Automatically Assessing the Oral Proficiency of Proficient L2 Speakers

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Pieter Müller[1], Febe de Wet[2,4], Christa van der Walt[3] & Thomas Niesler[1]

[1]Department of Electrical and Electronic Engineering, [2]Centre for Language and Speech Technology (SU-CLaST),
[3]Department of Curriculum Studies, Stellenbosch University, South Africa. [4]HLT Research Group, CSIR Meraka Institute, South Africa.

SU-CLaST Centre for Language and Speech Technology

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
  - Reading / writing ability not necessarily well correlated with oral ability

AIM: Develop automatic system for the large scale assessment of oral language proficiency

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 assessed by human raters - *Human Ratings*
- 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 taken over the telephone
  - Guided by spoken dialogue system
  - Calls made from dedicated phone in quiet surroundings
- Test taken by 120 students
  - Test set of 90 students
  - Development set of 30 students
- Test consisted of 7 tasks. We focus on two of these:

READING TASK

Subjects read sentences from a provided test sheet

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

REPEATING TASK

Subjects repeat sentences spoken by the system

EXAMPLE:
"Lecturers 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

Likert scale used to assess Reading Task:

Pronunciation

Educated SAE, accent barely discernable.	....	Accent clear but comprenensible	....	Some words/sounds mispronounced, distracting to listener.	....	Mispronunciation affects comprehension.

Likert scales used to assess Repeating Task:

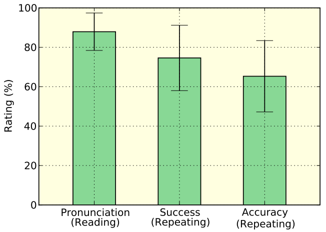
Success

Starts and completes repetition.	Some hesitation but then completes.	Starts but then gets into trouble.	A few words and then peters out.	Starts and then aborts attempt.	No start or attempt to repeat.

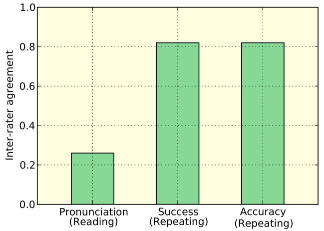
Accuracy

Correct repetition	Correct interpretation	Partially correct repetition or interpretation	Partially correct repetition of phrases	Repetition of some words but no coherence	No start or attempt to repeat.

Ratings awarded by raters



Agreement among raters



6. Automatic Assessment

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

Calculation of Automatic Indicators

Rate of Speech:

$$ROS = \frac{N_{SP}}{T_{Total}}$$

Accuracy:

$$Acc = \frac{N_C - N_I}{N_P}$$

Posterior Log Likelihood:

$$GOP(q_i) = \frac{|log(P(q_i|O_i))|}{N_F(O_i)}$$

Utterance Level Variants:
- GOP - All phones
- GOP_{SP} - Speech phones
- GOP_{SPC} - Speech context phones
- GOP_W - Normalised on word level

N_{SP} - Number of Speech Phones
T_{Total} - Total Duration
N_C - Number of Correct Phones
N_I - Number of Insertions
N_P - Number of Phones
q_i - i^{th} Phone
O_i - i^{th} Acoustic Segment
N_F - Number of Frames

Correlation between *Machine Scores* and *Human Ratings*

	Reading Task	Repeating Task	
	Pronunciation	Success	Accuracy
ROS	−0.46	−0.67	−0.65
Accuracy	–	−0.61	−0.63
GOP	0.02	0.39	0.35
GOP_{SP}	0.00	0.42	0.39
GOP_{SPC}	−0.02	0.45	0.41
GOP_W	−0.14	0.31	0.25

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