are deemed perfect. There are two observations that may affect the gradation accuracy of the prediction based on acoustic models. First, in the test set, some correctly pronounced words are recognized as word mispronunciations. Based on the statistics shown in Table 2, precision and recall values for the recognized word mispronunciations in the test data set are obtained as shown in Table 3:

Table 2. *Numbers of words in the test set belonging to the actual class versus the recognized class in terms of whether the corresponding words are mispronounced.*

		Actual class	
		Correct	*Misp.*
Recognized class	*Correct*	7512	2318
	Misp.	8093	13616

Table 3. *Precision and recall for recognized word mispronunciations in the test data set.*

Recognized word mispronunciations	*Precision*	*Recall*
	62.72%	85.45%

"Precision" means the actual word mispronunciations that are recognized as mispronounced over all recognized word mispronunciations. "Recall" is the actual word mispronunciations that are recognized as mispronounced over all actual word mispronunciations. Since our task is to grade mispronunciations, we focus on precision value. Precision of 62.72% means there are 37.28% of predicted gradations are given to actually correctly pronounced words, which results in a decrease in gradation accuracy. Second, actual word mispronunciations that are recognized as mispronounced may not get their actual phonetic transcriptions recognized exactly. This is the other factor that can affect the performance of gradation prediction based on acoustic models. However, as we can infer, better acoustic models can enhance gradation accuracy by improving precision and phonetic-level recognition accuracy, and ultimately make the performance get close to that of the prediction based on perfect transcriptions.

6. CONCLUSIONS

This paper presents our effort on gradation of word-level L2 English mispronunciations. Based on the crowdsouced "reliable" perceptual ratings of word mispronunciations and phonological modeling of mispronunciation gradation, we compare the predictions using perfect and imperfect phonetic transcriptions. The prediction based on manual transcriptions performs better; however, the prediction using acoustic models is more practical in real systems and the performance can be improved by refining the acoustic models.

7. 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).

8. REFERENCES

[1] R. Ellis, “Corrective Feedback and Teacher Development”, L2 Journal, Vol. 1, pp. 3-18, 2009.

[2] J. Surowiecki, “The Wisdom of Crowds”, Garden City: Doubleday. – Introduction, Kapitel 1 & 2 (pp. XI-XXI & 3-39), 2004.

[3] H. Wang and H. Meng, “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] H. Wang, X. J. Qian and H. Meng, “Predicting Gradation of L2 English Mispronunciations using Crowdsourced Ratings and Phonological Rules”, in Proc. of Speech and Language Technology in Education (SLaTE 2013), Grenoble, France, 30 - 31 August & 1 September, 2013.

[5] S. A. Kunath, and S.H. Weinberger, “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.

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

[7] M. Eskenazi, G.A. Levow, H Meng, G Parent and D. Suendermann, “Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment”, Wiley Publishing, (1st ed.), 2013.

[8] M.M. Shoukri, “Measures of interobserver agreement”, 2004.

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

[10] W.K. Lo, S. Zhang and H. Meng, “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.

[11] A.M. Harrison, W.K. Lo, X.J. Qian and H. Meng, “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.

[12] R. M. Kaplan and M. Kay, “Regular models of phonological rule systems”, Computational Linguistics, vol. 20, no. 3, pp. 331–378, 1994.

[13] M. Mohri, F.C.N. Pereira and M.D. Riley, “AT&T FSM Library v3.7”, http://www2.research.att.com/~fsmtools/fsm/, 1998.

[14] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech”, The Journal of the Acoustical Society of America, vol. 87, pp. 1738–1752, 1990.