Segmentation and accuracy-based scores for the automatic assessment of oral proficiency for proficient L2 speakers

Febe de Wet[1,3], Pieter Müller[1], Christa van der Walt[2] & Thomas Niesler[1]
[1]Department of Electrical and Electronic Engineering
[2]Department of Curriculum Studies, Stellenbosch University, South Africa
[3]HLT Research Group, CSIR Meraka Institute, South Africa.
pfdevmuller@gmail.com, {fdw,cvdwalt,trn}@sun.ac.za

***Abstract*—This paper reports on the automatic assessment of oral proficiency for advanced second language speakers. A spoken dialogue system is used to guide students through an oral test and to record their answers. Indicators of oral proficiency are automatically derived from the recordings and compared with human ratings of the same data. The proficiency indicators investigated here are based on the temporal properties of the students' speech as well as the their ability to repeat test prompts accurately. Results indicate that, both for segmentation as well as accuracy-based scores, the most simple scores correlate best with the humans' opinion on the students' proficiency. Combining different scores using multiple linear regression leads to marginally higher correlations. However, these improvements are too small to justify the associated increase in the computational complexity of the system.**

I. Introduction

Automatic language proficiency assessment can assist language teachers when student numbers make individual assessment infeasible. Even for limited student numbers, automatic assessment tools can contribute to the educational process by allowing more frequent testing. In addition, they offer consistent and impartial evaluation - both factors that are known to be problematic in human assessment. While text-based tests that assess reading and writing skills can already be automated by restricting them to multiple-choice questions, large-scale assessment of perceptual and oral skills remains a challenge.

This paper presents the most recent results obtained during the development of an automatic test for the assessment of the oral proficiency and listening comprehension of university students. Previous studies have shown that the majority of the students in our test population are very proficient in the L2 we are trying to assess (South African English) [1], [2]. As a consequence, our results differ from those reported by other researchers for students who are less proficient or are from more diverse populations.

More specifically, some of our previous experiments showed that applying non-linear transformations to the machine scores does not improve the correlation between the automatic measures and the human ratings of the data. In the same study it was found that there is no substantial difference between effectiveness of context independent (monophone) and context dependent (triphone) acoustic models to score learner speech automatically [1]. A related set of experiments, aimed at improving the correlation between the human rating of the data and automatically derived proficiency indicators based on posterior log-likelihood, revealed that even the most promising posterior scores correlate poorly with human assessments [2].

In this paper we elaborate on these investigations by evaluating the power of segmentation-based scores other than Rate of Speech (ROS), which was used in our previous experiments to predict human judgements of students' answers. We also investigate alternative accuracy-based scores. Each score's potential to predict human assessments of oral proficiency is evaluated by calculating the correlation[1] of the machine score with the human ratings applied to the data. In addition, we aim to determine whether a combination of scores is better at predicting human ratings than single scores.

II. Automatic test

Our application was implemented as a telephone-based test. The test consisted of a number of tasks in which students had to listen and respond, either by reading a sentence from a test sheet or by responding to an instruction. In this paper, we will focus on the following two tasks:

- **Reading task:** Students are provided with a list of sentences on a printed test sheet. The system randomly chooses six of these sentences, and instructs students to read each one in turn. For example, *"School governing boards struggle to make ends meet."*
- **Repeating task:** Students are asked to listen to sentences uttered by the system and to repeat the same sentence. For example, *"Student teachers do not get enough exposure to teaching practice."*

The sentences in the repeat 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 better working memory capacity in the second language would make it possible for them to repeat the sentences accurately.

A spoken dialogue system was developed to guide students through the test and to capture their answers. The system plays the test instructions, records the students' answers, and controls the interface between the computer and the telephone line. In a fully operational system, it would also control the flow of data to and from the automatic speech recognition (ASR) system, but in our set-up the students' answers were simply recorded for later, off-line processing.

120 students took the test as part of their oral proficiency assessment. The majority of the students speak Afrikaans as a first language and their proficiency in English varies from intermediate to advanced. Calls to the dialogue system 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 dialogue system, a printed copy of the test instructions was provided. No staff were present while the students were taking the test.

Feedback received immediately after completion of the test indicated that English-speaking students generally found the test manageable while the majority of Afrikaans students found it fairly

[1]All correlation values are expressed in terms of Spearman's rank correlation coefficients.

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.

III. Human assessment

Six teachers of English as a second or foreign language were asked to rate speech samples from the reading and repeating tasks in the test. For the reading task, each sentence was assessed on three separate scales in terms of degree of hesitation, pronunciation (including accent) and intonation. Figure 1 shows the assessment scales used for the reading task. Scores below three on the scale would indicate students who need additional language support. The raters were provided only with the scale but not the scores.

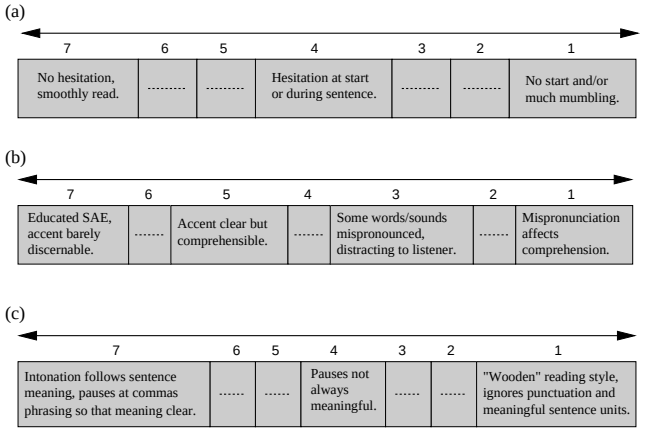

Fig. 1. Scales used to assess (a) degree of hesitation (b) pronunciation and (c) intonation in the reading test.

For the repeating task, the same process was followed. In this case, separate scales were designed to measure the success with which a repetition was formulated and the accuracy of the repetition, as shown in Figure 2. The scales were accompanied by precise descriptions for all the categories.

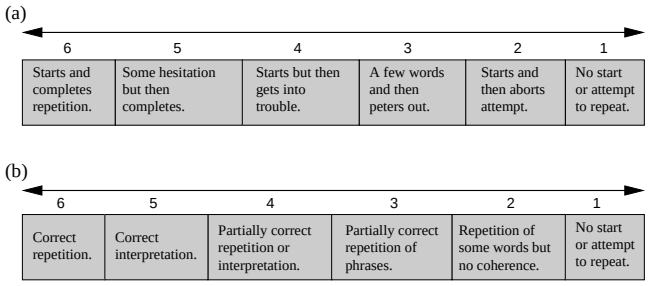

Fig. 2. Scales used to assess (a) degree of success and (b) accuracy in the repeating test.

Since the test was intended to measure listening comprehension, it seemed fair to distinguish correct *repetitions* from correct *interpretations*, since the latter would indicate that the students responded by interpreting what they heard. This kind of behaviour offers a glimpse into a speaker's working memory, which seems to reduce information into meaningful chunks in order to make sense of an incoming message.

In an initial training session, the scales were explained and some sample responses were played to clarify what was meant by categories such as "*Starts and then gets into trouble*" as opposed to "*A few words and then peters out*". The intra-rater correlations obtained using these assessment scales varied between 0.67 and 0.96 with an average value of 0.85 for the six raters who participated in the experiment.

Table I shows the correlation between the individual assessment criteria. Other researchers have reported correlations of more than 0.75 between human ratings of fluency and segmental quality for read speech [3]. In contrast, only degree of hesitation and intonation seem to show noteworthy correlation in our data. This difference can probably be ascribed to the fairly advanced L2 proficiency of most of the speakers in our test population. The relationship between fluency and pronunciation may be much weaker for these students than for less proficient speakers. The data in Table I also reveals a strong correlation between degree of success and repeat accuracy.

TABLE I
CORRELATION BETWEEN DIFFERENT ASPECTS OF PROFICIENCY EVALUATED IN THE READ AND REPEAT TASKS.

Task	Assessment criteria	Correlation
Read	Hesitation & pronunciation	0.40
Read	Hesitation & intonation	0.66
Read	Pronunciation & intonation	0.41
Repeat	Success & accuracy	0.89

IV. ASR system

The Hidden Markov Model Toolkit (HTK) version 3.4 was used for ASR [4]. The hidden Markov models (HMMs) used by the speech recogniser were trained on approximately 6 hours of telephone quality speech by English mother-tongue speakers. This data is part of the African Speech Technology (AST) corpus, and consists of phonetically and orthographically annotated speech gathered over South African fixed as well as mobile telephone networks [5]. Triphone HMMs were obtained by means of decision-tree state clustering and embedded Baum-Welsh re-estimation. The final set of triphone HMMs consisted of 4797 tied states based on a set of 52 phones, and a maximum of 8 Gaussian mixtures per HMM state.

Finite State Grammars (FSGs) were used for the automatic recognition of the *reading task*. It is expected that the students, who generally have good English reading skills, would make very few errors while reading prompts from a test sheet. Hence the use of a strict finite state grammar (FSG) is an appropriate recognition method for this task. For each prompt in the reading task an FSG was created allowing the desired utterance, as well as "*I don't know*" or simply "*don't know*". The branch allowing the desired utterance expects all words to be present in the correct order, but allows inserted silence, noise and hesitation sounds. These prompt-specific grammars were defined using extended Backus-Naur form (EBNF) notation and were parsed to lattice files that were used during recognition.

Unigram language models (LMs) were used for the automatic recognition of the *repeating task*. For this task, provision must be made for missing words, changes in word order, and the replacement of words or phrases with synonyms. This makes the use of a strict FSG less attractive than the use of a unigram language model, which places no restrictions on word order. A separate LM was therefore created for each prompt of the repeating task. Each LM consisted of a word loop, where the allowed words were obtained from the human transcriptions of the development set, as well as all words occurring in the prompt. Silence and noise sounds were again allowed between words. The word loop was unweighted, meaning that all word-to-word transitions had equal probability.

A sub-set of the data (30 speakers) was used to optimise the recogniser's word insertion penalty and language model scaling factor for these two configurations. The remainder of the data (90 speakers) was used as an independent test set.

V. Scores derived from segmentation information

Four segmentation-based scores were investigated in this study: rate of speech, articulation rate, phonation/time ratio and segment duration scores. These scores focus on the *temporal* features of speech, rather than on its acoustic characteristics, and are calculated from phone level alignments. A distinction is also made between the *speech phones* (those forming part of words) and *non-speech phones* (those forming part of silence or noise) in each utterance.

A. Rate of Speech

The *Rate of Speech* (*ROS*) of an utterance is defined in [3] 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.

B. Articulation Rate

Articulation Rate (*ART*) is similar to *ROS*, but does not take the duration of silence and noise in the utterance into account [6]. It is calculated using the total duration of speech phones in the utterance, T_{Speech}, rather than the total duration:

$$ART = \frac{M_{Speech}}{T_{Speech}}$$

C. Phonation/Time Ratio

The *Phonation/Time Ratio* (*PTR*) is the fraction of the utterance duration that consists of speech phones [6]. It is defined as:

$$PTR = \frac{T_{Speech}}{T_{Total}}$$

where T_{Speech} is the duration of all speech phones in the utterance and T_{Total} is the total duration of the utterance, ignoring leading or trailing silences.

D. Segment Duration Score

The *Segment Duration Score* (*SDS*) compares the duration of each phone in an utterance with the expected duration of that phone based on training data. It is based on the argument that the training data reflects the pronunciation expected from proficient speakers [7].

To allow for variations in speech rate between speakers, the duration of each phone is normalised by multiplication with the utterance *ROS*. This is done for the utterances to be evaluated, as well as for the training data. We define this normalised duration of a phone q_i as $f(q_i)$, where d_i is the duration of q_i:

$$f(q_i) = d_i \cdot ROS$$

The phone level *SDS* is defined as the probability of the normalised duration of the phone, given the type of phone:

$$SDS(q_i) = C_{q_i} \cdot p\big(f(q_i)|q_i\big)$$

The probability $p\big(f(q_i)|q_i\big)$ is based on a discrete distribution of normalised durations for the given phone, determined from the training data. The scaling factor C_{q_i} is associated with the probability distribution for the given phone, and is defined so that a phone duration corresponding to the peak of the probability distribution results in a phone level *SDS* score of 1. The omission of this scaling factor would result in uneven scoring between different phones. Lower scores would be assigned to phones with broader probability distributions, which do not have well-defined peaks, even when pronounced perfectly, i.e. with a normalised duration matching that of the distribution peak. By associating a scaling factor with each probability distribution, we can scale scores in such a way that all perfectly pronounced phones are assigned the same maximum score of 1, irrespective of the shape of their duration distributions. Finally, the utterance level *SDS* is defined as the average phone level *SDS* for the given utterance.

E. Results

Tables II and III show that the *Rate of Speech* scores are relatively well correlated with all human rating scales. *ROS* has the highest correlations of all segmentation based scores with the human ratings for *Intonation*, *Success* and *Accuracy*.

	Hesitation	Pronunciation	Intonation
ROS	0.54	0.48	**0.49**
ART	0.41	**0.50**	0.46
PTR	**0.64**	0.18	0.39
SDS	0.15	0.18	0.00

TABLE II
Correlation of scores derived from segmentation information with human ratings for the reading task.

	Success	Accuracy
ROS	**0.67**	**0.65**
ART	0.60	0.58
PTR	0.45	0.44
SDS	0.61	0.56

TABLE III
Correlation of scores derived from segmentation information with human ratings for the repeating task.

The performance of *ART* is similar to that of *ROS*, with a slight increase in correlation with the ratings for *Pronunciation*. *PTR* is best correlated with the human ratings for *Hesitation*, and yields the highest correlation with this scale of all the segmentation scores. The correlations between *PTR* scores and the other human rating scales are low compared to those of *ROS* and *ART*.

Finally, the correlations of the *SDS* scores with the three reading task scales are negligible. The correlations with the two repeating task scales are comparable with those of *ART*, but lower than those of *ROS*.

VI. Scores derived from repeat accuracy

We evaluated three scoring algorithms based on speech recognition accuracy: ASR Accuracy, ASR Correct and Weighted Correct. Because the students in our test population are generally highly proficient, they achieved a reading accuracy score of 100% for most prompts in the exercise. Reading accuracy is therefore not useful in scoring proficiency automatically. Repeat accuracy, on the other hand, is more variable and is considered here.

Results for two recognition strategies are presented, the first using a unigram language model, and the second an oracle FSG. Oracle FSG recognition is similar to recognition using an FSG grammar as described in Section IV. However, instead of creating a grammar for

each prompt based on the desired utterance, grammars are created for each utterance based on the human transcription of the actual utterance. As before, silence, noise and hesitation sounds are allowed between words[2]. The use of an oracle FSG is intended to obtain recognition output with a zero word-error rate, while including silence and noise where appropriate. The results for the oracle FSG grammar are included to indicate what the effect of better recognition accuracy would be on the performance of the machine scores.

A. ASR Accuracy

The score *ASR Accuracy* (Acc_{ASR}) is calculated using the HTK tool *HResults*, which uses a dynamic programming-based string alignment procedure to align the recogniser output with the reference transcription [4]. It counts the number of correctly aligned words (H), the number of insertions (I), and the number of words in the reference transcription (W). The score is then calculated as:

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

Note that this score is penalised by insertions. When the number of insertions exceeds the number of correctly recognised words, the score is negative.

B. ASR Correct

The score *ASR Correct* (Cor_{ASR}) indicates the percentage of reference transcription words present in the recogniser output [4], and is defined as:

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

In contrast to Acc_{ASR}, this score does not take insertions into account. It is also calculated by the HTK tool *HResults*.

C. Weighted Correct

When calculating Cor_{ASR}, all words in the reference transcription are regarded as equally important. However, it may be argued that, when human raters are assigning values to *Accuracy* or *Success*, they may penalise speakers less for missing certain unimportant words than for missing words that are more central to the semantic meaning of the target utterance.

To investigate this, we assign a rank to each word in the reference transcription as a measure of that word's semantic importance. In some cases, adjacent words are grouped together to form a phrase, which is then assigned a single rank. Each rank is associated with a *weight*, which represents the number of marks that will be awarded if the corresponding word or phrase occurs in the recogniser output.

The *Weighted Correct* ($Cor_{Weighted}$) score is defined as the percentage of marks that were awarded:

$$Cor_{Weighted} = \frac{\sum_{i=1}^{H} w_i}{\sum_{j=1}^{W} w_j} \times 100\%$$

where H is the number of correct words or phrases and W is the total number of ranked words or phrases in the reference transcription. The weight associated with the i^{th} correct word or phrase is indicated by w_i, and the weight associated with the j^{th} word or phrase in the reference transcription by w_j.

[2]While human transcriptions can be seen as the most accurate representation of the words in an utterance, non-speech events such as silence and noise are often not accurately transcribed.

The eight prompts for the repeating task were analysed[3] and a rank from 1 to 6 was assigned to each word or semantic group of words. Ranks were assigned by identifying the head of the sentence and elements that modify the head, as defined in [8]. A rank of 1 was assigned to words or phrases with the least semantic importance, and a rank of 6 to those with the most semantic importance.

The weights associated with the different ranks can be adjusted in an effort to approximate the relative importance human raters would attach to each rank. The more accurate the approximation, the stronger the correlation between the $Cor_{Weighted}$ scores and human ratings should be. We investigated four sets of weights based on four different mathematical relationships between the ranks. Where w_r is the weight associated with rank r, the four sets of weights are defined as:

Equal:	$w_r = 1$
Linear:	$w_r = r$
Quadratic:	$w_r = r^2$
Logarithmic:	$w_r = \log(r) + 1$

Ranks are numbered 1 to 6. A constant of 1 is added to the logarithmic weights to avoid a weight of 0 for the lowest rank.

D. Results

1) ASR Accuracy & ASR Correct: Table IV shows the correlation of the *ASR Accuracy* and *ASR Accuracy* scores with human ratings for the repeating task.

	Success		Accuracy	
	UG	Oracle	UG	Oracle
Acc_{ASR}	0.61	0.81	0.63	0.77
Cor_{ASR}	0.76	0.87	0.85	0.90

TABLE IV
CORRELATION OF Acc_{ASR} AND Cor_{ASR} SCORES WITH HUMAN RATINGS FOR THE SUCCESS AND ACCURACY RATING SCALES USING UNIGRAM (UG) AND ORACLE FSG RECOGNITION STRATEGIES.

When based on recognition performed using a unigram language model, the correlations obtained for *ASR Accuracy* are comparable with those obtained for *Rate of Speech* (Table III). However, the correlations between *ASR Correct* scores and the human ratings are the highest correlations between a machine score and human ratings for *Success* and *Accuracy* found in this study. Furthermore, the oracle FSG results in Table IV show that higher recognition accuracy may lead to even better correlations between the machine scores and the human ratings.

2) Weighted Correct: Table V shows the correlations between human ratings and *Weighted Correct* scores based on the four different weight sets described in Section VI-C. While better than those obtained for *ASR Accuracy*, the correlations are lower than those found for *ASR Correct*, which regards all words as equally important.

The weight sets with equal weights and logarithmically related weights resulted in the highest correlations. When using equal weights, the *Weighted Correct* score is similar to the *ASR Correct* score except for the grouping of some words into phrases.

As with *ASR Accuracy* and *ASR Correct*, the correlations are higher when based on recognition using an oracle FSG grammar, indicating that better recognition may improve results.

[3]Personal communication with Prof. C. van der Walt, Department of Curriculum Studies, Faculty of Education, Stellenbosch University, who designed the prompts for the automated test.

	Success		Accuracy	
Weight Set	UG	Oracle	UG	Oracle
Equal	**0.71**	0.86	**0.79**	0.90
Linear	0.62	0.80	0.73	0.84
Quadratic	0.47	0.68	0.59	0.71
Logarithmic	**0.70**	0.85	**0.79**	0.89

TABLE V
CORRELATION OF WEIGHTED CORRECT SCORES WITH HUMAN RATINGS FOR THE SUCCESS AND ACCURACY RATING SCALES, UNIGRAM (UG) AND ORACLE FSG RECOGNITION STRATEGIES, AND DIFFERENT WEIGHT SETS.

3) Optimised Weights: In an attempt to find an approximately optimal weight set, correlations were calculated for 4000 randomly-generated weights consisting of uniformly distributed integers between 0 and 100. Groups of six random values were generated at a time and sorted numerically to form weight sets.

Based on these 4000 random weight sets, correlations with human ratings for *Success* ranged between 0.29 and 0.73, and correlations with the *Accuracy* scale ranged between 0.40 and 0.81. Hence no better alternative to the straightforward application of *ASR Correct* was found.

VII. COMBINATION OF SCORES

In this section we describe the effects of combining different machine scores to predict human ratings, by using *multiple linear regression* (MLR). The regression models were trained and implemented using WEKA, a data mining software package developed at the The University of Waikato[4] [9].

MLR was used for five different configurations of target and predictor variables. Table VI shows the targets and predictor categories for each of these configurations. In each instance, human ratings are used as targets and machine scores as predictors. The results reported here were obtained with a set of 8 predictor variables for the reading task and 14 for the repeating task [10]. These predictors also included the posterior log-likelihood scores that have been evaluated in previous investigations, but that have not explicitly been described here [2].

Target		**Predictors**
Reading Task Human Ratings	*Hesitation*	Reading Task Machine Scores
	Pronunciation	
	Intonation	
Repeating Task Human Ratings	*Success*	Repeating Task Machine Scores
	Accuracy	

TABLE VI
TARGET AND PREDICTOR VARIABLES FOR MLR.

For each configuration, MLR models were trained for every possible combination of predictor variables. This allowed us to identify which combinations of predictors lead to the best performance, as well as combinations which have comparable success but require fewer predictors.

Due to the relatively small size of our corpus, leave-one-out cross validation was used to evaluate each combination of predictors. For N speakers, leave-one-out cross validation employs N different regression models. Each model is trained on $N-1$ speakers and used to predict the target associated with the N_{th} speaker. A total of N iterations are considered, leaving each speaker out in turn. This leads to a set of N predicted target values, each estimated using a separately trained model. In the scenario presented here, $N = 90$, corresponding to the 90 students in the test set.

The ability of each predictor combination to accurately estimate the target variable was evaluated by calculating the correlation between the actual target values and the predicted values.

Table VII presents the correlation coefficients obtained with the best MLR combinations for each target criterion. In each case, the highest correlation with the target variables achieved using a single predictor variable is presented as a baseline. The relative improvement achieved by combining this predictor with others is then indicated as a percentage. In [10] we report on the complete set of experiments that include all possible combinations of predictor variables and identify the most successful combination for each target and predictor configuration. In this study, we restrict the results to the most effective single predictor, followed by the combination of predictors that yielded the biggest improvement over the baseline.

Target criterion	Baseline		MLR		Improvement
	Predictor	Corr	Predictors	Corr	
Hesitation	PTR	0.64	PTR & ROS	0.68	8%
Pronunciation	ART	0.50	ART & SDS	0.53	13%
Intonation	ROS	0.49	ROS	0.49	0%
Success	Cor_{ASR}	0.76	Cor_{ASR} & ROS	0.82	9%
Accuracy	Cor_{ASR}	0.85	Cor_{ASR} & ROS	0.87	4%

TABLE VII
RESULTS FOR MLR PREDICTIONS OF HUMAN RATINGS FOR READING AND REPEATING TASK MACHINE SCORES.

From Table II, and repeated in Table VII, we see that the best single predictor for the human ratings of *Hesitation* was the *Phonation/Time Ratio* (PTR). Adding *ROS* to the MLR model improved the correlation by 8% from 0.64 to 0.68. No combination of three scores resulted in a stronger correlation than the grouping of *PTR* and *ROS*.

For the human ratings of *Pronunciation*, Tables II and VII show that the best performing single predictor was *Articulation Rate* (ART). Combining this with the *Segment Duration Score* (SDS) improved the correlation between predicted and actual human ratings by 13%, from 0.50 to 0.53. Once again, this was also the best performance of any combination of predictors for this target, and no further improvement was achieved by adding further predictors. The contribution made by *SDS* is surprising, considering that the correlation between SDS and *Pronunciation* ratings is only 0.18, as shown in Table II.

For the human ratings of *Intonation*, no combination of predictors was able to achieve better performance than the best single candidate, which was ROS.

For the repeating task, Table VII shows that, for both *Success* and *Accuracy*, the best single predictor is *ASR Correct*. In both cases the combination of this predictor with ROS by MLR leads to a score which is better correlated with the human rating. In particular, combining *ASR Correct* with (*ROS*) leads to a 9% increase in correlation for *Success* and a 4% increase in correlation for *Accuracy*. The addition of further scores (including posterior log-likelihood scores) to the combination of *ASR Correct* and *ROS* resulted in very small additional improvements in the correlation.

VIII. DISCUSSION

In general, the segmentation scores reported on in this paper are better correlated with human ratings than the posterior log-likelihood scores that were evaluated in previous studies [1], [2].

ROS scores are arguably the most simple to calculate, and correlate well with all human rating scales. *Articulation Rate*, which is closely

[4]Freely available online from www.cs.waikato.ac.nz/∼ml/weka/.

related to *ROS* did not perform better, except for a slight improvement in correlation with *Pronunciation* ratings. The *Phonation/Time Ratio* scores show promise as a predictor of human ratings for *Hesitation* in particular. The *Segment Duration Score* scores have no usable correlation with human ratings for the reading task, but are relatively well correlated with ratings for the repeating task.

In comparison with the segmentation based scores, each of the three accuracy-based scores presented correlates well with human ratings. The fact that *ASR Correct* performed better than *ASR Accuracy* leads us to suspect that human raters do not penalise students for word insertions in the way *ASR Accuracy* does. Instead, they appear to focus primarily on the number of correct words.

Even the highest correlation achieved by the *Weighted Correct* score (0.81, using and optimal weight set), is not as high as that attained by *ASR Correct*, 0.85. Of the mathematically calculated weight sets, those with equal weights and logarithmically related weights resulted in the highest correlations. This shows that the best results are achieved when *Weighted Correct* is most similar to *ASR Correct*, and leads us to believe that manually ranking words and phrases in a target utterance according to their semantic importance is not a promising method of improving automatic accuracy scoring.

The results in Section VII indicate that combining machine scores for the estimation of human ratings results in better predictions than the use of individual scores. In each instance, the greatest improvement was achieved by the first additional score.

Results for the combination experiments also showed that predictions of the reading task ratings relied mostly on segmentation based scores, while predictions of the repeating task ratings relied on scores derived from the repeat accuracy in conjunction with segmentation based scores. For both the repeating task rating scales, the best single predictor was *ASR Correct*, while the addition of *ROS* resulted in the greatest additional improvement in correlation. The high correlations of 0.82 and 0.87 obtained for *Success* and *Accuracy* respectively seem to confirm the feasibility of automated rating of a repeating task using the algorithms described in this study.

Linear regression is one of many methods that can be used to combine machine scores in order to predict proficiency ratings. For example, Franco et al. found non-linear approaches such as the use of artificial neural networks, distribution estimation and regression trees to be slightly more effective than linear regression [11]. However, we mitigate the possible negative effect of non-linearities on our correlation values by using Spearman's rank correlation rather than Pearson's correlation coefficient.

IX. Conclusions

The results of this study show that, for the reading task, the segmentation-based *Rate of Speech* (*ROS*) score correlates best with human ratings of oral proficiency, despite its apparent simplicity. For the repeating task, the *ASR Correct* score, which is based on the automatically-estimated repeat accuracy, delivers the highest correlation with the corresponding human ratings, and also the highest correlations found overall. Furthermore, experiments using an oracle language model indicate that this correlation can be further improved by increasing the accuracy of the automatic speech recogniser. In determining the accuracy-based score, our experiments show that all words should be viewed as equally important, and that attempts at weighting words in accordance with their perceived semantic significance are counterproductive. Finally, we show that the correlations between the human ratings and the machine scores can be further improved by making use of a combination of machine scores through the application of multiple linear regression. The best correlations between human and machine scores, especially for the repeating exercise, indicate that automatic oral proficiency assessment for our highly-proficient group of L2 speakers is technically feasible.

X. 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.

References

[1] F. De Wet, P. F. De V. Müller, C. Van der Walt, and T. R. Niesler, "Experiments in automatic assessment of oral proficiency and listening comprehension for bilingual South African speakers of English," in *Proceedings of the Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Cape Town, South Africa, 2008, pp. 67–72.

[2] 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 *Proceedings of SLaTE*, Warwickshire, UK, 2009, CD-ROM.

[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] 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.

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

[6] 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, pp. 989–999, 2000.

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

[8] J. Richards and R. Schmidt, *Longman Dictionary of Language Teaching and Applied Linguistics*. London, UK: Longman, 2002.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, 2009, available online at www.sigkdd.org, last accessed 2009/11/25.

[10] P. F. deV. Müller, "Automatic oral proficiency assessment of second language speakers of South African English," Stellenbosch University, Stellenbosch, South Africa, Master's Thesis, 2010.

[11] 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, Scotland, 2000, pp. 123–128.