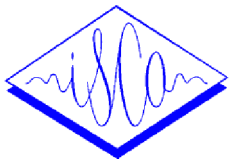
ISCA Archive
http://www.isca-speech.org/archive



Speech and Language Technology in Education (SLaTE2011)
Venice, Italy
August 24-26, 2011

Readability index as a design criterion for elicited imitation tasks in automatic oral proficiency assessment

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

[1]Human Language Technology Competency Area, CSIR Meraka Institute, Pretoria, South Africa.
[2]Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa.
[3]Department of Curriculum Studies, Stellenbosch University, South Africa.
fdwet@csir.co.za, pfdevmuller@gmail.com, {cvdwalt,trn}@sun.ac.za

Abstract

We investigate the effectiveness of using an accepted readability index, the Flesch Reading Ease (FRE) scale, to design prompts for an automatic oral proficiency assessment system. The prompts in question are uttered by the system, and must be repeated from memory by the test subjects, in the form of an elicited imitation exercise. The FRE scores for our prompts are shown to correlate well with the repeat accuracy ratings given to the same prompts by human judges. The prompts are also shown to perform well in the automatic assessment of oral proficiency by means of the elicited imitation tasks with more than one group of university students. We therefore conclude that a readability index, such as the FRE, can be used as a design criterion for prompts in an automated elicited imitation test.

Index Terms: elicited imitation task, readability index, Flesch Reading Ease scale, automatic oral proficiency assessment

1. Introduction

Proficiency in the language of instruction is key to a student's success at university. Most existing tests of language proficiency, and especially academic literacy proficiency, focus on reading and writing skills, because listening and speaking skills are more time-consuming and are difficult to assess consistently. Providing efficient and reliable assessments is particularly challenging when many students must be evaluated simultaneously and the results be available within a short space of time [2].

The application of automatic speech recognition (ASR) techniques to automatically assess oral proficiency and listening comprehension is one way in which these logistical problems can be obviated. Another appealing feature of automatic tests is that they provide a means to assess consistently and objectively, even for smaller groups for which human assessment would be feasible. This is in contrast to human assessments that are often plagued by inconsistency and subjectivity.

The automatic test used during this investigation evolved from a number of piloting experiments, and includes a reading and a repeating task. [1, 2]. A number of experiments were conducted with students who were enrolled for a post-graduate teaching certification course, as well as with undergraduate students. The majority of the students in question speak Afrikaans as a first language. The results consistently showed that the students did not find the reading task challenging, and obtained very high scores. This left little room for discrimination between different levels of proficiency. The repeating task was more useful because it yielded a wider range of scores and higher correlations between human and automatic assessments of the same data [2]. This study will therefore focus on the repeating task as an example of an elicited imitation exercise.

In addition to the ASR technology, implementing an automatic test including elicited imitation tasks requires a database of possible test items from which tests can be compiled automatically. With the exception of the work by Graham et al. [3], not much has been published on the design of elicited imitation tasks for the purpose of automatic assessment, although they are used implicitly by many automatic assessment systems, e.g. shadowing [4]. This paper investigates the feasibility of using a readability index, quantified in terms of the Flesch Reading Ease (FRE) scale, as a design criterion for elicited imitation tasks in automatic oral proficiency assessment.

2. Elicited imitation tasks

Elicited imitation (EI) tasks are constructed by recording sentences of varying lengths and complexity, and playing them to test subjects who must repeat what they heard. This is effective because oral proficiency is linked to the capacity of a speaker's working memory to store and retrieve information: more proficient speakers will be able to repeat more complex sentences more accurately [3]. The participant's working memory must be challenged by sentences that are slightly too long to be imitated by rote and therefore require the synthesis of meaning by 'chunking' parts of the sentence. In verbal interaction, the capacity of the working memory is crucial to construct meaning, and therefore the ability to imitate the sentences of a language demonstrates proficiency in that language [5]. In a second language, the capacity of the phonological working memory is constrained [2, 6]. Time pressure and limited access to the vocabulary and sound system of the L2 may impact negatively on the speaker's ability to chunk information when a sentence must be repeated. This deterioration can be measured by EI.

Since EI challenges the participants' working memory [3], it is a reasonable measure of 'global proficiency' [5]. It must however be kept in mind that fluent users of a language will sometimes repeat the sense of a sentence rather than the exact wording. Chaudron et al. speculate that such re-interpretation "suggests that an abstract message-meaning level can be accessed in EI" [5]. For this reason, the rating criteria for the EI part of the test allowed a correct repetition and a correct interpretation to be judged on equal terms.

Initially, Bley-Vroman and Chaudron were tentative about the usefulness of EI, maintaining that "[w]e regard it as premature to view elicited imitation as a proven method for inferring learner competence, because a considerable amount of research needs to be conducted to understand how performance under imitation conditions compares with other methods and with learner's underlying knowledge" [7]. By 2005, however, Chaudron et al. were more confident that EI can demonstrate

global (and not just oral) language proficiency [5].

In terms of oral proficiency assessment, in particular, EI has been found to outperform oral interviews and sentence completion exercises on measures of validity and reliability [8]. EI therefore appears to be a good foundation on which to base the design of oral proficiency tests.

3. Automatic oral proficiency testing

The automatic test used in this study consists of a spoken dialogue system, running on a desktop PC and administered in a university multi-media laboratory using headsets with directional, noise cancelling microphones. On average, the students took around 7 minutes to complete the test. During the test, students were prompted to read sentences from a test sheet as well as to repeat utterances produced by the system. As motivated in Section 1, the scope of this investigation is limited to the repeating task as an example of an elicited imitation exercise.

3.1. Prompt design

The level of sentence difficulty for this task was linked to the Flesch Reading Ease (FRE) scale in an attempt to find a consistent measure of complexity. The FRE score is a syllable- and sentence-based measure defined as:

$$FRE = 206.835 - 1.015 \cdot \frac{N_{wrd}}{N_{snt}} - 84.6 \cdot \frac{N_{syl}}{N_{wrd}}$$

where N_{snt}, N_{wrd} and N_{syl} are the total number of sentences, words and syllables respectively in the text under analysis. The FRE measures the complexity of a text, with higher FRE scores associated with easier material. For the purposes of this study, the FRE provided a consistent method of generating comparable sentences that are slightly too long to be imitated by rote, as required for elicited imitation tasks. Although readability measures do not measure syntactic complexity directly, they do provide a stable measurement of sentence and word lengths. This property proved an effective means of choosing samples that would require the participants to chunk the information in a sentence before repeating it.

Readability scores are often criticized, because they are generated at sentence level and do not take the overall structure of a text into account [9]. In the case of this project, however, the focus of imitation was on sentence level and the FRE scale was regarded as suitable.

For the purposes of initial and exploratory experimentation, the test prompts were divided into two sets: a fairly easy (*Repeat A*) and more challenging (*Repeat B*) set. The readability scores ranged from 65.7 to 85.2 (average = 70.5) for *Repeat A*, and from 46.6 to 57.7 (average = 50.8) for *Repeat B*. EI tasks depend on the participants' familiarity with vocabulary and grammatical structures for successful repetition [5]. Since most of the participants in this study are advanced users of English, the vocabulary was controlled by focusing on educational settings with which participants would be familiar, for example:

Repeat A

- I don't see useful teaching techniques in the schools. (FRE = 85.2)
- I learn nothing from other students' mistakes. (FRE = 66.9)

Repeat B

- Students' new teaching methods are scorned by experienced teachers. (FRE = 47.9)
- Teachers often resist change and don't want to see new methods, unfortunately. (FRE = 54.2)

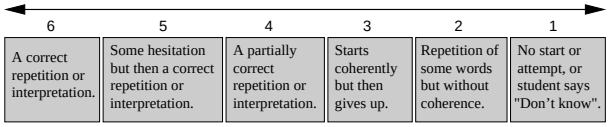


Figure 1: *Scale used by humans to rate repeated prompts.*

In total, there were seven sentences in *Repeat A* and eight in *Repeat B*. The system randomly selected three sentences from each set during the test, so that each student was prompted to repeat six utterances of which three were fairly easy and three were more challenging.

3.2. Test population

The test was taken by 55 first-year undergraduate students at Stellenbosch University. The students were divided into four groups in terms of their overall performance in their first year English Language and Literature module. The first group contained those students achieving an average mark between 40 and 49% (n=14), the second between 50 and 59% (n = 13), the third between 60 and 69% (n = 16), and the fourth between 70 and 75% (n=12). A mark of 40% is the lowest possible, while 75% represents the top mark among the 55 students. From these four groups, random selections were made to obtain a spread of participants from low to high scoring individuals.

4. Human assessment

Six teachers of English as a second or foreign language were asked to rate the responses of the 55 students, and each teacher rated at least three students twice. The intra-rater reliability for the majority of the teachers was above 0.9. Inter-rater agreement was determined in terms of two way, intraclass correlation coefficients (ICCs). The ICC values indicated that the evaluations of two of the teachers differed substantially from those of the other four (ICC < 0.2). The ratings of these two teachers were therefore not taken into account during the rest of the study. The inter-rater agreement for the four remaining teachers varied between 0.87 and 0.88.

Raters were asked to evaluate the extent to which students grasp the meaning of what they hear, and how well they repeat and/or rephrase what they heard, using the scale illustrated in Figure 1. These scales were drawn up [2] after listening to a sample of the recordings and differentiating among the various attempts at repetition and interpretation, as suggested by Upshur and Turner [10] and by Fulcher [11]. This process resulted in scales that are tailored to our particular population where, for example, participants repeated (interpreted) chunks of a sentence without maintaining syntactic coherence ("A partially correct repetition or interpretation") or they repeated (interpreted) only the first part of the sentence ("Starts coherently but then gives up"). The scale from 6 to 1 can be described as decreasing evidence of chunking, ending with "Repetition of some words but no coherence", where no chunking took place at all.

The raters were not provided with the numerical values indicated in Figure 1. These were used only to quantify the ratings for subsequent correlation with machine scores.

5. ASR system

The Hidden Markov Model Toolkit (HTK) version 3.4 was used for ASR [12]. The hidden Markov models (HMMs) used by the speech recogniser were trained on approximately 6 hours of telephone quality speech by English mother-tongue speak-

ers [13]. Triphone HMMs were obtained by means of decision-tree state clustering and embedded Baum-Welch re-estimation. The final set of triphone HMMs consisted of 4797 tied states based on a set of 52 phones, and 8 Gaussian mixtures per state.

Unigram language models (LMs) were used for the automatic recognition of the repeated prompts. For this task, provision must be made for missing words and changes in word order. A separate unigram LM was therefore created for each prompt of the repeating task. Each LM consisted of an unweighted word loop, with word-to-word transitions having equal probability. Silence and noise were allowed as insertions between words. The word insertion penalty and language model scale factor were optimised on a development set in previous experiments [2].

6. Automatic assessment

The output of an ASR system can be used in various ways to extract quantitative features from speech signals. Various techniques to derive these so-called *machine scores* from speech data have been reported on in the literature [14, 15]. In previous studies, we consistently found low correlation between posterior scores and human ratings [16, 2]. This investigation will therefore be restricted to scores derived from segmentation information and from repeat accuracy, because, to date, these have shown the highest correlation with human ratings [17].

6.1. Scores derived from segmentation information

The scores based on segmentation information focus on the *temporal* features of speech, rather than on acoustic characteristics, and are calculated from phone-level alignments. A distinction is made between the *speech phones* (those forming part of words) and *non-speech phones* (those forming part of silence or noise) in each utterance. In a previous study, four segmentation-based scores were investigated: rate of speech, articulation rate, phonation/time ratio and segment duration scores. The highest correlation between a segmentation based score and the human ratings of the same data was observed for rate of speech [17], to which and the scope of this study will consequently be restricted.

6.1.1. Rate of Speech

The *Rate of Speech* (*ROS*) of an utterance is defined in [18] as the number of speech phones per second, calculated using the number of speech phones in the utterance M_{Speech}, and the total duration of the utterance T_{Total}, in seconds:

$$ROS = \frac{M_{Speech}}{T_{Total}}$$

Any silences leading or trailing the utterance are ignored when determining the total duration.

6.2. Scores derived from repeat accuracy

Speech recognition accuracy can also be used as a score for automatic assessment. Two closely-related alternatives were evaluated in this study: ASR Accuracy and ASR Correct.

6.2.1. ASR Accuracy

The score *ASR Accuracy* (Acc_{ASR}) is calculated by determining a dynamic programming-based string alignment between the recogniser output and the reference transcription [12]. The number of correctly aligned words (H), the number of insertions (I), and the number of words in the reference transcription (W) determined from this alignment are used to calculate the score as follows:

$$Acc_{ASR} = \frac{H - I}{W} \times 100\%$$

where it is noted that this score is penalised by insertions.

6.2.2. ASR Correct

The score *ASR Correct* (Cor_{ASR}) reflects the percentage of reference transcription words that match and align with words in the recogniser output [12].

$$Cor_{ASR} = \frac{H}{W} \times 100\%$$

In contrast to Acc_{ASR}, this score is not influenced by insertions.

7. Results

7.1. Correlation between human ratings and FRE scores

Table 1 shows the correlation[1] between the per-utterance average of the human scores for the repeated prompts (as defined in Figure 1) and the difficulty of the corresponding utterance on the FRE scale. The results are shown for *Repeat A* and *Repeat B* separately as well as for the exercise as a whole (*Repeat*). The p-values associated with the correlations are also listed.

Table 1: *Correlation between average per-utterance human ratings and utterance difficulty according to the FRE scale for the repeating prompts.*

	Correlation	p-value
Repeat A	0.49	0.26
Repeat B	0.57	0.14
Repeat	0.86	0.00

According to the results in Table 1, the hypothesis that there is a correlation between the FRE score and the human ratings of the repeat accuracy of an utterance should be rejected if *Repeat A* and *Repeat B* are considered separately. However, for a combination of the easy and more challenging tasks, there is a high and significant correlation between the FRE scale levels and the human ratings. This seems to indicate that the FRE scale values can be used as a design criterion for elicited imitation exercises, provided that the exercises include utterances with varying levels of difficulty - as is indeed required by test designs that attempt to assess overall speaking proficiency [3].

7.2. Correlation between human ratings and machine scores

The correlations between the machine scores defined in Section 6 and the average ratings given to the same utterances by human raters are shown in Table 2. All the correlations in the table are significant at a 95% confidence level (p-values < 0.05).

Table 2 shows that, for *Repeat B*, the correlations between Cor_{ASR} and the human ratings of repeat accuracy are much higher than the corresponding values for Acc_{ASR}. In fact, the highest correlation between a machine score and the human ratings is observed for Cor_{ASR}. ROS exhibits the second highest correlation with the human ratings.

Table 2 also shows that the correlations associated with *Repeat B* are higher than those measured for *Repeat A*, with those for the whole test somewhere in between. *Repeat B* includes a much wider range of human ratings and corresponding machine

[1] All correlation values are Spearman's rank correlation coefficients.

Table 2: *Correlation between average per-utterance human ratings and corresponding machine scores.*

	ROS	Acc_{ASR}	Cor_{ASR}
Repeat A	0.47	0.42	0.42
Repeat B	0.65	0.49	0.84
Repeat	0.55	0.36	0.67

scores than *Repeat A*. Higher correlations are observed when scores span wider ranges of the assessment scale than when they are limited to a small interval. These trends and results are in good agreement with those from previous studies involving post-graduate students, indicating some degree of consistency of the test over different test populations [2, 17].

8. Discussion and conclusions

In this study we investigated the use of a readability index, quantified in terms of the Flesch Reading Ease (FRE) scale, as a design criterion for elicited imitation (EI) exercises in automatic proficiency assessment systems. The FRE provides an instrument whereby sentences can be measured consistently in terms of length (number of words) and lexical complexity (number of syllables). This allows for the generation of sentences which challenge the phonological and lexical working memory in a controlled manner. In order to distinguish among participants who were all advanced learners of English, it was necessary to confine this complexity to lie between certain defined limits. For EI exercises this is important, since varying levels of sentence complexity need to be represented to give a reliable picture of oral proficiency [3].

The evidence provided by the correlation between the FRE scores and the average of the scores given by the human judges indicates that readability index can indeed be a useful design criterion in this respect. Furthermore, the substantial difference in these correlations between the more narrowly-defined Repeat A and Repeat B tests, and the combined test, indicates that the use of the FRE becomes increasingly accurate as the prompt set includes a wider variety of sentence difficulty levels. Since the correlation between the machine scores and the human ratings is highest for the more difficult subset of prompts (Repeat B), it appears that, for a set of proficient speakers such as ours, it is advantageous to maximise the average level of sentence difficulty in the prompt set.

In terms of both test design and implementation, the EI approach may offer a solution to the subjectivity associated with oral proficiency assessment by human judges. Moreover, the use of the FRE score may allow the variation in the complexity of oral proficiency tests to be objectively controlled. At present, such tests range from oral proficiency interviews to extensive prepared speeches [8]. Future investigations will focus on the relationship between human ratings and written assessments of participants' language proficiency to determine the degree to which oral proficiency correlates with more open-ended measures of oral proficiency and with written products.

9. 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.

10. References

[1] 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, pp. 135–146, 2008.

[2] F. De Wet, C. Van der Walt, and T. R. Niesler, "Automatic assessment of oral language proficiency and listening comprehension," *Speech Communication*, vol. 51, pp. 864–874, 2009.

[3] C. R. Graham, D. Lonsdale, C. Kennington, A. Johnson, and J. McGhee, "Elicited imitation as an oral proficiency measure with ASR scoring," in *Proc. 6^{th} Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008, pp. 1604–1610.

[4] D. Luo, Y. Yamauchi, and N. Minematsu, "Speech analysis for automatic evaluation of shadowing," in *Proc. Interspeech 2010 Satellite Workshop on Second Language Studies: Acquisition, Learning, Education and Technology*, Tokyo, Japan, 2010.

[5] C. Chaudron, H. Nguyen, and M. Prior, "Manual for the Vietnamese elicited imitation test," 2005, http://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/10601/RN41m.pdf.txt?sequence=117, (accessed April 2011).

[6] 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.

[7] R. Bley-Vroman and C. Chaudron, "Elicited imitation as a measure of second-language competence," in *Research Methodology in Second-Language Acquisition*, E. E. Tarone, S. M. Gass, and A. D. Cohen, Eds. Lawrence Erlbaum Associates, 1994, pp. 245–360.

[8] G. Henning, "Oral proficiency testing: comparative validities of interview, imitation, and completion methods," *Language Learning*, vol. 33, no. 3, pp. 315–332, 1983.

[9] C. M. H. Shehadeh and J. B. Strother, "The use of computerized readability scores: Bane or blessing?" in *Proc. Annual Conference of the Society for Technical Communication*, vol. 41, 1994, p. 225.

[10] J. A. Upshur and C. E. Turner, "Constructing rating scales for second language tests," *English Language Teaching Journal*, vol. 49, no. 1, pp. 3–12, 1995.

[11] G. Fulcher, "The testing of L2 speaking," in *Encyclopedia of Language and Education: Language testing and assessment*, C. Clapham and D. Corson, Eds. Dordrecht: Kluwer Academic, 1997, vol. 17, pp. 75–85.

[12] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[13] J. C. Roux, P. H. Louw, and T. R. Niesler, "The African Speech Technology project: An Assessment," in *Proc. LREC*, Lisbon, Portugal, 2004, pp. I:93–96.

[14] Various Authors, "Special Issue on Language Learning," *Speech Communication*, vol. 30, no. 2-3, 2000.

[15] ——, "Special Issue on Spoken Language Technology for Education," *Speech Communication*, vol. 51, no. 10, pp. 831–1038, 2009.

[16] P. F. De V. Müller, F. De Wet, C. Van der Walt, and T. R. Niesler, "Automatically assessing the oral proficiency of proficient L2 speakers," in *Proc. SLaTE*, Warwickshire, UK, 2009, CD-ROM.

[17] F. De Wet, P. F. De V. Müller, C. Van der Walt, and T. R. Niesler, "Using segmentation and accuracy-based scores to automatically assess the oral proficiency of proficient L2 speakers," in *Proc. 21^{st} Annual Symposium of the Pattern Recognition Association of South Africa*, F. Nicolls, Ed., Stellenbosch, South Africa, 2010.

[18] 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.