Predicting Gradation of L2 English Mispronunciations Using ASR with Extended Recognition Network

Hao Wang, Helen Meng and Xiaojun Qian
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China
E-mail: {hwang, hmmeng, xjqian}@se.cuhk.edu.hk Tel: +852-39438327

***Abstract*—A CAPT system can be pedagogically improved by giving effective feedback according to the severity of mispronunciations. We obtained perceptual gradations of L2 English mispronunciations through crowdsourcing, conducted quality control to filter for *reliable ratings* and proposed approaches to predict gradation of word-level mispronunciations. This paper presents our work on making improvements using ASR with extended recognition network to the previous predicting approach to solve its limitations: 1. it is not working for those mispronounced words whose transcriptions are not immediately available; 2. perceptually differently articulated words with the same transcription have the same predicted gradation.**

I. INTRODUCTION

Research interest in the pedagogical effectiveness of computer-assisted pronunciation training (CAPT) technology is increasing in recent years. Based on the generation of corrective feedback in CAPT for L2 learning, methodologists suggest teachers focus their attention on a few error types rather than try to address all the errors [1] so that learners can prioritize their errors. One criterion for defining the priority is perceptual relevance – "subtle" mispronunciations without affecting intelligibility greatly may usually be tolerated by listeners; but those perceptually "serious" errors which hamper communication have to be corrected. Therefore, we are motivated to make a pedagogical improvement to a CAPT system by providing effective feedback through prioritizing detected mispronunciations according to their severity.

We collected perceptual gradations of word-level mispronunciations using crowdsourcing and conducted quality control to filter the crowdsourced results in terms of reliability by the WorkerRank algorithm [2]. Based on the crowdsourced *reliable* data, we proposed two approaches for predicting the gradation of word-level mispronunciations [3]. However, the proposed approaches have their limitations. In this paper, we illustrate the limitations of the previous approaches and propose an improved version using ASR with extended recognition network.

II. RELATED WORK

Our work collects human perceptual ratings of L2 English speech to develop some predictive model for human ratings according to different levels of severity of word-level mispronunciations. Related previous studies include:

Witt and Young [4] got a database of non-native utterances scored at both sentence and word levels on a scale of 1 to 4 by trained phoneticians. Kim et al. [5] conducted experiments using phonetically labeled nonnative French speech; a panel of five teachers of French was invited to rate the pronunciation quality of the selected phone segments on a scale of 1 (unintelligible) to 5 (native-like). The data in both of the above studies were mainly collected and used for performance evaluation rather than training for predictive scoring (as is done in our work).

The pronunciation quality in the above two work was labeled by a small number of experts. Another technique called crowdsourcing, which has been widely used for data collection and labeling in recent years, is a process of obtaining needed services, ideas or content by soliciting contributions from a large undefined group of people. In comparison with traditional methods for data collection and labeling, crowdsourcing is considerably more efficient, cost-effective and diversified. Amazon Mechanical Turk (AMT)[1] is one of the best known crowdsourcing platforms. It provides 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 [6].

Kunath and Weinberger [7] used AMT crowdsroucing service to collect English speech accent ratings from native English listeners. AMT Workers were asked to rate accentedness of the given speech on a five-point Likert scale (ranging from '1' for native accent to '5' for heavy, nonnative accent). The crowdsourced results were intended to be used as a training data set for an automatic accent evaluation system; but the paper did not give information about how to train such an automatic system

III. CROWDSOURCED MISPRONUNCIATION GRADATIONS

Our previous work [2] collected perceptual gradations of word-level mispronunciations in non-native English speech using the AMT crowdsourcing platform. This section presents a brief description of our crowdsourcing procedure and how we filter for *reliable* crowdsourced results.

A. L2 Corpus

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

The corpus we used is the Cantonese subset of the Chinese University Chinese Learners of English (CU-CHLOE) Corpus which contains speech recordings by 100 Cantonese speakers (50 males and 50 females) reading several types of carefully designed material as shown in TABLE I.

TABLE I
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 common representative examples of mispronunciations from Cantonese learners of English. Each of the 100 speakers read out all the 86 prompted texts containing 436 unique words from a total of 631 words.

B. Possible Gradation of Errors

We defined four grades of mispronunciation 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. 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.

C. Crowdsourcing Procedure and Reliable Results Selection

We published several distinct HITs, each of which contains a bunch of L2 English utterances for AMT Workers to rate; they were asked to rate every articulated word in the utterances according to the gradation criteria described in the previous subsection. Each distinct HIT was assigned to 3 AMT Workers.

To control the quality of the crowdsourced data, we use the WorkerRank algorithm [2] to identify and select *reliable Workers* and adopt their ratings as *reliable* ones, based on an assumption that *reliable Workers* will always provide *reliable ratings*.

For each articulated word in the corpus, we average across all its corresponding *reliable ratings* and consider the calculated average as the gradation score of that word; all these calculated average values are used in our further experiment on predicting gradations of word-level mispronunciations.

IV. PREDICTING MISPRONUNCIATION GRADATIONS

Our previous work [3] modeled the relationship between phonetic mispronunciations and the actual word-level gradations using an approach based on linear regression. The result showed reasonable correlation and agreement between human-labeled and machine-predicted gradations. This section presents a recap of the linear regression approach for predicting the gradation of word-level mispronunciations, and proposes an improved version by using ASR with extended recognition network.

A. Phonological Rules

Phonetic mispronunciation productions can be represented as context-dependent phonological rules of the form [3,8]:

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

which denotes that phone α is substituted by phone β, when it is preceded by phone σ and followed by 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.

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 [9], 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.

B. Approach based on Linear Regression

We modeled the gradation of a word mispronunciation as a linear combination of the gradation scores of all phonological rules occurring in that word mispronunciation [3]. This relationship can be expressed as:

$$G_w = \sum_r \big(G_r \cdot \delta(r)\big) + b, \qquad (1)$$

where G_w is the gradation score of a mispronounced word 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-level mispronunciation gradations can be expressed in a matrix form as follows:

$$\boldsymbol{G_w} = \boldsymbol{A G_r} + b\boldsymbol{e}, \qquad (2)$$

where $\boldsymbol{G_w}$ is a vector containing the gradation score of each articulated word, which is calculated by averaging across the corresponding *reliable ratings* of that word; $\boldsymbol{A}$ is a matrix

with binary elements A_{ij}, indicating whether the phonological rule j occurs in the word i; $\boldsymbol{G_r}$ is a vector that contains the gradation score of each rule in the system; $\boldsymbol{e}$ is the all-one vector.

Running least-square linear regression analysis for the training data set, a rule score vector $\boldsymbol{G_r}$ and the offset term b are obtained, and are used for predicting word-level mispronunciation gradations by Equation 1. In [3], we predicted the gradation of word-level mispronunciations using the given manual transcriptions of the test data set. Therefore, there are two limitations of this approach: 1. it is not working for those mispronounced words whose transcriptions are not immediately available; 2. perceptually differently articulated words with the same transcription have the same predicted gradation.

C. *Extended Recognition Network*

A standard recognition network (see Figure 1) can be built by using the canonical pronunciation of a word. This network is extended with derived phonological rules which are represented as finite state transducers [9,10]. This extended recognition network (see Figure 2) is used to generate all possible mispronunciations of a word according to the derived phonological rules.

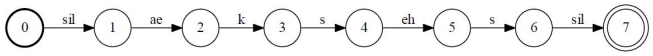

Figure 1. Standard recognition network of the word "access"

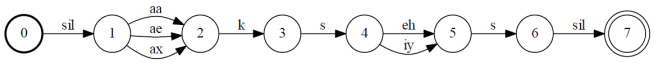

Figure 2. Extended recognition network of the word "access" by applying 3 phonological rules: ae → aa / # _ k, ae → ax / # _ k, eh → iy / s _ s.

D. *Using ASR with N-Best Recognized Transcriptions*

Using ERNs of the involved words, we can generate a dictionary that covers all the possible mispronunciations. An ASR system using this generated dictionary can solve the first limitation of the previous linear regression approach. Instead of outputting the best recognized transcription, we use an ASR system to output N-best recognized results, each of which has a probability (the probabilities of the recognized transcriptions for each articulated word are normalized such that they sum up to 1). This solves the second limitation that perceptually differently articulated words with the same transcription have the same predicted gradation. E.g., we have two utterances A and B by two speakers uttering the word "access". Utterances A and B have the same manual transcription although they are to-some-extent perceptually different. An ASR system is used to model the human perception by outputting expectedly different groups of recognized transcriptions with different probabilities for utterances A and B, since they have different speech signals.

Therefore, we can improve the previous linear regression approach as follows:

$$\overline{G_w} = \sum_m \left(P_m \cdot \sum_r \left(\overline{G_r} \cdot \delta(r)\right)\right) + \bar{b}, \tag{3}$$

where $\overline{G_w}$ is the gradation of a mispronounced word w; P_m is the normalized probability ($\sum_m P_m = 1$) of a recognized transcription m for the word w; $\overline{G_r}$ is the gradation score of the rule r; $\delta(r)$ is an indicator function, i.e. $\delta(r) = 1$ if r occurs in this transcription m, and $\delta(r) = 0$, otherwise; $\bar{b}$ is the offset term. The inner summation is taking over all r in the system; the outer summation is taking over all recognized transcriptions of this word w.

Multiple word-level mispronunciation gradations can be expressed in the following matrix form:

$$\overline{\boldsymbol{G_w}} = \boldsymbol{P}\boldsymbol{M}\overline{\boldsymbol{G_r}} + \bar{b}\boldsymbol{e}, \tag{4}$$

where $\overline{\boldsymbol{G_w}}$ is a vector containing the gradation score of each articulated word; $\boldsymbol{P}$ is a matrix with elements P_{ij} being the normalized probability of the recognized transcription j for the articulated word i; $\boldsymbol{M}$ is a matrix containing the elements M_{xy} indicating whether the phonological rule y occurs in the recognized transcription x; $\overline{\boldsymbol{G_r}}$ is a vector that contains the gradation scores of all rules; $\boldsymbol{e}$ is the all-one vector.

We obtain a rule score vector and an offset term by running least-square linear regression analysis for the training data set; and they are used to predict word-level mispronunciation gradation by Equation 3, e.g., for an articulated word "access", the system outputs 2-best recognized transcriptions: "ax s eh s" and "ax s eh sh" with normalized probabilities 0.64 and 0.36, respectively; we derive two phonological rules for the transcription "ax s eh s": "ae → ax / # _ k" and "k → 0 / ae _ s", and the corresponding gradation scores of these two rules obtained from the regression analysis are 0.74 and 0.44; we derive three phonological rules for the transcription "ax s eh sh": "ae → ax / # _ k", "k → 0 / ae _ s" and "s → sh / eh _ #" with the corresponding gradation scores 0.74, 0.44 and 0.07; thus, with the trained offset term $b = 1.38$, we calculate the gradation of this articulated word according to Equation 3 as: $0.64 \times (0.74 + 0.44) + 0.36 \times (0.74 + 0.44 + 0.07) + 1.38 \approx 2.59$.

V. EXPERIMENTS

A. *Procedure*

We split the corpus by speakers into disjoint training (25 males and 25 females) and test (25 males and 25 females) sets. 2,347 distinct context-dependent phonological rules are generated, which fully cover all phonetic mispronunciations in the training set. However, taking a closer look into the data, we find that many of the 2,347 generated rules have rare occurrences; these rules are probably due to misreading or guessed pronunciations for words unfamiliar to the speakers. Therefore, we prune those rules with rare (3 or less) occurrences in order to avoiding a big number of possibly noisy transcriptions generated from ERNs. After pruning, we have 765 rules. Using these rules, we build an ERN for each of the 436 distinct words in the corpus using AT&T Finite State Machine Library [11] and create a dictionary that covers a number of possible mispronunciations. In this experiment,

we use the tool HVite in HTK [12] to generate 2-best and 3-best outputs for testing; for each group of the outputs, we train a set of gradation scores of the phonological rules according to the approach described in the previous section. Each set of the trained scores is used to predict the gradation of the word-level mispronunciations in both training and test sets for comparison.

We calculate correlation and Cohen's weighted kappa [3,13,14] between human-labeled gradations (i.e. the average of the crowdsourced *reliable ratings* for each articulated word) and machine-predicted gradations by ASR with each of the 2-best and 3-best outputs for both training and test sets. To calculate the 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 III); a small number (less than 2% of total number of words) of the 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 shown in TABLE II.

TABLE II
EVALUATION RESULTS FOR 765 RULES ON BOTH TRAINING AND TEST SETS USING ASR WITH BOTH 2-BEST AND 3-BEST OUTPUTS, COMPARED WITH THE RESULTS USING THE PREVIOUS LINEAR REGRESSION APPROACH (BASED ON MANUAL TRANSCRIPTIONS).

765 rules	**2-best**		**3-best**		**Previous LR**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Correlation	0.520	0.388	0.520	0.383	0.765	0.644
Kappa	0.411	0.327	0.411	0.317	0.690	0.588

B. Discussion

The evaluation results in TABLE II show that there is hardly any notable difference between using 2-best and 3-best outputs in this experiment. The reason is that most of the first best recognized transcriptions are output with probabilities that are close to 1; thus the corresponding third best outputs have probabilities near 0, which make those third best outputs have negligible impact on the gradation results.

If we compare the results of the approach using ASR with the previous linear regression approach (based on manual transcriptions of all the words), the performance of the approach using ASR is not as good as the previous approach. The possible reasons are as follows: 1. the phonological rules we use in building ERNs do not have an expected coverage of mispronunciations; 2. the acoustic model we use in this experiment is not good enough to handle this task.

VI. CONCLUSIONS AND FUTURE WORK

This paper presents our work on predicting the gradation of word-level L2 English mispronunciations using our crowdsourced perceptual ratings. Based on our previous linear regression approach [3], we propose an improved version to solve the limitations of the previous approach by using ASR with extended recognition network.

The evaluation results illustrate that the correlation and agreement are not significant. Therefore, there is still big room for improvement on the performance of the predicting approach. In future work, we will try other set of phonological rules that have a better coverage of mispronunciations and use different acoustic models for testing. Furthermore, we will try other regression analysis to seek a better model.

ACKNOWLEDGMENTS

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

[2] 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.

[3] 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.

[4] S. Witt and S. Young, "Language Learning Based on Non-Native Speech Recognition", in Proc. of EUROSPEECH1997, pp. 633-636, 1997.

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

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

[7] 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.

[8] 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.

[9] 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.

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

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

[12] HTK, http://htk.eng.cam.ac.uk

[13] J. Cohen, "A coefficient of agreement for nominal scales", Educational and Psychological Measurement, 20 (1): 37–46, 1960.

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