Scoring an Oral Language Test using Automatic Speech Recognition

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

Brigham Young University
lonz@byu.edu
Outline

- Learner language
  - Elicited imitation (EI)
  - EI instruments
  - Scoring EI, correlation with other tests

- Automatic analysis
  - Automatic speech recognition (ASR)
  - Scoring EI data via ASR (native, non-native)
  - Ongoing efforts

- Conclusions, future work
Learner language
Oral language testing

- Difficult to do, human-intensive, expensive
  - Most L2 English programs do not systematically measure speaking outcomes
- Difficult to tell if we are making progress in teaching oral language skills
- Some standard methods:
  - Oral proficiency interviews
  - Simulated oral proficiency tests
  - Oral response elicitation tools
  - Elicited imitation tests
Purpose

- Develop a practical, valid and reliable instrument to measure oral language proficiency in L2

- The second and third stages of this process are the focus of this presentation
What is Elicited Imitation (EI)?
A short history of EI

- Early 1960’s: child L1 development, diagnosing language abnormalities
- 1970’s: L2 acquisition
- Recent years:
  - Psycholinguistic research into language competence and SLA processes
  - Indirect measurement of oral language proficiency
Chaudron's (2005) model

1) Subject hears utterance

2) Subject forms representation

3) Representation kept in STM

4) Subject forms sentence based on accessed representation
Designing EI proficiency tests

- Sentence Length
- Sentence Complexity
- Vocabulary levels
- Breadth of sampling of structures
Developing our EI test

- Pilot study (Forms A, B, C)
  - Sentences selected according to criteria (Chaudron et al, 2005)
  - Recorded in studio with both male and female voices
  - Tested on adult native speakers

- Refined instrument (Form D)
  - Best-performing items from pilot study
  - Compare to other testing methods
Pilot study items

- EI Test: three forms (A, B, C)
- 60 sentences / form (13 repeated on all three forms, 47 unique to each form)
- Sentence length: 3 to 24 syllables
- Wide variety of morphological and syntactic forms
- Variety of lexical items (81.3%=K1, 6.7%=K2, .23%=AWL, 11.6%=Off)
Pilot study subjects

- 223 learners of English in an IEP in U.S.
- 13 widely varying L1 backgrounds
- English proficiency: Novice to Advanced
- Ages 18 to 53, mean = 24.5, SD = 6.9
EI test administration

1. Orientation. Log on to computer
2. Hear sentences recorded by a NES
3. Response recorded as they imitate
4. Log off. Sound files saved to server
# Reliability (all three forms)

<table>
<thead>
<tr>
<th></th>
<th>Items / Persons</th>
<th>Person Raw-Score-to-Measure (RSM)</th>
<th>Cronbach ALPHA (KR-20) Person RSM Reliability</th>
<th>Item Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>58 / 78</td>
<td>.98</td>
<td>.97</td>
<td>.98</td>
</tr>
<tr>
<td>B</td>
<td>59 / 73</td>
<td>.99</td>
<td>.97</td>
<td>.98</td>
</tr>
<tr>
<td>C</td>
<td>60 / 72</td>
<td>.96</td>
<td>.96</td>
<td>.97</td>
</tr>
</tbody>
</table>
Sample best items (by form)

A When she went to Las Vegas, did she like the shows that she saw?
A Perhaps he works there.
A If her heart were to stop beating, we might not be able to help her.
A He has eaten lunch at that diner before.
A Have you eaten all your bread with out any honey at all?

B That woman should help her students.
B What did he eat for breakfast?
B They have grown their corn this year without watering.
B Does that woman help her students?
B We should have eaten breakfast by now.

C Why don't they ride their big horses?
C Won't they ride their big horses?
C Good cars will never break down.
C Why had he played games so well?
C I would eat ice cream every night.
C Do big ships sink in ice cold water?
Sample worst items (by form)

A The Cat jumped from the box to the kitchen counter.
A The situation in Iraq calls for diplomacy and sensitivity.
A Have you slept?
A Maybe she likes cats.
A She quickly jumped down.
A They play games.
A The situation in Iraq calls for diplomacy and sensitivity.

B How well had he played the guitar before he began playing in the band?
B We eat cookies.
B She is feeling better already
B How do big ships keep from sinking in stormy waters?
B Joe writes poetry.

C How do good children play baseball?
C They play games.
C Do they make sweaters?
C The child is acting obnoxious at preschool today.
C Are you feeling better?
Form A
Person / item map
### Form B

**Person / item map**

<table>
<thead>
<tr>
<th>Persons</th>
<th>MAP</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ 301</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ 300</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ 201</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T 501</td>
<td>201</td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>201</td>
<td></td>
</tr>
<tr>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>400</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5993</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4773</td>
<td>4922</td>
<td></td>
</tr>
<tr>
<td>4345</td>
<td>5805</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>4087</td>
<td>4832</td>
</tr>
<tr>
<td>3615</td>
<td>4213</td>
<td></td>
</tr>
<tr>
<td>3908</td>
<td>4953</td>
<td>5140</td>
</tr>
<tr>
<td>4968</td>
<td>S</td>
<td>001</td>
</tr>
<tr>
<td>3273</td>
<td></td>
<td>000</td>
</tr>
<tr>
<td></td>
<td>2562</td>
<td>3144</td>
</tr>
<tr>
<td>50</td>
<td>3367</td>
<td>3958</td>
</tr>
<tr>
<td>2402</td>
<td>3093</td>
<td>3361</td>
</tr>
<tr>
<td>3020</td>
<td>3033</td>
<td>3043</td>
</tr>
<tr>
<td>3349</td>
<td>3648</td>
<td>3963</td>
</tr>
<tr>
<td>2305</td>
<td>2449</td>
<td>2562</td>
</tr>
<tr>
<td>2448</td>
<td>2512</td>
<td>2778</td>
</tr>
<tr>
<td>40</td>
<td>1497</td>
<td>1953</td>
</tr>
<tr>
<td>3852</td>
<td></td>
<td>301</td>
</tr>
<tr>
<td>1108</td>
<td>3884</td>
<td></td>
</tr>
<tr>
<td>1918</td>
<td>3315</td>
<td>4425</td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>1486</td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>202</td>
<td>302</td>
</tr>
<tr>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1757</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T 500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>+ 200</td>
<td></td>
</tr>
</tbody>
</table>

<more> | <rare>

<less> | <frequent>
### Form C
### Person / item map

```plaintext
<table>
<thead>
<tr>
<th>Persons</th>
<th>MAP</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>+</td>
<td>201</td>
</tr>
<tr>
<td>90</td>
<td>+</td>
<td>201 501</td>
</tr>
<tr>
<td>80</td>
<td>+</td>
<td>400</td>
</tr>
<tr>
<td>70</td>
<td>+</td>
<td>401</td>
</tr>
<tr>
<td>60</td>
<td>+</td>
<td>101 101 201</td>
</tr>
<tr>
<td>50</td>
<td>+</td>
<td>100 303</td>
</tr>
<tr>
<td>40</td>
<td>+</td>
<td>100 303</td>
</tr>
<tr>
<td>30</td>
<td>+</td>
<td>400</td>
</tr>
<tr>
<td>20</td>
<td>+</td>
<td>000</td>
</tr>
<tr>
<td>10</td>
<td>+</td>
<td>203</td>
</tr>
</tbody>
</table>
```

*More items and persons are listed, but the table is truncated for brevity.*
Refined instrument (Form D)

- 60 best discriminating items from pilot study
- Sentence length 5 to 22 syllables
- Administered as in pilot study
Subjects (Form D)

- 156 ESL learners at a U.S. university
- 12 L1 backgrounds including Portuguese, Spanish, Korean, Japanese, Chinese, etc.
- English proficiency levels from Novice to Advanced
- Ages 18 to 55, mean = 24.3, SD = 6.8
Further speaking tests (Form D subjects)

- An informal 15-minute placement interview
- A 30 minute simulated oral proficiency test ECT
- A 30 minute computer elicited oral achievement test LAT
- An OPI administered by certified ACTFL testers (to a stratified random sample)
Scoring method 1

- Similar to Chaudron et al. (2005)
- Divide sentences into syllables
- Mark each syllable with 1 or 0
- Transcribe mistakes below target sentence.
- Scoring 0-4 -1 for each error
Scoring the Imitations

1. If she listens, she will understand. 4

2. Why had they liked peas so much? 3

3. Big ships will always make noise. 3

4. We should have eaten breakfast by now. 0

5. If her heart were to stop beating we might not be able to help her! 0
Scoring method 2

- Correct syllable count

- 1 point for each syllable repeated accurately

- 0 points for incorrect syllables
Form D reliability
57 items 154 persons

- Item reliability = .98
- Person raw score-to-measure (RSM) correlation = .96
- Cronbach ALPHA (KR-20) person (RSM) reliability = .96
## Pearson correlation across tests

<table>
<thead>
<tr>
<th></th>
<th>EI Traditional</th>
<th>EI Syllable</th>
<th>ECT L2 Speak</th>
<th>OPI</th>
<th>Oral Placem</th>
<th>LAT Speaking</th>
</tr>
</thead>
<tbody>
<tr>
<td>EI Traditional</td>
<td>1</td>
<td>.925(**)</td>
<td>.516(**)</td>
<td>.658(**)</td>
<td>.639(**)</td>
<td>.551(**)</td>
</tr>
<tr>
<td>EI Syllable</td>
<td>.925(**)</td>
<td>1</td>
<td>.465(**)</td>
<td>.648(**)</td>
<td>.691(**)</td>
<td>.414(**)</td>
</tr>
<tr>
<td>ECT L2 Speak</td>
<td>.516(**)</td>
<td>.465(**)</td>
<td>1</td>
<td>.432(**)</td>
<td>.577(**)</td>
<td>.442(**)</td>
</tr>
<tr>
<td>OPI</td>
<td>.658(**)</td>
<td>.648(**)</td>
<td>.432(**)</td>
<td>1</td>
<td>.660(**)</td>
<td>.652(*)</td>
</tr>
<tr>
<td>Oral Placem</td>
<td>.639(**)</td>
<td>.691(**)</td>
<td>.577(**)</td>
<td>.660(**)</td>
<td>1</td>
<td>.(a)</td>
</tr>
<tr>
<td>LAT Speaking</td>
<td>.551(**)</td>
<td>.414(**)</td>
<td>.442(**)</td>
<td>.652(*)</td>
<td>.(a)</td>
<td>1</td>
</tr>
</tbody>
</table>
Conclusions regarding EI

- We have presented large numbers of EI items to almost 400 ESL students
- Student responses to the elicited imitation task are very consistent
- Overall comparisons between EI scores and scores on other measures of oral language proficiency are promising
- The Chaudron et al (2005) scoring procedure works reasonably well; other procedures should be experimented with
Automatic analysis
Automatic speech recognition

- ASR for short
- Processing spoken language to extract its content
- Complex task: physics, engineering, mathematics, statistics, linguistics
- Accuracy ranges from barely tolerable to very good, depending on application
- Just becoming practical, viable in some domains for English
- Less well developed for most other language
- Increasingly more open-source availability
Conceptual overview of ASR
Language model specifications

PHONE MODEL

WORD MODEL

WORD CLASS

NAME = {John, Bill, Diego}

VERB = {painted, lifted, made}

DET = {his, a, my}

OBJECT = {chair, desk, table}

GRAMMAR

SILENCE → NAME → VERB → DET → OBJECT → SILENCE
Typical ASR components

(Zue & Cole 1995)
Sphinx

- ASR engine developed at CMU
- We manipulate three main components:
  - Recognizer: handles the signal processing
  - Grammar: specifies the language model
  - Linguist: manages lexis, phonology, etc.
EI/ASR: perhaps a bad idea

- ASR is still an evolving technology
- Technically complex integration with already complicated systems
- Speakers are non-natives, ASR models are trained on native speakers
- Granularity mismatch: human scores derived via syllable-level assessment, ASR handles whole words
EI/ASR: perhaps a good idea

- Humans can do the scoring task, so perhaps we can automate it
- Economic benefit: reduces scoring costs, turnaround
- Finite problem: expected input is known a priori
- Open-source technology
- Potential payoff: use for other languages
ASR for EI: our approach

- Convert EI files to appropriate format
- Test how well Sphinx does on native model items
  - Iteratively refine performance of integration
    - Try different grammars, phoneme mappings, language models, etc.
- Test how well Sphinx does on non-native subjects
  - Iteratively refine...
- Try unseen data
1st iteration (NS): full G, any WO

- All (and only) words from EI items; very small subset of English
- Word-level independence assumption
  
  \[\text{e.g.: } (i \mid \text{saw} \mid \text{her} \mid \text{run})^* \mid (\text{he} \mid \text{saw} \mid \text{me} \mid \text{jump})^* \mid \ldots\]

- Approach was limited
  - Dropped words, repeated words, any ordering
  - Processing perplexity (time)

- Still, we achieved 71% word-level accuracy (natives)
2nd iteration (NS): full G, WO enforced

- Full grammar as in 1st iteration
- No wildcard on individual word occurrences
e.g.: (i saw her run) | (he saw me jump) | ...
- Forces sentence-level ASR scoring: binary accept/reject based on analysis of each word
- Advantages:
  - Wholistic grade at the sentence level vs. the word level
  - Reasonable for the native speaker
  - Much faster, reduced perplexity
- Problem:
  - Repeated words are disallowed.
"i i saw her saw her run" → rejection,
"i i saw her i saw her run" → acceptance
- Achieved 81% accurate recognition rate
ASR improvement

- Developed **visualization tools** for finding problematic areas
- Listening to files for quality control
- List of sentences not recognized at all
- Grammar engineering
3rd iteration (NS): sentence G, WO enforced

- Each individual sentence has its own grammar
  - Only Possible because we know target utterance
e.g.: ( i saw her run )
- Similar to last iteration, but considerably reduced perplexity
- Achieved 84% accurate recognition rate
Acoustic model experimentation

- So far using Hub4 (broadcast news)
- Adjustments:
  - Files clipped at start
  - Simple grammar engineering for problem areas
- Results: low 90% accuracy
- Decided to try WSJ language model
- Improved results
  - 99.7% word-level accuracy for native speakers (men and women)
4th iteration: non-native speakers (NNS)

- Had tried small batches of NNS earlier (results were surprising to us)
- Remember: humans scored at syllable level
  - ASR scored at word level
  - Scoring system mapped between two levels
- Correlation with human scorers: 0.88 (averaged per subject)
Correlation with NNS scores
### 4th iteration: non-native speakers

<table>
<thead>
<tr>
<th>Sentence</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
<th>8th</th>
<th>9th</th>
<th>10th</th>
<th>Correctness</th>
</tr>
</thead>
<tbody>
<tr>
<td>he is running away</td>
<td>4</td>
<td>58</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td>67.44%</td>
<td></td>
</tr>
<tr>
<td>what did he eat for breakfast</td>
<td>6</td>
<td>58</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td>67.44%</td>
<td></td>
</tr>
<tr>
<td>when does he eat dinner</td>
<td>5</td>
<td>59</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td>68.60%</td>
<td></td>
</tr>
<tr>
<td>she is feeling better already</td>
<td>5</td>
<td>60</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td>69.77%</td>
<td></td>
</tr>
<tr>
<td>are you feeling better</td>
<td>4</td>
<td>59</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td>70.24%</td>
<td></td>
</tr>
<tr>
<td>I cried because the book was so sad</td>
<td>8</td>
<td>62</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
<td>70.45%</td>
<td></td>
</tr>
<tr>
<td>when he’s racing he’s very fast</td>
<td>6</td>
<td>62</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
<td>70.45%</td>
<td></td>
</tr>
<tr>
<td>she has been riding the horse</td>
<td>6</td>
<td>63</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
<td>70.79%</td>
<td></td>
</tr>
<tr>
<td>perhaps he works there</td>
<td>4</td>
<td>64</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
<td>71.91%</td>
<td></td>
</tr>
<tr>
<td>we are going to walk slowly</td>
<td>6</td>
<td>62</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td>72.09%</td>
<td></td>
</tr>
<tr>
<td>why is the boy eating donuts</td>
<td>6</td>
<td>62</td>
<td>85</td>
<td></td>
<td></td>
<td></td>
<td>72.94%</td>
<td></td>
</tr>
<tr>
<td>they are going to work hard</td>
<td>6</td>
<td>62</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td>73.81%</td>
<td></td>
</tr>
<tr>
<td>when does she study math</td>
<td>5</td>
<td>62</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td>73.81%</td>
<td></td>
</tr>
<tr>
<td>she speaks English</td>
<td>3</td>
<td>64</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td>74.42%</td>
<td></td>
</tr>
<tr>
<td>sometimes they go to town</td>
<td>5</td>
<td>69</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
<td>77.53%</td>
<td></td>
</tr>
<tr>
<td>we eat cookies</td>
<td>3</td>
<td>67</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td>77.91%</td>
<td></td>
</tr>
<tr>
<td>what do you do all day</td>
<td>6</td>
<td>69</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td>80.23%</td>
<td></td>
</tr>
<tr>
<td>she walked the dog today</td>
<td>5</td>
<td>72</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
<td>80.90%</td>
<td></td>
</tr>
<tr>
<td>have you slept</td>
<td>3</td>
<td>72</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
<td>80.90%</td>
<td></td>
</tr>
<tr>
<td>I can make cookies</td>
<td>4</td>
<td>70</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td>81.40%</td>
<td></td>
</tr>
<tr>
<td>she watches tv in the commons room</td>
<td>7</td>
<td>72</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
<td>81.82%</td>
<td></td>
</tr>
<tr>
<td>maybe she likes cats</td>
<td>4</td>
<td>73</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
<td>82.02%</td>
<td></td>
</tr>
<tr>
<td>is john going to the party</td>
<td>6</td>
<td>74</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
<td>83.15%</td>
<td></td>
</tr>
<tr>
<td>joe writes poetry</td>
<td>3</td>
<td>218</td>
<td>259</td>
<td></td>
<td></td>
<td></td>
<td>84.17%</td>
<td></td>
</tr>
<tr>
<td>she quickly jumped down</td>
<td>4</td>
<td>77</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
<td>86.52%</td>
<td></td>
</tr>
<tr>
<td>they played games</td>
<td>3</td>
<td>228</td>
<td>259</td>
<td></td>
<td></td>
<td></td>
<td>88.03%</td>
<td></td>
</tr>
<tr>
<td>are you done</td>
<td>3</td>
<td>90</td>
<td>91</td>
<td></td>
<td></td>
<td></td>
<td>98.90%</td>
<td></td>
</tr>
</tbody>
</table>
5th iteration: non-native speakers (NNS) syllable-level scoring

- Grammar and lexicon were re-engineered with syllables only
  - Or’s and kleene star: 0.84
  - Forced ordering: 0.85
  - Better than what we expected

- Note: no syllable-level language model
Syllable-level correlation (NNS)
Future work

- Refine EI instrument w/r/t ASR capabilities
- More tests to score (Form D, unseen)
- Work towards forced alignment
- Develop non-native acoustic models
- Refine linguistic issues in ASR process (contractions, false starts, etc.)
- Other languages (EI instrument, ASR)
Future work

- Explore student responses vs. EI variables (e.g. sentence length, sentence complexity, vocabulary)
- Examine responder variables (e.g. working memory, native language, age, ...)
- Experiment with new ways of scoring and weighting items (annotating level/type of errors)
- Use new analysis tools to examine factors which contribute to learner responses
- Develop computer adaptive speaking tests similar to those used currently in reading and listening
Conclusions

- **ASR**: best to use acoustic model as intended (i.e. word level)
- **OTS Sphinx**: excellent results for NS
- **Given that EI is workable, we can assure concomitant Sphinx performance**
- **EI/ASR was quite impervious to L1, age, gender effects**
Questions?