UNIVERSITY OF CAMBRIDGE

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

Kate Knill

Speech Research Group, Machine Intelligence Lab

Cambridge University Engineering Dept

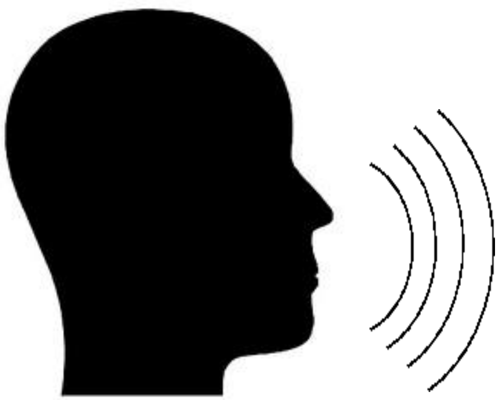
11 May 2016

Cambridge ALTA
Institute for Automated Language Teaching and Assessment

Cambridge ALTA Instititute

- 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

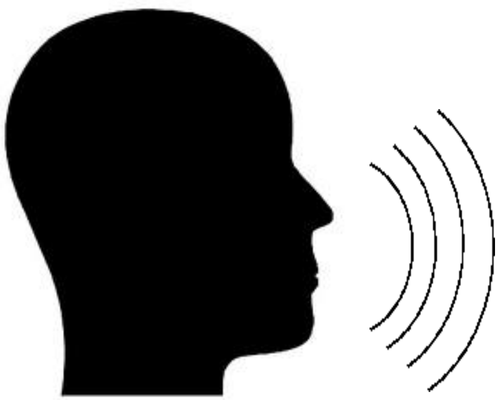


Speaker Characteristics
Environment/Channel

Pronunciation
Prosody

Message Construction | Message Realisation | Message Reception

Spoken Communication



Speaker Characteristics
Environment/Channel

Pronunciation
Prosody

Message Construction | Message Realisation | Message Reception

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?

Spoken Language Versus Written

ASR Output

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

Spoken Language Versus Written

ASR Output

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 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 } /

Spoken Language Versus Written

ASR Output

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 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

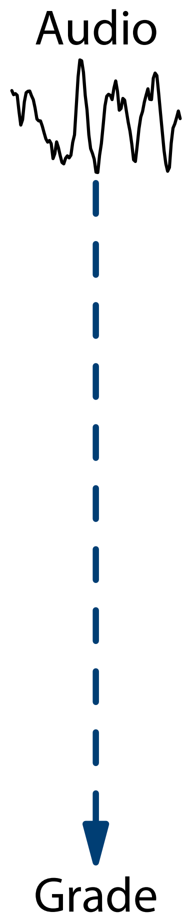
D. Interpret and Discuss Chart/Slide: example above

E. Answer Topic Questions: 5 questions about organising a meeting

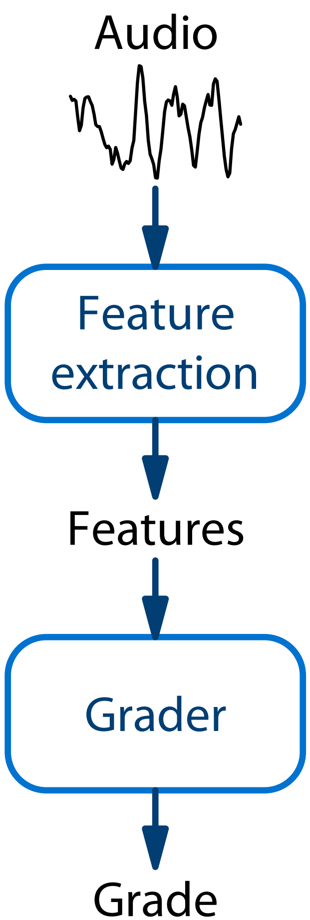
Common European Framework of Reference (CEFR)

Level	Global Descriptor
C2	Fully operational command of the spoken language
C1	Good operational command of the spoken language
B2	Generally effective command of the spoken language
B1	Limited but effective command of the spoken language
A2	Basic command of the spoken language
A1	Minimal command of the spoken language

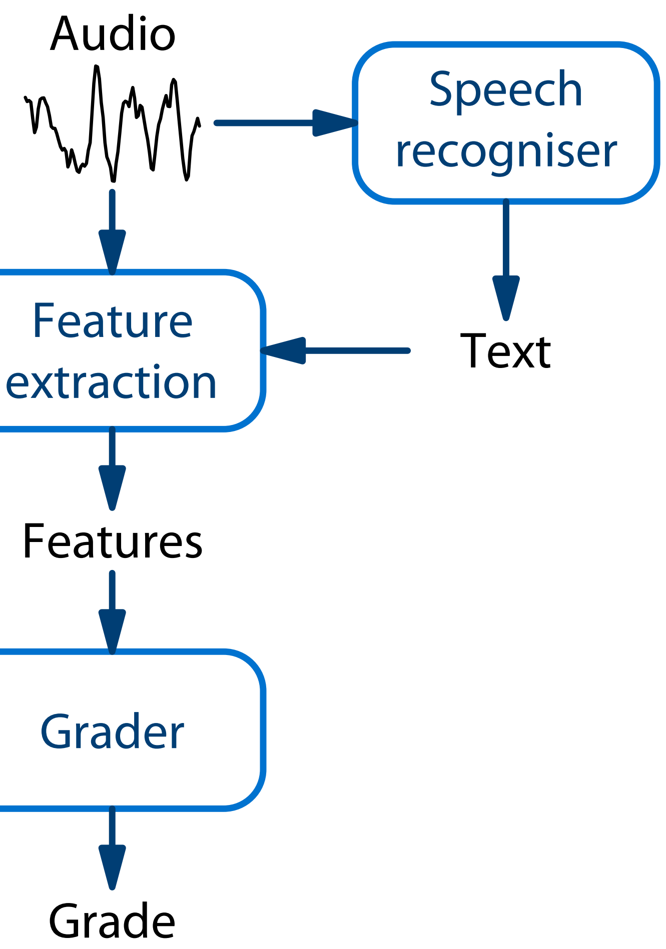
Automated assessment of one speaker



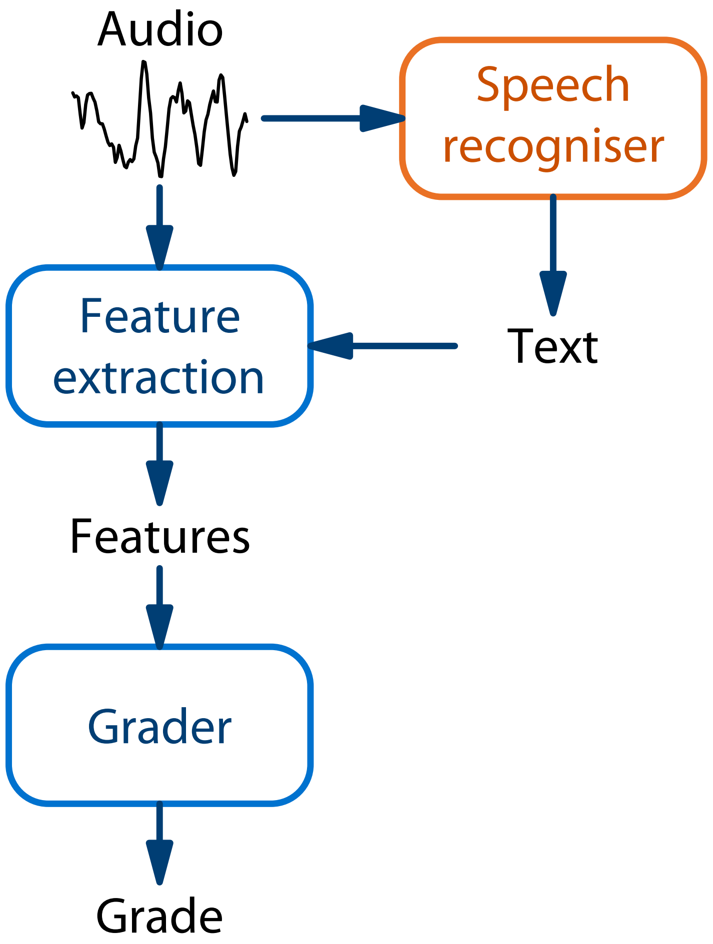
Automated assessment of one speaker



Automated assessment of one speaker



Outline



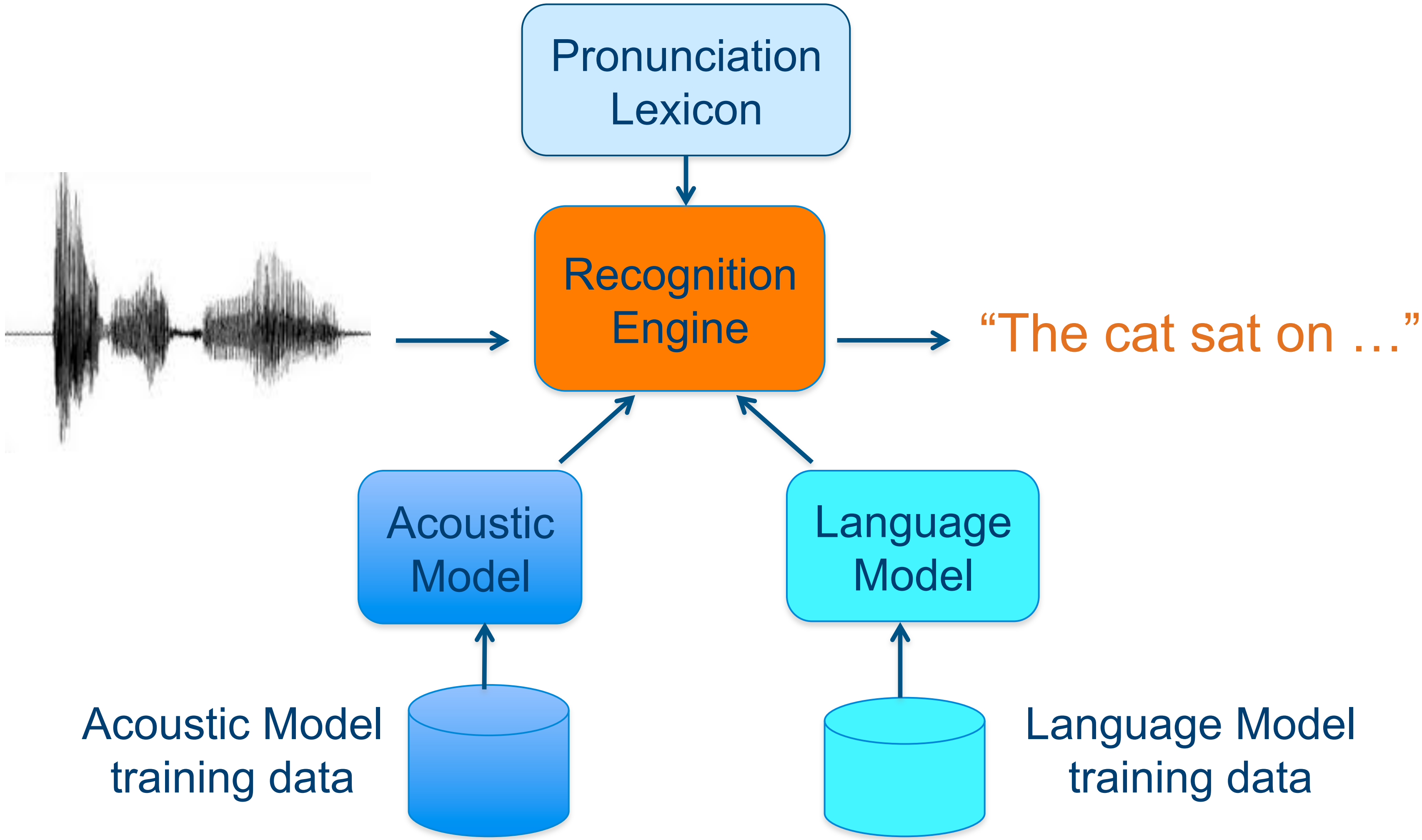
Audio
Speech recogniser
Feature extraction
Text
Features
Grader
Grade

Speech Recognition Challenges

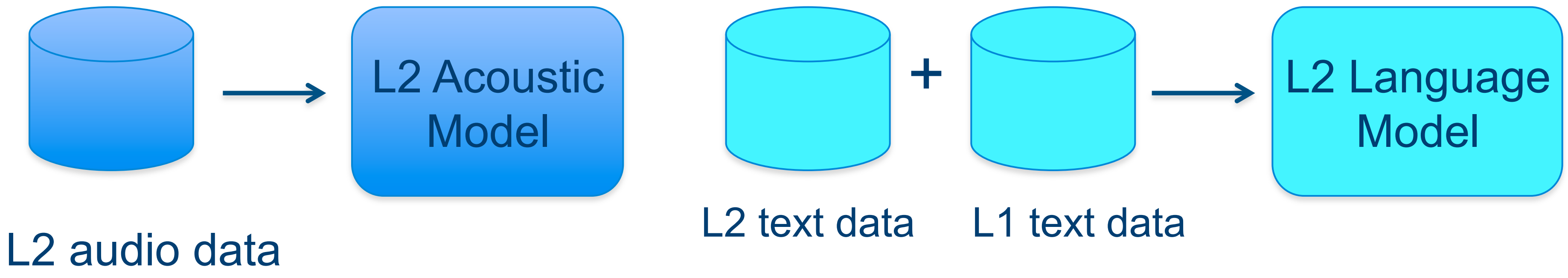
- Non-native ASR highly challenging
  - Heavily accented
  - Pronunciation dependent on L1
- Commercial systems poor!
- State-of-the-art CUED systems

Training Data	**Word error rate**
Native & C-level non-native English	54%
BULATS speakers	30%

Automatic Speech Recognition Components

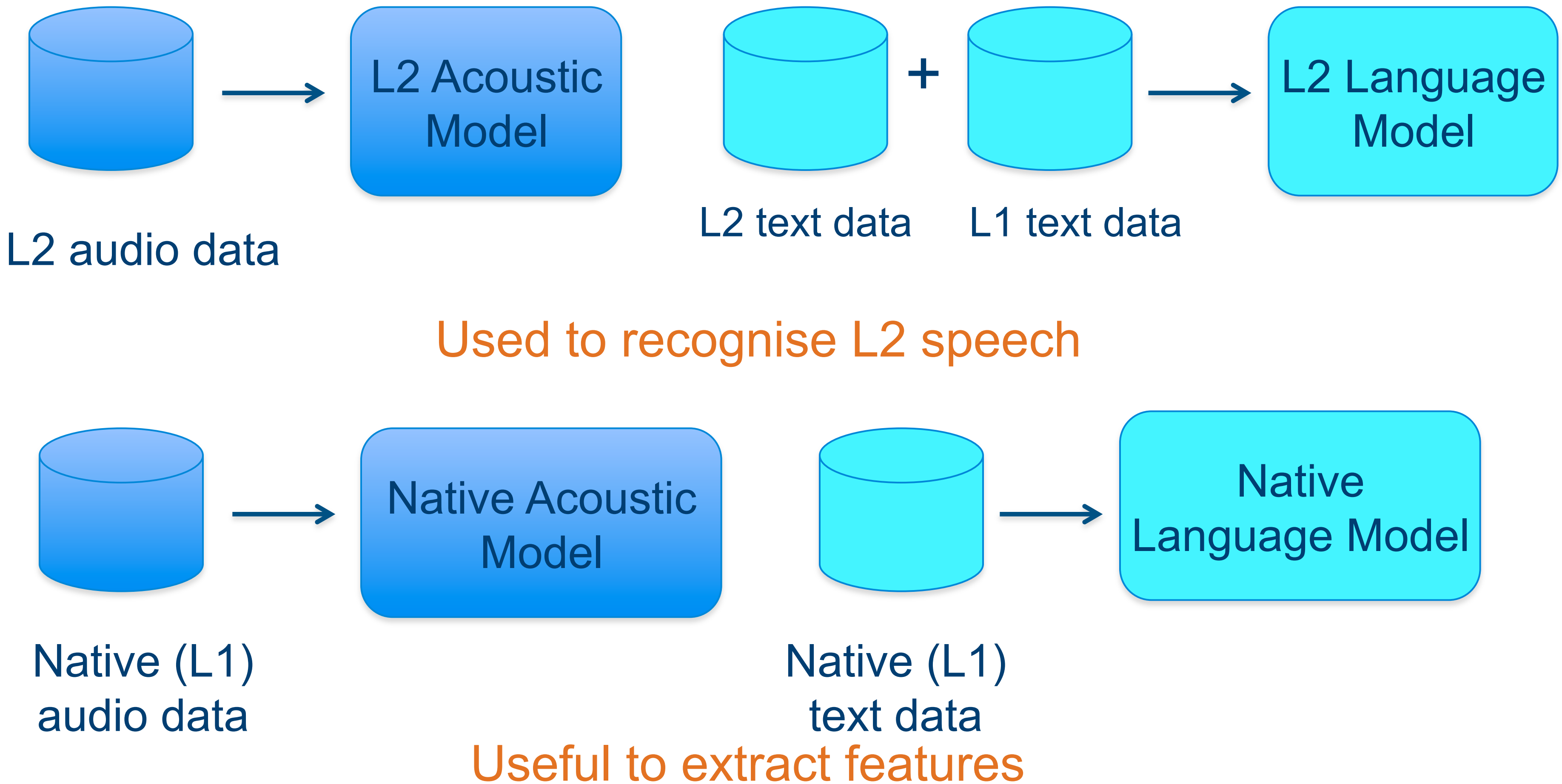


Forms of Acoustic and Language Models

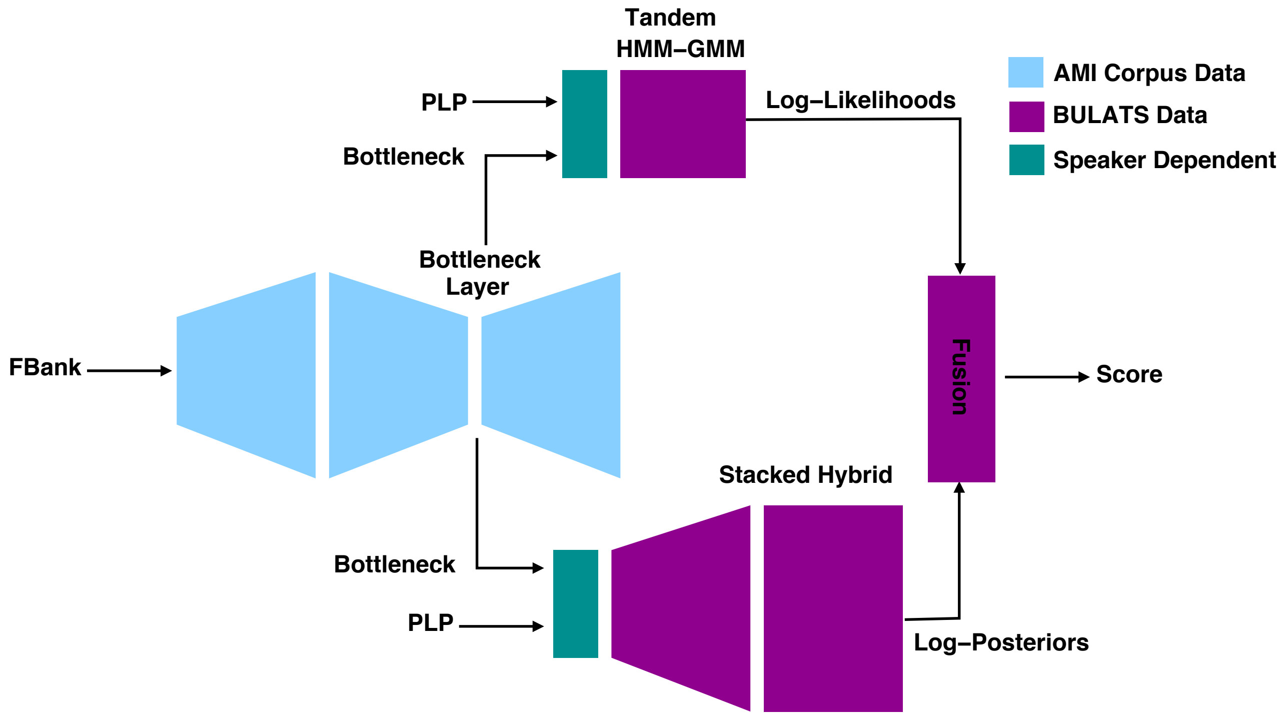


Used to recognise L2 speech

Forms of Acoustic and Language Models



Speech Recognition System

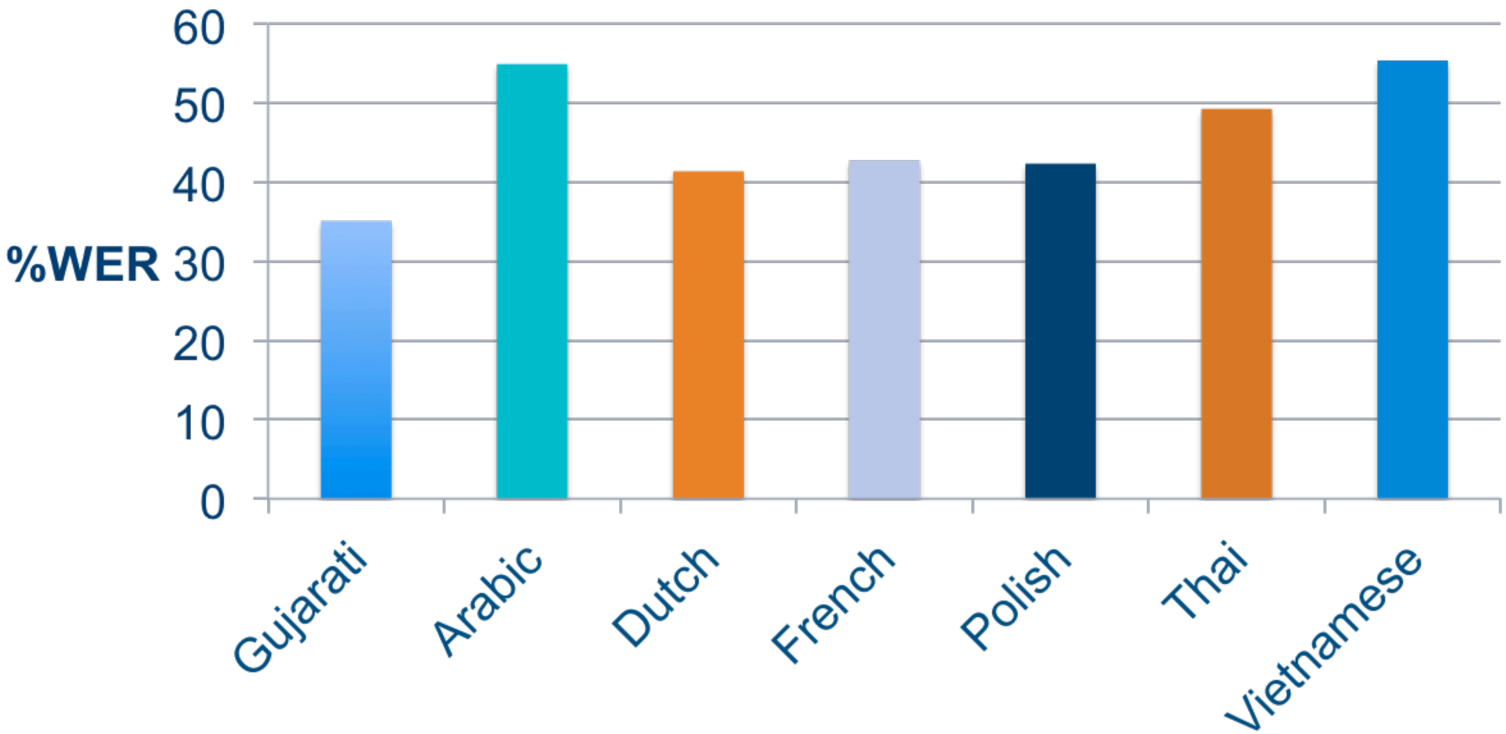


- Joint decoding - frame-level combination

$$L(o_t \mid s_i) = \lambda_T L_T(o_t \mid s_i) + \lambda_H L_H(o_t \mid s_i)$$

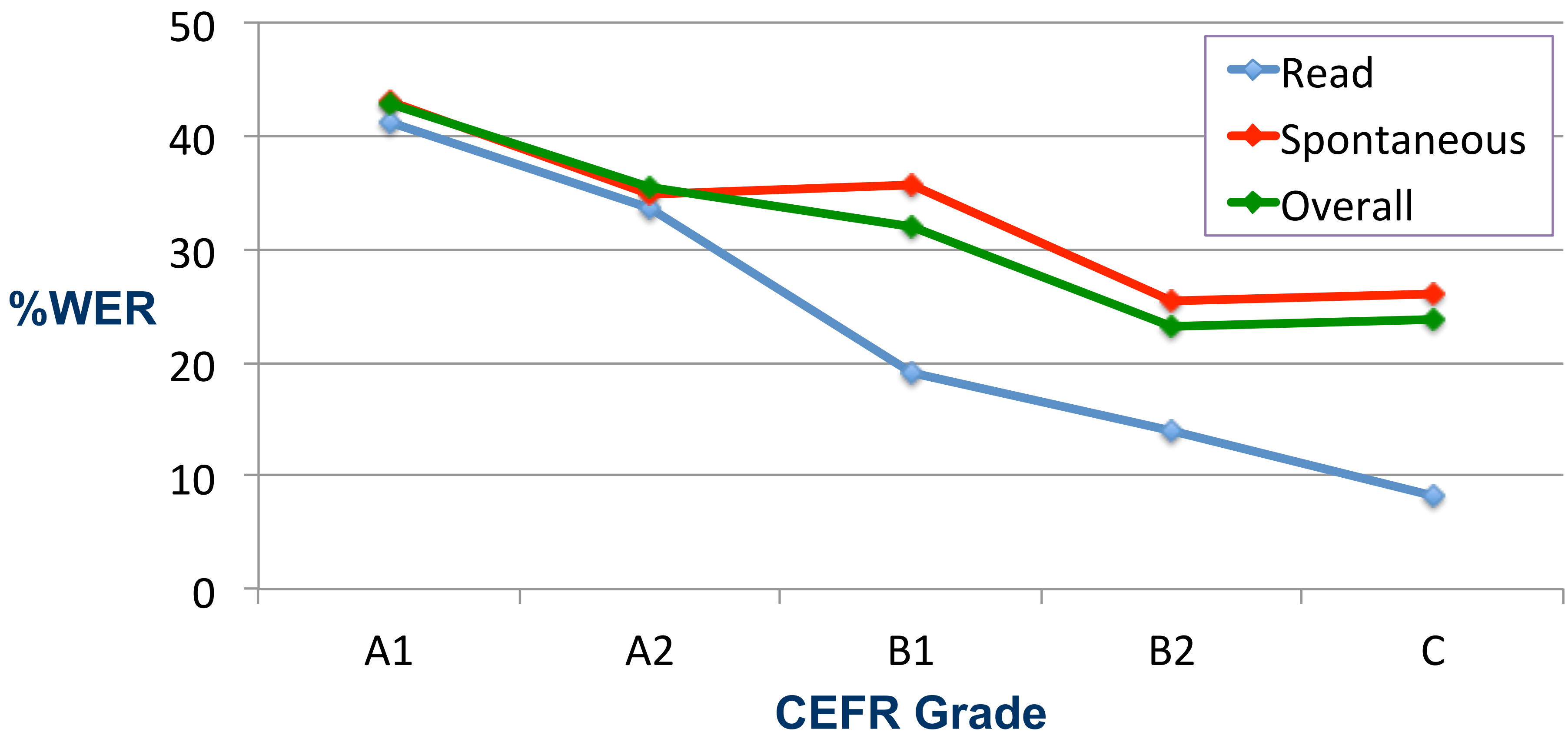
Recognition Rate vs L1

- Acoustic models trained on English data from Gujarati L1

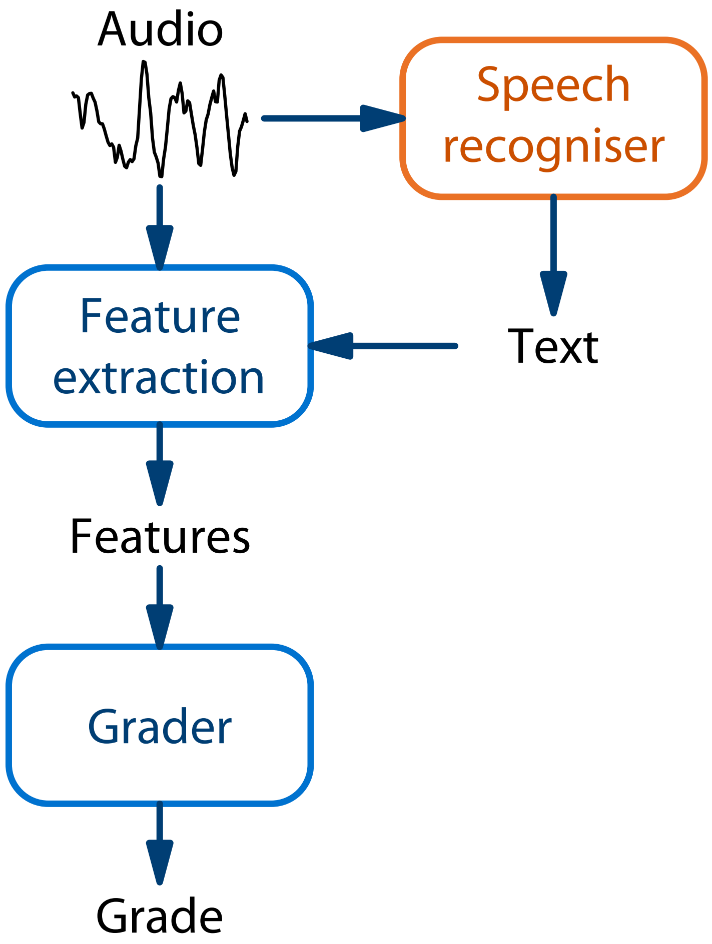


scored against crowd-sourced references

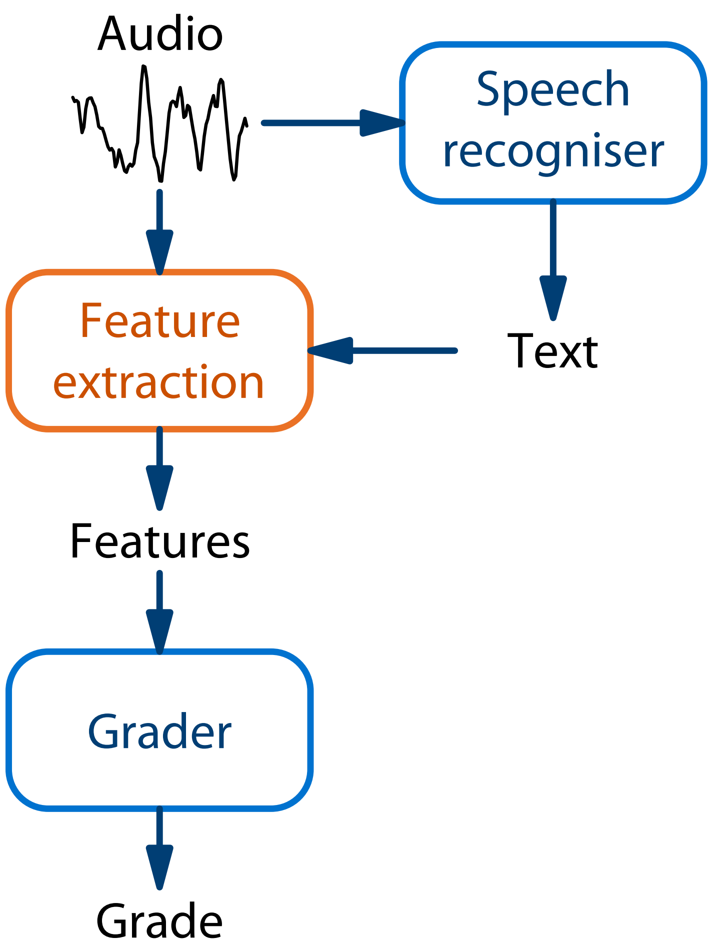
Recognition Error Rate vs Learner Progression



Outline


Audio
Speech recogniser
Feature extraction
Text
Features
Grader
Grade

Outline

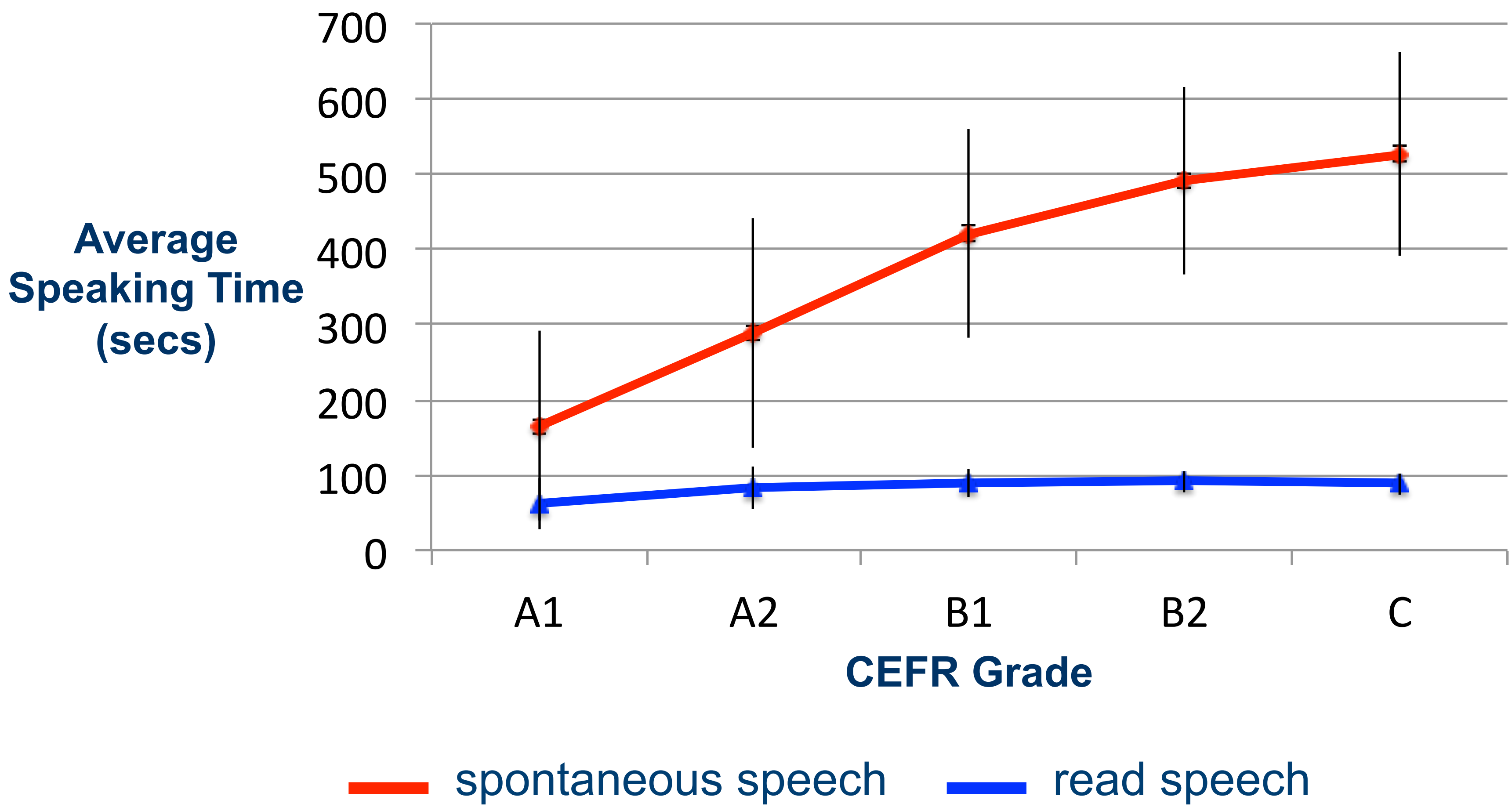


Audio
Speech recogniser
Text
Feature extraction
Features
Grader
Grade

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)

Speaking Time vs Learner Progression



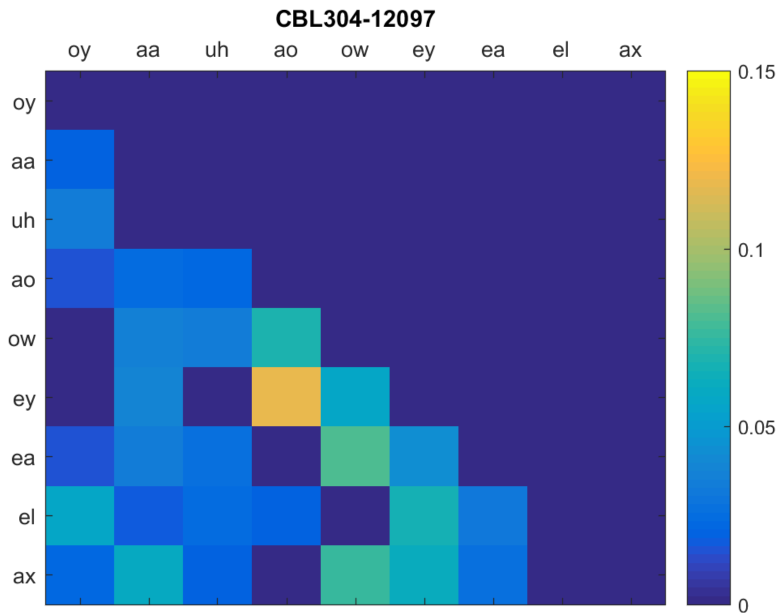
Pronunciation Features

- 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

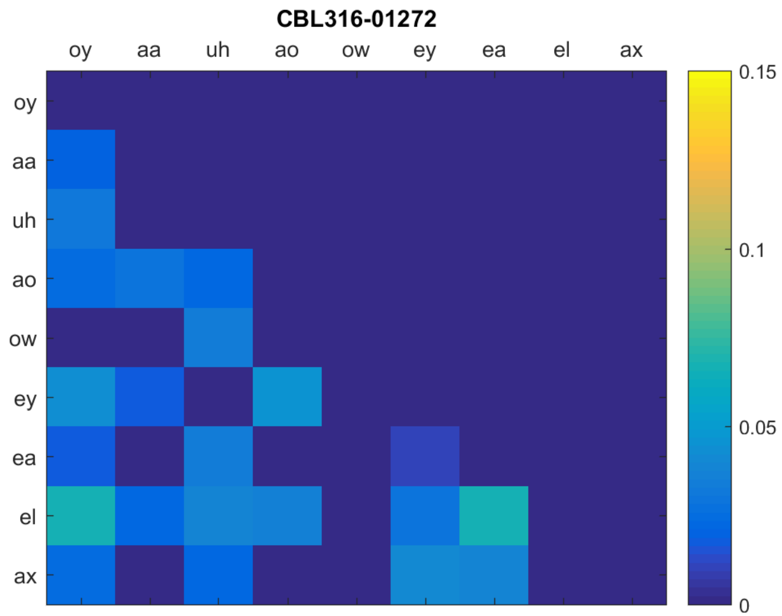
$$JSD(p_1(x), p_2(x)) = \frac{1}{2}\left[KL(p_1(x) \| p_2(x)) + KL(p_2(x) \| p_1(x)\right]$$

$$KL(p_1(x) \| p_2(x)) = \frac{1}{2}\left(tr(\Sigma_2^{-1}\Sigma_1 - \mathrm{I}) + (\mu_1 - \mu_2)^T \Sigma_2^{-1}\right)\left(\mu_1 - \mu_2\right) + \log\left(\frac{\left|\Sigma_2^{-1}\right|}{\left|\Sigma_1^{-1}\right|}\right)$$

Pronunciation Features vs Learner Progression



Candidate Grade A1



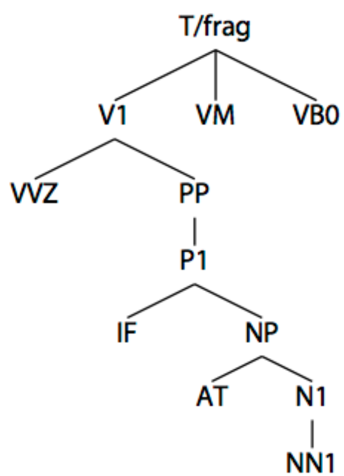
Candidate Grade C2

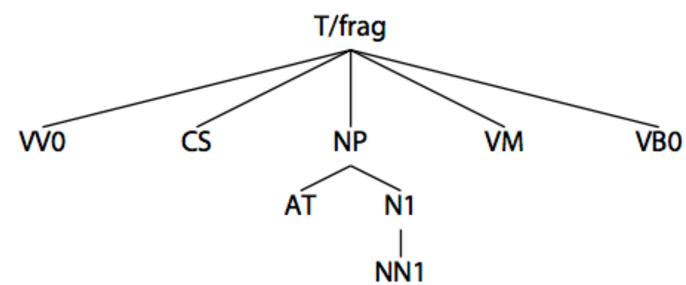
- Pattern of distances different between candidates of different levels
- Correlation with score: mis-pronounced phones higher K-L distance
    - opposite of expectation that poor speakers have more overlap

Statistical Parser Features

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

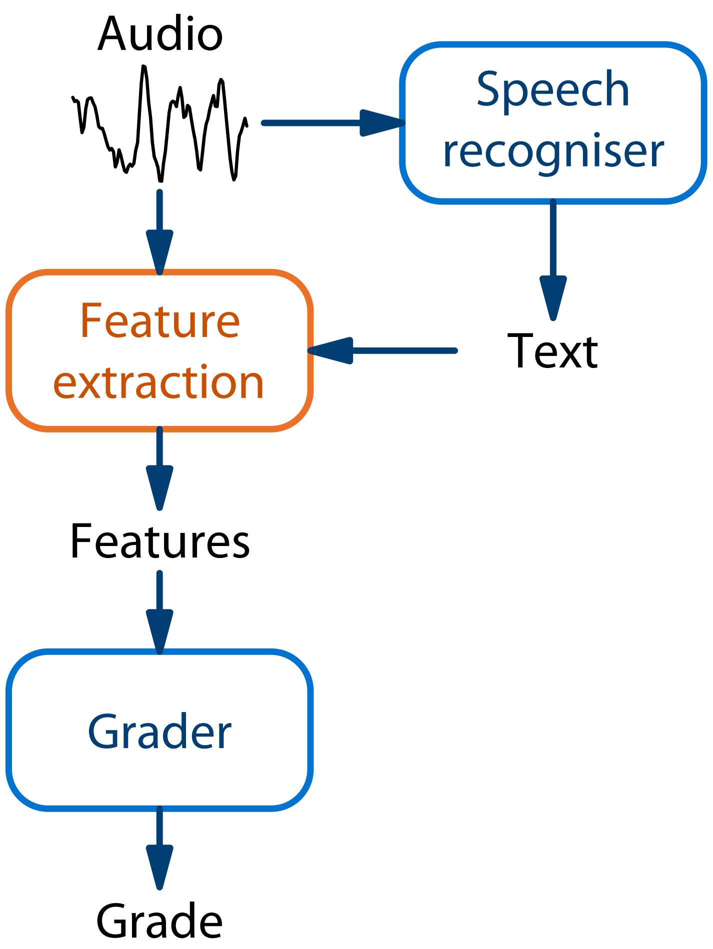
Correct (DTAL)	Speech recognition
advocates for the supplier must be	advocate so the supplier must be





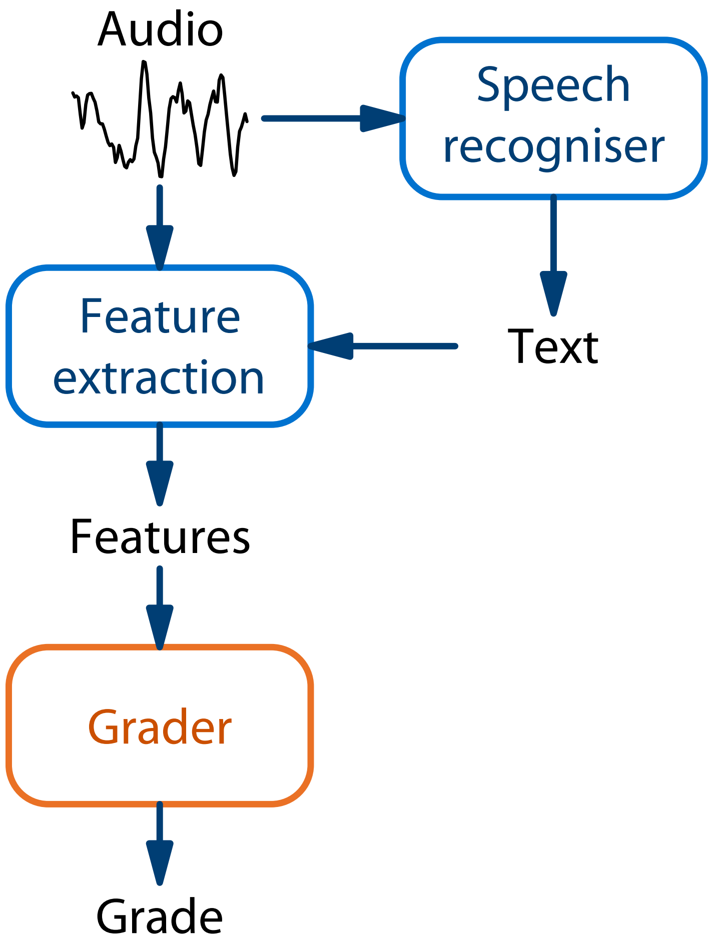
- Smaller subtrees and leaves are fairly robust

Outline



Audio
Speech recogniser
Feature extraction
Text
Features
Grader
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

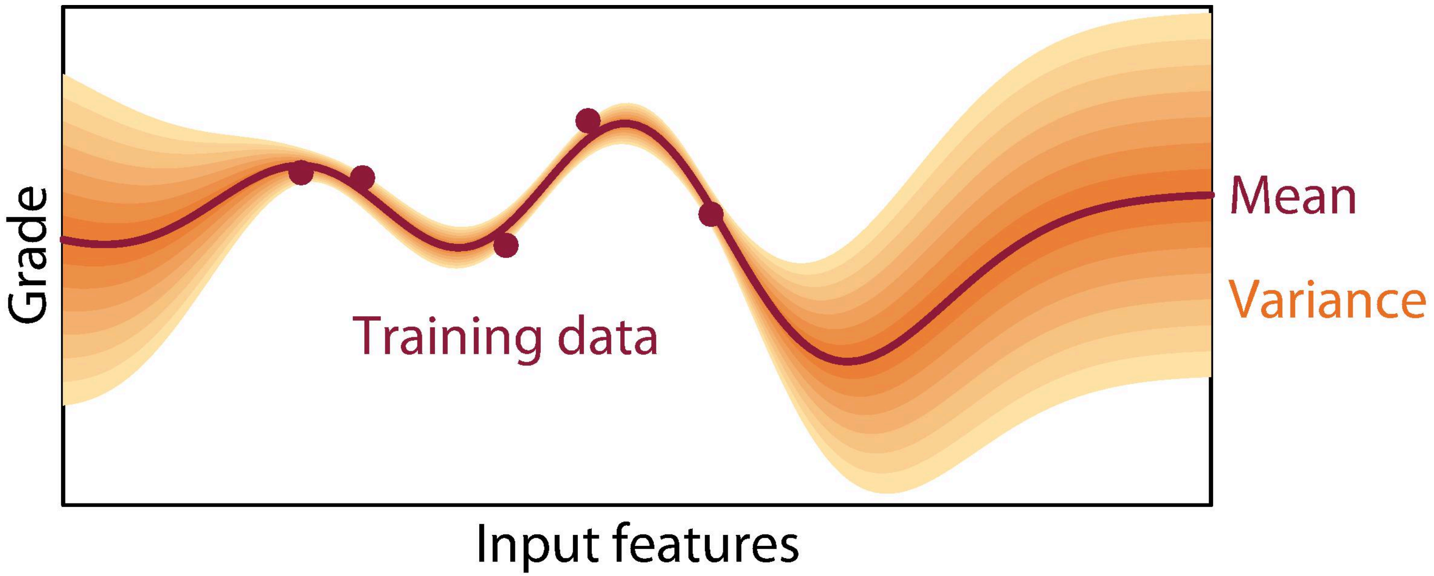
  ✔ more consistent and potentially always available

  ✖ 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

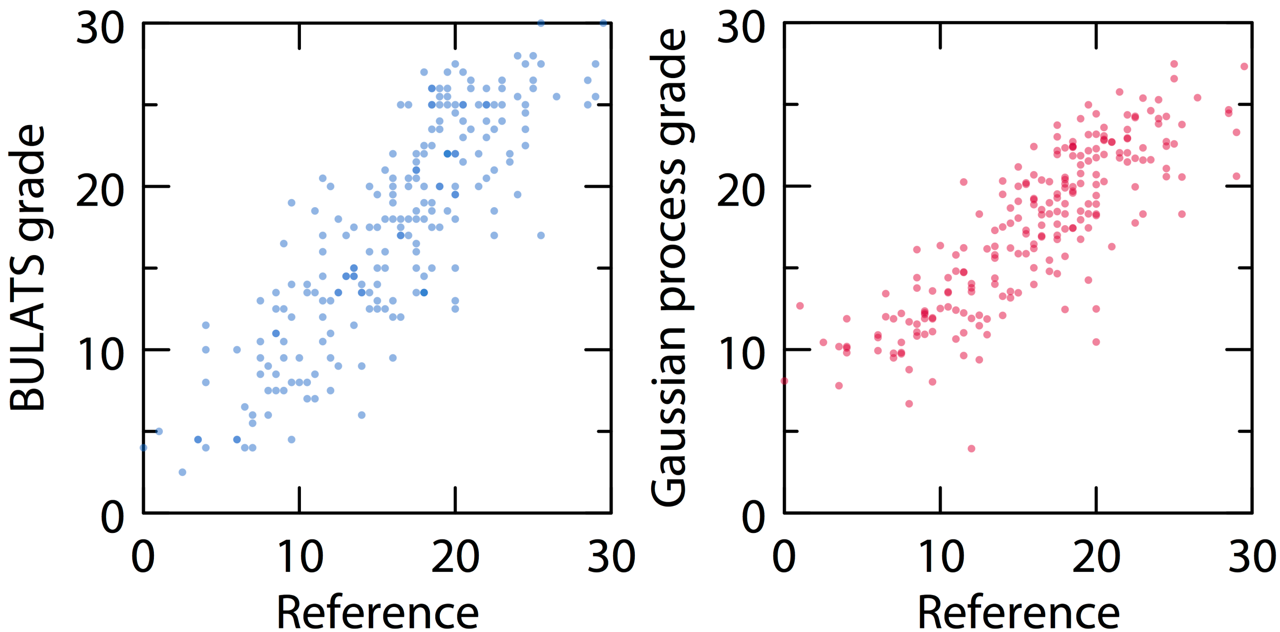
Gaussian Process Grader



- 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

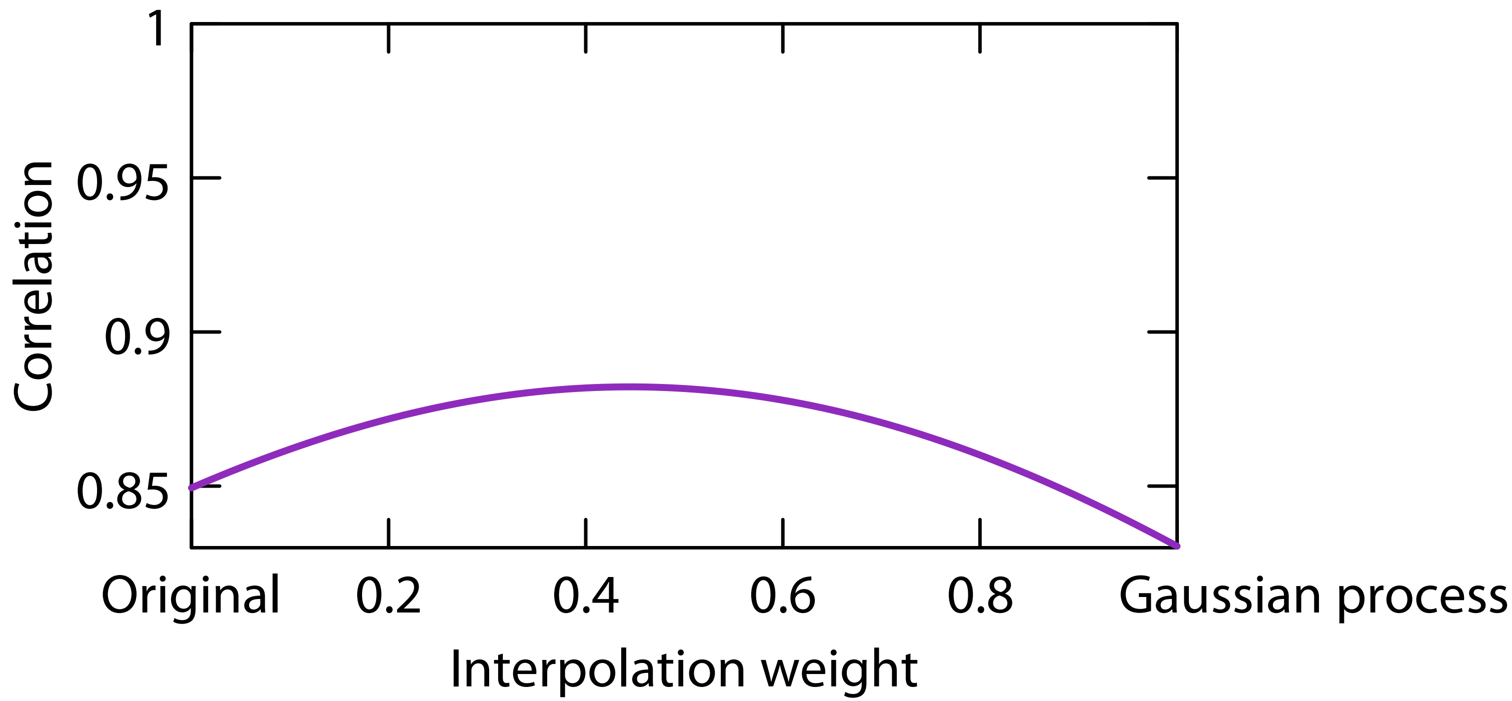


Graders	Pearson Correlation
Human experts	0.85
Automatic GP	0.83 – 0.86

Effect of Grader Features

Grader	Pearson Correlation with Expert Graders
Standard examiners	0.85
Automatic baseline	0.83
+ Pronunciation	0.84
+ RASP	0.85
+ Confidence	0.83
+ RASP + Confidence	**0.86**
Pronunciation features	0.82

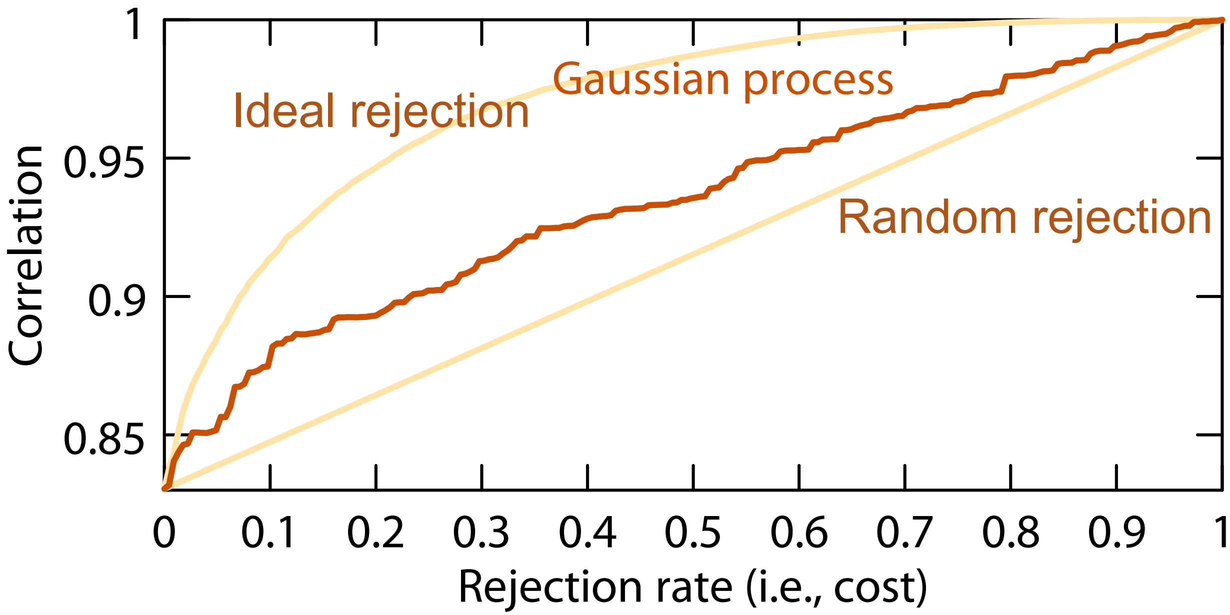
Combining Human and Automatic Graders



- 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

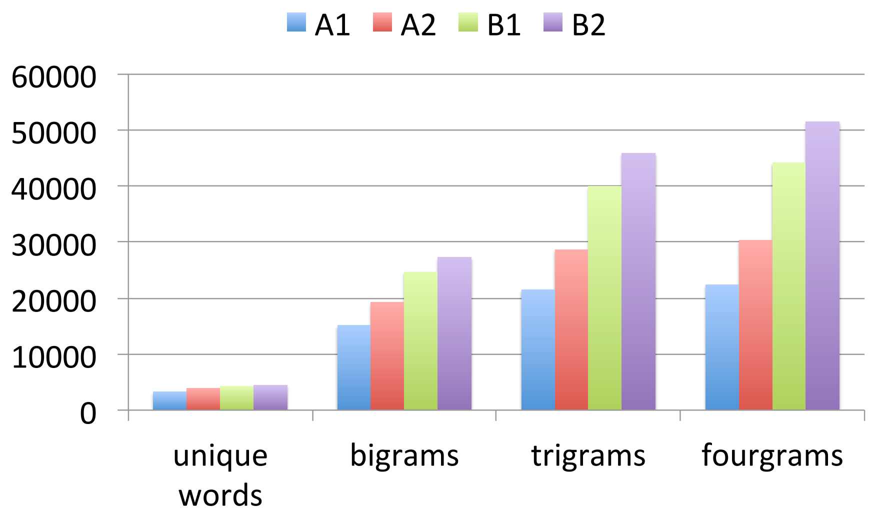
- Standard (BULATS) graders handle standard speakers very well
  - non-standard (outlier) speakers less well handled
  - use Gaussian Process variance to automatically detect outliers

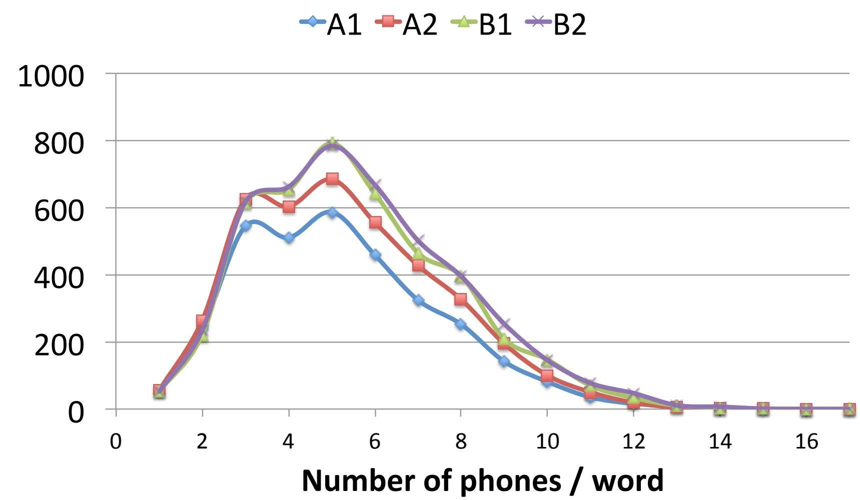


- Back-off to human experts - reject 10%: performance 0.83 ➔ 0.88

Assessing Communication Level

- Ignore high-level content and communication skills currently”



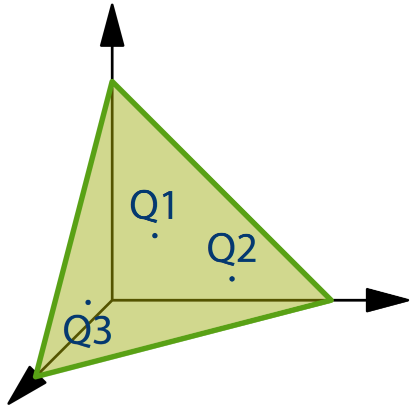


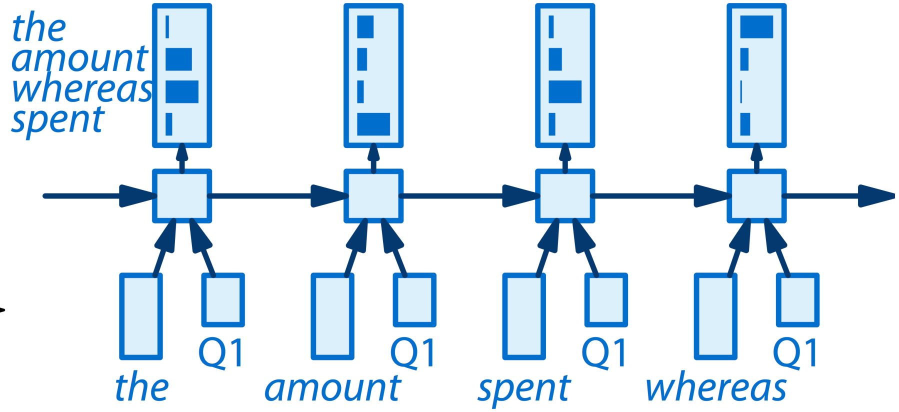
- Language complexity is related to proficiency
  - Future work – look into e.g.
    - McCarthy’s use of chunks “I would say”, “and then”
    - Abdulmajeed and Hunston’s “correctness analysis”

Assessing Content

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

Topic space:





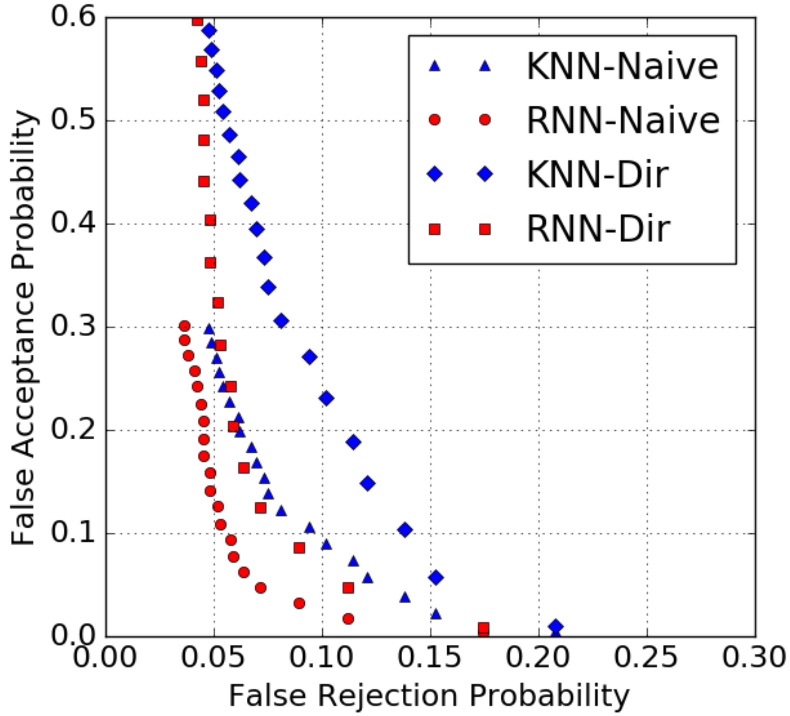
- Train a Recurrent Neural Network Language Model for each question
  - assess whether the response is consistent with example answers

Topic Classification

System	HL-dim	Training Data	% Error
KNN	-	SUP	20.8
RNNLM	100		17.5
RNNLM	200	Semi-SUP	9.3

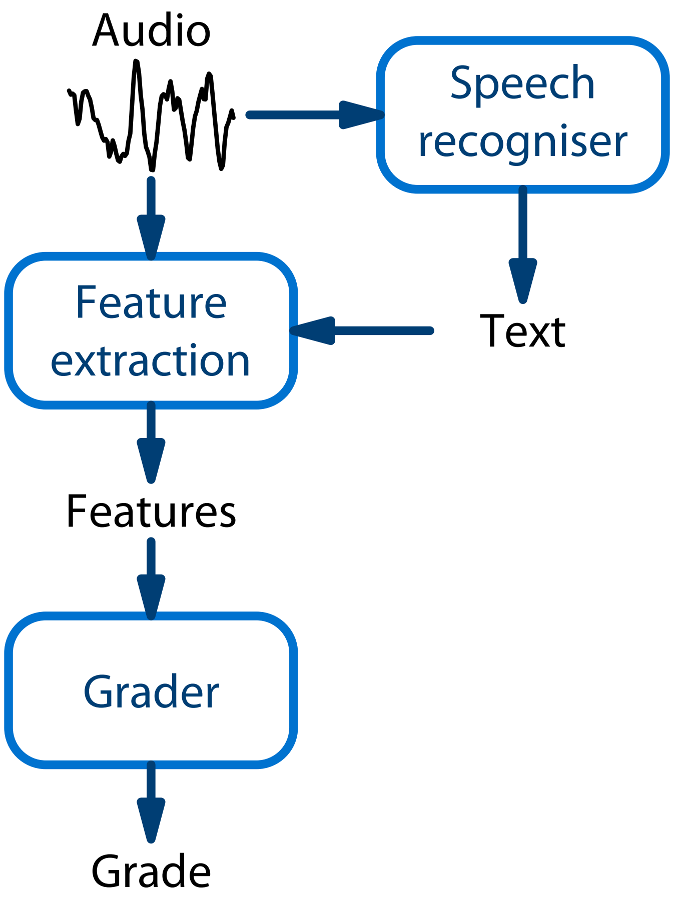
- Experiment details
  - 280-D LSA topic space
  - Supervised (SUP): 490 speakers, 2x crowd-sourced transcriptions
  - Semi-supervised (Semi-SUP): + 10005 speakers, ASR transcriptions
- Increasing quantity of data helps even though high %WER
- RNNLM can handle large data sets unlike K-Nearest Neighbour (KNN)

Off-Topic Response Detection



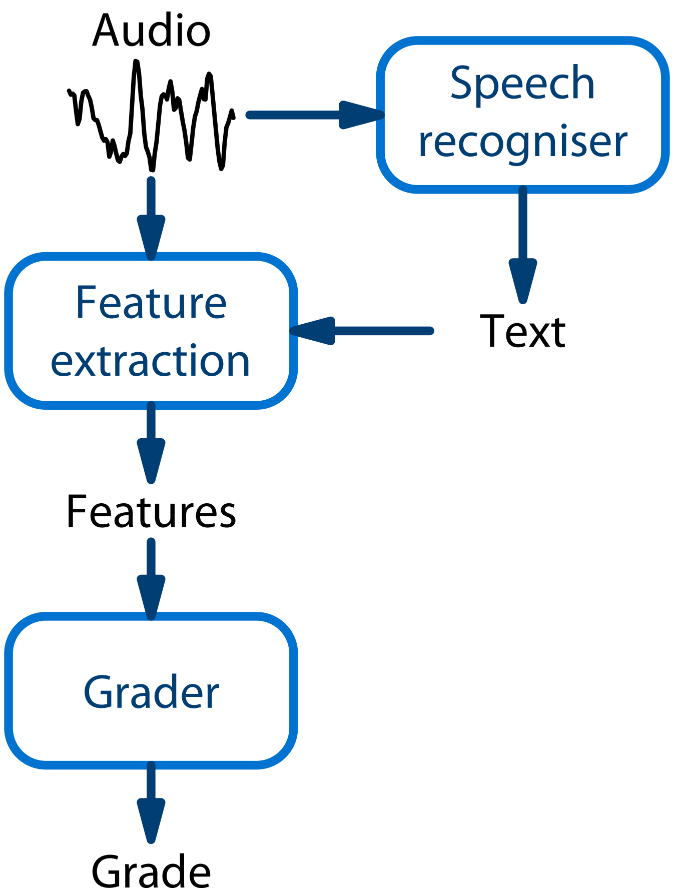
- Synthesised pool of off-topic responses
  - Naïve – select incorrect response from any section
  - Directed – select incorrect response from same section

Spoken Language Assessment



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

Spoken Language Assessment

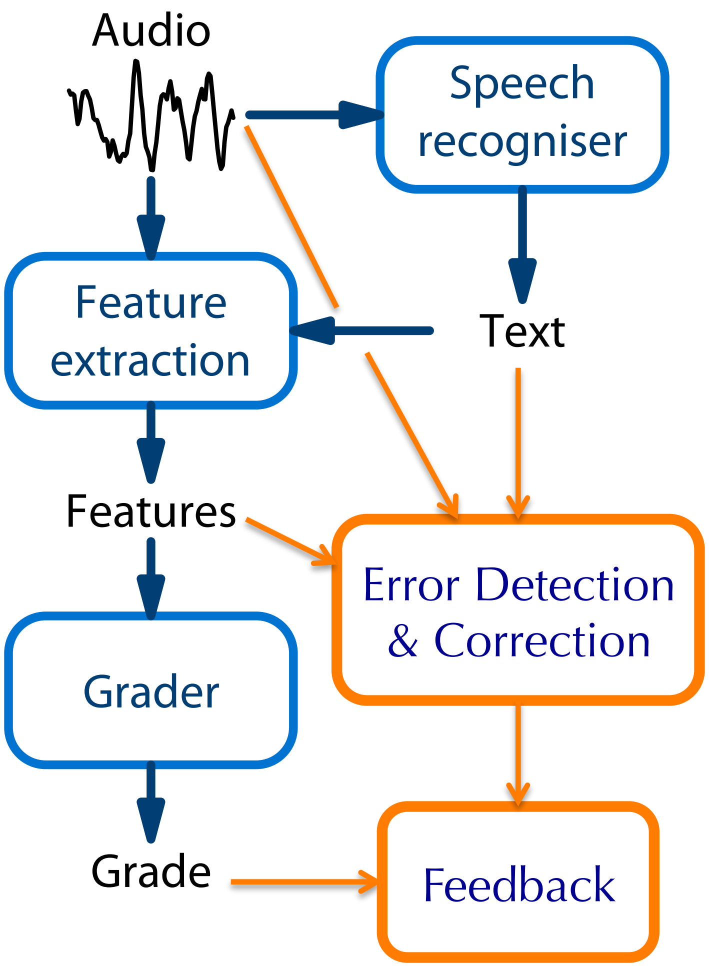


- Automatically assess:
  - Message realisation
    - Fluency, pronunciation

  Achieved (with room for improvement)
  - Message construction
    - Construction & coherence of response
    - Relationship to topic

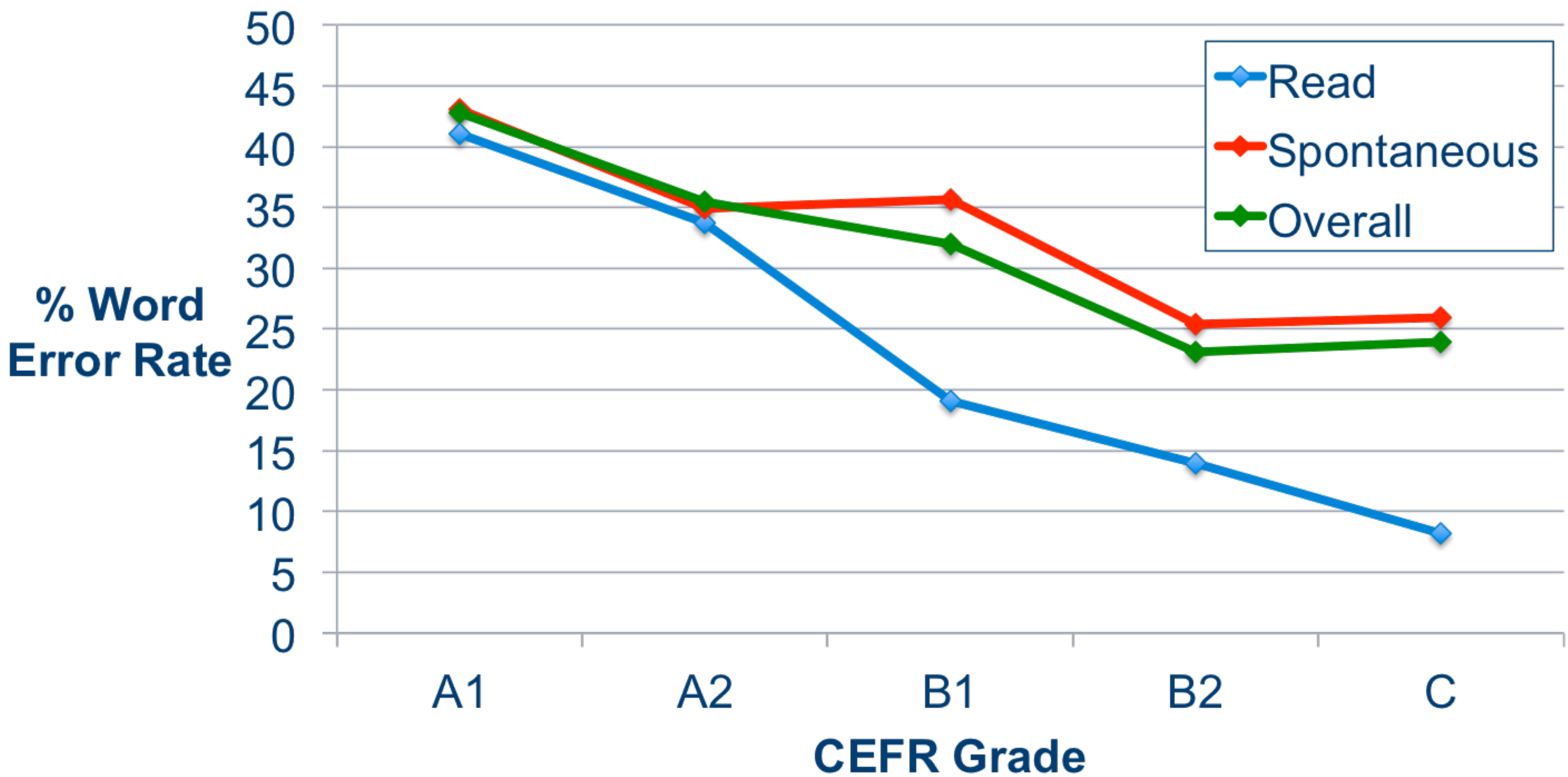
  Unsolved – active research areas

Spoken Language Assessment and Feedback

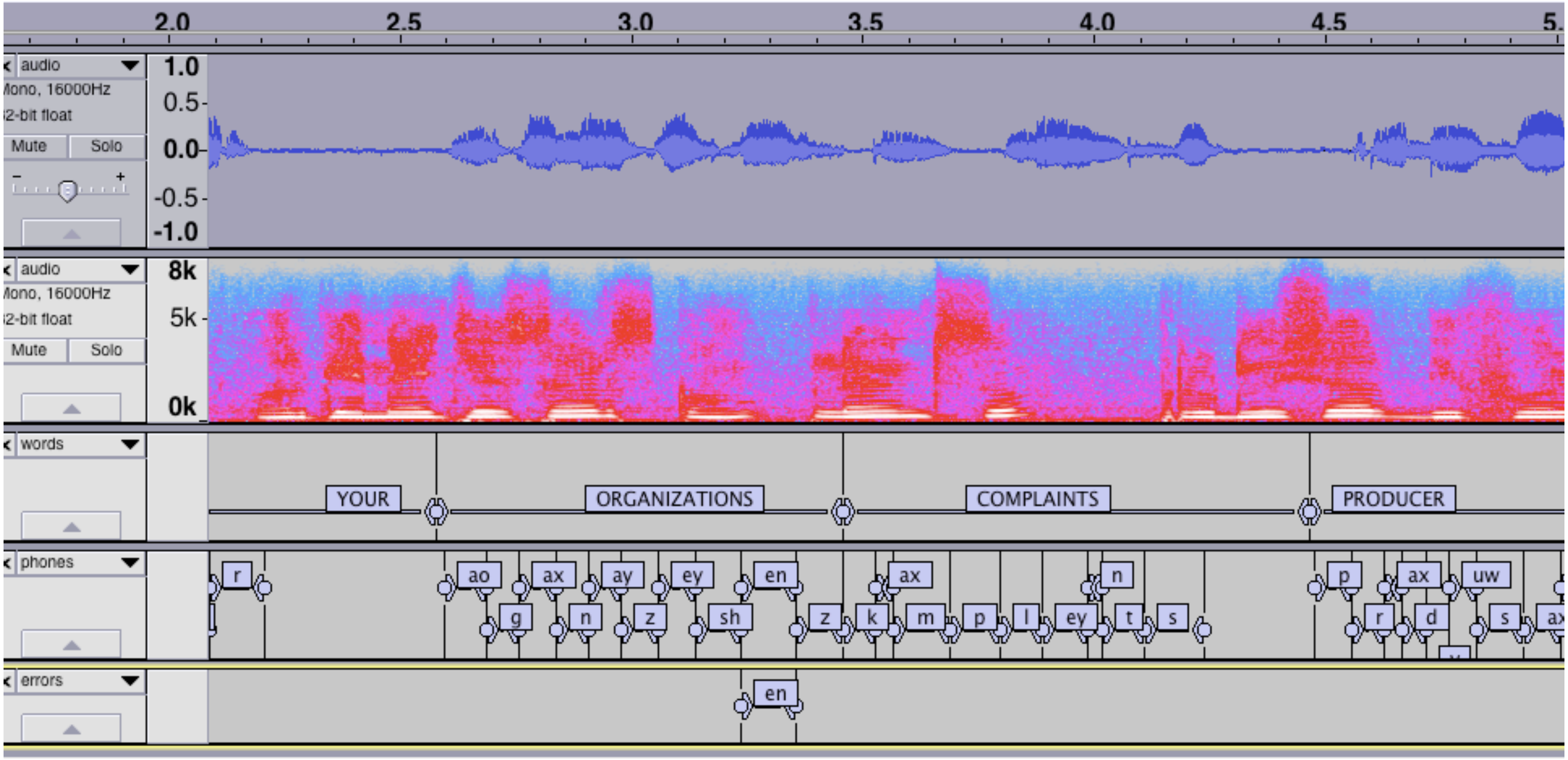


- Automatically assess:
  - Message realisation
    - Fluency, pronunciation
  - Message construction
    - Construction & coherence of response
    - Relationship to topic
- Provide feedback:
  - Feedback to user: realisation, construction
  - Feedback to system: adjust to level

Recognition Error Rate Versus Learner Progression



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?