Elicited Imitation for Prediction of OPI Test Scores

Kevin Cook
Brigham Young University
Department of Computer Science
kevincook13@gmail.com

Jeremiah McGhee, Deryle Lonsdale
Brigham Young University
Department of Linguistics
{jlmcghee,lonz}@byu.edu

Abstract

Automated testing of spoken language is the subject of much current research. Elicited Imitation (EI), or sentence repetition, is well suited for automated scoring, but does not directly test a broad range of speech communication skills. An Oral Proficiency Interview (OPI) tests a broad range of skills, but is not as well suited for automated scoring. Some have suggested that EI can be used as a predictor of more general speech communication abilities. We examine EI for this purpose. A fully automated EI test is used to predict OPI scores. Experiments show strong correlation between predicted and actual OPI scores. Effectiveness of OPI score prediction depends upon at least two important design decisions. One of these decisions is to base prediction primarily on acoustic measures, rather than on transcription. The other of these decisions is the choice of sentences, or EI test items, to be repeated. It is shown that both of these design decisions can greatly impact performance. It is also shown that the effectiveness of individual test items can be predicted.

1 Introduction

1.1 Background

Learning to speak a second language is an important objective for many people. Assessing progress in oral proficiency is often expensive and time-consuming. The development of automated systems promises to significantly lower costs and increase accessibility.

Elicited imitation (EI) has been used for nearly half a century to measure abnormal language development (Fujiki and Brinton, 1987) and the performance of second language learners (Chaudron et al., 2005; Vinther, 2002). As a method for assessing oral proficiency it consists of a person listening to a test item, typically a full sentence, and then doing their best to repeat it back correctly. This method is also referred to as sentence repetition, or more simply as repeats. One motivation for using EI, as opposed to some other form of test, is that it is relatively inexpensive to administer. An EI test can be effectively scored by non-experts in a relatively short amount of time. It is also well suited for automated scoring (Graham et al., 2008), since correct responses are predictable.

1.2 Motivation

The language skills directly measured by an EI test are those involved in repeating back what one has just heard. In order to directly measure a broader set of language skills, other tests must be used. One of these is the Oral Proficiency Interview (OPI).

The OPI is face-to-face interview conducted to assess language proficiency. The interview tests different types of relevant skills and lasts for about 30 minutes. Additionally, a validated OPI requires a second review of a recording created during the initial interview with arbitration if necessary. This process is expensive ($150 U.S.) and time-consuming with a turn-around of several weeks before finalized results are received.

A fully automated OPI test does not seem to be practical. This is especially the case when the in-

Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications, pages 30–37,
Portland, Oregon, 24 June 2011.

terpersonal aspects of a face-to-face interview are considered. There have been several efforts to automatically score the type of speech which might be spoken by an OPI test-taker, spontaneous non-native speech (Zechner and Xi, 2008). It has been shown that current automatic speech recognition (ASR) systems, used to transcribe such speech, have error rates which make it challenging to use transcripts for testing purposes.

The argument has been made that although EI does not directly measure communicative skills, such as the ability to converse with another person, it can be used to infer such skills (Henning, 1983). Part of the theory behind EI is that people typically are not able to memorize the sounds of an utterance the length of a full sentence. Rather, people build a mental model of the meaning of an utterance, and are then able to remember the model. People who cannot understand the utterance are not able to build a mental model, and are therefore unable to remember or repeat the utterance. If it is true that EI can be used to infer more general speech communication abilities, even if only to a limited extent, then EI may be useful for predicting test scores which are designed to directly measure that ability.

Bernstein et al. (2000) describe a system which elicits short predictable responses, such as readings, repeats (EI), opposites, and short answers, for automated testing. A similar system is discussed later in Bernstein et al. (2010). It is evident that EI is used in these systems, as part of a greater whole. The argument is made that although the skills directly tested are limited, the scores produced may be useful for inferring more general language abilities. It is shown that automated scores correlate well with scores from conventional tests, such as the OPI. One aspect which may not be as clear is the role that EI plays as compared to other methods used in the automated test.

We are interested in the use of a fully automated EI test as a means to predict more general ability in spoken language communication. Since the OPI test is specifically designed to measure such general ability we use it as a gold standard, in spite of the fact that we do not expect it to be a perfect measure. We are interested in learning the extent to which OPI scores can be predicted using an EI test. We are also interested in learning how to design an automated system such that prediction of OPI scores is most effective. We evaluate system performance based on how highly correlated OPI score predictions are with actual OPI scores.

Several design decisions must be made in the development of such a system. One, is which method to use for converting spoken responses to OPI score predictions. Another, is the choice of sentences, or EI test items, to be repeated. We address both of these issues.

There are at least two approaches to scoring spoken responses. One, is to score based on transcriptions, generated by a speech recognizer. Another, is to score based on acoustic measures alone, such as pronunciation and fluency (Cincarek et al., 2009). The primary difference between these two approaches is what is assumed about the textual content of a spoken response. Acoustic measures are based on the assumption that the textual content of each spoken response is known. Speech recognition is based on the assumption that the content is not known. We explore the effect of this assumption on OPI prediction.

The selection of effective EI test items has been the subject of some research. Tomita et al. (2009) outline principles for creating effective EI test items. Christensen et al. (2010) present a tool for test item creation. We explore the use of OPI scores as a means to evaluate the effectiveness of individual test items.

2 Related Work

The system described by Bernstein et al. (2010) uses EI as part of the automated test. Sentences range in length from two to twenty or more syllables. If fewer than 90% of natives can repeat the sentence verbatim, then the item is not used. An augmented ASR system is used which has been optimized for non-native speech. The ASR system is used to transcribe test-taker responses. Transcriptions are compared to the word string recited in the prompt. Word errors are counted and used to calculate a score. Fluency and pronunciation of spoken responses are also scored.

Graham et al. (2008) report on a system which uses EI for automated assessment. Results show that automated scores are strongly correlated with man-

ual EI scores. ASR grammars are specific to each test item. Our work is based on this system.

Müller et al. (2009) compare the effectiveness of reading and repeating (EI) tasks for automated testing. Automated scores are compared with manual scores for the same task. It is found that repeating tasks provide a better means of automatic assessment than reading tasks.

3 Experiments

In this section we describe experiments, including both an OPI test and an automated EI test. We detail the manner of automated scoring of the EI test, together with the method used to predict OPI scores.

3.1 Setup

We administer an ACTFL-OPI (see www.actfl.org) and an automated EI test to each of 85 English as a Foreign Language learners of varying proficiency levels. This group of speakers (test-takers) is randomly divided into a 70%/30% training/testing split, with 60 speakers forming the training set and the remaining 25 forming the test set. Training data consists of OPI scores and EI responses for each speaker in the training set. Test data consists of OPI scores and EI responses for each speaker in the test set.

An OPI is a face-to-face interview conducted by a skilled, certified human evaluator. (We do not expect that this interview results in an ideal evaluation of oral proficiency. We use the OPI because it is designed to directly test speech communication skills which are not directly tested by EI.) OPI proficiency levels range across a 10-tiered nominal scale from Novice Low to Superior. We convert these levels to an integer score from 1 to 10 ($NoviceLow = 1$, $Superior = 10$).

The EI test consists of 59 items, each an English sentence. An automated system plays a recording of each sentence and then records the speaker's attempt to repeat the sentence verbatim. A fixed amount of time is allotted for the speaker to repeat the sentence. After that fixed time, the next item is presented, until all items are presented and all responses recorded. The choice of which items to include in the test is somewhat arbitrary; we select those items which we believe might work well, given past experimentation with EI. We expect that improvement could be made in both the manner of administration of the test, and in the selection of test items.

Responses are scored using a Sphinx 4 (Walker et al., 2004) ASR system, version 1.0 beta 4, together with the supplied 30-6800HZ WSJ acoustic model. ASR performance is affected by various system parameters. For our experiments, we generally use default parameters found in configuration files for Sphinx demos. The ASR system has not been adapted for non-native speech.

3.2 Language Models

We vary the language model component of the ASR system in order to evaluate the merit of assuming that the content of spoken responses is known. Speech recognizers use both an acoustic model and a language model, to transcribe text. The acoustic model is used to estimate a probability corresponding to how well input speech sounds like output text. The language model is used to estimate a probability corresponding to how well output text looks like a target language, such as English. Output text is determined based on a joint probability, using both the acoustic and the language models. We vary the degree to which it is assumed that the content of spoken responses is known. This is done by varying the degree to which the language model is constrained to the text of the expected response.

When the language model is fully constrained, the assumption is made that the content of each spoken response is known. The language model assigns all probability to the text of the expected response. All other output text has zero probability. The acoustic model estimates a probability for this word sequence according to how well the test item is pronounced. If the joint probability of the word sequence is below a certain rejection threshold, then there is no output from the speech recognizer. Otherwise, the text of the test item is the output of the speech recognizer. With this fully constrained language model, the speech recognizer is essentially a binary indicator of pronunciation quality.

When the language model is fully unconstrained, there is no relationship between the language model and test items, except that test items belong to the English language. In this case, the speech recognizer functions normally, as a means to transcribe spoken responses. Output text is the best guess of the ASR

system as to what was said.

A partially constrained language model is one that is based on test items, but also allows variation in output text.

We perform experiments using the following five language models:

1. **WSJ20K** The 20K word Wall Street Journal language model, supplied with Sphinx.

2. **WSJ5K** The 5K word Wall Street Journal language model, supplied with Sphinx.

3. **EI Items** A custom language model created from the corpus of all test items.

4. **Item Selection** A custom language model constraining output to any one of the test items.

5. **Forced Alignment** A custom language model constraining output to only the current test item.

The first two language models, WSJ20K and WSJ5K, are supplied with Sphinx and have no special relationship to the test items. The training corpus used to build these models is drawn from issues of the Wall Street Journal. These models are fully unconstrained.

The third model, EI Items, is a conventional language model with the exception that the training corpus is very limited. The training corpus consists of all test items; no other text is included in the training corpus. The fourth model, Item Selection, is not a conventional language model. It assigns a set probability to each test item as a whole. That probability is equal to one divided by the total number of test items. Such a simple language model is sometimes referred to as a grammar (Walker et al., 2004; Graham et al., 2008). Both the EI Items and Item Selection models are partially constrained. The Item Selection model is much more highly constrained than the EI Items model.

The last model, Forced Alignment, is fully constrained. It assigns all probability to item text. These five language models are chosen for the purpose of evaluating the effectiveness of constraining the language model to the text of the expected response.

i	Item
I	Number of items
s	Speaker (test-taker)
S	Number of speakers
x_{is}	Score for item i, speaker s
$y_{\,s}$	Predicted OPI score for speaker s
o_s	Actual OPI score for speaker s
$MSE_{\,i}$	Mean squared error for item i

Figure 1: Notation used in this paper.

3.3 Scoring

Each response is scored using a two-step process. First, the spoken response is transcribed by the ASR system. Second, word error rate (WER) is calculated by comparing the transcription to the item text. WER is converted to an item score x_{is} for item i and speaker s in the range of 0 to 1 using the following formula:

$$x_{is} = \begin{cases} 1 - \frac{WER}{100} & \text{if } WER < 100\% \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

A list of notation used in this paper is shown in Figure 1.

3.4 Prediction

In order to avoid over-fitting, a simple linear model is trained (Witten and Frank, 2005) to predict an OPI score y_s, given items scores x_{is} together with model parameters a and b. The mean of item scores for speaker s is multiplied by parameter a. This product plus parameter b is the OPI score prediction: (I is the total number of items.)

$$y_s = \frac{1}{I} \sum_i x_{is} \cdot a + b \quad (2)$$

Correlation is calculated between predicted and actual OPI scores for all speakers in the test set.

4 Results

Correlation for each of the language models using all 59 test items is shown in Figure 2. Correlation for both of the unconstrained language models was relatively poor. Performance improved significantly as the language model was constrained to the expected response. These results suggest that it is effective

to assume that the content of spoken responses is known.

Fully constraining the language model to the text of the expected response results in an item score which is a binary indicator (because, in this case, WER is either 100% or 0%) of how well the spoken response sounds like the expected response. In this case, prediction is based on the output of the acoustic model of the speech recognizer, an acoustic measure. Prediction is not based on transcription, since a specific transcription is assumed prior to processing the spoken response. When the language model is fully unconstrained, an item score is an indicator of how well ASR transcription matches the text of the expected response. In this case, prediction is based on transcription, the speech recognizer's best guess of which words were spoken. Results indicate that correlation between predicted and actual OPI scores improves as prediction is based on acoustic measures, rather than on transcription.

Language Model	Constrained	Corr.
WSJ20K	Not	0.633
WSJ5K	Not	0.600
EI Items	Partial	0.737
Item Selection	Partial	0.805
Forced Alignment	Full	0.799

Figure 2: Correlation with OPI scores, for all 5 language models, using all 59 test items. Language models are unconstrained, partially constrained, or fully constrained to the text of the expected response.

4.1 Item MSE

The effectiveness of individual test items is explored by defining a measure of item quality. If each item score x_{is} were ideally linearly correlated with the actual OPI score o_s for speaker s then the equality shown below would hold: (o_s is an integer from 1 to 10. x_{is} is a real number from 0 to 1.)

$$IDEAL \implies o_s = x_{is} * 9 + 1 \quad (3)$$

We calculate the difference between this ideal and the actual OPI score:

$$(x_{is} * 9 + 1) - o_s \quad (4)$$

This difference can be seen as a measure of how useful the item is as a predictor OPI scores. For better items, this difference is closer to zero. The mean of the squares of these differences for a particular item, over all S speakers in the training set, is a measure of item quality MSE_i:

$$MSE_i = \frac{1}{S} \sum_s ((x_{is} * 9 + 1) - o_s)^2 \quad (5)$$

Because we expect improved results by assuming that the content of expected responses is known, we use the Forced Alignment language model to calculate an MSE score for each test item. A sample of items and their associated MSE are listed in Figure 3.

MSE	Item text
9.28	He should have walked away before the fight started.
10.48	We should have eaten breakfast by now.
...	
14.53	She dove into the pool gracefully, and with perfect form.
14.68	If her heart were to stop beating, we might not be able to help her.
...	
25.78	She ought to learn Spanish.
26.09	Sometimes they go to town.

Figure 3: Sample EI items with corresponding MSE scores.

Item MSE scores are used to define various subsets of test items, better items, worse items, and so on. Better items have lower MSE scores. These subsets are used to compute a series of correlations for each of the five language models. First, correlation is computed using only one test item. That item is the item with the lowest (best) MSE score. Then, correlation is computed again using only two test items, the two items with the lowest MSE scores. This process is repeated until correlation is computed using all test items. Results are shown in Figure 4. These results show even more convincingly that OPI prediction improves by assuming that the content of spoken responses is known.

4.2 OPI Prediction

Figure 4 also gives an idea of how effectively EI can be used to predict OPI scores. Correlation over 0.80 is achieved using the Forced Alignment language

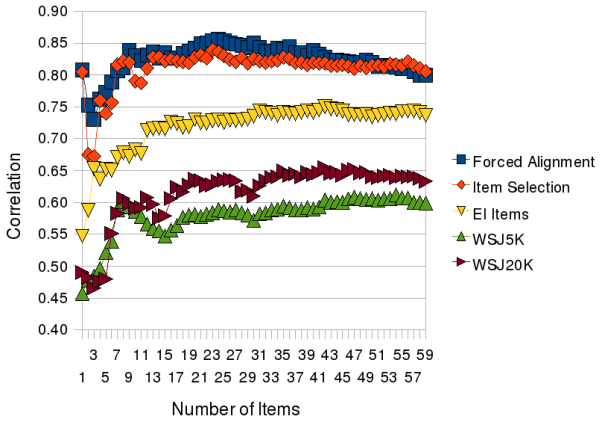

Figure 4: Correlation with OPI scores, for all 5 language models, using varying numbers of test items.

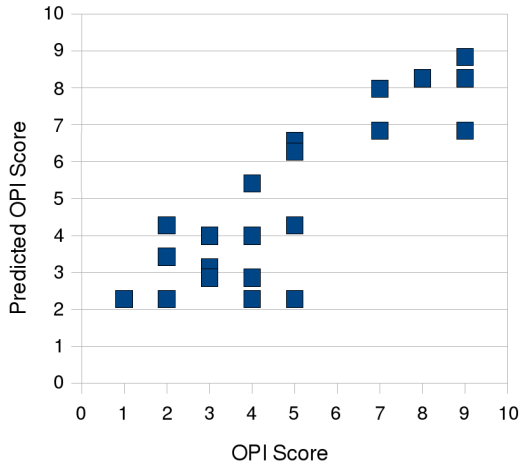

Figure 5: Plot of predicted OPI scores as a function of actual OPI scores, using the Forced Alignment language model and the best 24 test items.

model for all but 7 of the 59 subsets of test items. Correlation is over 0.84 for 11 of the subsets (best 20 - best 31). Correlation is above 0.85 for 3 subsets (best 23 - best 25). Predicted OPI scores correlate strongly with actual OPI scores.

Figure 5 shows a plot of predicted OPI scores as a function of actual OPI scores, using the Forced Alignment language model and only the best 24 test items. Correlation is 0.856. Interestingly, two of the outliers (OPI=5, predicted OPI=2.3) and (OPI=4, predicted OPI=2.3) were for speakers whose responses contained only silence, indicating those participants may have experienced technical difficulties or may have been uncooperative during their test session. The inferred model used to calculate OPI predictions for Figure 5 is shown below:

$$y_s = \frac{1}{I} \sum_i x_{is} * 6.8 + 2.3 \quad (6)$$

(Given this particular model, the lowest possible predicted OPI score is 2.3, and the highest possible predicted score is 9.1. The ability to predict OPI scores 1 and 10 is lost, but the objective is to improve overall correlation.)

4.3 Item Selection

To see more clearly the effect that the choice of test items has on OPI prediction, we compute a series of correlations similar to before, except that the order of test items is reversed: First, correlation is computed using only the test item with the highest (worst) MSE score. Then, correlation is computed again using only the two worst items, and so on. This series of correlations is computed for the Forced Alignment language model only. It is shown together with the original ordering for the Forced Alignment language model from Figure 4.

These two series are shown in Figure 6. The series with generally high correlation is computed using best items first. The series with generally low correlation is computed using worst items first. At the end of both series all items are used, and correlation is the same. As mentioned earlier, correlation using only the best 24 items is 0.856. By contrast, correlation using only the worst 24 items is 0.679. The choice of test items can have a significant impact on OPI score prediction.

Figure 6 also shows that the effectiveness of individual test items can be predicted. MSE scores were calculated using only training data. Correlations were calculated for test data.

4.4 Rejection Threshold

Since the Forced Alignment language model is found to be so effective, we experiment further to learn more about its behavior. Using this language model, item scores are either zero or one, depending upon whether ASR output text is the same as item text, or there is no output text. If joint probability, for a spoken response, is below a certain rejection threshold, no text is output. We perform experiments

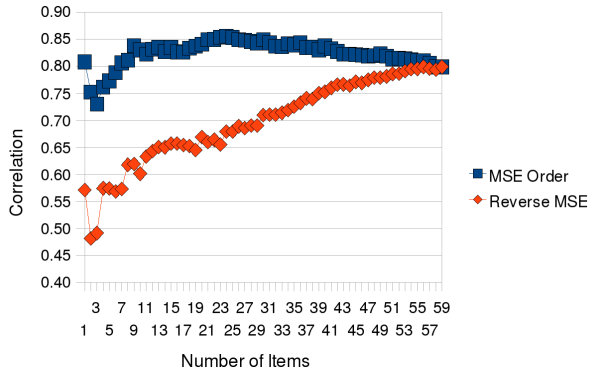

Figure 6: Correlation with OPI scores, showing the difference between best and worst items, using the Forced Alignment language model.

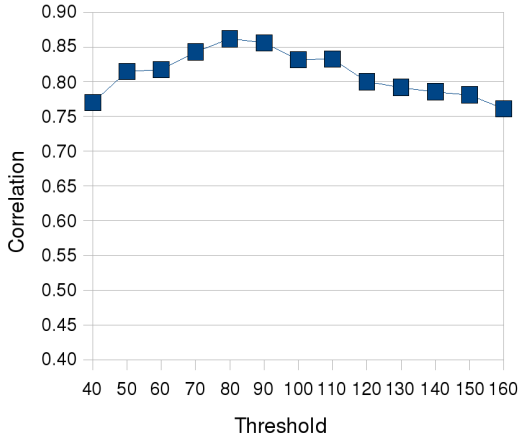

Figure 7: Correlation with OPI scores versus rejection threshold.

to see how sensitive OPI predictions are to the setting of this threshold.

Any ASR system parameter which affects probability estimates of word sequences can affect the rejection threshold. We make the arbitrary decision to vary the Sphinx $relativeBeamWidth$ parameter. For all previous experiments, the value of this parameter was fixed at $1E-90$. The $wordInsertionProbability$ parameter, which also affects the rejection threshold, was fixed at $1E-36$.

Correlation is computed for various values of the $relativeBeamWidth$ parameter. Results are shown in Figure 7. Good results are obtained over a wide range of rejection thresholds. Correlation peaks at $1E-80$. OPI prediction does not appear to be overly sensitive to the setting of this threshold.

5 Discussion

We conclude that a fully-automated EI test can be used to effectively predict more general language ability than those abilities which are directly tested by EI. Such an EI test is used to predict the OPI scores of 25 test-takers. Correlation between predicted and actual OPI scores is strong.

Effectiveness of OPI score prediction depends upon at least two important design decisions. One of these decisions is to base prediction primarily on acoustic measures, rather than on transcription. The other of these decisions is the choice of sentences, or EI test items, to be repeated. It is shown that both of these design decisions can greatly impact performance. It is also shown that the effectiveness of individual test items can be predicted.

We quantify the effectiveness of individual test items using item MSE. It may be possible to use item MSE to learn more about the characteristics of effective EI test items. Developing more effective test items may lead to improved prediction of OPI test scores. In this paper, we do not attempt to address how linguistic factors (such as sentence length, syntactic complexity, lexical difficulty, and morphology) affect test item effectiveness for OPI prediction. However, others have discussed similar questions (Tomita et al., 2009; Christensen et al., 2010).

It may be possible that a test-taker could learn strategies for doing well on an EI test, without developing more general speech communication skills. If test-takers were able to learn such strategies, it may affect the usefulness of EI tests. Bernstein et al. (2010) suggest that, as yet, no conclusive evidence has been presented on this issue, and that automated test providers welcome such research.

It is possible that other automated systems are found to be more effective as a means for testing speech communication skills, or as a means for predicting OPI scores. We expect this to be the case. The purpose of this research is not to design the best possible system. Rather, it is to improve understanding of how such a system might be designed. It is shown that an EI test can be used as a key component of such a system. Strong correlation between actual and predicted OPI scores is achieved without using any other language testing method.

Acknowledgments

We would like to thank the Brigham Young University English Language Center for their support. We also appreciate assistance from the Pedagogical Software and Speech Technology research group, Casey Kennington, and Dr. C. Ray Graham.

References

Jared Bernstein, John De Jong, David Pisoni, and Brent Townshend. 2000. Two experiments on automatic scoring of spoken language proficiency. In P. Delcloque, editor, *Proceedings of InSTIL2000 (Integrating Speech Technology in Learning)*, pages 57–61.

Jared Bernstein, Alistair Van Moere, and Jian Cheng. 2010. Validating automated speaking tests. *Language Testing*, 27(3):355–377.

Craig Chaudron, Matthew Prior, and Ulrich Kozok. 2005. Elicited imitation as an oral proficiency measure. Paper presented at the 14th World Congress of Applied Linguistics, Madison, WI.

Carl Christensen, Ross Hendrickson, and Deryle Lonsdale. 2010. Principled construction of elicited imitation tests. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Tobias Cincarek, Rainer Gruhn, Christian Hacker, Elmar Nth, and Satoshi Nakamura. 2009. Automatic pronunciation scoring of words and sentences independent from the non-native's first language. *Computer Speech and Language*, 23(1):65 – 88.

Martin Fujiki and Bonnie Brinton. 1987. Elicited imitation revisited: A comparison with spontaneous language production. *Language, Speech, and Hearing Services in the Schools*, 18(4):301–311.

C. Ray Graham, Deryle Lonsdale, Casey Kennington, Aaron Johnson, and Jeremiah McGhee. 2008. Elicited Imitation as an Oral Proficiency Measure with ASR Scoring. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 1604–1610, Paris, France. European Language Resources Association.

Grant Henning. 1983. Oral proficiency testing: comparative validities of interview, imitation, and completion methods. *Language Learning*, 33(3):315–332.

Pieter Müller, Febe de Wet, Christa van der Walt, and Thomas Niesler. 2009. Automatically assessing the oral proficiency of proficient L2 speakers. In *Proceedings of the ISCA Workshop on Speech and Language Technology in Education (SLaTE), Warwickshire, UK*.

Yasuyo Tomita, Watuaru Suzuki, and Lorena Jessop. 2009. Elicited imitation: Toward valid procedures to measure implicit second language grammatical knowledge. *TESOL Quarterly*, 43(2):345–349.

Thora Vinther. 2002. Elicited imitation: a brief overview. *International Journal of Applied Linguistics*, 12(1):54–73.

Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel. 2004. Sphinx-4: A flexible open source framework for speech recognition.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.

Klaus Zechner and Xiaoming Xi. 2008. Towards automatic scoring of a test of spoken language with heterogeneous task types. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 98–106. Association for Computational Linguistics.