1. Background

- Oral proficiency tests are an important aspect of language skill assessment.
- Human assessment:
  - Labour intensive
  - Is often very subjective
- Reading and writing skill tests can be computerised, but human assessment is impractical.
- Reading / writing ability not necessarily well correlated with oral ability.

2. Context

- Students at the Stellenbosch University Education Faculty must enrol in a language module appropriate to their level of proficiency.
- Progress must be monitored regularly thereafter.
- Large number of students per staff member makes human assessment impractical.
- Students have high L2 proficiency.
- English as second language rather than foreign language.

3. Method

- Students took oral test and responses were recorded.
- Responses were scored using ASR system - Machine Scores.
- Correlation between Human Ratings and Machine Scores were calculated.
- Good correlations indicate scoring algorithms with the potential to accurately predict human assessments.

4. Test Design

- Test consisted of 7 tasks. We focus on two of these:
  - READING TASK
  - REPEATING TASK

  **EXAMPLE**
  "Many participants asked if this was the best way forward"

  "Learners who are out of touch with school practice have unrealistic expectations"

5. Human Assessment

- For each student, 3 randomly selected reading and 3 repeating responses were assessed by human raters.
- 6 raters, teachers of English as a second language.
- Each student was assessed by 3 raters.
- Allows calculation of inter-rater correlation.
- Each rater assessed 5 students twice.
- Average intra-rater correlation of 0.85.

6. Automatic Assessment

- ASR system used speaker independent cross-word triphone HMMs trained on 6h of phonetically annotated telephone speech.
- Reading Task recognition used fixed finite state grammar and Repeating Task recognition used equal probability unigram language model.

**Calculation of Automatic Indicators**

**Correlation between Machine Scores and Human Ratings**

- Rate of Speech: $R_{OS}$
- Total Duration: $T_{Total}$
- Number of Speech Phones: $N_{SP}$
- Number of Correct Phones: $N_{C}$
- Number of Insertions: $N_{I}$
- Number of Deletions: $N_{D}$

**Accuracy**

\[ A_{acc} = \frac{N_{C}}{N_{P}} \]

**Pronunciation**

\[ p_{acc} = \frac{N_{C}}{N_{P}} \]

**Success**

\[ s_{acc} = \frac{N_{C}}{N_{P}} \]

**Accuracy**

\[ A_{acc} = \frac{N_{C}}{N_{P}} \]

**Rate of Speech**

\[ R_{OS} = \frac{N_{SP}}{T_{Total}} \]

**Total Duration**

\[ T_{Total} \]

**Number of Speech Phones**

\[ N_{SP} \]

**Number of Correct Phones**

\[ N_{C} \]

**Number of Insertions**

\[ N_{I} \]

**Number of Deletions**

\[ N_{D} \]

**Expanding the Assessment of Oral Language Proficiency**

- Develop automatic system for the large scale assessment of oral language proficiency.

7. Conclusions

- Rate of Speech appears to be the most promising feature for predicting human assessments of proficiency.
- Posterior Log Likelihood scores show little correlation with pronunciation ratings.
  - Where Posterior Log Likelihood scores are employed, they are best calculated based only on speech phones in the context of other speech phones.
- For proficient L2 speakers, repeating prompts spoken by the system appears to be a better test of oral proficiency than reading prompts from a test sheet.
- Reading Task must be more challenging to be useful for assessing our proficient speaker population.