Automatic large-scale oral language proficiency assessment

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

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

Abstract

We describe first results obtained during the development of an automatic system for the assessment of spoken English proficiency of university students. The ultimate aim of this system is to allow fast, consistent and objective assessment of oral proficiency for the purpose of placing students in courses appropriate to their language skills. Rate of speech (ROS) was chosen as an indicator of fluency for a number of oral language exercises. In a test involving 106 student subjects, the assessments of 5 human raters are compared with evaluations based on automatically-derived ROS scores. It is found that, although the ROS is estimated accurately, the correlation between human assessments and the ROS scores varies between 0.5 and 0.6. However, the results also indicate that only two of the five human raters were consistent in their appraisals, and that there was only mild inter-rater agreement.

Index Terms: automatic oral proficiency assessment, rate of speech (ROS), computer assisted language learning (CALL).

1. Introduction

Assessment of a student's entrance level language skills for the purpose of placement into appropriate language programmes, or for syllabus design, is often restricted to reading and writing proficiency tests. Listening and speaking skills are frequently not properly appraised because they either require specialised equipment or labour intensive procedures. In addition, the assessment of oral skills is generally highly subjective, and efforts that enhance inter-rater reliability further increase the labour intensiveness of the assessment process. The assessment of reading and writing comprehension skills, on the other hand, can be automated by means of computerised multiple choice tests, which have vastly reduced time and manpower requirements for their administration.

Studies agree, however, that good results in a written test are not necessarily good predictors of corresponding results in an oral test [1]. Hence language proficiency cannot be accurately assessed without considering spoken and listening skills. One of the best-known procedures for the assessment of spoken communication is the oral proficiency interview, such as the one developed by the American Council on the Teaching of Foreign Languages (ACTFL) [2]. However, even with such a standardized test, it is not only difficult to achieve consistency among raters, but also among the different scales and measures used to describe the performance of the examinees [3]. Although existing attempts to improve objectivity in oral proficiency assessment have been criticised, these methods remain the primary means of student assessment and curriculum development.

This study describes an attempt to develop an automated system for the assessment of oral language proficiency to improve the level of objectivity, reduce the associated manual workload, and allow speedy availability of the test results. To do this, we investigate the use of automatic speech recognition (ASR) for the automated computer-based assessment of listening and speaking skills. This is in line with the wider perception that "the computerized delivery of tests has become an appealing and a viable medium for the administration of standardized L2 tests in academic and non-academic institutions" [4].

The system is developed within the very specific context of the Education Faculty at Stellenbosch University, where new students are required to obtain a language endorsement on their teaching qualification. For English, this means in practice that students have to enrol for a module appropriate to their level of proficiency, and their progress is monitored regularly thereafter. With a current ratio of between 100 and 200 students per university staff, this is only feasible by placing the greatest emphasis on computerised multiple-choice reading and writing tests. Since students regard oral proficiency as an important component of their teaching abilities, they are not happy with the focus on writing and reading skills and the infrequency of oral assessment is regarded with much suspicion. A technological solution may not only lighten the heavy workload of staff, but also provide a transparent and more objective metric with greater acceptance among students.

2. ASR-based oral proficiency assessment

The feasibility of limited spoken communication between humans and machines by means of ASR has added a new dimension to computer assisted language learning (CALL). Exercises that require speech production such as reading, repeating and speaking about specific topics can be included in ASR-enhanced CALL systems. Currently such systems fall into two categories: (i) systems that provide synchronous feedback on pronunciation quality (e.g. [5, 6]) and (ii) systems that provide global assessment of oral language proficiency on the basis of a few spoken sentences (e.g. [7, 8]). Both are quite different from the Computerized Oral Proficiency Instrument (COPI) developed by the ACTFL, where the computer reacts to the examinee's input, but the speech is recorded and later rated by human examiners.

Automatic assessment systems are designed to predict human ratings of oral proficiency in terms of measures such as fluency, intelligibility and overall pronunciation quality [9, 8, 10]. Various automatic measures have been investigated and it has been shown that they correlate differently with different aspects of human rating [7]. Among the most promising indicators of human ratings are the so-called posterior, duration, and rate of speech (ROS) scores [10]. Because the system proposed here is still in the initial phases of development, it was decided to restrict the scope of the research to investigating the correlation between human ratings of read, repeated and spontaneous speech and automatically derived ROS scores.

3. Test development

The goal of the test was to assess listening and speaking skills limited to the specific context of school education. The test was therefore designed to elicit performances of the specific language behaviors that we wished to assess, rather than a test where students are required to complete real-life tasks. There was no attempt to mimic real life communication except in the sense that the test content related to teaching and learning in a school environment.

A phone-in test was chosen for our application because it requires a minimum of specialised equipment and allows flexibility in terms of the location from which the test may be taken. Moreover, in previous years on-line telephone assessments using human judges were found to give a good indication of oral and aural proficiency. The automated test included a number of open-ended questions. The students' responses to these were recorded and could be assessed at a later stage if necessary, for example in borderline cases.

3.1. Test design

The test was designed to include instructions and tasks that require comprehension of spoken English and elicit spoken responses from students. Oral proficiency was tested by means of three different types of questions:

1. **Reading task**. Subjects were asked to read sentences printed on the provided test sheet. Example: "School governing boards struggle to make ends meet."
2. **Repeat task**. Subjects were asked to repeat a sentence once it had been read to them. Example: "Student teachers do not get enough exposure to teaching practice."
3. **Open-ended task**. Subjects were asked to respond spontaneously to a general question. Example: "What is your biggest fear when you go into a classroom?"

The complete test also includes a variety of other questions which, for example, test the subjects' grasp of appropriateness and formality in language usage, and the extent of passive and active vocabulary. However, we will focus on the results obtained for the above three tasks in the remainder of this paper.

3.2. Test implementation

A spoken dialogue system (SDS) was developed to guide students through the test and capture their answers. To make the test easy to follow, voice prompts were designed and recorded using different voices for test guidelines, for instructions and for examples of appropriate responses. The SDS plays the test instructions, records the students' answers, and controls the interface between the computer and the telephone line. In operational systems the SDS also controls the flow of data to and from the ASR system, but in the system described here, the students' answers were simply recorded for later off-line processing.

3.3. Test administration

A total of 106 students took the test as part of their oral proficiency assessment during the first academic semester of 2006. Calls to the SDS were made from a telephone in a private office reserved for this purpose. Each student also completed a short questionnaire collecting information regarding their home language, academic performance, and opinion of the language course. 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.

Feedback received immediately after completing the test indicated that English-speaking students generally found the test manageable while the majority of Afrikaans students found it fairly challenging. Most students found the instructions clear and found that the paper copy of the test provided adequate guidelines and extra security in a stressful situation.

4. Human & automatic test evaluation

Once all students had completed the test, their recorded replies to the questions were transcribed orthographically by human annotators. The group of 106 students was then divided into two groups: a development set of 16 speakers and a test set of 90 speakers. Data from the development set was used to optimise the ASR system parameters. The remaining 90 students' responses were subsequently assessed by human raters as well as by the ASR system. The following sections compare these automatic scores with the human judgements.

4.1. Evaluation by human raters

Five teachers of English as a second or foreign language were asked to rate speech samples from the read, repeat, and open-ended tasks in the test. In addition, they were requested to give each student an *overall impression* mark. The raters did not know the students whom they were rating. Each rater assessed 45 students and each student was assessed by at least two human raters. In order to measure intra-rater consistency, five students were presented twice to each rater. Each rater therefore performed 50 ratings: 45 unique and 5 repeats.

The literature on the role played by human judges and the instruments that they use for oral assessment is vast. We have relied heavily on overview studies, such as [3] and studies that focus on advanced students of English, such as [1]. This last consideration was very important for our study, since many of the students who took part are home language speakers of Afrikaans, but are nevertheless fluent in English. Decisions about the use of assessment criteria were made on the basis of Jesney's report to the Language Research Centre at the University of Calgary, where she finds that the use of Likert scales are appropriate specifically for the assessment of accentedness [3]. After considering a variety of assessment rubrics and grids a decision was made to use a 5-point scale for all four tasks but to vary the assessment criteria depending on the focus of each task. Table 1 summarises the scale's extremities for each task.

4.2. Evaluation by ASR-based automatic rater

In South Africa, ASR is a relatively new research field and the resources that are required to develop applications are limited. A recent initiative collected telephone speech databases in South African English, isiZulu, isiXhosa, Sesotho and Afrikaans [11]. Prototype speech recognisers were subsequently developed for each of these languages, and the current study makes use of the standard South African English ASR system. The system is based on context-dependent hidden Markov phone models and was trained on approximately six hours of telephone speech data.

Other studies have found that the rate of speech (ROS) is one of the best indicators of speech fluency [7]. We calculate

Table 1: *Summary of the Likert scales and associated assessment criteria used by human raters for each task. Only the extremities of the scales are shown.*

Task	Score	Corresponding assessment criterion
Reading	5	Pronunciation, intonation and rhythm almost mother-tongue.
	1	Speech difficult to understand and poorly articulated.
Repeat	5	Repetition was accurate and prompt.
	1	No attempt was made to repeat.
Open-ended	5	Confident and completely fluent reply.
	1	Only a feeble attempt to formulate meaningful contribution.
Overall	5	Fluent and correct use of English, easy to understand.
	1	Poor production of English, extremely difficult to understand.

the ROS according to Equation (1), as proposed in [9]:

$$ROS = \frac{N_p}{T_{sp}} \qquad (1)$$

where N_p is the number of speech phones in the utterance, and T_{sp} is the total duration of speech in the utterance, including pauses.

For each sentence in the the reading task, a BNF grammar was constructed allowing two options: the target utterance and “I don’t know”. Filled pauses, silences and speaker noises were permitted between words by the grammar. Recognition system parameters were chosen such that the correlation between the ROS values derived from the manual and automatic transcription of the data was optimal on the development set.

For the repeat and open-ended tasks a unigram language model with uniform probabilities for all words was derived from the manual transcriptions of the development data. A language model was constructed for each sentence of the repeat task, while the responses to all the open-ended questions were pooled for a common language model. Recognition parameters for the repeat task were subsequently chosen by optimising the correlation between word accuracies as well as the ROS values between the manual and automatic transcriptions of the data. A similar strategy was followed for the open-ended task except that word accuracy was not taken into account because there are no model answers to open-ended questions.

5. Experimental results

5.1. Performance of human raters

Table 2 gives an overview of the average (across all tasks) intra-rater correlations that were determined for the human raters. These values were derived from the scores assigned to the five students that each rater assessed twice.

Table 2: *Intra-rater correlations for human raters.*

Rater	Intra-rater correlation
1	0.32
2	0.74
3	0.30
4	0.73
5	0.40

According to the data in Table 2, raters 2 and 4 are consistent in their judgements, the others are not. This indicates that, on average, the consistency of human assessment is rather low.

Figure 1 shows the average (across all raters) intra-task correlation for the four judgements made by the human raters. The last column indicates the average value for all four judgements. The figure shows that the raters’ judgements agreed most for the repeat task and least for the open-ended questions. It is also apparent that, overall, there is not a strong consensus among the human raters.

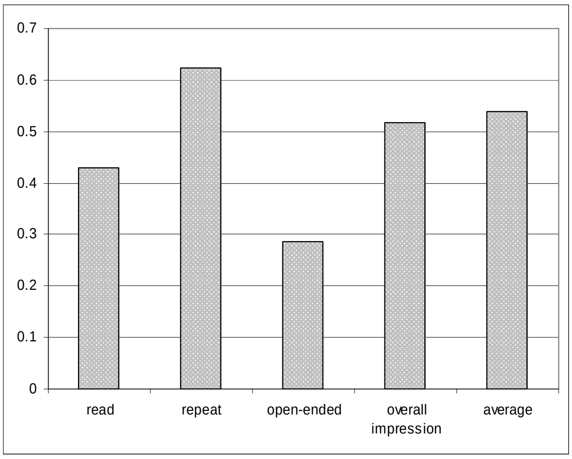

Figure 1: *Average intra-task correlation for the read, repeat, open-ended and overall impression judgements made by the human raters. The last bar corresponds to the average correlation for all four judgements.*

Figure 2 illustrates the average score (across all raters) given for each task. The highest mark that could be awarded in each section was five and the lowest mark one.

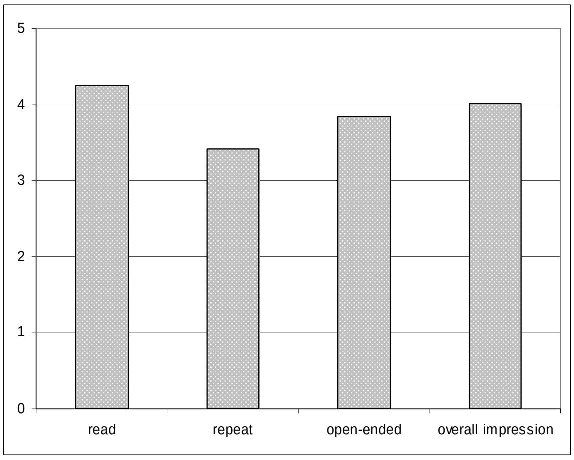

Figure 2: *Average scores awarded by the human raters for the read, repeat and open-ended tasks. The last bar corresponds to the overall impression marks.*

Figure 2 shows that the human raters gave the students fairly high marks. In fact, a score of one was assigned only once. On average, the students received the highest marks for the reading task and the lowest marks for the repeat task. It is interesting to note that the *overall impression* marks are almost as high as those for the reading task, even though the marks for the two other tasks are lower.

5.2. Performance of the automatic rater

The average ROS values that were measured for the read, repeat and open-ended tasks are shown in Table 3. The observation that the highest value corresponds to the reading task and the lowest to the open-ended task is in good agreement with what one would intuitively expect - given the level of difficulty of the three tasks.

Table 3: *Average ROS scores based on the automatic test-data transcriptions, and the correlation between these ROS values and the ROS derived from manual test data transcriptions.*

Task	Average ROS	Correlation
Read	6.0	0.98
Repeat	5.0	0.94
Open-ended	4.8	0.86

The correlation between the ROS values derived from the manual and automatic transcriptions of the test data for the 90 test subjects are also listed in Table 3. The values in the table indicate that the automatic system's ability to segment the speech into phones compares very well with its human counterpart, especially for the read and repeat tasks.

5.3. Correlation between human and automatic raters

Table 4 gives the correlation between the human raters' scores and the corresponding automatically derived ROS values per task. The last row of Table 4 indicates the correlation between the *overall impression* marks assigned by the human raters and the average value of the ROS values for the read, repeat and open-ended tasks.

Table 4: *Correlation between average scores assigned by human raters and corresponding ROS values.*

Task	Correlation
Read	0.52
Repeat	0.58
Open-ended	0.48
Overall impression	0.56

The correlation between ROS and the human ratings of fluency (the main emphasis of the reading task) is lower than the corresponding values reported in [9]. However, in general the correlation between the human and the automatic raters shown in Table 4 compare favourably with those reported in similar studies [10, 6]. It is also interesting to note that these correlation values show the same trend as the intra-class correlation values illustrated in Figure 1, where the highest value was observed for the repeat task and the lowest for the open-ended task. In this regard the automatic rater seems to be behaving like its human counterparts.

6. Discussion and conclusion

One aspect of our experimental setup that has become apparent during our analysis is that the 5-point Likert scale available to the human raters was used very unevenly, with awarded scores of generally 3, 4 or 5 and rarely 1 or 2. This restricts the resolution of our analysis and may have negatively affected the correlation between these values and the ROS scores. In future, a finer scale must be adopted to mitigate this loss in resolution.

We were also surprised by how inconsistent the human ratings usually were, and how weak the agreement between raters was overall. From the point of view of objectivity and consistency, the automatic system therefore shows clear promise.

Currently we are addressing these issues, as well as considering the inclusion of additional features over and above the ROS scores. We are also improving the accuracy of our ASR system to allow more flexible treatment of the open-ended questions, for example by significantly expanding the recognition vocabulary.

7. Acknowledgements

This research was supported by the Fund for Research and Innovation in Learning and Teaching at Stellenbosch University.

8. References

[1] S. Sundh, “Swedish school leavers' oral proficiency in english,” Ph.D. dissertation, Uppsala University, Uppsala, 2003.

[2] http://www.actfl.org/, (accessed 16/03/2007).

[3] K. Jesney, “The use of global foreign accent rating in studies of L2 acquisition,” Language Research Centre, University of Calgary, Tech. Rep., 2003.

[4] M. Chalhoub-Deville, “Language testing and technology: past and future,” *Language, Learning & Technology*, vol. 5, no. 2, p. 95, 2001.

[5] A. Neri, C. Cucchiarini, and H. Strik, “ASR corrective feedback on pronunciation: does it really work?” in *Proceedings of Interspeech*, Pittsburgh, USA, 2006, pp. 1982–1985.

[6] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, J. Butzberger, R. Rossier, and F. J. Cesari, “The SRI EduSpeak^{TM} system: Recognition and pronunciation scoring for language learning,” in *Proceedings of InSTILL 2000*. Dundee: University of Abertay, 2000, pp. 123–128.

[7] 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.

[8] J. Bernstein, J. de Jong, D. Pisoni, and B. Townshend, “Two experiments on automatic scoring of spoken language proficiency,” in *Proceedings of InSTILL 2000*. Dundee: University of Abertay, 2000, pp. 57–61.

[9] C. Cucchiarini, H. Strik, and L. Boves, “Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology,” *Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 989–999, 2000.

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

[11] J. C. Roux, P. H. Louw, and T. R. Niesler, “The African Speech Technology project: An assessment,” in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 2004, pp. I:93–96.