Elicited Imitation as an Oral Proficiency Measure with ASR Scoring

C. Ray Graham, Deryle Lonsdale, Casey Kennington, Aaron Johnson, and Jeremiah McGhee

Brigham Young University
Provo, UT, USA 84602
{ray_graham,lonz}@byu.edu, {casey.r.kennington,ajmagnifico,jlmcghee}@gmail.com

Abstract

This paper discusses development and evaluation of a practical, valid and reliable instrument for evaluating the spoken language abilities of second-language (L2) learners of English. First we sketch the theory and history behind elicited imitation (EI) tests and the renewed interest in them. Then we present how we developed a new test based on various language resources, and administered it to a few hundred students of varying levels. The students were also scored using standard evaluation techniques, and the EI results were compared to more traditionally derived scores. We also sketch how we developed a new integrated tool that allows the session recordings of the EI data to be analyzed with a widely used automatic speech recognition (ASR) engine. We discuss the promising results of the ASR engine's processing of these files and how they correlated with human scoring of the same items. We indicate how the integrated tool will be used in the future. Further development plans and prospects for follow-on work round out the discussion.

1. Background

1.1. Oral proficiency and testing

Acquiring a language is a difficult and time-consuming process, and language learners need frequent (if not continual) feedback on their skills development. Various methods have been developed for testing oral language acquisition: oral proficiency interviews (OPI's), simulated oral proficiency interviews (SOPI's), oral response elicitation tools, and elicited imitation (EI) tests. However, oral language testing is difficult to do and systematic testing is expensive, so most language acquisition programs do not systematically measure speaking outcomes for their programs or students.

In recent years foreign language learners across the entire world are beginning to recognize that, in order for language learning to be practical, it must include oral fluency. Thus the demand for the development of speaking ability in foreign language learning is driving interest in developing more efficient techniques for measuring oral language learning outcomes. In the past it has been impractical to systematically measure speaking outcomes in foreign language programs because testing methods were time consuming and expensive. For the most part learners had to be tested individually by highly trained interviewers and the resulting scores were so broadly defined that progress made over a semester of instruction was hard to detect.

1.2. Elicited imitation

The present paper describes a study in which we explore the use of an oral language testing technique which promises to be inexpensive, efficient, and reliable. Its development was begun almost a half century ago and it has been used for measuring normal and abnormal language development in native speakers (Ervin-Tripp, 1964; Fujiki and Brinton, 1987; Devescovi and Caselli, 2007) as well the development of speaking skill in foreign languages (Vinther, 2002; Chaudron et al., 2005). It consists of having learners listen to and repeat, to the best of their ability, utterances of varying lengths and complexities in the language being acquired.

This technique for assessing oral language has been somewhat controversial because of its apparent lack of validity. Practitioners have been slow to accept it since it is hard for them to see how repeating sentences orally can measure something as complex as oral language proficiency. However, research is confirming that the technique measures a construct similar to that of oral language proficiency in a highly reliable manner. For example in a study involving the evaluation of three oral language assessment techniques—elicited imitation, oral interview, and sentence completion—Henning (1983) found that the elicited imitation technique outperformed both of the other techniques on measures of validity and reliability.

An account of how elicited imitation works is given by Bley-Vroman and Chaudron (1994) as follows:

- The subject hears the input and processes it, forming a representation (in memory).
- The resulting representation includes information at various levels (including meaning).
- The representation must be kept in short-term memory.
- The subject formulates (and produces) a sentence based on the accessed representation. (There may also be monitoring of the phonetic plan, comparing it to the model.)

Since short-term or working memory is limited, the retention of a representation there is, by most accounts, dependent upon the number of units being processed. As the length of utterances becomes greater, it necessitates the chunking of information into successively larger units in order that the representation may be retained in working memory until it is repeated. It is believed that language competence is what facilitates this chunking process. The more proficient the speaker is in the language in question, the more efficient are the automatic formulations of representations and the more accurate the reconstruction of utterances. So, if learners are presented with a variety of sentences that vary in length and syntactic complexity, their

ability to understand the sentence and then reconstruct them through their interlanguage system, will vary according to their overall speaking proficiency.
The more proficient the speaker the longer and more complex will be the sentences which he or she can repeat accurately. Thus the elicited imitation technique promises to provide an efficient and reliable method, albeit somewhat indirect, of measuring second language speaking proficiency. Figure 1 shows some sample EI sentences of varying complexity.

```
She speaks English.

Perhaps he works there.

Does that woman help her students?

When I was a teenager I would go to town every day.

I hope that she likes the play because if she does
   we'll have a party.

Hesitating before she spoke her next line, the actress
   reached the pinnacle of her nervousness.
```

Figure 1: Some sample EI sentences.

2. The Elicited Imitation Study

The first part of this research involved development and refinement of an elicited imitation instrument, which proceeded in two phases. In this section we sketch the process for both phases.

2.1. The pilot study

The first phase involved a pilot study where three separate EI tests (Forms A, B, and C) were developed in parallel. Each form had sixty items (i.e. sentences), each chosen in accord with the criteria established in previous literature (Chaudron et al., 2005). These included a wide variety of morphological and syntactic structures involving variables such as sentence length, sentence complexity, vocabulary levels, and breadth of sampling structures. For example, the items in each form ranged in length from three syllables to twenty-four syllables. 13 items were repeated on all forms, and 47 sentences were unique to each form.
High-quality recordings of these stimulus sentences in the three forms were made in a studio with both male and female voices, and the forms were tested on adult native speakers. Subsequently the three forms were presented in parallel to 232 ESL learners in an intensive English program (IEP) in the U.S. The students represented 13 widely varying first-language (L1) backgrounds, and proficiency levels from novice to advanced. Their ages ranged from 18 to 53 years (mean$= 24.5$, s.d.$= 6.9$).
Subjects listened to the stimulus sentences via computers with microphone headsets and recorded their responses, saving their sound files to a server. These were retrieved and each sentence was scored for accuracy independently by two separate human raters. Each rater used two systems for scoring each item, one using a four-point scale per standard recommendations (Chaudron et al., 2005) and the other by simply counting the total number of syllables repeated correctly.
Figure 2 shows some sample scored items. Associated with each (pseudo-)syllable is either a 1 (meaning the syllable was pronounced) or a 0 (meaning it wasn't). The final score for each item depends on how many misses there were in that item. Where different words were used, they were also annotated parenthetically.
Item analyses were performed on these scores and reliability coefficients were computed for each form. Table 1 shows the very encouraging results.
Figure 3 shows two person/item maps from IRT analyses, one for Form A and one for Form C. On the left side of each, subject scores are presented on a standard scale with more proficient learners at the top and less proficient ones at the bottom. Item scores are presented on the right side of each, with difficult items at the top and easier items at the bottom. Mean scores for persons and items are marked with an "M" on either side of the middle line. Test difficulty can be ascertained by observing the distribution of the points up the scale.
More details on the pilot study and a deeper analysis can be found elsewhere (Graham, 2006).

Form	Items	Persons	Person RSM	Cronbach Alpha RSM	Item Reliability
A	58	78	.98	.97	.98
B	59	73	.99	.97	.98
C	60	72	.96	.96	.97

Table 1: Reliability for items from the three pilot study forms (RSM=Raw-Score-to-Measure).

2.2. The refined test

From the 60 best-discriminating items in the pilot study we created a new refined test, called Form D. The selected sentences ranged in length from five syllables to twenty-two syllables. Form D was administered to 156 adult ESL learners in the same IEP program. They came from twelve L1 backgrounds, their English proficiency levels ranged from novice to advanced, and their ages ranged from 18 to 55 (mean$= 24.3$, s.d.$= 6.8$). On average the learners took from seven to ten minutes to complete the test.
On separate occasions within a few days of the elicited imitation test, these subjects were also given additional speaking tests administered by qualified examiners. These included:

- an informal 15-minute placement interview,
- a 30-minute simulated computer administered oral proficiency test (ECT),
- a 30-minute computer elicited oral achievement test (LAT), and
- an oral proficiency interview (OPI) administered by certified ACTFL testers to a stratified random sample of 40 of the 156 participants.

The utterances from Form D were scored by two humans as described above for the pilot study. In addition, these EI test results were correlated with the outcomes of these other testing modalities, as explained below.

```
   1    1     1     1     1     1    1   1      1
1. If  she   lis   tens, she  will  un  der   stand.                          Score=4

   1    0   1    1     1    1  1
2. Why had they liked peas so much?                                         Score=3

    1     1    0    1     1    1     1
3. Big ships  will  al   ways make noise.                                    Score=3
             (are)

     0   1     0   1    0    1      1   0    1
4.  We should have ea  ten  break  fast by now.                             Score=0
   (They)          (eat)             (right)

   1   1   1     0    0   1   1    1   1   0     1  0   0   0   1   1    1
5. If her heart were to stop beat ing  we might not be  a  ble  to help her! Score=0
               (will)(be)                          (will)(being)
```

Figure 2: Scoring some sample EI sentences.

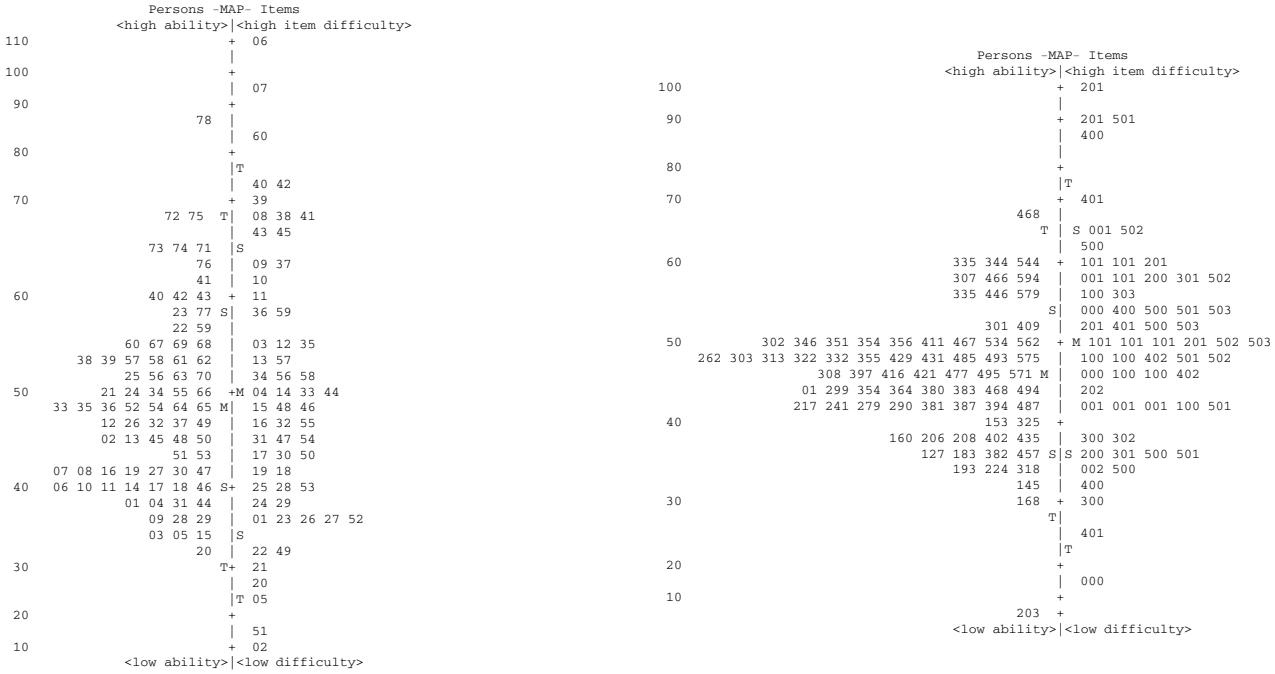

Figure 3: IRT analyses for Form A (a) and Form C (b).

2.3. Analysis of results

Table 2 presents the reliability statistics for the item analysis performed on the outcomes of Form D. One item was repeated correctly by all of the participants and two items were not repeated correctly by anyone, so these were removed, hence the 57 items analyzed. Two subjects had incomplete data and were removed from the analysis.

Form	Items	Persons	Person RSM	Cronbach Alpha RSM	Item Reliability
D	57	154	.98	.96	.96

Table 2: Reliability for items in Form D, the refined test (RSM=Raw-Score-to-Measure).

It was clear from the pilot study and from Form D that some EI sentences perform well in assessing students' abilities, and others don't. Whether an item will in fact perform well or not is not *a priori* obvious. To illustrate, the items listed in Figure 4(a) performed well whereas those in Figure 4(b) did not.

Table 3 shows how well the EI scoring compares with that of the various human-administered testing instruments. The 0.92 correlation between two different methods of scoring the EI test suggest that either method works about equally as well.

Notice also that correlations between the EI test scores are on the same order as intercorrelations among the other four more conventional methods of measuring oral language proficiency. In particular, the EI correlates with the OPI as well or better than the informal placement interview and the two computerized speaking tests, which require much more training and time to administer and score.

Our work in developing and administering the EI instrument involved presenting large numbers of EI items to almost 400 ESL students, whose responses were very consistent. Furthermore, overall comparisons between EI scores and scores on other measures of oral language proficiency were very promising. In our work EI was shown to be a

When she went to Las Vegas, did she like the shows that she saw?

Perhaps he works there.

If her heart were to stop beating, we might not be able to help her.

Had you ever flown that high before?

Good cars will never break down.

(a)

Have you slept?

Maybe she likes cats.

We eat cookies.

He should have walked away before the fight started.

How do good children play baseball?

Chris has yelled louder than ten sheep.

(b)

Figure 4: Well-performing EI sentences (a) and poorly-performing ones (b).

	EI Traditional	EI Syllable	ECT Speaking	OPI	Oral Placemt.	LAT Speaking
EI Traditional	1	.925	.516	.658	.639	.551
EI Syllable	.925	1	.465	.648	.691	.414
ECT Speaking	.516	.465	1	.432	.577	.442
OPI	.658	.648	.432	1	.660	.652
Oral Placemt.	.639	.691	.577	.660	1	—
LAT Speaking	.551	.414	.442	.652	—	1

Table 3: Pearson correlations of the various oral language measures used in the study. Cases where subjects took mutually exclusive tests are indicated with —.

highly reliable way of measuring a single trait of oral language use. However, exactly what that trait is and the degree to which it correlates with other measures of oral language proficiency could be further elucidated with subsequent work. Still, specific items can be shown to consistently detract from or add to the reliability of the measure, and it is these items that interest us most particularly. Finally, our work has shown that the scoring procedure developed by Chaudron et al. (2005) appears to work reasonably well, although other procedures should be experimented with.

Obvious advantages of the EI technique over conventional methods of oral language testing include that:

- the test can be administered to multiple learners at the same time,
- it can be administered in a conventional computer lab without the assistance of a highly trained oral interviewer, and
- it can be scored rather efficiently by a reasonably proficient speaker of the target language.

This last advantage is made even more interesting by our attempts at developing an automatic scoring procedure using speech recognition technology. This research will be described in the following section.

3. Speech recognition

Automatic speech recognition (ASR) involves processing spoken language to extract its content. It is a complex task combining physics, engineering, mathematics, statistics, and linguistics. The current state of the art has produced accuracy ranges from barely tolerable to very good, depending on the particular application. In this regard ASR is just becoming practical and viable in some domains for English, though generally it is less well developed for most other languages. Though there are notable commercial enterprises involved in ASR development, the technology is becoming increasingly more available in open-source repositories. Our prior work (Lonsdale et al., 2005) has focused on several ASR applications from dialogue to language pedagogy.[1]

Conceptually, ASR involves taking an input acoustic signal (pre-digitized if necessary) and sampling it at regular intervals. The samples are then analyzed for features that are salient for downstream processing. The properties of each sample are sent through a classifier to ascertain which language sounds (or phones) best match the sample in question.

For this project we used the Sphinx ASR engine (Lee, 1989) which was developed at CMU. We manipulated three main components:

- the recognizer, which handles the signal processing;
- the grammar, which specifies the language model, and
- the linguist, which manages lexical and phonological properties of the words in the language (in this case English)

3.1. ASR for EI

The applicability of ASR for EI is an interesting question that to our knowledge has not received any attention in the current research literature.

There are *a priori* a few considerations which may lead one to suspect that scoring EI tests with ASR might perhaps be problematic:

- ASR is still an emerging technology.

[1] For more details see http://psst.byu.edu.

- Automating the task involves a nontrivial integration with already complex systems.

- The speakers in EI tests are *non-native* speakers with (sometimes heavy) L1 accents, whereas ASR models are tuned and trained for recognition of native speakers.

- There is a granularity mismatch in the data since EI scores are done at the syllable level whereas ASR scores are computed at the word level.

On the other hand, several considerations make the EI/ASR nexus compelling:

- Since humans can score EI following strict scoring criteria, it is reasonable to expect that one could automate the task.

- The expected input for any given test sentence is already known, so the ASR task is much more constrained.

- The ASR task can be developed with open-source technology.

- There is a sizable potential economic benefit if the test can be delivered on a large scale, short turnaround time, and at low cost when compared to human scoring.

- The procedure can be applied to score learners of other languages, provided ASR models are available for those languages.

In an effort to explore the tradeoffs just mentioned, we initiated research in developing an ASR capability for scoring the EI sessions. To summarize, it involved:

1. converting the files to an appropriate format;

2. testing how well Sphinx scores first the native model utterances and then iteratively refining this capability;

3. testing how accurately the ASR engine scores on non-native subjects;

4. iteratively refining the ASR engine on non-native subject recordings; and then

5. trying the system on unseen data and comparing the results to human evaluation scores.

The iterative refinement process involved trying out different recognizer parameters, grammar and lexical specifications, and language models. We discuss each iteration in improving the system's performance on (first) native model utterances and (then) non-native subject utterances. Word recognition rates ranged first from the low 70%s for native models to eventually the high 80%s for non-native subjects as a result of our improvements to the ASR system.

3.1.1. Processing native utterances

Our testing of the ASR performance on the native model speakers proceeded incrementally. We briefly summarize the stages of development undertaken in this phase.

First, to minimize grammar engineering at the onset, our grammars consisted of simply all words used in the EI form in any order, thus assuming word-level independence. No other constraints or adaptations were made to the ASR engine or the knowledge sources. This allowed for rapid system development and a conservative baseline to compare future work on. The inputs were scored at the sentence level on an binary accept/reject basis. The result was a 71% recognition accuracy rate. Of course this left room for improvement, but we were somewhat surprised that the first attempts were this promising. Because of the unconstrained nature of the grammar, the perplexity was high and thus we had reason to believe that the results could be improved on.

We next proceeded to develop a full grammar where all sentences forced the system to recognize the words in the correct order. This reduced perplexity considerably and hence the scoring was much faster. One drawback with this type of analysis was that the sentence had to occur *in toto*, so that in:

```
i i saw her saw her run
i i saw her i saw her run
```

the former utterance would be rejected but the latter would be accepted since the whole sentence is uttered in one chunk. Given this setup the system achieved an 81% recognition accuracy.

We also developed visualization tools to help analyze the scoring data and thus help find problematic items and difficult areas within them. It also became clear that some of the files were clipped prematurely at the beginning, resulting in lower scores until we padded the sound files with leading silence, which helped noticeably.

The next level of effort involved forcing the system to use a fully specified grammar, but only for the sentence in question, when processing an item file. Note that this is only possible when the model utterance is known *a priori*, which is the case for EI tests (but not for typical ASR applications, e.g. speech transcription). This yielded an accuracy of just above 90%.

Up to this point we had been using the Hub4 acoustic model, which is trained on broadcast news. Replacing the model with the Wall Street Journal (WSJ) model boosted the word-level accuracy score to 99.7% for men and women, with a 93% accuracy rate at the sentence level.

3.1.2. Processing non-native utterances

Encouraged by the results for the native model speakers, we proceeded to evaluate how well the system worked on the non-native subject utterance files, which were scored by human judges. This process also involved several iterations.

For the first iteration we just computed the match between sentence-level ASR on the pilot test data we had been working with. This attempt, on Forms A, B, and C, resulted in a 0.88 correlation with the human scores.

However, we knew that we would eventually need to develop a scoring system that would take into consideration the granularity mismatch in scoring. Recall that ASR is

scored at the word or sentence level whereas the human judges provided syllable-level scores. We therefore developed a scoring system that maps from the syllable level to the word level and computes correlations accordingly. Even with this indirection in the scoring mechanism, the system achieved a 0.85 correlation. Note that this is in spite of the fact that the ASR system has no syllable-level language model. The slight loss in this rubric for measurement was tolerable since the scoring approach is much more ecologically valid.

Finally, we upgraded the grammar by strategically introducing more wildcards. This resulted in a final correlation for forms A, B, and C of 0.90 with the human scores. Figure 5 shows a scatterplot of the scores obtained during this development cycle for scoring all items used in the EI pilot study. In a perfectly correlated system the points would all lie along the diagonal.

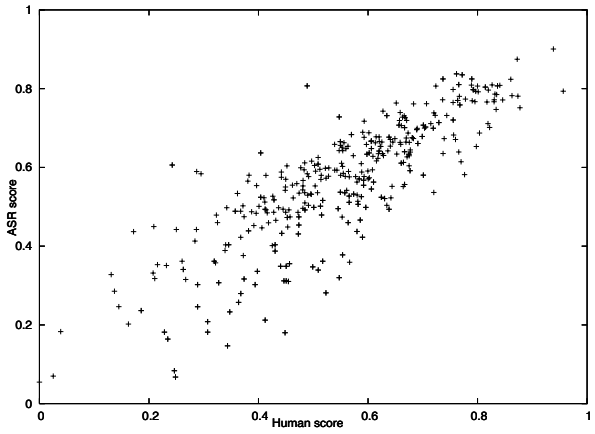

Figure 5: Correlation of ASR scores with human scores for Forms A, B, and C (development data).

Of course, a valid test of our work would need to be carried out on held-out or unseen data. Having refined the system with the pilot data (i.e. Forms A, B, and C) we proceeded to test it on Form D. Some work was required to reformat the human scores for this form, which was annotated in a slightly different manner. The acoustic data was also recorded using different tools and thus had to undergo another conversion process. Still, the data for Form D was for all intents and purposes unseen. The scoring obtained by the system on Form D achieved a 0.83 correlation with the human scores for this form. Figure 6 shows a scatterplot of these results.

As a final check on our results we also ran other validation tests. We consolidated Forms A, B, C, and D together and selected random subsets of data from each form, creating new test sets. On all of these (sub)sets we achieved correlations of between 0.85 and 0.88. Though technically the data in these sets was no longer unseen, we view these additional results as encouraging support for the results we have been achieving, since they preclude any possible effects due to temporal or scoring practice factors across the forms.

4. Future work

We anticipate future work in several directions.

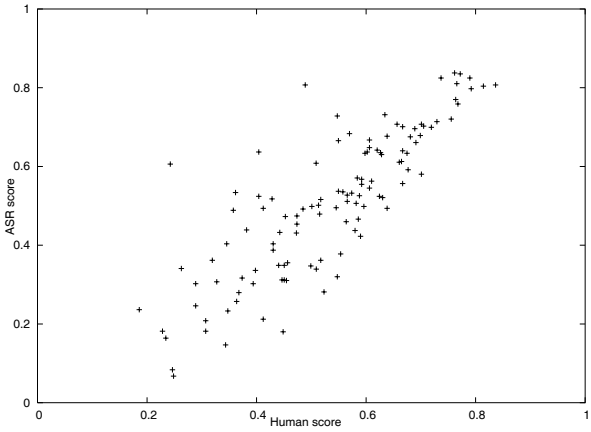

Figure 6: Correlation of ASR scores with human scores for Form D (unseen data).

First, we are in the process of refining the EI instrument, culling out the sentences that do not perform well in the ASR scoring. This will help produce future tests that will be even more amenable to ASR scoring.

We intend to carry out further exploration with the interrelationships between student responses and EI variables such as sentence length, complexity, and vocabulary. This includes developing extensive criteria for complexity at all levels: lexical, phonological, morphological, syntactic, and semantic. Using these criteria we will be able to better create a wide range of new EI items covering the spectrum of difficulty. Indeed, it might even be possible to semi-automate this process and produce an interactive tool for EI instrument development.

Further examination of responder variables such as working memory, native language, and age still need to be carried out. We expect that such features, in combination with the raw EI scores, will be useful in eventually applying machine learning or some other form of classification methodology to better score the students' responses.

Many of the ASR-related issues that remain include speech performance difficulties including false starts, the use of contractions, filled pauses, long hesitations, and poor recording quality on some responses.

We also have several hundred more EI tests that have already been administered and that still need to be scored by humans.

We intend to investigate EI with other languages. Of course, this will require the appropriate ASR infrastructure for those languages. The process will involve developing an EI instrument for that language and verifying that it correlates well with human scores, integrating the student recordings with an ASR engine developed for that language, and executing the process as described in this paper.

Another ASR development possibility is to train up one or more non-native acoustic models for the ASR component. This would improve scoring non-native speech. However this task seems unlikely in the near term since there is still a paucity of annotated corpora that could be used to train up a recognizer for this purpose.

We are also working toward using forced alignment (Li et al., 2005) as a diagnostic tool. This will be helpful in better

identifying items and passages where ASR scoring did not perform well.

Finally, we intend to pursue the development of EI instruments and ASR models for testing L2 learner abilities for other languages besides English.

Ultimately our goal is to develop a run-time adaptive speaking test that can be deployed for EI-based proficiency scoring, similar to those that are currently in use for evaluating reading and listening comprehension. If adjustments could be made in real time, the system could adjust selection of EI items based on the subject's performance, thus calibrating the test for a more exact evaluation.

5. Acknowledgements

We would like to thank Meghan Eckerson, Dan Rasband, Ben Millard, Ross Hendrickson, and Kevin Cook for linguistic and programming support on this project. We also appreciate the BYU English Language Center for its support in carrying out the various language testing activities.

6. References

R. Bley-Vroman and C. Chaudron. 1994. Elicited imitation as a measure of second-language competence. In E.E. Tarone, S. Gass, and A.D. Cohen, editors, *Research methodology in second language acquisition*, pages 245–261. Lawrence Erlbaum, Hilldale.

C. Chaudron, M. Prior, and U. Kozok. 2005. Elicited imitation as an oral proficiency measure. Paper presented at the 14th World Congress of Applied Linguistics, Madison Wisconsin.

A. Devescovi and M.C. Caselli. 2007. Sentence repetition as a measure of early grammatical development in Italian. *International Journal of Language and Communication Disorders*, 42(2):187–208.

S. Ervin-Tripp. 1964. Imitation and structural change in children's language. In E.H. Lenneberg, editor, *New directions in the study of language*, pages 163–189. M.I.T Press, Cambridge, MA.

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

C. R. Graham. 2006. An analysis of elicited imitation as a technique for measuring oral language proficiency. In Yi ju Chen and Yiu nam Leung, editors, *Selected Papers from the Fifteenth International Symposium on English Teaching*, pages 57–67, Taipei, Taiwan. English Teachers' Association.

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

K-F. Lee. 1989. *Automatic Speech Recognition: The Development of the SPHINX System*. Kluwer Academic Publishers, Boston, MA.

S.W. Li, H.T. Lin, and H.Y. Chen. 2005. How speech/text alignment benefits web-based learning. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, pages 259–260, New York, NY. ACM Press.

D. Lonsdale, C.R. Graham, and R. Madsen. 2005. Learner centered language programs: Integrating disparate resources for task-based interaction. In Panayiotis Zaphiris and Giorgos Zacharis, editors, *User Centered Computer Aided Language Learning*, pages 116–132. Information Science Publishing, Hershey, PA.

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