Southern African Linguistics and Applied Language Studies 2008, 26(1): 135–146
Printed in South Africa — All rights reserved

SOUTHERN AFRICAN LINGUISTICS
AND APPLIED LANGUAGE STUDIES
ISSN 1607–3614
doi: 10.2989/SALALS.2008.26.1.11.426

Oral proficiency assessment: the use of automatic speech recognition systems

Christa van der Walt[1]*, Febe de Wet[2] and Thomas Niesler[3]
[1] *Department of Curriculum Studies, Stellenbosch University, Private Bag X1, Stellenbosch 7602, South Africa*
[2] *Stellenbosch University Centre for Language and Speech Technology, Department of African Languages, Stellenbosch University, Private Bag X1, Stellenbosch 7602, South Africa*
[3] *Electric and Electronic Engineering, Stellenbosch University, Private Bag X1, Stellenbosch 7602, South Africa*
* *Corresponding author, e-mail: cvdwalt@sun.ac.za*

Abstract: The development and assessment of oral proficiency and listening comprehension is one of the most problematic aspects in language teaching, especially when the majority of test-takers are non-standard users of English. The main problems concern the feasibility of such testing and the need for reliable scoring. As far as feasibility is concerned, oral proficiency tests of communicative proficiency require much time as well as repeated assessment in the course of a semester or a year. To obtain a reliable score it is necessary to have more than one examiner, preferably also users of English as an additional language, to assess the same task. This paper will describe an attempt to use automatic speech recognition systems to obtain an objective score for oral proficiency. The process of test development and the subsequent digitalisation of speech, trialling and evaluation will be discussed with specific reference to a course that leads up to a language endorsement required by teacher trainees in South Africa.

Introduction

The assessment of entrance-level language skills with a view to placement in a specific module or for syllabus design is often restricted to tests of reading and writing proficiency. The assessment of students' language proficiency typically requires results within a few days, which also means that a multiple choice format is used. Listening and speaking skills are neglected because they either require specialised equipment (in the case of listening skills) or labour-intensive procedures (in the case of speaking skills). A further problem with the assessment of speaking skills is that it is generally subjective and efforts to enhance inter-marker reliability will again increase the labour intensiveness of the assessment process.

The problematic nature of assessing language learners' communication abilities (in speech and writing) is well known. The development of marking rubrics for extended writing and procedures to ensure inter-marker reliability form the basis of many doctoral theses and academic articles. A wide variety of instruments and procedures have been developed for the assessment of spoken communication. One of the best known is the oral proficiency interview (with concomitant assessment levels and training manuals) such as the one developed by the American Council on the Teaching of Foreign Languages (ACTFL)[1]. In an overview article on assessment criteria for measuring accent in particular, Jesney (2003) discusses the problems of achieving consistency not only among raters but also among the different scales and measures used to describe the performance of the examinees.

Although all attempts to create a degree of objectivity in the assessment of oral proficiency have been criticised, students are still being assessed and decisions for advancement or curriculum development are still taken on the basis of these assessments. As Kenyon *et al.* (2001), indicate, in the absence of a 'common metric' teachers and researchers will continue to build on existing scales and rubrics (like the ACTFL guidelines). In many teacher-training courses students are, in fact, encouraged to develop their own assessment criteria and to tailor them to specific tasks and the contexts within which they are executed.

In this paper an attempt to standardise an oral proficiency test is described within a very specific context and for a very specific purpose. Since it is important to build some degree of objectivity into an oral proficiency test, for the sake of lecturers as well as students, the quest for a feasible, economic and reliable test of listening and speaking skills in this study led to the use of an automatic speech recognition (ASR) system. The context is that of the education faculty at Stellenbosch University where students are required (as at other South African universities) to obtain a language endorsement on their teaching qualification, in the case of English also known as a 'Big E' or 'small e'. What this means in practice is that students have to enrol for a module which develops their English language proficiency so that they are able to either teach their subjects in English (the 'Big E') or use English in professional contexts (the 'small e'). Although such language endorsements are no longer required by the Western Cape Department of Education, principals of schools still trust these endorsements and insist on them for permanent appointment and promotion.

The problem with the management of these courses is that students need to be identified and supported according to their specific language proficiency levels. Students who are fluent users of English generally follow the 'Big E' course. The focus here is twofold: students develop an awareness of teaching limited English proficiency learners in the first place and, secondly, their ability to use subject-specific terminology in teaching and assessment is developed. Weaker students follow the 'small e' course where attempts are made to strengthen their language proficiency by developing professional language skills (writing reports, conducting meetings, etc.).

The process whereby the placement of students in the 'Big E' or the 'small e' course takes place is regarded with much suspicion by students and there are usually complaints. Students are not excluded from following the 'Big E' course: there are certain mechanisms in place that make it possible for them to change to this course after the first semester. In an effort to make the initial division between the two groups more transparent, objective and less time consuming, lecturers have turned to the use of technology — in line with Chalhoub-Deville's (2001) perception that 'the computerised delivery of tests has become an appealing and a viable medium for the administration of standardised L2 tests in academic and non-academic institutions'. The assessment of reading comprehension skills can be done by means of a multiple choice test delivered on the university intranet, but students point out, quite rightly, that their spoken proficiency should be taken into account as well since their future profession would require a high level of oral proficiency. Furthermore, as Sundh (2003) shows, studies on the correlation between oral and written proficiency agree that good results in a written test do not necessarily predict good results in an oral test. For these reasons it was necessary to investigate ways in which speaking and listening skills could be tested in as objective a manner as possible. This would be useful not only for initial testing but also for practice and assessment later in the semester.

The use of ASR systems for oral proficiency assessments

ASR technology enables humans to speak to computers. In countries such as the USA, UK, Japan, and many European countries, research in the field of ASR has already resulted in applications such as dictation systems and voice-operated telephone services. In South Africa, ASR is a relatively new research area and the resources that are required to develop applications are limited. One of the initiatives to gather South African speech data was the African Speech Technology (AST) project at Stellenbosch University, funded by the Department of Science and Technology (Roux *et al.*, 2004). During the project, telephone speech databases were collected in South African English, isiZulu, isiXhosa, Sesotho and Afrikaans. A prototype speech recogniser was subsequently developed for each of these languages. For the purposes of this study the use of these recognition systems was utilised for the assessment of students' oral proficiency.

The possibility of spoken communication between humans and computers has added a new dimension to computer-assisted language learning (CALL) systems. ASR technology makes it possible to include exercises that require speech production, such as reading, repeating and speaking about specific topics, in CALL applications. ASR-enhanced CALL systems usually fall into one of two main categories: (1) systems that provide synchronous feedback on the quality of pronunciation while students are working in a CALL environment[2]; and (2) systems that provide global assessment of oral language proficiency on the basis of a few sentences that are read or said by the student via a telephone connected to a computer.[3] Both these types of applications are quite different from the Computerized Oral Proficiency Instrument (COPI) developed by ACTFL, where the computer reacts to the examinee's input, but the speech is recorded and rated by human judges.

The main aim in this study was to find some kind of objective testing method that would also be efficient in terms of time, energy and purpose of the test, which is to test and place students appropriately. In the course of the semester the initial data would be supplemented by more open-ended classroom assessments. A phone-in test seemed most appropriate to implement in this case because it is fairly 'low-tech' (in the sense that students only make a telephone call) and telephone assessment mimics a real-life communication situation. At the beginning of the year, when these assessments were to take place, the university computer centres are usually fully booked and the organisational energy required to get all students to a central point outside the education faculty building seemed inordinate in relation to the output that would be required.

Automatic assessment systems are designed to predict human ratings of oral proficiency in terms of measures such as fluency, intelligibility and overall pronunciation quality (Bernstein *et al*., 2000; Cucchiarini *et al*., 2000b; Neumeyer *et al*., 2000). Various automatic measures have been investigated and it has been shown that they correlate differently with different aspects of human rating (Cucchiarini *et al*., 2000a). A number of studies have shown that the most promising indicators of human ratings are the so-called posterior, duration and rate of speech scores (Neumeyer *et al*., 2000). In a study on read speech a correlation of more than 0.9 was found between human ratings of fluency and rate of speech (ROS) (Cucchiarini *et al*., 2000a). Based on these findings it was decided to restrict the scope of this pilot investigation to studying the correlation between human ratings of read, repeated and spontaneous speech and automatically derived ROS scores.

Developing the pilot project

The funding for this project provided only for a trialling period that ended in May 2006. There were four distinct phases:

Phase 1: Test development

The overarching goal of the test was to assess listening and speaking skills of limited scope within a specific context (school education). As such, the test can be described as a performance test which Brown (2004: 92) defines as 'any tests that are designed to elicit performances of the specific language behaviors that the testers wish to assess' rather than a task-based test where students are required to complete 'real-life' tasks. This means that 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.

It is self-evident that the test should include instructions and tasks that require comprehension of spoken English and elicit spoken responses from students. To give the test validity it was necessary to create an 'educational' context; in other words, to use test items related to teaching and a school environment. Furthermore, since schools are often very hierarchical institutions, it seemed important to assess students' awareness of socially appropriate language use. There is some evidence from a Finnish study (Pietilla, 1999) that advanced users of English at university level tend to perform equally well as far as grammar use and pronunciation are concerned, but that their sociopragmatic skills are not well developed. A good performance in grammar and pronunciation do not necessarily predict a good performance in sociopragmatic tasks either. These findings strengthened the argument that such tasks should be included in the test.

The construction of the test depended on specific skills that would distinguish students with the potential to follow the 'Big E' course successfully from those who need more explicit language development. The skills that were regarded as important were the following:

Oral proficiency for the Postgraduate Certificate in Education and for purposes of teaching a subject in English will include at least the following:
Ability to recognise and produce appropriate language
Ability to read aloud
Ability to understand and make meaning
Passive vocabulary of academic words
Active vocabulary of academic words
Native-like intonation and stress patterns
Correct pronunciation
Fluency
Ability to respond meaningfully

Within the limits of a phone-in test that depends on the current ASR system, it seemed feasible to test the skills indicated with a 'yes' without the intervention of human raters:

Oral proficiency for the Postgraduate Certificate in Education and for purposes of teaching a subject in English will include at least the following:	**What a phone-in test can conceivably do:**
Ability to recognise and produce appropriate language	Yes
Ability to read aloud	Yes
Ability to understand and make meaning	Yes
Passive vocabulary of academic words	Yes
Active vocabulary of academic words	No
Native-like intonation and stress patterns	No
Correct pronunciation	Yes
Fluency	Yes
Ability to respond meaningfully	No

On the basis of this list a test was developed in which students had to listen and respond to fairly detailed instructions, they had to respond and produce language at word, sentence and above sentence level while demonstrating a feeling for appropriacy and formality.

The test was designed with 7 sections:

(A) Recognising whether a response is appropriate (answering only 'yes' or 'no')
(B) Reading sentences as instructed
(C) Repeating sentences as instructed
(D)Repeating one of two responses in terms of appropriateness
(E) Recognising the correct academic word from two alternatives
(F) Building sentences
(G) Talking spontaneously about a topic

Sections A, D and E are tasks in which a choice must be made from a set of alternatives, sections B, C and F require reading and repetition of previously heard sentences, while section G is an open-ended task. The open-ended task was included with the idea that the lecturer could listen to the recorded speech and assess it at a later stage or as the need arose, e.g. for borderline cases.

Phase 2: Test implementation

In this phase the test had to be embedded in a spoken dialogue system (SDS). The procedure is for the SDS to play the instructions that guide students through the test and record their answers. The system also controls the interface between the computer and the telephone line used during the test. In operational systems the SDS also controls the flow of data to and from the ASR system. However, in the pilot system described here, the students' answers were simply recorded and processed off-line.

Test instructions were converted into a dialogue that explained to students how to respond. Three people acted out the following roles:

- 'Instructor' — explained each section and gave instructions
- 'Example' — provided example utterances
- 'Student' — gave possible answers

The voices were recorded, digitised and incorporated into the SDS. The instructions would look like this:

Instructor:	Section A: Appropriate language. In this section you have to say whether the language use in a number of situations is appropriate or not. For example, think about the following situation at school:
User:	A new member of staff is introduced to the principal and says, 'Hi! Howzit?'
Instructor:	I will ask: Do you think this utterance is appropriate? Please say yes or no. And then you say:
Student:	'No.'

Phase 3: Trial tests

Since the test was only in its trialling phase it was necessary to ask volunteers to do the test. Several announcements were made in classes and, with the promise of a R20 remuneration, 30 students tried out the test. They were also asked to complete a short questionnaire about their home language, their academic performance and their opinion of the Big E / small e course.

Oral instructions were given to the students before they dialled the number that connected them to the computer and the test. In addition to the instructions given by the SDS, a printed copy of the main test instructions was provided during the test, following the example set by the Ordinate test. Since students had to do an exercise where they were required to read certain sentences aloud, they had to have a set of sentences in hard copy. For the rest of the test only the headings of the various sections were provided which meant that the majority of the instructions depended on students' ability to listen and understand. There were no lecturers present while the students were taking the test.

After the test the students had to complete a second questionnaire commenting on the test itself: whether the instructions were clear, whether they found the test difficult and whether the paper copy of the test helped them. Since the results of this questionnaire are not the main focus of this paper, it is sufficient to say that English-speaking students 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.

Phase 4: Rating the students' answers

Once the students had completed the test the real assessment started as the recorded data was transcribed manually (see '*Automatic rater*' below) and the ASR system fine-tuned and its performance tested for the assessment task. For this trialling phase the students' responses were assessed by human raters as well as the ASR system. The aim was to use the correlation — or lack thereof — between the two types of evaluation to determine which automatic scores would be good predictors of human judgement.

Human raters

Eight lecturers were asked to rate speech samples from the read (section B), repeat (section C) and open-ended (section G) tasks in the test. In addition, they were requested to give each student an 'overall impression' mark. The lecturers are all teachers of English (as a second and a foreign language) at Stellenbosch University and they did not know the students they were rating. Assessing language proficiency is part of their job and something they do regularly. The students were divided into four groups and each group was assessed by two different raters. To test for inter-rater consistency, three students were the same for all four groups, i.e. three students were rated by all the lecturers. Intra-rater consistency was tested by presenting two students twice to each lecturer, but using different speech samples.

The role played by human judges and the instruments that they use for oral assessment is well documented. For the purposes of this project much use was made of overview studies, such as those by Jesney (2003), and studies that focus on advanced students of English, such as the study by Sundh (2003). Sundh's focus was very important, because many of the students who took part in this study are home-language speakers of Afrikaans, but they are fluent in English and can therefore be regarded as advanced students of 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 it was found that the use of Likert scales are appropriate specifically for the assessment of accentedness (Jesney, 2003). After studying a wide variety of assessment rubrics and grids a decision was made to use a five-point scale for all four tasks, but to vary the assessment criteria depending on the focus of each task (see Appendix A).

Automatic rater

The 'standard' South African English ASR system that was developed during the AST project was used to assess the data automatically. The system is based on hidden Markov phone models and was trained on approximately six hours of telephone data. A subset of the orthographically transcribed data recorded during the trial tests was used to optimise the system's parameter settings for this application.

During the automatic assessments, the recogniser was used in two different ways. For the 'correct alternative'-type tasks the recogniser was used to transcribe the students' answers automatically. By comparing these transcriptions to the model answers provided by the lecturer, the students' answers could be marked automatically. For the read and repeat tasks the recogniser was used to determine the rate at which the students were speaking. Rate of speech (ROS) was calculated as:

$$\text{ROS} = \frac{\text{number of phonemes}}{\text{total duration of speech + pauses}}$$

Recognising the fluent, spontaneous speech in the answers to section G of the test is beyond the capabilities of the current version of the ASR system. However, because the data was transcribed manually, it was also possible to derive ROS values for the open-ended section of the test. In this task ROS could give a real indication of fluency, since students had to come up with their own ideas in answer to a general question on their ambitions or the challenges of the teaching profession.

Results

Human raters

Table 1 gives an overview of the intra and inter-rater correlations that were determined for each of the eight lecturers who were asked to assess the students' answers.

According to the data in Table 1, the intra-rater correlations for three of the five judges are below 0.5, indicating poor consistency. The table also shows that the inter-rater correlation varies between 0.82 (lecturer 5) and 0.34 (lecturer 3). It was decided to remove lecturers 1, 3 and 8 from the group of human raters because of the low correlation between their judgements and those of the other lecturers. Lecturer 3 was consistent with his own judgements, but his ratings did not correlate well with those of the other raters, while lecturers 1 and 8 had very low intra-rater correlation scores. When the age, gender, years of experience and home-language background of the raters were studied, no single factor could be isolated to explain why these three did not judge as well as the others. Age ranged from 25 to 55; the only male in the group of raters was among them but since he was the only male, it is not possible to deduce that gender may account for these differences. As far as experience is concerned, all the raters had been teaching between 2 and 30 years. The judges whose scores were disregarded had between 2 and 10 years' experience, which does not distinguish them from the group as a whole either. Both home and second-language speakers acted as raters and those whose scores were disregarded included one second-language and two home-language lecturers.

Figure 1 gives an overview of the remaining five lecturers' judgements of the students' answers to sections B, C and G of the test. The fourth value in the figure relates to the 'overall impression' marks the lecturers were asked to give. The highest mark that could be awarded in each section was one and the lowest mark five.

Figure 1 shows that the human raters gave the students fairly high marks. On average, they received the highest marks for the reading task (section B) and the lowest marks for the repeat task (section C). It is interesting to note that the 'overall impression' marks are almost as high as the marks for the reading task, even though the marks for the two other tasks are lower.

Automatic rater

The two sets of bars in Figure 2 compares the results for the 'correct alternative'-type tasks (sections A, D and E) based on the manually created orthographic transcriptions and the automatic transcriptions obtained from the ASR system.

Figure 2 shows a good correspondence between the marks derived from the human and the automatic transcriptions of the students' answers. Where the two sets of results differ (section E), the 'error' made by the ASR system benefits the students. The standard deviation in the human and automatic marks is almost exactly the same.

Figure 3 compares the average ROS values obtained from the human transcriptions of Sections B, C, and F of the test with those obtained automatically from the output of the ASR

Table 1: Intra- and inter-rater correlations for the eight human raters

Lecturer	Intra-rater correlation	Inter-rater correlation
1	0.26	**0.41**
2	0.30	0.62
3	0.72	**0.34**
4	0.66	0.75
5	0.57	0.82
6	0.64	0.76
7	0.76	0.68
8	0.23	**0.42**

system. The figure also shows the average of the ROS values derived from the human transcriptions of Section G.

As Figure 3 shows, the ROS values calculated from the human and automatic transcriptions are well matched. Once again, the standard deviation in the two sets of scores is almost identical. The ROS values derived from the human transcriptions of Section G were included in the figure to show that they are close to the values corresponding to the repeat task. The ROS values for the repeat task could therefore be used as an indication of the ROS a student would have been able to achieve in a free speech task.

Correlation between human and automatic raters

Table 2 gives an overview of the correlation between the human raters' scores and the corresponding ROS values per section. For the 'read and repeat' tasks the ROS values were calculated from

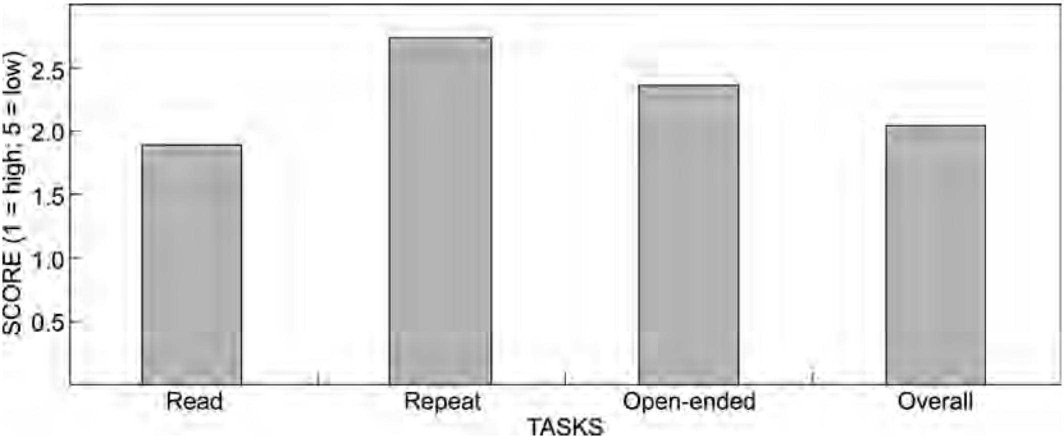

Figure 1: Average scores awarded by the human raters for the read, repeat and open-ended tasks. The last bar corresponds to the 'overall impression' marks

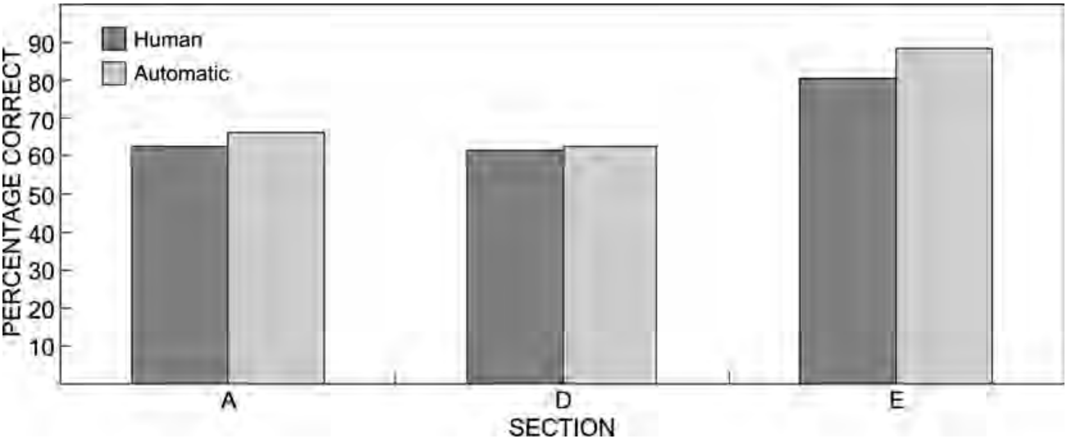

Figure 2: Students' marks for the 'choose the correct alternative'-type tasks derived from human and automatic transcriptions of the data

automatic transcriptions of the data. The ROS values for the open-ended task were derived from the human transcriptions of Section G. The value in the last row of Table 2 is the correlation between the 'overall impression' marks assigned by the human raters and the average value of the ROS values for Sections B, C and G of the test.

In contrast to the findings of other studies, the correlation between ROS and the human ratings of fluency (the main emphasis of the reading task) is low: –0.22. However, the correlation between the human and the automatic raters are higher for the other tasks and compare favourably with those reported in similar studies (Franco *et al*., 2000; Neumeyer *et al*., 2000). The highest correlation is observed between the 'overall impression' marks and the average ROS.

Discussion

The group of students in this project was small and the results are perhaps less interesting than the process of developing and implementing the test. The initial aim of the project, which was to develop a test that is economic in terms of time and human resources and that can be scored objectively, was reached with moderate success.

Most of the students were high achievers, which would account for their volunteering to take part. This meant, however, that their scores did not show much variation. Judges found it difficult to distinguish among them using criteria that were meant to separate weak from good students, rather than good from excellent students (see criteria in Appendix A).

It is clear that the ASR system is open to the same objections as any other computer-assisted language testing method. In his criticism of the use of computers in language testing, Norris (2001) cautions:

Table 2: Correlation between the scores assigned by the lecturers and ROS values

Task	Correlation: Human scores and ROS
Read	–0.22
Repeat	–0.61
Open-ended	–0.59
Overall impression	–0.74

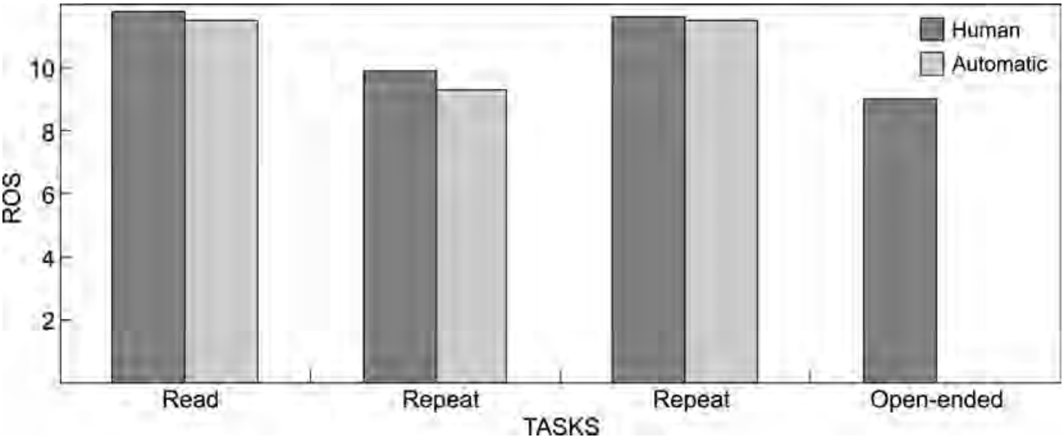

Figure 3: Average ROS values derived from the human and automatic transcriptions of sections B, C, F and G of the test

> [L]anguage test developers need to begin their deliberations about speaking assessment not by asking what computers are capable of doing, but rather by asking (a) what kinds of interpretations actually need to be made about L2 speaking abilities; (b) what kinds of evidence a test will need to provide in order to adequately inform those interpretations; and (c) what kinds of simulation tasks will provide the required evidence?

In constructing the test these questions were answered systematically by working from the interpretations that needed to be made about students' oral proficiency within a very short time and for a very specific reason followed by the process of designing the kinds of tasks that would provide the evidence by means of the ASR system. The types of tasks were limited and there was no attempt to use the results as a full indication of students' competence in English.

A lasting impression of this pilot project remains the inconsistency of human raters and confirms the urgency of finding a more consistent assessment mechanism as part of a battery of tests to determine language proficiency. In cases where students were unhappy with the results, the open-ended part of the test (Section G) would be available for lecturers to listen to and compare to the computer-generated results.

Although the subjectivity of human raters is generally acknowledged in the literature (Oller, 1979; Ellis, 2003), language teachers seem to expect that, with experience, they will develop some objectivity in assessing oral ability. This faith, supported by Oller (1979: 393), seems to be misplaced in the light of the results of this pilot project. If three of the eight lecturers involved in language teaching could differ so widely from their colleagues, the search for more objective and reliable assessment techniques *must* be taken more seriously. In his discussion of communicative language testing, Bachman (1990) criticises the way in which Oller relativises reliability in language testing and calls for more objective assessment methods.

The ASR system had virtually no problems with the 'correct alternative'-type tasks and it will be useful to elaborate on them so as to exploit the system's capabilities. The correlation with human assessments on these tasks was very good. The 'read and repeat' tasks were and remain the most challenging tasks to assess automatically. For this study the rate of speech was determined and used as an indication of fluency. However, in the case of the exercise where students had to remember the constituent parts of a sentence and build a coherent sentence from the jumbled parts, it was much more difficult to draw conclusions on the basis of rate of speech only. If the machine only looked at rate of speech, any fluent sentence, whether it had anything to do with the cued clues or not, would be regarded as evidence of fluency, whereas the purpose was also to test the students' ability to make meaning (albeit to a limited degree).

Prospects for further use

Without much change the system can currently act as recording equipment, providing the opportunity for lecturers to listen to students' answers to the open-ended task and also to get a second examiner to check their assessments. However, it would be ideal if the tasks could be split up and extended so that: (1) students have more opportunities to do shorter versions of the test in the course of the year, and (2) the items could be randomised so that each student gets a different test. The test described in this project took between 15 and 20 minutes to complete. In a class of approximately 120 students who must be assessed within two days, this is not always feasible.

It must be emphasizsed that the application of an automatic speech recognition system cannot replace more qualitative measures such as an oral proficiency interview or prepared speeches. What this system offers is an admittedly rough but reliable and consistent measure that can be used to create a type of yardstick to which more subjective assessments can be compared.

Notes

[1] See the ACTFL website for more detail: htto://www.actfl.org/.

[2] See http://ww.eduspeak.com/ and http://www.auralog.com/.

[3] See http://www.ordinate.com/.

References

Bachman L. 1990. *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bernstein J, De Jong J, Pisoni D & Townshend B. 2000. Two experiments on automatic scoring of spoken language proficiency. *Proceedings of InSTILL 2000*. Dundee: University of Abertay.

Brown JD. 2004. Performance assessment: Existing literature and directions for research. *Second Language Studies* **22**(2): 91–139.

Chalhoub-Deville M. 2001. Language testing and technology: Past and future. *Language, Learning & Technology* **5**(2): 95. Infotrac OneFile. Thomson Gale. University of Stellenbosch. 5 September 2006. http://find.galegroup.com/itx./infomark.do?&contentSet=IAC

Cucchiarini C, Strik H & Boves L. 2000a. Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication* **30**: 109–119.

Cucchiarini C, Strik H & Boves L. 2000b. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America* **107**(2): 989–999.

Ellis R. 2003. *Task-based language learning and teaching.* Oxford: Oxford University Press.

Franco H, Abrash V, Precoda K, Bratt H, Rao R, Butzberger J, Rossier R & Cesari F. 2000. The SRI EduSpeak^{TM} system: Recognition and pronunciation scoring for language learning. *Proceedings of InSTILL 2000*. Dundee: University of Abertay.

Jesney K. 2003. *The use of global foreign accent rating in studies of L2 acquisition*. Report prepared for the Language Research Centre, University of Calgary.

Kenyon DM, Malabonga V & Carpenter H. 2001. Response to the Norris commentary. *Language, Learning & Technology* **5**(2): 106.

Neumeyer L, Franco H, Digalakis V & Weintraub M. 2000. Automatic scoring of pronunciation quality. *Speech Communication* **30**: 83–93.

Norris JM. 2001. Concerns with computerized adaptive oral proficiency assessment. *Language, Learning & Technology* **5**(2): 99.

Oller JW. 1979. *Language tests at school.* London: Longman.

Pietilä P. 1999. L2 Speech: Oral proficiency of students of English at university level. Publications from the Department of English **19**, University of Turku.

Roux JC, Louw PH & Niesler TR. 2004. *The African Speech Technology Project: An assessment*. Proceedings of the 4th International Conference on Language Resources and Evaluation, Vol. 1: 93–96. Lisbon: ELRA.

Sundh S. 2003. *Swedish school leavers' oral proficiency in English*. Dissertation for the degree of Doctor of Philosophy. Uppsala: Uppsala University.

Appendix A: Rating criteria for human raters

Task 1: Pronunciation and phrasing

(A) Correct pronunciation of individual sounds and very good intonation and rhythm. An L2 accent may be discernible.

(B) Very good pronunciation. Good intonation and rhythm. Some influence of native language clearly detectable.

(C) Good pronunciation of most of the individual sounds and acceptable intonation. L2 accent evident.

(D) Individual sounds poorly articulated. Some utterances difficult to understand, putting strain on the listener.

(E) Difficult to understand the pronunciation. First language intonation. Poor articulation of individual sounds.

Task 2: Listening comprehension

(A) Accurate and prompt production of sentences. One or two slight hesitations.

(B) Mostly accurate production. Hesitation obvious.

(C) Tentative production but two or more sentences constructed correctly.

(D) Only one sentence constructed correctly. Few attempts made.

(E) No attempt.

Task 3: Communication and fluency

(A) Comfortable and completely fluent. Performs with ease. Long and well-formed sentences.

(B) Extended contributions at natural speed and tempo. Easy to understand. Occasional hesitations and pauses to search for words.

(C) Easy to understand. Sometimes produces longer and fairly coherent flow of language.

(D) Hesitant contribution. Incomplete sentences.

(E) Little attempt to formulate meaningful contribution. Gives up quickly.

Global evaluation

(A Fluent and overall correct use of English; extremely easy to understand.

(B Very good production with minor inaccuracies. Easy to understand.

(C) Acceptable language in spite of errors. Listener struggles to understand in one or two places.

(D) Not showing enough competence to pass a professional L2 user of language. Errors in pronunciation and grammar interfere with comprehension.

(E) Poor production of English in many respects. Extremely difficult to understand.