Machine Learning of Level and Progression in Second/Additional Language Spoken English

Kate Knill
Speech Research Group, Machine Intelligence Lab
Cambridge University Engineering Dept

11 May 2016
Cambridge ALTA Institute

- Virtual institute at University of Cambridge
- Computing, Linguistics, Engineering, Language Assessment
- Sponsorship from Cambridge English Language Assessment
- Work presented was done at CUED – thanks to:
  - Mark Gales, Rogier van Dalen, Kostas Kyriakopoulos, Andrey Malinin, Mohammad Rashid, Yu Wang
Spoken Communication

- Message Construction
- Pronunciation
- Prosody
- Speaker Characteristics
- Environment/Channel
- Message Realisation
- Message Reception
Spoken Communication

- Spoken communication is a very rich communication medium.
Spoken Communication Requirements

• Message Construction should consider:
  • Has the speaker generated a coherent message to convey?
  • Is the message appropriate in the context?
  • Is the word sequence appropriate for the message?
Spoken Communication Requirements

message construction should consider:
- Has the speaker generated a coherent message to convey?
- Is the message appropriate in the context?
- Is the word sequence appropriate for the message?

message realisation should consider:
- Is the pronunciation of the words correct/appropriate?
- Is the prosody appropriate for the message?
- Is the prosody appropriate for the environment?
Spoken Communication Requirements

• Message Construction should consider:
  • Has the speaker generated a coherent message to convey?
  • Is the message appropriate in the context?
  • Is the word sequence appropriate for the message?

• Message Realisation should consider:
  • Is the pronunciation of the words correct/appropriate?
  • Is the prosody appropriate for the message?
  • Is the prosody appropriate for the environment?
okay carl uh do you exercise yeah actually um i belong to a gym down here gold’s gym and uh i try to exercise five days a week um and now and then i’ll get it interrupted by work or just full of crazy hours you know
okay carl uh do you exercise yeah actually um i belong to a gym down here gold’s gym and uh i try to exercise five days a week um and now and then i’ll get it interrupted by work or just full of crazy hours you know
okay carl uh do you exercise yeah actually um i belong to a gym down here gold’s gym and uh i try to exercise five days a week um and now and then i’ll get it interrupted by work or just full of crazy hours you know

Meta-Data Extraction Markup

Speaker1: / okay carl {F uh} do you exercise /  
Speaker2: / {DM yeah actually} {F um} i belong to a gym down here /  
/ gold’s gym / / and {F uh} i try to exercise five days a week {F um} /  
/ and now and then [REP i’ll + i’ll] get it interrupted by work or just full of crazy hours {DM you know} /  

Written Text

Speaker1: Okay Carl do you exercise?  
Speaker2: I belong to a gym down here, Gold’s Gym, and I try to exercise five days a week and now and then I’ll get it interrupted by work or just full of crazy hours.
Example of a test of communication skills

A. Introductory Questions: where you are from
B. Read Aloud: read specific sentences
C. Topic Discussion: discuss a company that you admire

D. Interpret and Discuss Chart/Slide: example above
E. Answer Topic Questions: 5 questions about organising a meeting
<table>
<thead>
<tr>
<th>Level</th>
<th>Global Descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td>Fully operational command of the spoken language</td>
</tr>
<tr>
<td>C1</td>
<td>Good operational command of the spoken language</td>
</tr>
<tr>
<td>B2</td>
<td>Generally effective command of the spoken language</td>
</tr>
<tr>
<td>B1</td>
<td>Limited but effective command of the spoken language</td>
</tr>
<tr>
<td>A2</td>
<td>Basic command of the spoken language</td>
</tr>
<tr>
<td>A1</td>
<td>Minimal command of the spoken language</td>
</tr>
</tbody>
</table>
Automated assessment of one speaker

Audio

Grade
Automated assessment of one speaker

Audio

Feature extraction

Features

Grader

Grade
Automated assessment of one speaker

1. Audio
2. Feature extraction
3. Speech recogniser
4. Text
5. Features
6. Grader
7. Grade
Speech Recognition Challenges

- Non-native ASR highly challenging
  - Heavily accented
  - Pronunciation dependent on L1
- Commercial systems poor!
- State-of-the-art CUED systems

<table>
<thead>
<tr>
<th>Training Data</th>
<th>Word error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native &amp; C-level non-native English</td>
<td>54%</td>
</tr>
<tr>
<td>BULATS speakers</td>
<td>30%</td>
</tr>
</tbody>
</table>
Automatic Speech Recognition Components

- Pronunciation Lexicon
- Recognition Engine
  - Acoustic Model
    - Acoustic Model training data
  - Language Model
    - Language Model training data
- "The cat sat on …"
Forms of Acoustic and Language Models

L2 audio data → L2 Acoustic Model

L2 text data + L1 text data → L2 Language Model

Used to recognise L2 speech
Forms of Acoustic and Language Models

L2 audio data → L2 Acoustic Model

L2 text data + L1 text data → L2 Language Model

Native (L1) audio data → Native Acoustic Model

Native (L1) text data → Native Language Model

Used to recognise L2 speech

Useful to extract features
Joint decoding - frame-level combination

\[ L(o_t \mid s_i) = \lambda_T L_T(o_t \mid s_i) + \lambda_H L_H(o_t \mid s_i) \]
Recognition Rate vs L1

- Acoustic models trained on English data from Gujarati L1 scored against crowd-sourced references
Outline

Audio

Speech recogniser

Feature extraction

Text

Features

Grader

Grade
Outline

Audio

Speech recogniser

Feature extraction

Text

Features

Grader

Grade
Baseline Features

• Mainly fluency based:

• Audio Features: statistics about
  • fundamental frequency (f0)
  • speech energy and duration

• Aligned Text Features: statistics about
  • silence durations
  • number of disfluencies (um, uh, etc)
  • speaking rate

• Text Identity Features:
  • number of repeated words (per word)
  • number of unique word identities (per word)
Pronunciation Features

• Hypothesis: poor speakers are weaker at making phonetic distinctions
  • less proficient – phone realisation closer to L2
  • more proficient – phone realisation closer to L1

• Statistical approach – learn phonetic distances from graded data
  • single multivariate Gaussian of K-L divergence per phoneme pair
  • 1081 phoneme pairs

\[ JSD(p_1(x), p_2(x)) = \frac{1}{2} \left[ KL(p_1(x) \parallel p_2(x)) + KL(p_2(x) \parallel p_1(x)) \right] \]

\[ KL(p_1(x) \parallel p_2(x)) = \frac{1}{2} \left( tr(\Sigma_2^{-1} \Sigma_1 - I) + (\mu_1 - \mu_2)^T \Sigma_2^{-1} (\mu_1 - \mu_2) + \log \left( \frac{\Sigma_2^{-1}}{\Sigma_1^{-1}} \right) \right) \]
Pronunciation Features vs Learner Progression

- Pattern of distances different between candidates of different levels
- Correlation with score: mis-pronounced phones higher K-L distance
  - opposite of expectation that poor speakers have more overlap

**Candidate Grade A1**

**Candidate Grade C2**
Statistical Parser Features

- Parser features from RASP system improve grades for written tests
- Problem: *speech recognition accuracy*

<table>
<thead>
<tr>
<th>Correct (DTAL)</th>
<th>Speech recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>advocates for the supplier must be</em></td>
<td><em>advocate so the supplier must be</em></td>
</tr>
</tbody>
</table>

- Smaller subtrees and leaves are fairly robust
Outline

Audio → Speech recogniser

Feature extraction

Features

Grader

Text

Grade
Outline

Audio

Speech recogniser

Feature extraction

Text

Features

Grader

Grade
Uses of Automatic Assessment

• Human graders
  ✔ very powerful ability to assess spoken language
  ✖ vary in quality and not always available

• Automatic graders
  ✔ more consistent and potentially always available
  ✖ validity of the grade varies and limited information about context
Uses of Automatic Assessment

• Human graders
  ✓ very powerful ability to assess spoken language
  ✖ vary in quality and not always available

• Automatic graders
  ✓ more consistent and potentially always available
  ✖ validity of the grade varies and limited information about context

• Use automatic grader
  • for grading practice tests/learning process
  • in combination with human graders
    • combination: use both grades
    • back-off process: detect challenging candidates
Currently have 1000s candidates to train grader
- limited data compared to ASR frames (100,000s frames)
- useful to have confidence in prediction

Gaussian Process is a natural choice for this configuration
Form of Output

<table>
<thead>
<tr>
<th>Graders</th>
<th>Pearson Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human experts</td>
<td>0.85</td>
</tr>
<tr>
<td>Automatic GP</td>
<td>0.83 – 0.86</td>
</tr>
</tbody>
</table>
## Effect of Grader Features

<table>
<thead>
<tr>
<th>Grader</th>
<th>Pearson Correlation with Expert Graders</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard examiners</td>
<td>0.85</td>
</tr>
<tr>
<td>Automatic baseline</td>
<td>0.83</td>
</tr>
<tr>
<td>+ Pronunciation</td>
<td>0.84</td>
</tr>
<tr>
<td>+ RASP</td>
<td>0.85</td>
</tr>
<tr>
<td>+ Confidence</td>
<td>0.83</td>
</tr>
<tr>
<td>+ RASP + Confidence</td>
<td>0.86</td>
</tr>
<tr>
<td>Pronunciation features</td>
<td>0.82</td>
</tr>
</tbody>
</table>
Combining Human and Automatic Graders

- Interpolate between human and automated grades
- higher correlation i.e. more reliable grade produced
- Content checking can be done by the human grader
Detecting Outlier Grades

- Standard (BULATS) graders handle standard speakers very well
- non-standard (outlier) speakers less well handled
- use Gaussian Process variance to automatically detect outliers

![Graph showing rejection rate vs correlation]

- Back-off to human experts - reject 10%: performance 0.83 → 0.88
Assessing Communication Level

- Ignore high-level content and communication skills currently

- Language complexity is related to proficiency
  - Future work – look into e.g.
    - McCarthy’s use of chunks “I would say”, “and then”
    - Abdulmajeed and Hunston’s “correctness analysis”
Assessing Content

- Grader correlates well with expert grades
  - features do not assess content – primarily fluency features

- Train a Recurrent Neural Network Language Model for each question
  - assess whether the response is consistent with example answers
## Topic Classification

### Experiment details
- 280-D LSA topic space
- Supervised (SUP): 490 speakers, 2x crowd-sourced transcriptions
- Semi-supervised (Semi-SUP): + 10005 speakers, ASR transcriptions

### Increasing quantity of data helps even though high %WER
- RNNLM can handle large data sets unlike K-Nearest Neighbour (KNN)

### System Comparison

<table>
<thead>
<tr>
<th>System</th>
<th>HL-dim</th>
<th>Training Data</th>
<th>% Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>-</td>
<td>SUP</td>
<td>20.8</td>
</tr>
<tr>
<td>RNNLM</td>
<td>100</td>
<td></td>
<td>17.5</td>
</tr>
<tr>
<td>RNNLM</td>
<td>200</td>
<td>Semi-SUP</td>
<td>9.3</td>
</tr>
</tbody>
</table>

---

[Table with System, HL-dim, Training Data, % Error columns]
Off-Topic Response Detection

- Synthesised pool of off-topic responses
  - Naïve – select incorrect response from any section
  - Directed – select incorrect response from same section
Spoken Language Assessment

- Automatically assess:
  - Message realisation
    - Fluency, pronunciation
  - Message construction
    - Construction & coherence of response
    - Relationship to topic
Spoken Language Assessment

- Automatically assess:
  - Message realisation
    - Fluency, pronunciation
  - Message construction
    - Construction & coherence of response
    - Relationship to topic

Achieved (with room for improvement)

Unsolved – active research areas
Spoken Language Assessment and Feedback

- Automatically assess:
  - Message realisation
    - Fluency, pronunciation
  - Message construction
    - Construction & coherence of response
    - Relationship to topic

- Provide feedback:
  - Feedback to user: realisation, construction
  - Feedback to system: adjust to level
Time Alignment and Pronunciation Feedback
Conclusions

- Automated machine-learning for spoken language assessment
  - important to keep costs down
  - able to be integrated into the learning process

- Current level – assessment of fluency
  - ongoing research into assessing communication skills:
    - appropriateness and acceptability

- Error detection and feedback is challenging
  - high precision required in detecting where errors have occurred
  - supplying feedback in appropriate form for learner
Questions?