Learning Pronunciation and Accent from The Crowd

Frederick Liu, Jeremy Chiaming Yang and **Jane Yung-jen Hsu**
Department of Computer Science and Information Engineering
National Taiwan University
{b99902007, r98922038, yjhsu}@csie.ntu.edu.tw

Abstract

Learning a second language is becoming a more popular trend around the world. But the act of learning another language in a place removed from native speakers is difficult as there is often no one to correct mistakes nor examples to imitate. With the idea of crowd sourcing, we would like to propose an efficient way to learn a second language better.

Introduction

There are several aspects in learning a language, with the two most essential aspects being the ability to write and speak it. (Bernstein et al. 2010) discussed proof writing with the crowd. In our study, we focus on the speaking abiltiy. Although there have been some computer applications for practicing speaking, the learning effect is also constrained by the quantifiable data. Learning the proper pronunciation and accent can be hard to improve without a suitable environment prividing unquantifiable imformation.

In our work, we developed a workflow similar to the crowdsourcing pattern "Find-Fix-Verify" in (Bernstein et al. 2010). We further enhanced the verify step with an ensemble system involving two steps. The Two-step validation utilizes both the crowd and computing power. The diversity of crowds gives us the advantage of pointing out pronunciation mistakes precisely and adapting accents all over the world. We started our work with English learning and leveraged Amazon Mechanical Turk where 92% of workers are English speakers(Ross et al. 2010).

Work-Flow

We developed a platform powered by crowd sourcing for non-native speakers to upload their text and record a recording of their text for correct pronunciation (see Figure 1). The recording will be corrected by a native speaker. The whole process can be divided into three major steps: Sentence Recording, Sentence Correction and Corrected Sentence Validation.

Recording Stage

In this stage, users record their sentences they want to be corrected and upload the corresponding texts.

Correction Stage

For the correction stage, we focused on the correction of pronunciation of English words instead of phrases or the whole sentence. This is mainly due to two reasons. First, words can avoid the ambiguity that phrases bring. Second, users may not be able to concentrate on long sentences and may lead to a poor result. In this stage, users are asked to listen to the recording of a sentence and tick the words that are pronounced incorrectly or different from their own pronunciation. Users then record their pronunciation of the words they ticked. Users can submit without clicking a word if they do not think there is a mistake.

Two-step Validation Stage

Step1: Validation by Speech Processing

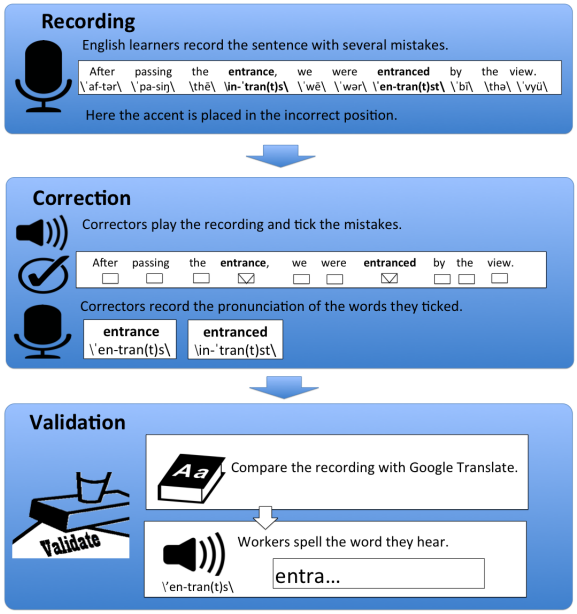

Figure 1: Work Flow

We used the speech file of Google Translate as a baseline and compared the similarity of the recorded word with it. In order to calculate the similarity of the two recordings, we extracted MFCC features and calculated the similarity with dynamic time wrapping. The result is represented as a score in the range of [0,1]. Score lower than a threshold will be eliminated from the correction list of the crowd.

Step2: Validation by Crowd

Users in this step are asked to listen to 5 recordings generated from the correction stage and type the word they heard. We calculate the Levenshtein distance between the word that workers typed and the word being corrected previously. The distance is normalized in the range of [0,1] as a score of the recording. Score higher than a threshold will be selected as the final correction.

The results are displayed to the users as in Figure 2.

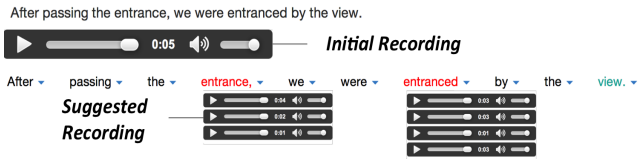

Figure 2: User Interface

Experiment on Amazon Mechanical Turk

In our experiment, we released 30 recordings of different sentences recorded by two non-native speakers while each sentence contains one to three pronunciation errors.(see Table 1). Table 2 provides our experiment settings.

Type	Example	Correct	Wrong
accent	entrance(n.)	\ˈen-trən(t)s\	\en-ˈtrən(t)s\
vowel	certificate (n.)	\sər-ˈti-fi-kət\	\sər-ˈti-fə-kāt\

Table 1: Error type

Stage	# of HITs	cost/HIT	# of data
Correction	100	$0.1	189
Validation by Crowd	97	$0.05	459

Table 2: Experiment settings

The validation stage include two steps: validation by speech processing and validation by crowd. According to Figure 3, the first step eliminates 21.2% of the recording from the correction stage. These recordings had issues regarding silence, back ground noise or simply an incorrect word. This step helps reduce the workload of the next step. The second step, validation by crowd reduced the number recordings by another 36.4% of the toal recordings. After these two steps, 42.4% of the corrected recording are left to the user. Through these two steps, we would leave the recordings of suggested pronunciation to the users with relatively high quality. We experimentally evaluate our result with Table 3. We classify the results into three groups: corrections made by U.S. and Canada, India and the third group being U.S., Canada and India together. The corrections made by the Americans and Canadiens were better under our settings because we regard the American pronunciation as our ground truth. However, for those seeking to learn British English might prefer people from the UK do the corrections.

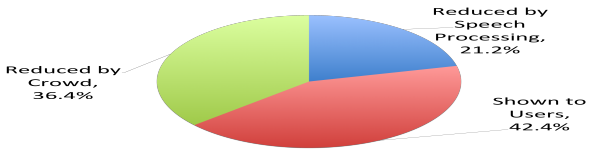

Figure 3: Percentage Reduced by Validation

	US+CA	IN	All
Accuracy (Acc)	93.95%	84.88%	89.04%
True Positive Rate (TPR)	80.90%	33.33%	54.82%
False Positive Rate (FPR)	27.27%	61.70%	44.04%

Table 3: Statistics

Conclusion and Future Work

We experimented our idea on MTurk with 30 recorded English sentences. Through our experiment, we noticed that the accuracy of workers differs from their nationalities. Since the recorders of the sentences learned their English based on American English, the result seems reasonable. The result of the experiment showed the potential our platform can bring. We would like to allow users to select the nationality of the workers correcting their recordings so the users can decide the accent and pronunciation they would like to learn.

Acknowledgment

This work is supported in part by National Taiwan University, Intel Corporation and Ministry of Science and Education under Grants NTU103R7501, NSC102-2911-I-002-001, NSC 101-2627-E-002 -002 and MOST 103-3113-E-002-008.

References

Bernstein, M. S.; Little, G.; Miller, R. C.; Hartmann, B.; Ackerman, M. S.; Karger, D. R.; Crowell, D.; and Panovich, K. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 313–322. ACM.

Ross, J.; Irani, L.; Silberman, M.; Zaldivar, A.; and Tomlinson, B. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, 2863–2872. ACM.