Automatically assessing the oral proficiency of proficient L2 speakers

Pieter Müller[1]*, Febe de Wet*[2]*, Christa van der Walt*[3] *& Thomas Niesler*[1]

[1]Department of Electrical and Electronic Engineering,
[2]Centre for Language and Speech Technology (SU-CLaST),
[3]Department of Curriculum Studies, Stellenbosch University, South Africa.
pfdevmuller@dsp.sun.ac.za, {fdw,cvdwalt,trn}@sun.ac.za

Abstract

We consider the automatic assessment of oral proficiency for advanced second language speakers. A spoken dialogue system is used to guide students through a reading and a repeating exercise and to record their responses. Automatically-derived indicators of proficiency that have proved successful in other studies are calculated from their speech and compared with human ratings of the same data. It is found that, in contrast to the findings of other researchers, posterior scores correlate poorly with human assessments of the reading exercise. Furthermore, the repeating exercise is found both to be more challenging and to provide a better means of automatic assessment than the reading exercise for our test population.

1. Introduction

This paper describes progress in the development of an automated system for the assessment of oral language proficiency for large groups of students. The system is intended for use by the Education Faculty at Stellenbosch University, where students must be regularly assessed in terms of their second language (L2) proficiency. Most of these students speak Afrikaans as a first language (L1) and their English proficiency varies from intermediate to advanced. With between 100 and 200 students per university staff member, this evaluation is currently based heavily on computerised multiple-choice reading and writing tests. However, it is well established that such written tests are not necessarily good indicators of oral proficiency [1].

A factor which sets this study apart from others is the high L2 proficiency of the test population. In other studies, this proficiency varies to a much greater degree [2, 3, 4, 5]. Our research therefore focuses on students who speak English as a *second* language rather than as a *foreign* language.

Previous research has shown that, for this type of test population, posterior scores derived at the utterance level do not give a good indication of pronunciation quality [6]. This is in contrast to their reported utility in other studies [2, 4, 5]. This paper investigates the performance of posterior scores more closely in an effort to determine reasons for this discrepancy. Furthermore, we attempt to determine which automatically-derived scores can be used as meaningful indicators of proficiency for advanced English learners.

2. Computerised test development

A telephone-based test was implemented because it requires a minimum of specialised equipment and allows flexibility in terms of the location from which the test may be taken. Past experience at the Faculty of Education has indicated that on-line telephone assessments using human judges give a fair indication of oral and aural proficiency.

2.1. Test design

The test was designed to include instructions and tasks that require comprehension of spoken English and elicit spoken responses from students. In this paper we will focus on two of the seven tasks that comprise the test, namely the *reading* and the *repeating* tasks. For a detailed description of the complete test, the reader is referred to [7].

- **Reading task:** The system randomly chooses six from a printed list of 12 sentences, and instructs students to read each one in turn. For example, *"Many participants asked if this was the best way forward."*
- **Repeating task:** Students are asked to listen to and then repeat sentences played by the system. For example, *"Lecturers who are out of touch with school practice have unrealistic expectations."*

The first task is familiar to students since reading aloud is taught as a basic skill at secondary school level. The construction of the repeating task is based on the hypothesis that phonological working memory capacity influences oral production in first language users [8, 9] and even more so in second language learners [10, 11]. In terms of this hypothesis, second language learners will find it harder to produce the target language in face-to-face communication because of time pressure in conjunction with limited access to the L2 vocabulary and sound system.

The sentences in the repeating task ranged from fairly simple (e.g. *It is boring to sit and watch teachers all day.*) to longer and more complex sentences where the subject is a separate clause (e.g. *How parents' interests and hopes are accommodated is crucial to the success of a school.*). In the case of advanced learners, it was assumed that their working memory capacity in the second language would make it possible for them to repeat the sentences accurately.

2.2. Test implementation

A spoken dialogue system (SDS) guided students through the test and captured their answers. For clarity, different voices were used for test guidelines, for instructions and for examples of appropriate responses. The SDS also controlled the interface between the computer and the telephone line, but the calculation of proficiency indicators was done off-line.

2.3. Test administration

One hundred and twenty students took the test as part of their quarterly assessment. Calls to the SDS were made from a telephone located in a private office reserved for this purpose. Oral instructions were given to the students before the test. In addition to the instructions given by the SDS, a printed copy of the test instructions was provided. No staff were present while the students were taking the test.

3. Human assessments

Teachers of English as a second or foreign language were asked to rate speech samples drawn from the reading and repeating tasks. The raters were not personally acquainted with the students. A subset of 90 students was selected from the group of 120 who took the test. This subset was chosen to represent male and female as well as Afrikaans and English mother tongue speakers in accordance with the composition of the student population at the Faculty of Education. Given the large numbers, it was not feasible to have each utterance rated. Instead, three examples of each student's reading and repeating responses were randomly chosen to be judged by the raters. All raters attended a training session on the use of the rating scales, at which example utterances and associated ratings were presented.

Six raters each assessed 45 students and each student was assessed by three human raters. In order to measure intra-rater consistency, five students were presented twice to each rater. Each rater therefore performed a total of 50 ratings.

The current investigation focuses specifically on the correlation between posterior scores and human ratings. The scope of the study will therefore be restricted to those aspects of the human ratings that showed the highest correlation with posterior scores in previous experiments [6], i.e. *pronunciation* for the reading task and *success* and *accuracy* for the repeating task. Human assessment was done by means of Likert scales. Eight different levels were defined for pronunciation, six for success and five for accuracy. The upper and lower extremities of the scales are illustrated in Figure 1. The complete scales are discussed at length in [6].

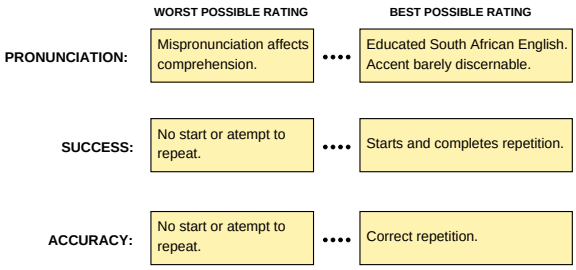

Figure 1: Upper and lower extremities of the assessment scales used by the human raters.

3.1. Results: Human assessments

Two-way random, intra-class correlation coefficients were calculated to determine intra-rater reliability. The correlation ranged between 0.67 and 0.96, with an average of 0.85 for the six raters. These values compare favourably with those reported in other studies [3, 12].

The average rating (calculated across all raters) and its standard deviation for each of the three chosen scales are shown in Figure 2. It is evident from the high averages that the assigned ratings were concentrated in the upper part of the rating scale. This occurred despite previous deliberate efforts to broaden the scale in a bid to obtain a wider spread of assigned ratings [6]. The figure also shows that, on average, students performed better in the reading task (pronunciation) than in the repeating task (success and accuracy). The high average and small standard deviation of the pronunciation ratings is an indication that the students did not find the reading task challenging.

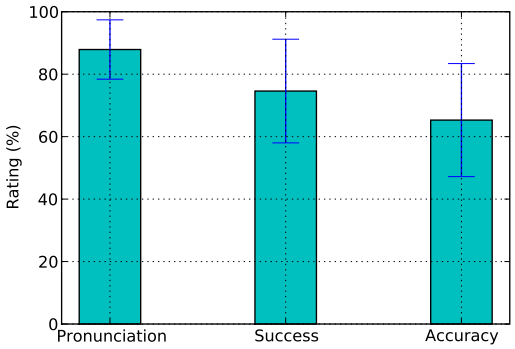

Figure 2: Average ratings assigned by human judges for pronunciation, success and accuracy. Standard deviations of the mean are shown.

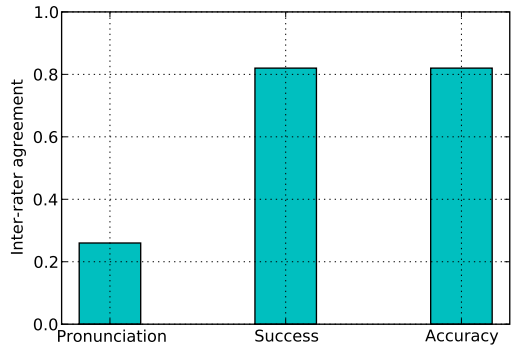

Figure 3: Inter-rater agreement for pronunciation, success and accuracy.

Figure 3 illustrates the inter-rater agreement for the rating scales pronunciation, success and accuracy. By comparing pronunciation with success and accuracy, we see that the inter-rater agreement was higher for the repeating task than for the reading task. The values shown in Figure 3 for the read speech are lower than those reported in [3] and [4], but are similar to those reported in [5]. The small standard deviation in ratings for the reading task shown in Figure 2 is one possible cause for this low inter-rater agreement. The raters appear to be less consistent in their assessments when there is little variation in proficiency because they are often in disagreement about how such small differences should be evaluated. In studies where higher inter-rater agreement was measured, the speaker populations were more diverse in terms of L2 proficiency.

4. ASR-based assessment

Numerous studies on the role of ASR in language learning applications have been published in the last decade e.g. [2, 3, 4, 5]. A common aim of most studies is the identification of parameters that can be automatically derived from speech data and that correlate well with human judgements of oral proficiency.

4.1. ASR system

A corpus consisting of approximately six hours of phonetically-annotated, South African English (L1), telephone speech data was parameterised as mean-normalised Mel-frequency cepstral coefficients (MFCCs) and their first and second differentials. A set of speaker-independent cross-word triphone HMMs was subsequently obtained using decision-tree state clustering and embedded Baum-Welsh re-estimation, resulting in a total of 4797 clustered states [13]. Each triphone model employed eight Gaussian mixtures per state and diagonal covariance matrices. On a separate test set, using a bigram language model, the triphone models had a phone recognition accuracy of 73%.

The students' responses to the test were transcribed orthographically by human annotators. The data that was assessed by the human raters was used as an independent test set (90 speakers). The remainder of the data (30 speakers) was used as a development test set.

For each sentence in the the reading task, a finite-state grammar was constructed allowing two options: the target utterance and "I don't know". Students were instructed to say "I don't know" if they were unsure about how to respond to a test item. Filled pauses, silences and speaker noises were permitted between words by the grammar. The recogniser's word insertion penalty was chosen to ensure optimal correlation between the rate of speech (ROS) values derived from the manual and automatic transcriptions of the development test set.

For the repeating task, two methods of recognition were used. The *goodness of pronunciation* (see Section 4.2) was calculated using the same configuration used for the reading task. However, the indicators *rate of speech* and *transcription accuracy* (also described in Section 4.2) used a unigram language model derived from the manual transcriptions of the development test set, with equal probabilities for all words. A separate language model was constructed for each sentence of the repeating task. The recogniser's word insertion penalty and language model factor were chosen to maximise the correlation between recognition accuracy as well as the ROS values derived from the manual and automatic transcriptions of the development test set.

4.2. Automatically-derived proficiency indicators

Many indicators of oral proficiency that can be automatically derived from speech data have been proposed in the literature. We have chosen three that have been reported to perform best by several authors, namely rate of speech, posterior scores, and transcription accuracy.

4.2.1. Rate of speech

Previous studies have found that, for read speech, the rate of speech (ROS) is one of the best indicators of fluency [3, 5]. In our experiments ROS was calculated according to Equation (1), as described in [3].

$$ROS = \frac{N_p}{T_{sp}} \tag{1}$$

Here N_p denotes the number of speech phones in the utterance, and T_{sp} the duration of speech, including pauses.

4.2.2. Posterior scores

Posterior HMM likelihood scores have been shown to be an effective automatic measure of pronunciation quality. As an example of this general class of scores, we used "goodness of pronunciation" (GOP) as proposed in [2]. The GOP score of phone q_i is defined as the frame-normalised logarithm of the posterior probability $P(q_i|O)$, where O refers to the acoustic segment uttered by the speaker.

$$GOP(q_i) = \frac{|log(P(q_i|O))|}{NF(O)} \tag{2}$$

In equation (2), $NF(O)$ corresponds to the number of frames in acoustic segment O. A GOP score was determined for each phone in an utterance. Utterance level scores were subsequently obtained by averaging these phone scores.

A previous study on this data found no useful correlation between GOP scores and the human ratings for pronunciation [6]. In an effort to improve this correlation, three variations of the GOP score algorithm were evaluated:

- GOP_SP is calculated using only the scores of speech phones, thus excluding silence and other non-speech sounds.
- GOP_SPC is calculated using only the scores of speech phones in the left and right context of other speech phones. This excludes phones in the context of silence or other non-speech sounds [4].
- GOP_W is calculated by first determining the time-normalised GOP score for each word, and then averaging these word-level scores for the utterance [12].

4.2.3. Transcription accuracy

Because highly restrictive finite-state grammars were used for the reading task, the recognition accuracy obtained for the read responses was very high and therefore not used as a proficiency indicator. Repeat accuracy, on the other hand, was considered as a proficiency indicator, as is also proposed in [5]. The ASR output for the repeated utterances was compared to the target prompts and accuracy was subsequently calculated according to Equation 3, as described in [13].

$$Accuracy = \frac{H - I}{N} \times 100\% \tag{3}$$

In Equation 3, H is the number of correctly recognised words, I is the number of insertion errors and N is the total number of words in an utterance.

5. Correlation between human and ASR-based assessment

Table 1 gives the correlation between the ratings given by the human raters and the automatically-derived proficiency indicators for both the reading task and the repeating task. Spearman rank correlation coefficients were used because the data is ordinal. The automatically-derived proficiency indicators were calculated for the same material evaluated by the human raters.

For the reading task, the ROS scores show the strongest correlation with human ratings, while the GOP scores show almost no correlation. Given the definition of our rating scales, GOP is expected to show a positive correlation with human ratings. It is clear that the word level GOP score, GOP_W, performs particularly poorly.

Proficiency indicator	Reading	Repeat	
	Pronunciation	Success	Accuracy
ROS	-0.46	-0.71	-0.68
Accuracy	-	0.68	0.69
GOP	0.02	0.48	0.58
GOP_SP	0.00	0.52	0.62
GOP_SPC	-0.02	0.56	0.64
GOP_W	-0.14	0.41	0.50

Table 1: Correlation between human ratings and automatic scores.

The poor correlation between GOP scores and human ratings of the read material supports the observation made in [3], where the weakest correlation between human and automatic scores was measured for likelihood ratios. This trend seems to indicate that posterior scores derived at the utterance level do not provide meaningful information on the pronunciation quality of proficient L2 speakers. When calculated using only the 5 highest and 5 lowest rated students, the correlation of GOP_SPC with human ratings rises to 0.18. This indicates that there is some correlation between GOP scores and human ratings for pronunciation, but that the variation in pronunciation quality in our data is too small for the GOP score to be a good indicator.

For the repeating task, both ROS and accuracy are well correlated with the human ratings for success and accuracy. Although the GOP scores are not as well correlated with the human ratings in the repeating task as ROS or accuracy scores, the correlations are consistently higher than for the reading task. The correlation with both success and with accuracy ratings was improved by restricting the GOP calculation to speech phones, and improved even further by considering only speech phones in the context of other speech phones. Normalising the GOP score on the word-level performed poorer than the GOP scores normalised on a the phone-level GOP scores. This observation has also been made by other researchers [12].

6. Discussion and conclusions

Despite being proposed specifically as a measure to predict human ratings of pronunciation [2], GOP shows little correlation with the pronunciation ratings in our study. We believe that a possible cause is the high mean and small standard deviation of the pronunciation ratings (Figure 2), resulting from the high oral proficiency of the test population and the relatively low difficulty level of the reading task. We therefore conclude that posterior scores such as GOP appears to be less effective at predicting pronunciation ratings for proficient L2 speakers than for foreign speakers. Where GOP scores are employed, they are best calculated based only on speech phones in the context of other speech phones. GOP scores normalised at the word-level do not correlate better with human ratings than GOP scores normalised at the utterance-level.

For the student population under investigation our results show that the reading task is not sufficiently challenging. When the assessment exercises are too easy, students make few mistakes. This results in a narrow range of assigned scores by the human judges using the intrinsically discrete Likert scales. Consequently, the discretisation of the human assessments by the scales becomes too coarse to allow meaningful correlations to be discovered.

For the repeating task, the human ratings are better correlated with automatically-derived indicators than for the reading task. We conclude that, for proficient L2 speakers, a repeating exercise is a better indicator of oral proficiency than a reading exercise. For the reading task to be effective, its level of difficulty would have to be increased significantly.

7. Acknowledgements

This research was supported by an NRF Focus Area Grant for research on *English Language Teaching in Multilingual Settings* as well as NRF grants TTK2007041000010 and GUN2072874 and the "Development of Resources for Intelligent Computer-Assisted Language Learning" project sponsored by the NHN.

8. References

[1] S. Sundh, "Swedish school leavers' oral proficiency in English," Ph.D. dissertation, Uppsala University, Sweden, 2003.

[2] S. M. Witt, "Use of speech recognition in computer-assisted language learning," Ph.D. dissertation, Department of Engineering, University of Cambridge, UK, 1999.

[3] C. Cucchiarini, H. Strik, and L. Boves, "Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms," *Speech Communication*, vol. 30, pp. 109–119, 2000.

[4] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech Communication*, vol. 30, pp. 83–93, 2000.

[5] T. Cincarek, R. Gruhn, C. Hacker, E. Nöth, and S. Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-native's first language," *Computer Speech and Language*, vol. 23, pp. 65–88, 2009.

[6] F. De Wet, C. Van der Walt, and T. R. Niesler, "Automatic assessment of oral language proficiency and listening comprehension," *Speech Communication*, 2009, available online at doi:10.1016/j.physletb.2003.10.071.

[7] C. Van der Walt, F. De Wet, and T. R. Niesler, "Oral proficiency assessment: the use of automatic speech recognition systems," *Southern African Linguistics and Applied Language Studies*, vol. 26, no. 1, pp. 135–146, 2008.

[8] M. Daneman, "Working memory as a predictor of verbal fluency," *Journal of Psycholinguistic Research*, vol. 20, no. 6, pp. 445–464, 1991.

[9] G. Wigglesworth, "An investigation of planning time and proficiency level on oral test discourse," *Language Testing*, vol. 14, no. 1, pp. 85–106, 1997.

[10] N. C. Ellis and S. Sinclair, "Working memory in the acquisition of vocabulary and syntax: Putting language in good order," *Quarterly Journal of Experimental Psychology*, vol. 49, no. A, p. 234250, 1996.

[11] J. S. Payne and B. M. Scott, "Synchronous CMC, working memory, and L2 oral proficiency development," *Language Learning & Technology*, vol. 9, no. 3, pp. 35–54, 2005.

[12] O. D. Deshmukh, S. Joshi, and A. Verma, "Automatic pronunciation evaluation and classification," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 1721–1724.

[13] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book, version 3.2.1*. Cambridge University Engineering Department, 2002.