Automatic Large–Scale Oral Language Proficiency Assessment

Febe de Wet[†], Christa van der Walt[‡] and Thomas Niesler[*]

[†]Centre for Language and Speech Technology (SU–CLaST) [‡]Department of Curriculum Studies

[*]Department of Electronic Engineering, University of Stellenbosch, South Africa

UNIVERSITEIT STELLENBOSCH UNIVERSITY

SU-CLaST Centre for Language and Speech Technology

1. Background

- Oral proficiency tests are an important aspect of language skill assessment
  - Listening skills
  - Speaking skills
- Human assessment is the usual approach
  - Highly labour–intensive
  - Very subjective
- Reading and writing skills can be computerised
  - Fast
  - Reduced manpower
  - Objective and transparent
- However, good results in written tests are not necessarily reliable predictors of corresponding performance in oral tests

AIM Develop automatic system for the assessment of oral language proficiency

2. Context

- Students at Stellenbosch University's Education Faculty must enrol in a language module appropriate to their level of proficiency
- Progress must be monitored regularly thereafter
- 100–200 students per staff member: human assessment is impractical

3. Test design

Reading task

Subjects read a sentence from a provided test sheet

EXAMPLE:

"School governing boards struggle to make ends meet."

Repeat task

Subjects repeat a sentence spoken by the system

EXAMPLE:

"Student teachers don't get enough exposure to teaching practice."

Open–ended task

Subjects respond spontaneously to a general question

EXAMPLE:

"What is your biggest fear when entering a classroom?"

4. Test administration

- 106 students completed test in the 1st semester 2006
- Spoken dialogue system guides students through test and captures replies for subsequent assessment
- Calls made from a dedicated telephone in quiet surroundings
- English mother–tongue speakers generally found test manageable; Afrikaans–speaking students found it challenging
- All 106 tests transcribed orthographically by human experts
- Reserve 16 as development data, use remaining 90 as test set
- Test set is assessed automatically as well as by human raters

5. Human assessment

- 5 human raters, each teaches English as a foreign language
- Each student was assessed by at least 2 raters
- Each rater assessed 50 tests, 5 of which were repetitions to test for intra–rater consistency

5–point Likert scale as used by human raters

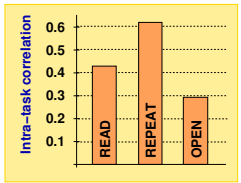
Task	Score	Assessment criteria
Read	5	Pronunciation, intonation and rhythm almost mother–tongue
	1	Speech difficult to understand and poorly articulated
Repeat	5	Repetition accurate and prompt
	1	No attempt was made to repeat
Open	5	Confident and fluent reply
	1	Only a feeble attempt was made to formulate a reply

Consistency of raters

Human rater	Intra–rater correlation
1	0.32
2	0.74
3	0.30
4	0.73
5	0.40

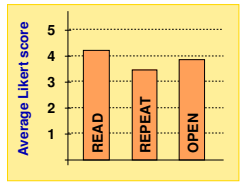
- Only raters 2 and 4 show fair consistency

Agreement among raters



- Raters' assessments agree most for repeat task

Marks awarded by raters



- Fairly high marks awarded
- Range of scale not evenly exploited

6. Automatic assessment

- ASR system uses speaker–independent cross–word triphone HMMs with 8 mixtures/state trained on approximately 6h of telephone speech
- Rate of speech (ROS) was taken as a measure of fluency:

$$\text{ROS} = \frac{N_p}{T_{sp}}$$

N_p: speech phones / utterance

T_{sp}: duration of utterance (including pauses)

- **Reading task:** Finite–state grammar allowing correct utterance with pauses and speaker noises between words
- **Repeat task:** Zerogram LM (one per utterance) using words in manual transcription of development data
- **Open–ended task:** Single zerogram LM obtained by pooling development transcriptions

Predicting human assessments using estimated ROS

Task	Average ROS	Correlation between est. ROS and true ROS	Correlation between est. ROS and human scores
Read	6.0	0.98	0.52
Repeat	5.0	0.94	0.58
Open	4.8	0.86	0.48

7. Conclusions

- Human raters are surprisingly inconsistent
- Correlation between automatic scores and human assessments is not particularly high, but compares well with other published figures
- Rate of speech (ROS) is promising as a ***consistent*** measure of fluency
- Other automatic measures are being investigated

Presented at: Interspeech 2007, 27–31 August 2007, Antwerp, Belgium.