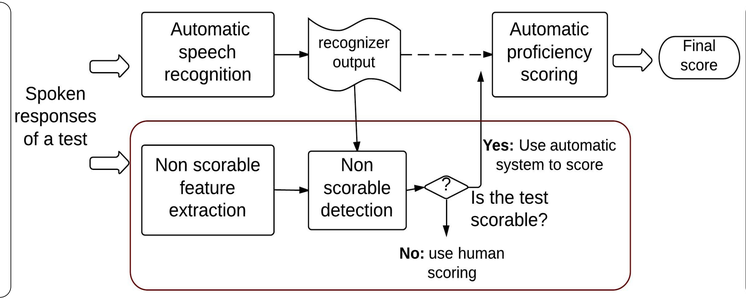
Syllable and language model based features for detecting non-scorable tests in spoken language proficiency assessment applications

Angeliki Metallinou (angeliki.metallinou@pearson.com), Jian Cheng (jian.cheng@pearson.com)
Knowledge Technologies, Pearson, Menlo Park, California

Introduction

- **Non-scorable test:** Can't be reliably scored automatically
  - Noisy, unintelligible, non-English, off-topic, etc.
- Propose new features for non-scorable detection
  - Exploit similarities between different information sources
- Achieve 21% rel. performance increase
  - When combining our features with existing ones



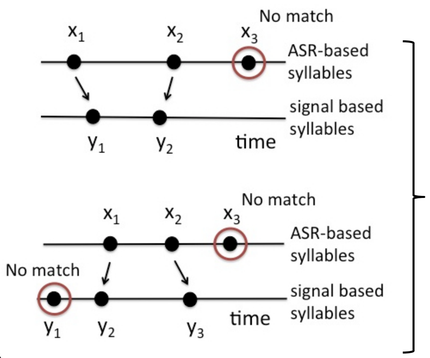
Data

- Automatic proficiency Assessment scoring of K-12 students
  - Test contains both repeat and open-ended tasks
- 6000 spoken tests: 4800 train, 1200 test
  - Tests double graded by professionals (scale: 0-14 points)
- **Define non-scorable test**: | human - machine grade | > 3 points
  - 308 tests are non-scorable (~5%)

Proposed Features

Syllable Based

- Human and machine scores often differ because of ASR errors
- How can we detect such cases?
  - Inconsistency between ASR info and pitch
- Estimate syllables using ASR result
  - Approx. number/location of vowels
- Estimate syllables from pitch/energy
- *If estimates don't match, the test may be unscorable*



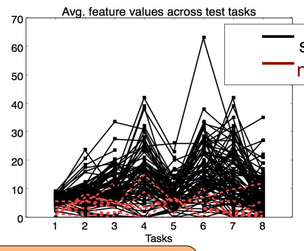
From the two estimates, we extract various *similarity based features*
- Sequence length difference
- Sequence lengths
- Number of syllable pairs
- Unpaired syllables
- Avg., max, min distance of pairs, etc.

Language Model Based

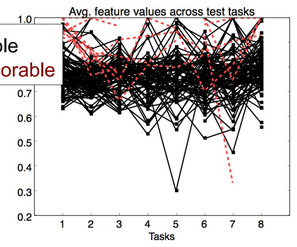
- Explore different LMs
- Task constrained word bigram LM
- Proposed task independent phone bigram LM
  - Can handle off-topic or non English words
- Estimate ASR similarity using edit distance
  - *Dissimilarity may indicate non-scorable*

Phoneme level similarity features based on edit dist.
- Edit distance (normalized)
- Difference of insertions, deletions, substitutions
- Sequence lengths, and difference, etc.

Number of syllable pairs across tasks of a test



Edit distance (normalized) across tasks of a test



Confidence Based

- Use proposed phone bigram LM to extract ASR confidence scores
  - 3 variants of confidence score *
  - recognition log-likelihoods
- Similar features also extracted from word LM

* described in [Cheng and Shen, 2011]

Combine response-level features at test level
- Average features separately over repeat, open-ended tasks
  - Features may slightly vary between the two tasks
- For responses with undefined features:
  - Include percentage of responses where feature is defined

Existing Features

- We extract state-of-the-art features for non-scorable and off-topic detection

Summary of 'Base' feature set

Feature Type	Description
Signal derived	Max and min energy, nonzero pitch frames, avg. pitch, SNR
ASR derived	Number of spoken words, pauses and hesitations, utterance durations, speech rate, avg. interword pause duration, leading pause duration
	ASR log-likelihood, avg. LM likelihood, phonemes pruned, word lattice confidence, perc. of low confidence words and phonemes
	Repeat: number of insertions, deletions, substitutions, perc. of recognized prompt words Open-ended: number of recognized key words
Indicator	Number of zero pitch frames >threshold, while ASR recognizes silence

Experiments and Results

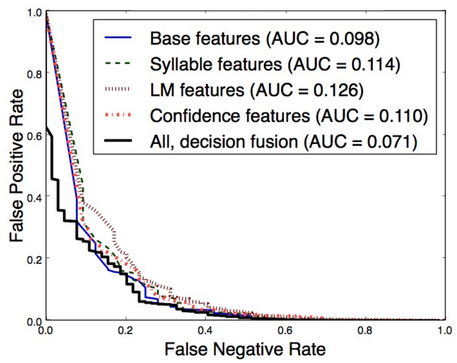
Experimental Setup

- Random forest classifiers for non-scorable detection
  - using different feature sets
- Estimate the ROC curves
- Minimize Area Under Curve (AUC)
- 5-fold cross validation over all 6000 tests
- Repeat experiment 10 times
  - report avg. and std. dev of AUC over the 10 runs

Detection Results

features	AUC (avg ± std.dev)
Base	0.102 ± 0.007
Syllable	0.122 ± 0.011
LM	0.123 ± 0.008
Confidence	0.106 ± 0.011
Classifier Decision Combination	
Base+Syllable	0.087 ± 0.008
Base+LM	0.085 ± 0.007
Base+ Confidence	0.084 ± 0.007
All	**0.081 ± 0.006**

ROC curves



Top 10 selected features

Feature set	Description	
Syllable	diff_length_nrm (avg,r)	diff_length_nrm (av,op)
	min_pair_dist (avg,op)	diff_length (avg,r)
	n_pairs_nrm (avg, op)	diff_length (avg,op)
	avg_pair_dist (avg,r)	min_pair_dist (avg,r)
	n_pairs_nrm (avg,r)	max_pair_dist (avg,op)
LM	edit_dist_nrm (avg,r)	diff_length_nrm (avg,r)
	n_insert_nrm (avg,r)	edit_dist_nrm (avg,op)
	diff_length_nrm(avg,op)	n_insert_nrm (avg,op)
	n_subst_nrm (avg,op)	min_length (avg,op)
	min_length (avg,r)	n_subst (avg,op)
Confidence	avg_conf_pLM(avg,op)	min_lglik_pLM (avg,r)
	min_lglik_pLM (avg,op)	max_lconf_pLM (avg,r)
	min_conf_pLM (avg,op)	std_lglik_pLM (avg,op)
	min_lglik_wLM (avg,r)	min_conf_pLM (avg,r)
	std_lglik_pLM (avg,r)	avg_lglik_pLM (avg,r)

Conclusions

- Proposed syllable and LM-based features for non-scorable detection
  - Estimate syllable locations
  - Propose task-independent phone LM
- Features lead to improvement in AUC when combined with existing ones
  - 0.102 → 0.081 (21 % rel. reduction)
- Our final system combines 4 random forest classifiers
  - one using existing features
  - three using the proposed features