Sentence selection for automatic scoring of Mandarin proficiency

Jiahong Yuan[1], Xiaoying Xu[2], Wei Lai[2], Weiping Ye[2], Xinru Zhao[2], Mark Liberman[1]

[1]Linguistic Data Consortium
3600 Market St., Suite 810
Philadelphia, PA 19104, USA

[2]Beijing Normal University
19 Xinjiekou Wai Street
Haidian district, Beijing 100875, China

Jiahong@ldc.upenn.edu, xuxiaoying2000@bnu.edu.cn, laiwei_0508@126.com
yeweiping@bnu.edu.cn, xrzhao@bnu.edu.cn, myl@ldc.upenn.edu

Abstract

A central problem in research on automatic proficiency scoring is to differentiate the variability between and within groups of standard and non-standard speakers. Along with the effort to improve the robustness of techniques and models, we can also select test sentences that are more reliable for measuring the between-group variability. This study demonstrated that the performance of an automatic scoring system could be significantly improved by excluding "bad" sentences from the scoring procedure. The experiments on a dataset of *Putonghua Shuiping Ceshi* (Mandarin proficiency test) showed that, compared to all available sentences, using only best-performed sentences improved the speaker-level correlation between human and automatic scores from r = .640 to r = .824.

1 Introduction

Automatic scoring of spoken language proficiency has been widely applied in language tests and computer assisted language learning (CALL) (Wang et al., 2006; Zechner et al., 2009; Streeter et al., 2011). A central problem in this research area is to differentiate the variability between and within groups of standard and non-standard speakers. One way to tackle the problem is, as done in most previous studies, to improve the robustness and reliability of techniques and models. There is also another way to look at the problem: not every sentence is equally good for revealing a speaker's language proficiency. The purpose of this study is to demonstrate that, given an automatic scoring technique, we can significantly improve the performance of the technique by selecting well-performed sentences (with respect to the given technique) as input for scoring.

Most of the automatic scoring systems rely on automatic speech recognition (ASR). The common practice is to build HMM-based acoustic models using a large amount of "standard" speech data. To assess an utterance, pronunciation scores such as log likelihood scores and posterior probabilities are calculated by performing speech recognition (or forced alignment if the sentence is known) to the utterance based on the pre-trained acoustic models (Franco et al., 1997; Neumeyer et al., 2000; Witt and Young, 2000; Yan and Gong 2011; Hu et al., 2015). Prosody scores, e.g., duration, F_0, and pauses, have also been shown important (Cucchiarini et al., 2000; Nava et al., 2009). These individual scores are combined with statistical models such as linear regression, SVM, and neural network to produce an overall score for the test utterance (Franco et al., 2000; Ge et al., 2009).

The performance of model-based automatic scoring systems much depends on the amount and quality of the training data. For the purpose of this study, we adopted a simple, comparison-based approach. This approach is to measure the goodness of a test utterance by directly comparing it to a standard version of the same sentence and calculating the distance between the two (Yamashita et al., 2005; Lee and Glass, 2013).

2 Data

We used a dataset of *Putonghua Shuiping Ceshi* (PSC) from Beijing Normal University. PSC is the national standard Mandarin proficiency test in China, which is taken by several million people each year. The test consists of four parts: The first two parts are to read 100 monosyllabic and 50 disyllabic words; the third part is to read an article of 300 characters, randomly selected from a pool of 60 articles; and the last part is to speak freely on a given topic. The four parts are graded separately with a numeric score, and the total score (out of 100 points) is converted to a categorical proficiency level.

Our dataset consists of recordings of ~800 college students at Beijing Normal University who took the PSC test in 2011 and the grades they received on the test. We only used the part of article reading in this study. The students who read an article being selected for less than 9 other students (i.e., the total number of students reading that article is less than 10) were excluded. The final dataset contains 630 speakers reading 42 articles. Each student was graded by two examiners. The distribution of the examiners' scores on this part (out of 30 points, averaged by two examiners' scores) is shown in Figure 1. The correlation between the two examiners' scores on this part is r = 0.819.

As a demonstration, two professional voice talents have recorded the 60 articles in PSC (one male and one female, each read 30 articles). We used their spoken articles as a reference standard to which the students' were compared.

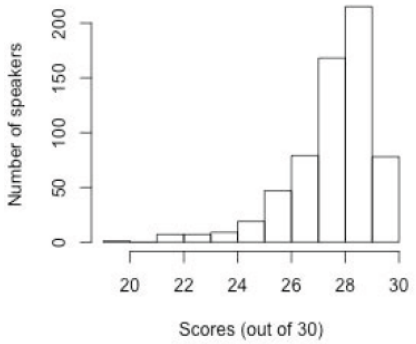

Fig. 1. Distribution of human scores in the dataset.

3 Method

Using a state-of-the-art Mandarin forced aligner (Yuan et al., 2014), we extracted utterances (delimited by a punctuation mark in the text) from the spoken articles and also obtained phonetic boundaries in the utterances. All utterances from a speaker share the same proficiency score, which is the average of the two examiners' scores the speaker received on the test.

In the dataset, every sentence has at least 10 utterance versions, each from a different speaker, plus one standard version. The goodness of a sentence to be used for automatic scoring is measured by the correlation between the distances of the students' utterances from the standard version and the utterances' proficiency scores, as shown in Figure 2. We expect negative correlations for "good" sentences: a greater difference from the standard version should result in a lower proficiency score.

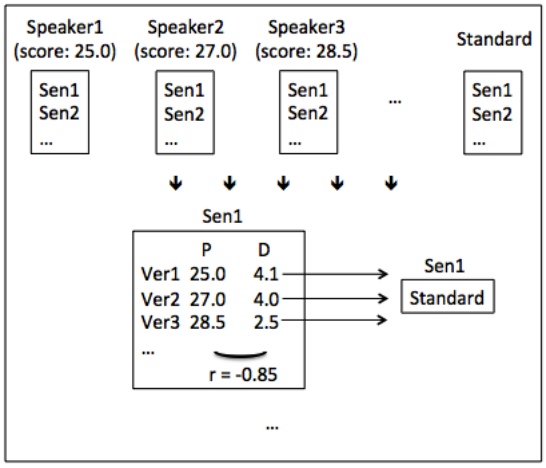

Fig. 2. Paradigm for measuring sentence goodness.

The distance between an utterance and its standard version was calculated, respectively, on three acoustic dimensions: duration, F_0, and spectrum. For each of the distance measures, an experiment was conducted using the top 10%, 20%, …, 100% sentences to obtain a distance score for every speaker, i.e., the average distance of all utterances of the speaker. The correlation between the speakers' distance scores and their human-graded proficiency scores are reported to show the effect of sentence selection.

Finally, we combined the three distance scores based on duration, F_0, and spectrum, plus a statistic of pauses, to build an automatic scoring system, and compared the performance of the system between using all available sentences and using best-performed sentences only.

4 Experiments and results

4.1 Sentence selection based on duration

The distance on duration between a test utterance and its standard version was calculated from the root mean square difference between paired segments (syllables, phones, or words) in the utterances, as shown in (1). Segment durations were derived from forced aligned boundaries.

$$D_{dur} = \sqrt{\frac{\sum_{i=1}^{n} (d_{test,i} - d_{ref,i})^2}{n}} \quad (1)$$

where $d_{test,i}$ is the duration of the i*th* segment in the test utterance, $d_{ref,i}$ is the duration of the i*th* segment in the standard utterance, and n is the total number of segments in an utterance.

To remove the effect of speaking rate on the duration distance, the segment durations in the test utterance were normalized in a way that the

total duration of the test utterance (excluding pauses) is the same as that of the standard one, as shown in (Norm 1.1):

$$d_{test,i} = d_{test,i} * \frac{\sum_{k=1}^{n} d_{ref,k}}{\sum_{k=1}^{n} d_{test,k}} \quad \text{(Norm 1.1)}$$

Figure 3 shows the correlation (-1*r) between the speakers' duration distance scores and their proficiency scores when using all sentences, top 90%, top 80%, …, and top 10% sentences (as described in Section 3). We can see that the correlation increases when excluding more "bad" sentences from being used for calculating the duration distance scores. With respect to the performance of different types of segments, syllables and words are better than phones.

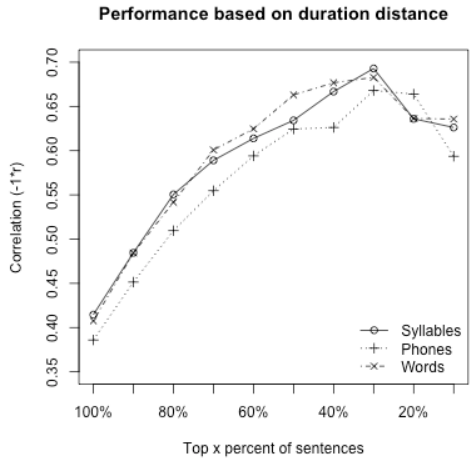

Fig. 3. Duration distance: different types of segments.

Another way to normalize the segment duration is to transform the durations to Z-scores per spoken article, as shown in (Norm 1.2).

$$d_{test,i} = \frac{d_{test,i} - \mu_{test,article}}{\sigma_{test,article}}$$
$$d_{ref,i} = \frac{d_{ref,i} - \mu_{ref,article}}{\sigma_{ref,article}} \quad \text{(Norm 1.2)}$$

where μ is the mean of the durations of all segments in the spoken article; σ is the standard deviation of the durations.

Figure 4 compares the performance of the two normalization methods (Norm 1.1 and Norm 1.2), as well as the performance of using unnormalized durations (Raw). Syllable durations were used for the comparison. From Figure 4, we can see that the normalization using z-scores per article (Norm 1.2) outperforms the normalization based on per utterance pair (Norm 1.1). Both normalizations significantly improved the correlation, compared to using unnormalized durations.

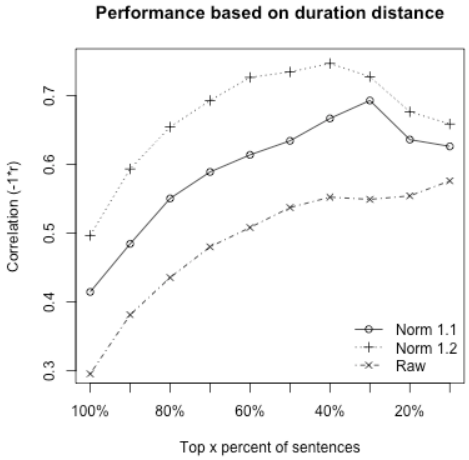

Fig. 4. Duration distance: different normalizations.

4.2 Sentence selection based on F_0

The F_0 contours of the utterances were extracted using *esps/get_f0* with a 10 ms frame rate. The contours were linearly interpolated to be continuous over the unvoiced segments, and smoothed by passing them (both forward and reverse to avoid phase distortion, *filtfilt*) through a Butterworth low-pass filter with normalized cutoff frequency at 0.1.

The distance on F_0 between a test utterance and its standard version was calculated from the root mean square difference between F_0s in paired syllables. Because the number of F_0s in a syllable is determined by the syllable duration, we normalized the number of F_0s in each pair of syllables with Python spline interpolation (*scipy.interpolate.UnivariateSpline, smoothing factor = 0.001)*, for which the number of F_0s in the standard syllable was used as the normalized number. After the normalization, the distance was calculated using all F_0s in an utterance.

The values of F_0s were also normalized to remove the effects of pitch range (e.g., female is higher than male). *Z*-scores were used for the normalization, calculated both per utterance (Norm 2.1) and per article (Norm 2.2).

Figure 5 shows the correlation (-1*r) between the speakers' F_0 distance scores and their proficiency scores for the two normalizations, (Norm 2.1) and (Norm 2.2). We can see that the correlation improves when excluding more "bad" sentences, which is the same as the result on dura-

tion. With regard to the two normalization methods, the per-utterance normalization (Norm 2.1) outperforms the per-article normalization (Norm 2.2).

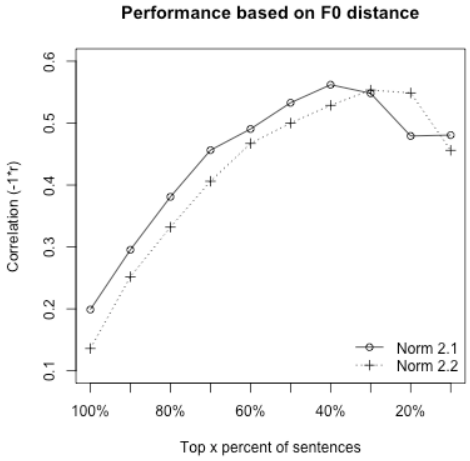

Fig. 5. F_0 distance: different normalizations.

4.3 Sentence selection based on spectrum

Dynamic Time Warping (DTW) was used to calculate the spectral distance between a test utterance and its standard version. The feature vector consists of the standard 39 PLP coefficients, of which the 13 static ones were zero-meaned per utterance. As shown in Figure 6, the correlation increases when excluding more "bad" sentences, which is the same as the results on both duration and F_0.

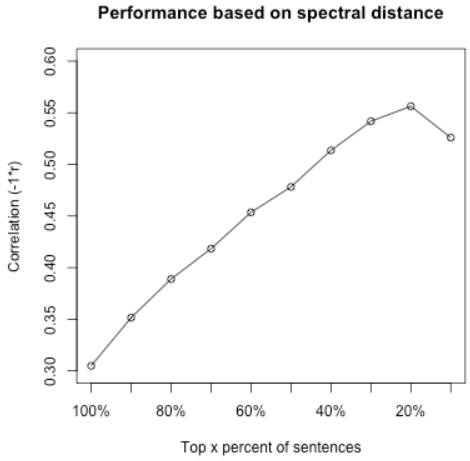

Fig. 6. The performance of spectral distance.

4.4 Combining distance scores

In this section, we investigate the combination of different distance scores. A statistic of pause was also included, which is the average number of pauses per utterance for a speaker. A SVM regression model was trained to predict human graded scores from the calculated distance scores at the speaker level. We employed 5-fold cross validation to separate training and test data. The correlations between model-predicted scores and human scores on the test data are reported in Table 1, for both using all available sentences and using only the best-performed sentences, determined by the experiments above.

Distance scores used	All sentences	Best sentences
D	.495	.747
F_0	.173	.562
S	.296	.514
D + F_0	.526	.786
D + F_0 + S	.566	.804
D + F_0 + S + P	.640	.824

D: syllable duration, normalized per article;
F_0: normalized per utterance; S: spectrum; P: pauses

Table 1: Speaker-level correlations between SVM-predicted and human scores.

From Table 1 we can see that compared to using all available sentences, using only best-performed sentences significantly improved the performance. When all the three distance scores as well as the pause statistic are combined, the correlation increased from .640 to .824, which is comparable to the correlation (r = .819) between the two examiners' scores. We should note that, however, the human scores used in the experiments are the averages of the two examiners' scores, and that although training and test data were separated in building SVM models for score combination, all data have been used to determine best-preformed sentences.

5 Conclusion

We proposed a method to select well-performed sentences for automatic scoring of spoken language proficiency. Our experiments demonstrated that the speaker-level correlation between human and machine scores could be significantly improved when excluding "bad" sentences from automatic scoring. Continuing research is needed to understand the linguistic factors that determine the goodness of a sentence for automatic proficiency scoring, and to understand the speech characteristics that differentiate the variability between and within groups of standard and non-standard speakers.

References

Catia Cucchiarini, Helmer Strik, and Lou Boves. 2000. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America*, 107(2):989-99.

Horacio Franco, Leonardo Neumeyer, Vassilios Digalakis, and Orith Ronen. 2000. Combination of Machine Scores for Automatic Grading of Pronunciation Quality. *Speech Communication*, 30(2-3):121-130.

Horacio Franco, Leonardo Neumeyer, Yoon Kim, and Orith Ronen. 1997. Automatic pronunciation scoring for language instruction. *Proceedings of ICASSP 1997*, pp. 1471-1474.

Fengpei Ge, Fuping Pan, Changliang Liu, Bin Dong, Shui-duen Chan, Xinhua Zhu, and Yonghong Yan. 2009. An SVM-Based Mandarin Pronunciation Quality Assessment System. In: *Advances in Intelligent and Soft Computing,* Vol. 56, pp. 255-265.

Wenping Hu, Yao Qian, Frank K. Soong, and Yong Wang. 2015. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, 67:154-166.

Ann Lee and James Glass. 2013. Pronunciation assessment via a comparison-based system. *Proceedings of SLaTE 2013*, pp. 122-126.

Emily Nava, Joseph Tepperman, Louis Goldstein, Maria Luisa Zubizarreta, and Shrikanth S. Narayanan. 2009. Connecting rhythm and prominence in automatic ESL pronunciation scoring. *Proceedings of Interspeech 2009*, pp. 684-687.

Leonardo Neumeyer, Horacio Franco, Vassilios Digalakis, and Mitchel Weintraub. 2000. Automatic scoring of pronunciation quality. *Speech Communication*, 30(2-3):83-93.

Lynn Streeter, Jared Bernstein, Peter Foltz, and Donald DeLand. 2011. *Pearson's automated scoring of writing, speaking, and mathematics* (White Paper). Retrieved from http://researchnetwork.pearson.com

Ren-Hua Wang, Qingfeng Liu, and Si Wei. 2006. Putonghua Proficiency test and evaluation. In: *Advances in Chinese Spoken Language Processing*, pp. 407-429.

Silke Maren Witt and Steve Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2-3):95-108.

Yoichi Yamashita, Keisuke Kato, and Kazunori Nozawa. 2005. Automatic Scoring for Prosodic Proficiency of English Sentences Spoken by Japanese Based on Utterance Comparison. *IEICE Transactions on Information and Systems*, E88-D(3):496-501.

Ke Yan and Shu Gong. 2011. Pronunciation Proficiency Evaluation based on Discriminatively Refined Acoustic Models. *International Journal of Information Technology and Computer Science*, 3(2):17-23.

Jiahong Yuan, Neville Ryant, and Mark Liberman. 2014. Automatic phonetic segmentation in Mandarin Chinese: boundary models, glottal features and tone. *Proceedings of ICASSP 2014*, pp. 2539-2543.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic Scoring of Non-Native Spontaneous Speech in Tests of Spoken English. *Speech Communication*, 51(10):883–895.