Experiments in automatic assessment of oral proficiency and listening comprehension for bilingual South African speakers of English

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

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

Abstract

We describe ongoing research into the automatic assessment of listening comprehension and oral language proficiency of South African L2 English speakers. Proficiency indicators are extracted from the speech signals by means of an automatic speech recognition system, and compared with assessments of the same speech by human experts. By means of carefully designed assessment scales, we are able to achieve high intra-rater correlations for the human scores. We show that, in accordance with the findings of other authors, rate of speech (ROS) is the most successful among the automatically derived measures that were evaluated. We also determine the effect of including context dependency in the speech recogniser's acoustic models, and investigate the effect which reciprocal transformations have on the correlations with human scores. Our results provide no evidence to support the hypothesis that context independent acoustic models yield better proficiency indicators when dealing with non-native speech. We also find that the use of the reciprocal of ROS does not lead to consistently better correlations.

1. Introduction

Assessment of a student's entrance level language skills for the purpose of placement into appropriate language programmes is often restricted to reading and writing proficiency tests. Listening and speaking skills are frequently not assessed because they either require specialised equipment or labour intensive procedures. In addition, the assessment of oral skills is generally highly subjective, and efforts that enhance inter-rater reliability further increase the labour intensiveness of the assessment process. The assessment of reading and writing comprehension skills, on the other hand, can be automated by means of computerised multiple choice tests, which reduce the time and manpower requirements for their administration. However, research has shown that good results in written tests are not necessarily good predictors of corresponding results in an oral test [1].

This study describes progress in an ongoing effort to develop an automated system to assess the listening comprehension and oral language proficiency of large numbers of students. The system will operate within the specific context of the Education Faculty at Stellenbosch University, where new students are required to obtain a language endorsement on their teaching qualification. For English, this means that students must enrol for a language module appropriate to their level of proficiency, and that their progress must be monitored regularly thereafter. With a current ratio of between 100 and 200 students per university staff member, this is only feasible when placing a major emphasis on computerised multiple-choice reading and writing tests. However, students regard oral proficiency as an important component of their teaching abilities. Consequently, they object to an exclusive focus on writing and reading skills, and regard the infrequency with which their oral skills are assessed with much suspicion. A technological solution may not only lighten the heavy workload of staff, but also provide a more transparent and objective metric with greater acceptance among students.

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

2. Computerised test development

The goal of the computerised test was to assess listening and speaking skills limited to the specific context of secondary school education. The test was therefore designed to evaluate language behaviour that is specifically relevant to this domain. There was no attempt to mimic natural human dialogue except in the sense that the test content relates specifically to teaching and learning in a school environment.

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

2.1. Test design

The test was designed to include instructions and tasks that require comprehension of spoken English and elicit spoken responses from students. In this paper we will focus on two of the seven tasks that comprise the test, namely the *reading* and the *repeating* tasks. For a detailed description of the complete test, the reader is referred to [6].

- **Reading task:** Students are provided with a list of 12 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 practise."*

The first task is a familiar one to students since reading aloud is a task they had to complete successfully for their fi-

nal school examinations. Moreover, the students could rely on the printed test sheet, which helped nervous candidates to relax.

The construction of the repeating task is based on the hypothesis that phonological working memory capacity influences oral production in first language users [7, 8] and even more so in second language learners [9, 10]. In terms of this hypothesis, second language learners will struggle to produce the target language in face-to-face communication because of time pressure in conjunction with limited access to vocabulary and the L2 sound system.

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

2.2. Test implementation

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

2.3. Test administration

A number of students volunteered to test the SDS in a pilot experiment. 120 students subsequently 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 SDS were made from a telephone located in a private office reserved for this purpose. Oral instructions were given to the students before the test. In addition to the instructions given by the SDS, a printed copy of the test instructions was provided. No staff were present while the students were taking the test.

3. Human assessments

Teachers of English as a second or foreign language were asked to rate speech samples from the read and repeat tasks in the test. The raters were not personally acquainted with the students they rated.

A subset of 90 students was selected from the group of 120 who took the test. Students were chosen to represent male and female as well as Afrikaans and English mother tongue speakers in accordance with the composition of the student population at the Faculty of Education. Students were chosen to ensure a balanced test population with regard to mother tongue and gender. Given the large number of students, it was not feasible to have each utterance of every student rated. Three examples of each student's read and repeat responses were randomly chosen to be judged by the raters.

Six raters each assessed 45 students and each student was assessed by three human raters. In order to measure intra-rater consistency, five students were presented twice to each rater. Each rater therefore performed 50 ratings: 45 unique and 5 repeats.

Before judging the students, the raters attended a training session on the use of the rating scales. Example utterances and their respective ratings were also presented.

For the reading task, each sentence was assessed on three separate scales in terms of degree of hesitation, pronunciation (including accent) and intonation, as shown in Figure 1. The scales were conceptualised as a continuum on which certain points are described, with the possibility to mark points between two descriptions.

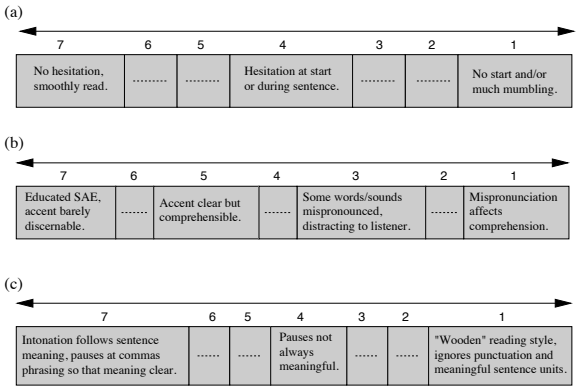

Figure 1: Scales used to assess (a) degree of hesitation (b) pronunciation and (c) intonation in the reading test.

The numbers above the scales are meant to guide the eventual mark allocation, providing numerical information that can be used to grade students. Scores below three on the scale would indicate students who need additional language support. However, numerical scores were not included on the scales supplied to the raters in order to avoid pre-conceptions about student grades.

For the repeating task, a different set of scales was designed in order to measure the success with which a repetition was formulated and the accuracy of the repetition, as shown in Figure 2. In this case the scales contained precise descriptions for all the categories.

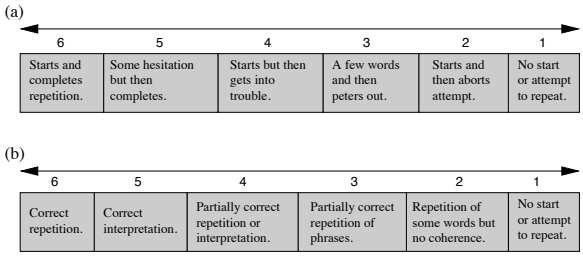

Figure 2: Scales used to assess (a) degree of success and (b) accuracy in the repeating test.

During the pilot test it had become clear that students did not necessarily repeat each sentence accurately, but were nevertheless able to comprehend it and could repeat a rendition that reflected the meaning of the original sentence. Since the

test was also 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.

3.1. Results: Human assessments

Table 1 shows the intra-rater correlations[1] that were obtained using these scales. These values are much higher than those obtained in our previous study based on global assessments [6]. This seems to indicate that the more detailed assessment guidelines introduced in this study assist the human raters in allocating marks more consistently.

Rater	Intra-rater correlation
1	0.83
2	0.94
3	0.81
4	0.96
5	0.67
6	0.91

Table 1: Intra-rater correlations for human raters.

Figures 3(a) and 3(b) illustrate the inter-rater agreement for the read and repeat tasks, respectively.

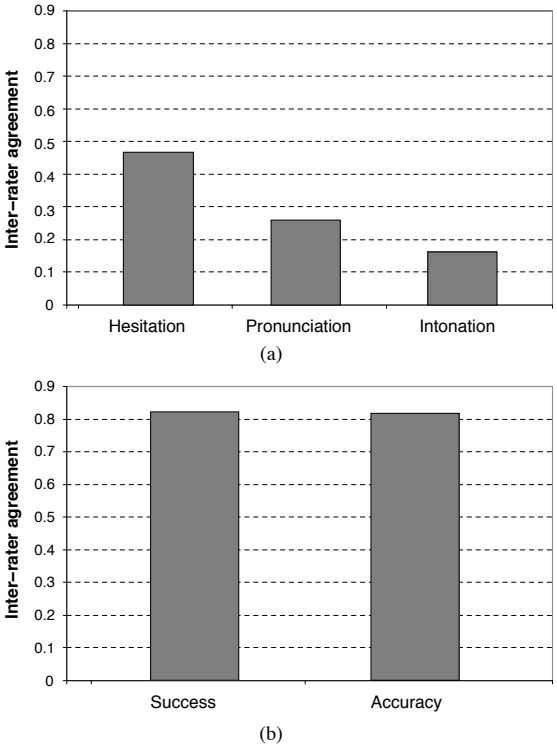

Figure 3: Inter-rater agreement for the read (a) and repeat (b) tasks.

[1]The correlations are two-way random, intra-class correlation coefficients and were calculated using Statistica [11].

The three bars in Figure 3(a) indicate the values for the three scales shown in Figure 1, and the two bars in Figure 3(b) the values for the scales in Figure 2. By comparing Figures 3(a) and 3(b) we see that the inter-rater agreement was higher for the repeat than for the read task. The raters clearly disagree in their assessment strategies for the pronunciation and intonation aspects of the reading task.

The values shown in Figure 3(a) for the read speech are lower than those reported in [3] and [4], but are similar to those reported in [5]. The fact that our test population is fairly homogeneous in terms of proficiency could explain this observation. The raters appear to be less consistent in their assessments when there is little variation in proficiency. In studies where higher inter-rater agreement was measured, the speaker populations were more diverse in terms of L2 proficiency. Furthermore, in our experiments the human judges also rated fewer utterances per speaker, i.e. two or three as opposed to 10 in [3] and 30 in [4].

The average score (percentages calculated across all raters) for the read and repeat tasks are shown in Figures 4(a) and 4(b). The standard deviation around the mean values is indicated by the vertical lines in the figures. Figure 4 shows that students performed better in the reading task than in the repeating task and that, on average, they were given good marks. In previous studies we observed that only the top part of the assessment scales were used by the judges, especially for the reading task [6]. Despite our efforts to 'broaden' the assessment scales in this experiment, the lower extremes of the scales were again rarely chosen.

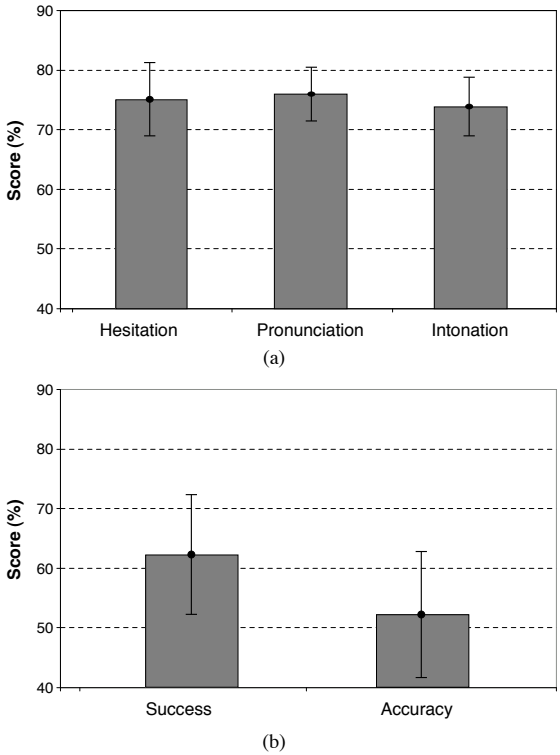

Figure 4: Average scores for the read (a) and repeat (b) tasks. The standard deviation around the mean is indicated by the vertical line in each bar.

4. ASR-based assessment

Numerous studies on the role of ASR in language learning applications have been published in the last decade e.g. [2, 3, 4, 12]. However, the investigations reported on in the literature differ in terms of several aspects of experimental design. As a result, it is difficult to make direct comparisons between studies. Nevertheless, a common aim of most studies in this field is the identification of parameters that can be automatically derived from speech data and that correlate well with human judgements of oral proficiency.

4.1. ASR system

ASR is a relatively new research field in South Africa and the resources that are required to develop applications are limited. During the *African Speech Technology* project, telephone speech databases were compiled for South African English, isiZulu, isiXhosa, Sesotho and Afrikaans [13]. Prototype speech recognisers were subsequently developed for each language, and this study makes use of the standard South African English ASR system.

A training set consisting of approximately six hours of phonetically-annotated telephone speech data was parametrised as Mel-frequency cepstral coefficients (MFCCs) and their first and second differentials. Cepstral mean normalisation (CMN) was applied on a per-utterance basis.

A set of 52 speaker-independent monophone hidden Markov models (HMMs) with three states per model and 64 mixtures per state was trained on this data by embedded Baum-Welsh re-estimation using the HTK tools [14]. A set of speaker-independent cross-word triphone HMMs was then obtained using decision-tree state clustering, resulting in a total of 4797 clustered states. Each triphone model employed eight Gaussian mixtures per state and diagonal covariance matrices. The phone recognition accuracies of the monophone and the triphone models on a separate test set and using a bigram language model are shown in Table 2.

Model	Phone Recognition Accuracy (%)
Monophone	58.4
Triphone	73.0

Table 2: Phone recognition accuracies measured for the monophone and triphone acoustic models.

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

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

For the repeat task, a unigram language model with equal probabilities for all words was derived from the manual transcriptions of the development test set. A separate language model was constructed for each sentence of the repeat task. The recogniser's word insertion penalty and language model factor were chosen to maximise the correlation between recognition accuracy as well as the ROS values derived from the manual and automatic transcriptions of the development test set.

4.2. Automatically derived proficiency indicators

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

4.2.1. Rate of speech

Previous studies have found that, for read speech, rate of speech (ROS) is one of the best indicators of fluency [3, 12]. In our experiments ROS was calculated according to Equation (1), as proposed in [15].

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

The quantity N_p denotes the number of speech phones in the utterance, while T_{sp} is the total duration of speech in the utterance, including pauses.

The correlation between the ROS values derived from the manual and automatic transcriptions of the test data were 0.98 and 0.94 for the read and repeat data, respectively. These values indicate that the automatic system's ability to segment the speech into phones compares very well with its human counterpart.

4.2.2. Goodness of pronunciation

As an example of the general class of posterior HMM likelihood scores [4, 12], we used the "goodness of pronunciation" (GOP) proposed in [2]. The GOP score of phone q_i is defined as the frame-normalised logarithm of the posterior probability $P(q_i|O)$, where O refers to the acoustic segment uttered by the speaker.

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

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

Some authors claim that less detailed native models, like monophone HMMs, perform better for non-native speakers than detailed native models like triphone HMMs [16, 17]. Others report very small differences between the results obtained with monophone and triphone models for non-native speakers [18]. In this study we investigate the influence of model complexity on automatically derived proficiency indicators by deriving GOP scores from monophone (GOPmono) as well as cross-word triphone (GOPxword) HMMs.

4.2.3. Transcription accuracy

Because highly restrictive finite-state grammars were used for the reading task, the recognition accuracy obtained for the read responses was in all cases very high and therefore not used as a proficiency indicator. Repeat accuracy, on the other hand, was considered as a proficiency indicator, as is also proposed in [12].

This accuracy was determined by comparing the ASR output for the repeated utterances to the orthographic transcriptions of the sentences students were prompted to repeat during the test. Accuracy was subsequently calculated according to Equation 3, as proposed in [14]:

$$Accuracy = \frac{H - I}{N} \times 100\% \quad (3)$$

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

4.2.4. Nonlinear transformations

Research has shown that it is possible to improve the correlation between automatically derived indicators and human ratings by using a non-linear combination of several machine scores. However, it was found that these improvements are often due mainly to the non-linear transformation of a single indicator that already correlates well with human ratings [19]. One such non-linear transformation is the reciprocal, and experiments have suggested that using $\frac{1}{\text{ROS}}$ instead of ROS leads to a slightly higher correlation with human ratings [12]. We will establish whether this is true for our experimental conditions in the following section.

4.3. Results: ASR-based assessment

The ASR system judged the same material previously evaluated by the human raters. The average ROS, accuracy and GOP values that were measured for the read and repeat tasks are shown in Table 3.

	ROS	Accuracy	GOP
Read	11.94	-	3.88
Repeat	9.81	50.16	4.05

Table 3: Average ROS, accuracy and GOP scores for the test data.

The observation that average ROS is higher for the reading task than for the repeat task is in agreement with what one would intuitively expect, given the level of difficulty of the tasks. The correlations measured between the ROS, accuracy and GOP scores are listed in Table 4.

Task	Score pair	Correlation
Read	ROS & GOP	0.01
Repeat	ROS & Accuracy	0.75
Repeat	ROS & GOP	-0.44
Repeat	Accuracy & GOP	-0.40

Table 4: Correlation between ROS, accuracy and GOP scores for the read and repeat tasks.

Table 4 shows that repeat accuracy correlates strongly with ROS and to a lesser extent with the GOP scores. In contrast, there is no correlation between ROS and the GOP scores for the read data and only a weak correlation for the repeat data. This observation seems to indicate that ROS is not related to the acoustic properties of the data. ROS and GOP scores could therefore be used to evaluate different aspects of speech.

5. Correlation between human and ASR-based assessment

Table 5 gives the correlation[2] between the scores given by the human raters and the automatically derived proficiency indicators for the reading task. The highest correlation in Table 5 is observed between degree of hesitation and ROS. To the extent that degree of hesitation is an indicator of fluency, this result is consistent with what has been reported in the literature [15].

Indicator	Hesitation	Pronunciation	Intonation
ROS	0.53	0.46	0.49
1/ROS	0.55	0.45	0.51
GOPmono	0.03	0.18	0.01
GOPmono/ROS	0.37	0.19	0.39
GOPxword	0.11	0.13	0.05
GOPxword/ROS	0.43	0.19	0.36

Table 5: Correlation between human and automatic scores for the reading task.

The GOP scores show almost no correlation with the human judgements of the read material. This result is similar to the observation made in [3], where the weakest correlation between human and automatic scores was measured for likelihood ratios. This trend seems to indicate that posterior scores derived at the utterance level do not provide meaningful information on pronunciation. Discriminating between GOP scores for vowels and consonants or deriving phone-specific GOP scores for a number of "problematic" phones may improve the GOP scores' correlation with the human data.

Table 6 shows the correlation between the scores the human raters assigned for the repeat task and those derived automatically using the ASR system.

Indicator	Success	Accuracy
Accuracy	0.68	0.69
ROS	0.71	0.68
1/ROS	0.71	0.66
GOPmono	0.31	0.32
GOPmono/ROS	0.60	0.57
GOPxword	0.40	0.40
GOPxword/ROS	0.67	0.63

Table 6: Correlation between human and automatic scores for the repeat task.

ROS as well as accuracy correlate well with the human scores. However, it should be kept in mind that these variables are also strongly correlated with each other for the repeat task (Table 4). The GOP scores are only poorly correlated to the human ratings of the repeat data, but the correlations are consistently higher than those in Table 5.

The results in Tables 5 and 6 show that there is no consistent improvement in correlation when using the reciprocal of ROS as a proficiency indicator. This is in contrast to the results reported in [12], where small improvements were observed. The two tables also show that the assertion made in [16, 17] that monophone acoustic models are more appropriate than triphone models when dealing with non-native speech is not borne out by our experiments. The small improvement observed for using

[2] Spearman rank correlation coeffi cients were derived (using Statistica [11]) because the data in question is ordinal.

context sensitive models (Table 6) is consistent with the results reported in [18].

6. Discussion and conclusion

We have described progress made in our effort to develop an automated system to assess the listening comprehension and oral language proficiency of South African L2 English speakers. Despite our revised and more specific rating scales, we found that the scores allocated by the human raters for the reading task still fall within a narrow range of high marks. We believe that this narrow range led to the associated relatively poor inter-rater agreement, and possibly also the low correlations with the automatically derived indicators. To improve this, we will attempt to increase the difficulty of the read sentences in future implementations of the test, in order to achieve a greater spread of human scores. For the repeat task, the spread of the scores was considerably greater as were their correlation between the automatically derived indicators, especially ROS.

Using the reciprocal of ROS instead of ROS as an indicator showed no consistent improvement in the correlation with the human scores, in contrast with other published research. This probably indicates that the relationship between the scores' distributions in our study is different to the relationships observed in other studies. Other non-linear transformations, such as neural networks and distribution estimation, have also been reported to improve the correlation with human ratings to a greater degree [20]. The effect of these alternative and more flexible transformations on our data will be investigated in future research.

When comparing the effectiveness of context independent (monophone) and context dependent (triphone) acoustic models, we found that the triphones performed slightly better in the repeat task, and there was no consistent difference for the read task. Thus, the finding that context independent models show superior performance for the automatic assessment of non-native speech does not hold for our experimental situation.

7. Acknowledgements

This research was supported by the Fund for Innovation and Research into Teaching and Learning at Stellenbosch University, an NRF Focus Area Grant for research on *English Language Teaching in Multilingual Settings* and the "Development of Resources for Intelligent Computer-Assisted Language Learning" project sponsored by the NHN.

8. References

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

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

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

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

[5] T. Cincarek, "Pronunciation scoring for non-native speech," Master's thesis, Institut für Informatik, Friedrich-Alexander Universität, Erlangen-Nürnberg, Germany, 2004.

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

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

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

[9] 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, pp. 234250, 1996.

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

[11] StatSoft Inc., Ed., *STATISTICA 8.0*, www.statsoft.com, 2008.

[12] C. Hacker, T. Cincarek, R. Gruhn, S. Steidl, E. Nöth, and H. Niemann, "Pronunciation feature extraction," in *Proceedings of 27th DAGM Symposium*, August 2005, pp. 141–148.

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

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

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

[16] T. Schultz and K. Kirchhoff, *Multilingual speech processing*, chapter Other challenges: non-native speech, dialects, accents and local interfaces, Academic Press, 2006.

[17] X. He and Y. Zhao, "Model complexity optimisation for non-native English speakers," in *Proceedings of Eurospeech*, Aalborg, Denmark, 2001, pp. 1461–1463.

[18] O. Ronen, L. Neumeyer, and H. Franco, "Automatic detection of mispronunciation for language instruction," in *Proceedings of Eurospeech*, Rhodes, Greece, 1997.

[19] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, "Combination of machine scores for automatic grading of pronunciation quality," *Speech Communication*, vol. 30, pp. 121–130, 2000.

[20] 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, 2000, University of Abertay, pp. 123–128.