Prompt-based Content Scoring for Automated Spoken Language Assessment

Keelan Evanini[1], Shasha Xie[2], and Klaus Zechner[1]
[1]Educational Testing Service, [2]Microsoft

Listening. Learning. Leading.®

1. Introduction

- Spoken language proficiency tasks eliciting *spontaneous speech* are necessary for valid scores
- Automated scoring systems have focused on pronunciation, prosody, and fluency
- *Prompt-based* materials can be used to score content

2. Prompt-based Materials

- **Listening passage (L):** recorded lecture or dialogue containing information relevant to the test question
- **Reading passage (R):** article or essay containing additional information relevant to the test question
- **Sample response (S):** sample response provided by the test designers containing the main ideas expected in a model answer

3. Baseline Content Features

- Sim_i: the similarity score between the words in the spoken response and a content model trained from responses receiving score *i*
- Cosine similarity (CVA) and PMI used
- Expensive to train (requires pre-scored responses)

4. Prompt-based Content Features

Feature	Description
stimulus_cosine	cosine similarity between the spoken response and each of the stimulus materials
token_overlap *type_overlap*	the number of lexical tokens / types that occur in both the spoken response and each of the stimulus materials (normalized by response length)
token_unique *type_unique*	the number of word tokens / types that occur in both the spoken response and one or two of the materials, but do not occur in the remaining material(s)

5. Data and Methodology

- Spoken responses to *TOEFL iBT®*
- 4 prompts, 60 seconds per response
- Scored by expert raters on 4-point scale
- Spoken responses processed using ASR system
- ASR output used to compute content features

6. Baseline Feature Correlations

Feature Set	Feature	*r*
CVA	Sim_1	0.091
	Sim_2	0.186
	Sim_3	0.261
	Sim_4	**0.311**
PMI	Sim_1	0.191
	Sim_2	0.261
	Sim_3	0.320
	Sim_4	**0.361**

7. Feature Correlations (Prompt-based)

Feature Set	Feature	*r*
stimulus_cosine	**L**	**0.384**
	R	0.176
	S	**0.384**
token_overlap	L	0.022
	R	0.096
	S	**0.121**
type_overlap	**L**	**0.426**
	R	0.142
	S	0.128

Feature Set	Feature	*r*
token_unique	L'RS	0.116
	L'RS'	0.162
	LR'S	0.219
	LR'S'	**0.337**
type_unique	L'RS	0.140
	L'RS'	0.166
	LR'S	0.259
	LR'S'	**0.450**

8. Score Prediction (Methodology)

- 9 baseline speaking proficiency features extracted using *SpeechRater*^{SM}
- linear regression models: 794 training, 395 evaluation.

Feature Category	Features
Fluency	normalized number of silences > 0.15 sec, normalized number of silences > 0.495 sec, average chunk length, speaking rate, normalized number of disfluencies
Pronunciation	normalized Acoustic Model score from forced alignment using a native speaker AM, average normalized phone duration difference compared to a reference corpus
Prosody	mean deviation of distance between stressed syllables
Grammar	Language Model score

9. Score Prediction (Results)

Feature Set	response *r* (N=395)	speaker *r* (N=97)
Baseline proficiency features	0.607	0.687
+ *type_overlap*	0.612	0.701
+ *token_overlap*	0.615	0.700
+ *token_unique*	0.616	0.695
+ *stimulus_cosine*	0.630	0.716
+ *type_unique*	0.658	0.761
+ CVA	0.665	0.762
+ all prompt-based	0.677	0.779
+ PMI	0.723	0.818
+ CVA and PMI	0.723	0.818
+ all content	0.742	0.838

10. Summary

- Prompt-based content scoring is a viable alternative to approach using pre-scored responses
- Improvement in correlation over baseline proficiency features
- Features measuring overlap with listening materials perform best