Spoken Communication

Message Construction  Message Realisation  Message Reception

Speaker Characteristics
Environment/Channel

Pronunciation
Prosody
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 i’ ll get it interrupted by work or just full of crazy hours you know

Meta-Data Extraction (MDE) Markup

Speaker1: / okay carl {F uh} do you exercise /
Speaker2: / {DM yeah actually} {F um} i belong to a gym down here /
    {F uh} and 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 } /
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 (MDE) 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.
Business Language Testing Service (BULATS)
Spoken Tests

• 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
Automated Assessment of One Speaker

Audio

Grade
Automated Assessment of One Speaker

Audio

Feature extraction

Features

Grader

Grade
Automated Assessment of One Speaker

Audio

Speech recogniser

Feature extraction

Text

Features

Grader

Grade
Outline

Audio

Speech recogniser

Feature extraction

Text

Features

Grader

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

- Acoustic Model
  - Acoustic Model training data
- Language Model
  - Language Model training data
- Recognition Engine
  - Pronunciation Lexicon
  - “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

L2 audio data

Used to recognise L2 speech

Native (L1) audio data -> Native Acoustic Model

Native (L1) text data -> Native Language Model

Useful to extract features
Deep Learning for Speech Recognition

- Fusion of HMM deep neural network and Gaussian mixture models
- trained on BULATS data
Outline

Audio

Feature extraction

Features

Grader

Grade

Speech recogniser

Text
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)
Speaking Time Versus Learner Progression

Average Speaking Time (secs)

CEFR Grade

spontaneous speech
read speech
Pronunciation Features

- **Hypothesis:** poor speakers are weaker at making phonetic distinctions
- **Statistical approach** – learn phonetic distances from graded data
**Pronunciation Features**

- **Hypothesis:** poor speakers are weaker at making phonetic distinctions
- **Statistical approach** – learn phonetic distances from graded data

- Pattern of distances different between candidates of different levels
Outline

Audio

Speech recogniser

Feature extraction

Text

Features

Grader

Grade
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
Gaussian Process Grader

- 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>
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

\[
\begin{array}{c|c|c|c|c|c|c|c}
\text{Rejection rate (i.e., cost)} & 0 & 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 0.7 & 0.8 & 0.9 & 1 \\
\text{Correlation} & 0.85 & 0.9 & 0.95 & 1 & 0.85 & 0.9 & 0.95 & 1 & 0.85 & 0.9 & 0.95 & 1 \\
\end{array}
\]

- Back-off to human experts
- Reject 10%: performance 0.83 $\rightarrow$ 0.88
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
Spoken Language Assessment

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

Audio

Feature extraction

Speech recogniser

Text

Grader

Grade
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
Recognition Error Rate Versus Learner Progression

% Word Error Rate

CEFR Grade

- Read
- Spontaneous
- Overall
• Lightly supervised:
  • No pronunciation labelling required – trained just on grades
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
Thank You

• Acknowledgement: members of CUED MIL ALTA team:
  • Rogier van Dalen, Kostas Kyriakopoulos, Andrey Malinin, Yu Wang