UNIVERSITY OF CAMBRIDGE

Towards Learning Orientated Assessment for Non-native Learner Spoken English

Kate Knill

Mark Gales, Kostas Kyriakopolous, Edie Lu, Potsawee Manakul, Andrey Malinin, Anton Ragni. Linlin Wang, Yu Wang

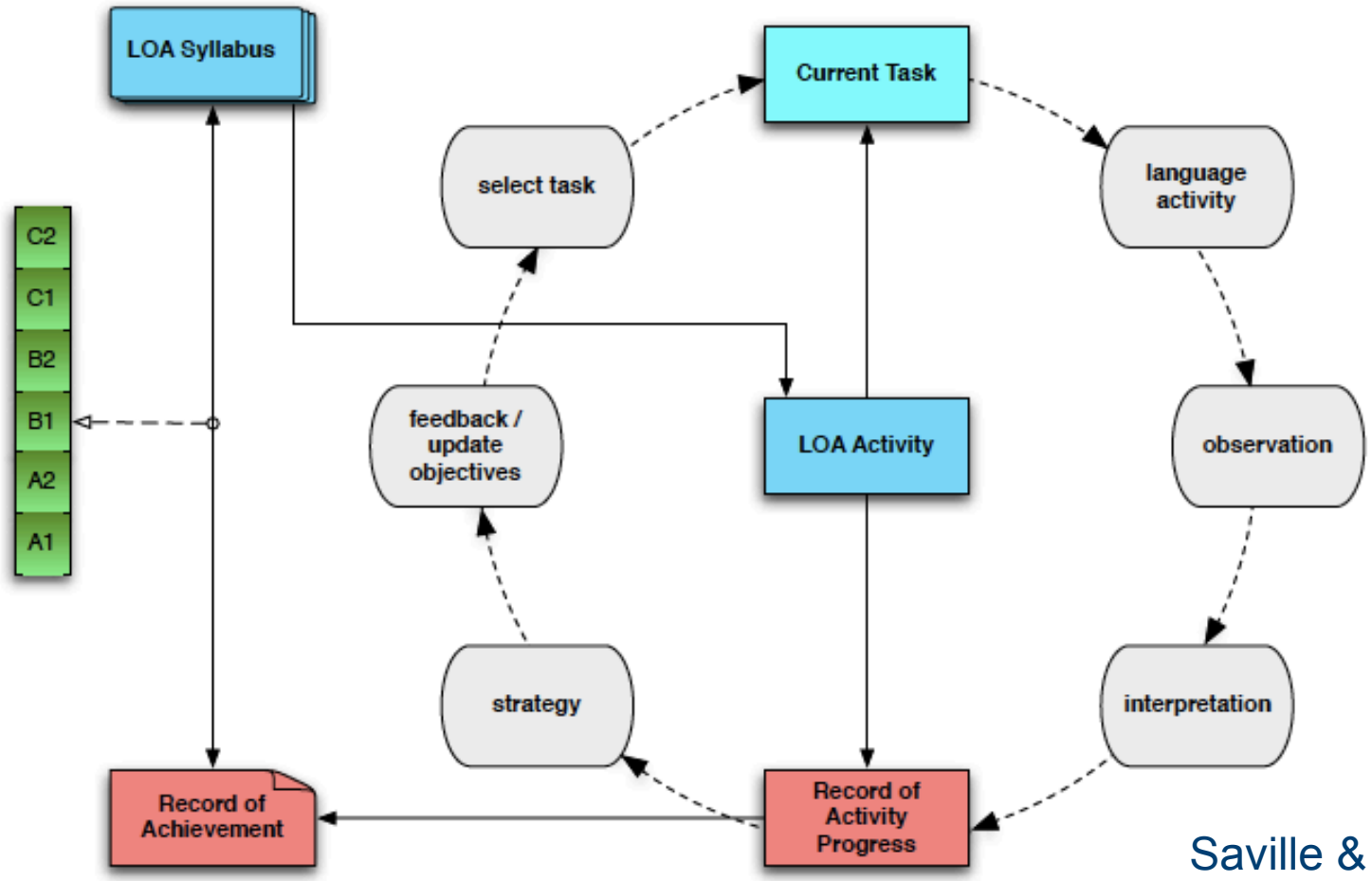
ALTA Institute, Cambridge University Engineering Department

6 March 2019

Cambridge ALTA

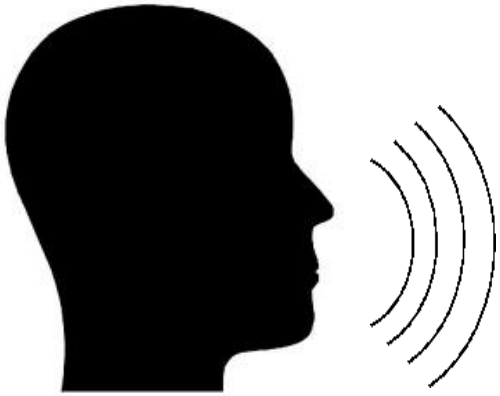
Institute for Automated Language Teaching and Assessment

Learning Oriented Assessment



Saville & Jones

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?
- 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

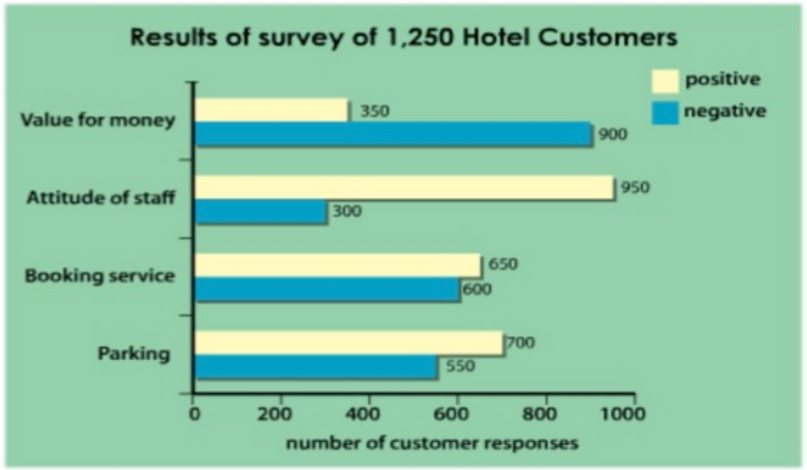
- Message Construction should consider:
  - Has the speaker generated a coherent message to convey?
  - Is the message appropriate in the context?
  - Is the word sequence appropriate for the message?
- Message Realisation should consider:
  - Is the pronunciation of the words correct/appropriate?
  - Is the prosody appropriate for the message?
  - Is the prosody appropriate for the environment?

Spoken Communication Assessment

- Construct
  - Read speech
  - Free speaking – short response to prompt/question; 20s-1min responses
  - Conversation – with examiner or between candidates with examiner

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)

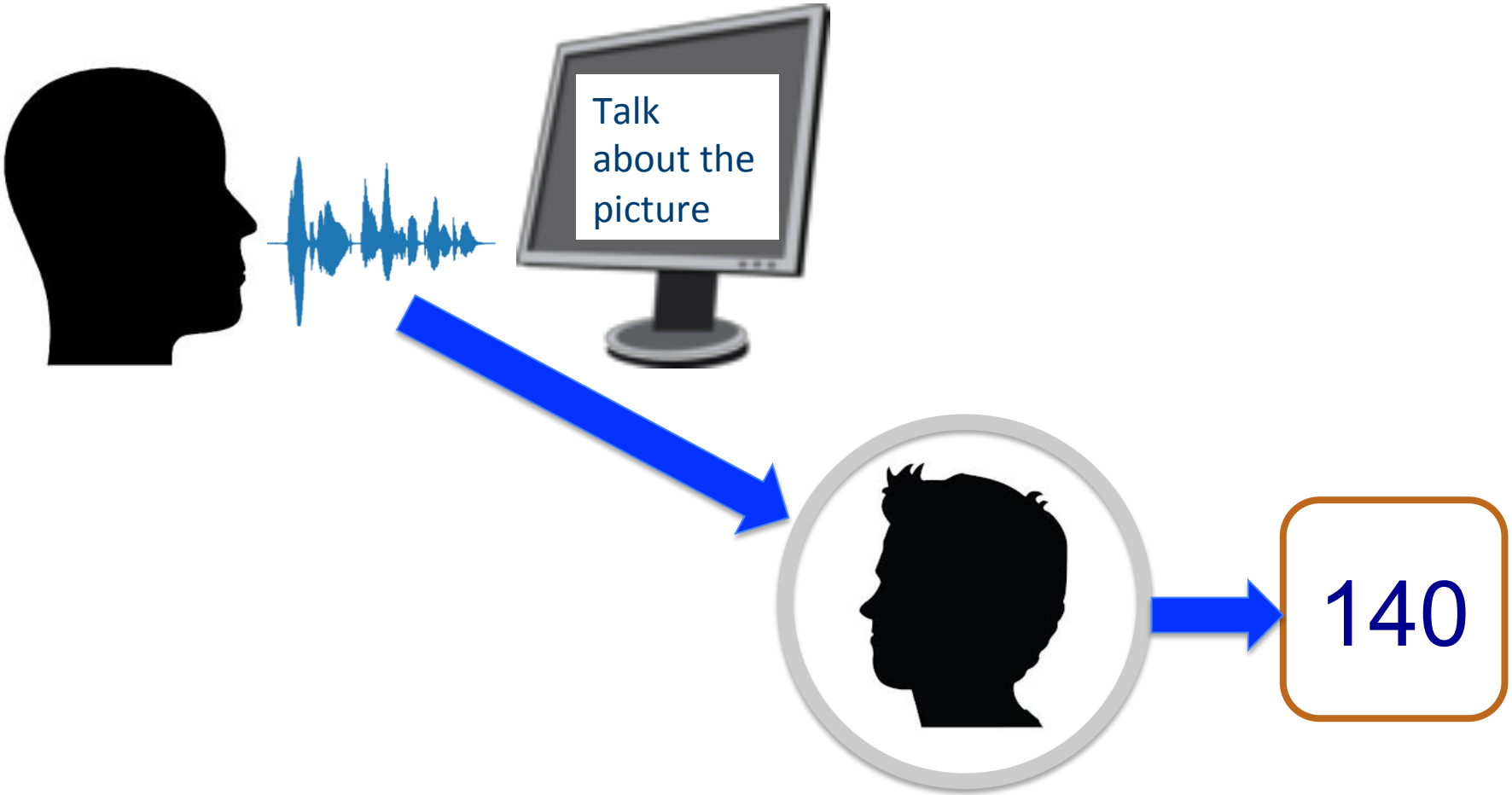
Level	Global Descriptor
C2	Fully operational command of the spoken language
C1	Good operational command of the spoken language
B2	Generally effective command of the spoken language
B1	Limited but effective command of the spoken language
A2	Basic command of the spoken language
A1	Minimal command of the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

Candidate Speech Examples

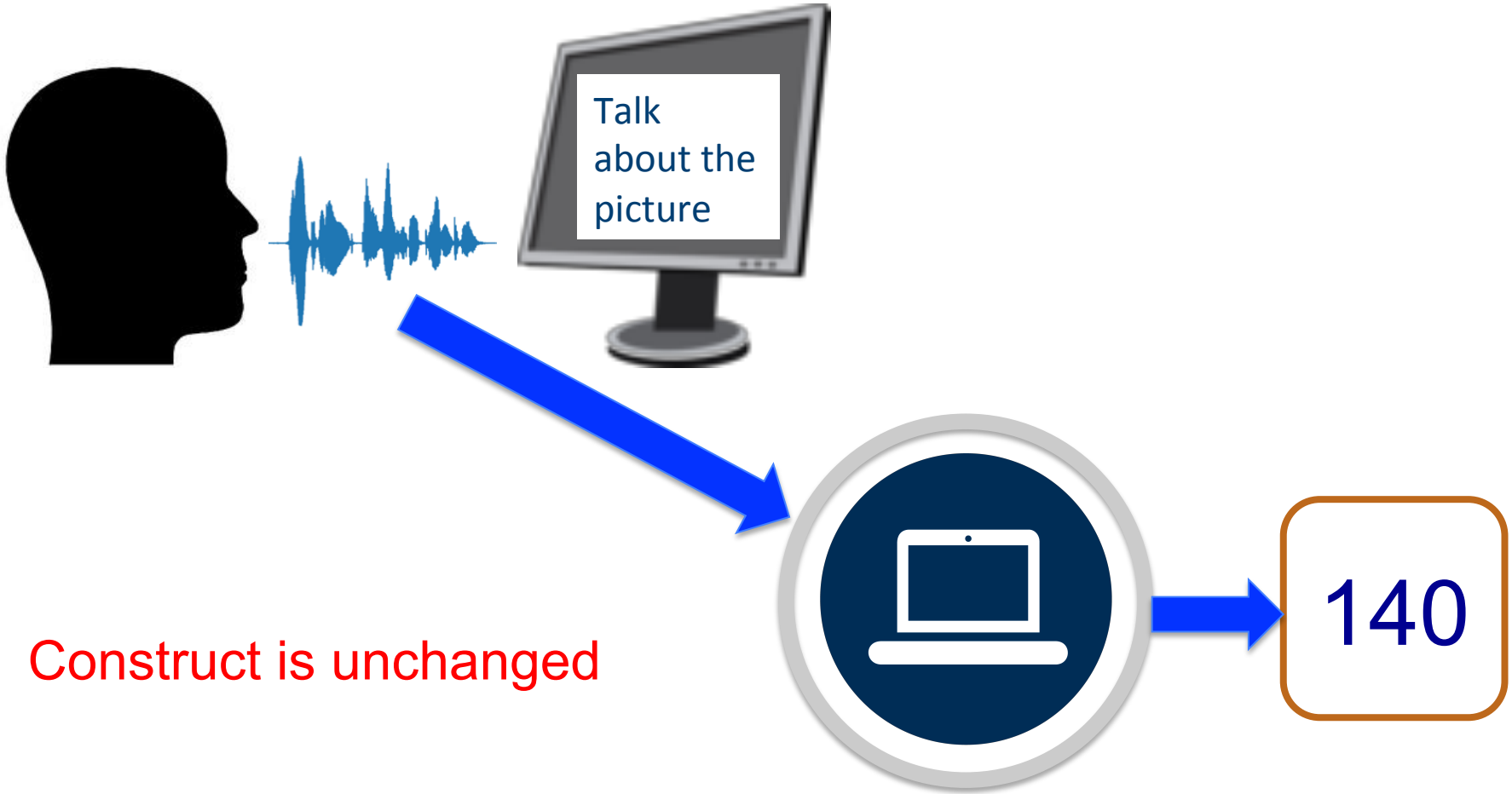
- Level B2

- Level A2

Human Examiner Marking

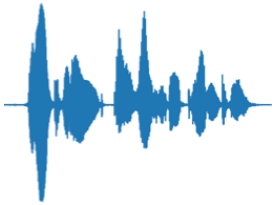


Speaking Auto-marking



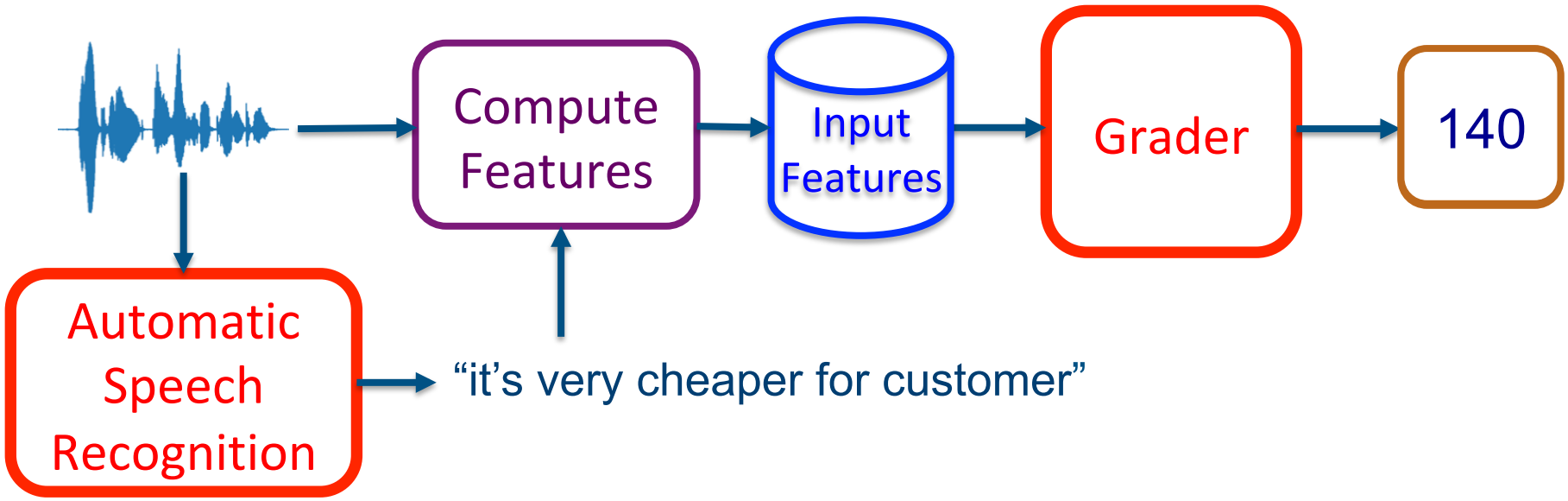
Construct is unchanged

Auto-marker

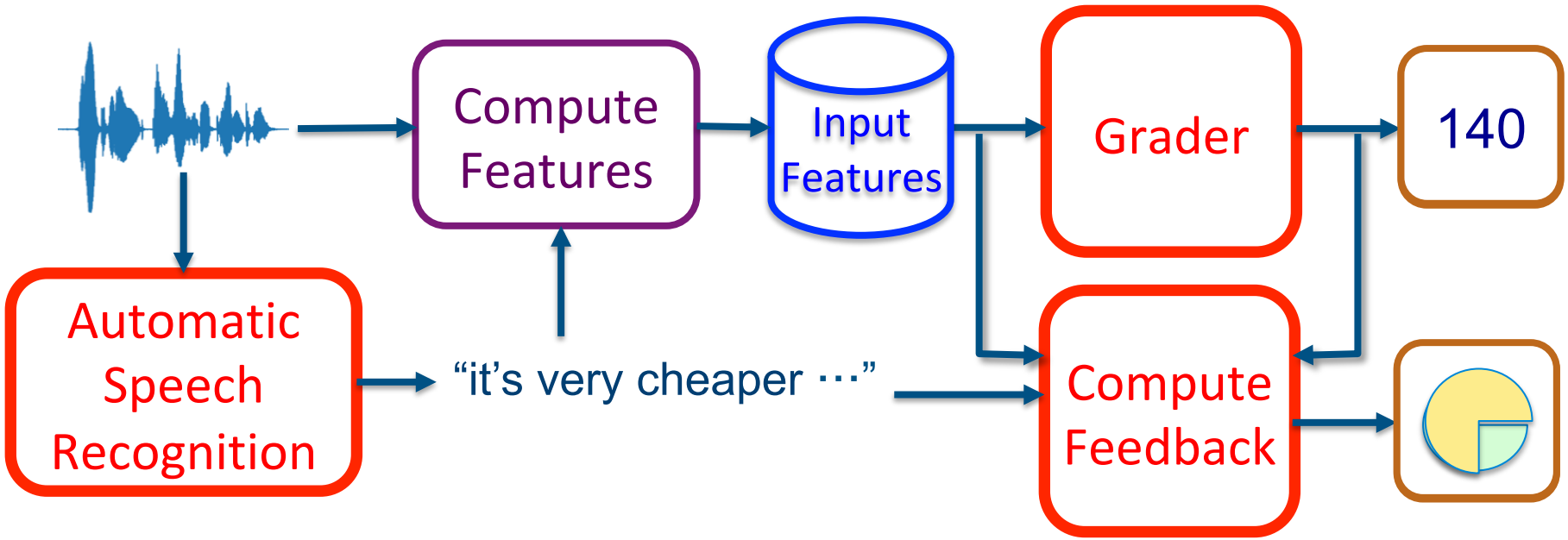


?

Auto-marker


Compute Features
Input Features
Grader
140
Automatic Speech Recognition
“it’s very cheaper for customer”

Auto-marker and Feedback



Auto-marker: Automatic Speech Recognition (ASR)

Automatic Speech Recognition

→ “it’s very cheaper for customer”

?

Speech Recognition Challenges

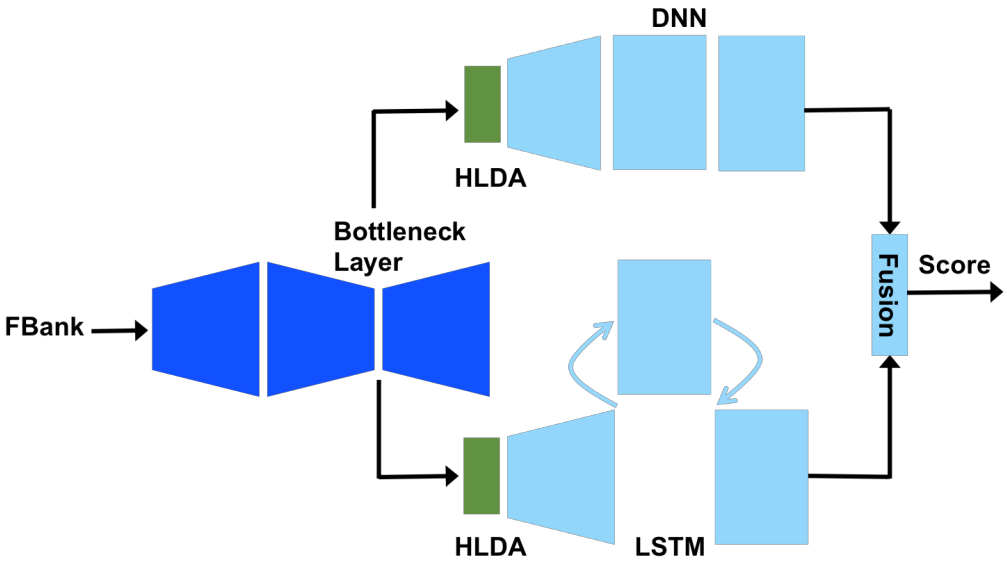
- Non-native ASR highly challenging
  - Heavily accented
  - Pronunciation dependent on L1
- Commercial systems poor!
- 2015 CUED systems

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

Evaluation set

- 225 non-native speakers from BULATS test
  - 6 L1s: Arabic, Dutch, French, Polish, Thai, Vietnamese
  - Approx. equal distribution over CEFR proficiency levels A1-C
- ASR: long free speaking sections only (C-E)
  - Manual transcriptions (1x annotator)
- Grading: all sections
  - Expert marks
- Pronunciation and grammatical error detection
  - Manual annotations

ASR



- Trained on real examination data from over 30 L1s
- Transcriptions − automatic merging 2x crowd-sourced trans; ASR
- Graphemic lexicon
- Kaldi training (sMBR/LF-MMI) and decoding

ASR on Non-native Speech (1)

AM	LM	% WER
300	300	25.5
		24.5
400	400	24.4

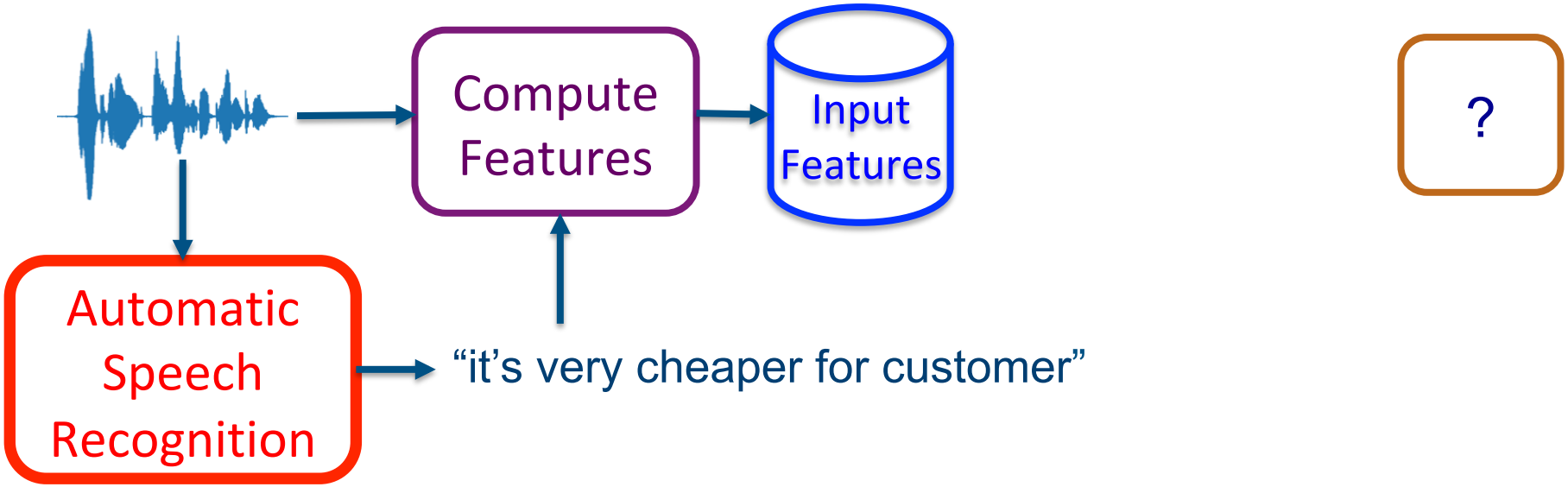
- Joint DNN+LSTM AM + trigram LM
- AM and LM trained on BULATS data only
- Hours of data varied
- LM trained on transcriptions, interpolated with BNE

ASR on Non-native Speech (2)

AM	LM	% WER
TDNN-F	trigram	23.4
+AMI		22.7
+TS Ensemble		22.3
+ i-vectors		21.3
	RNNLM+succ	19.5

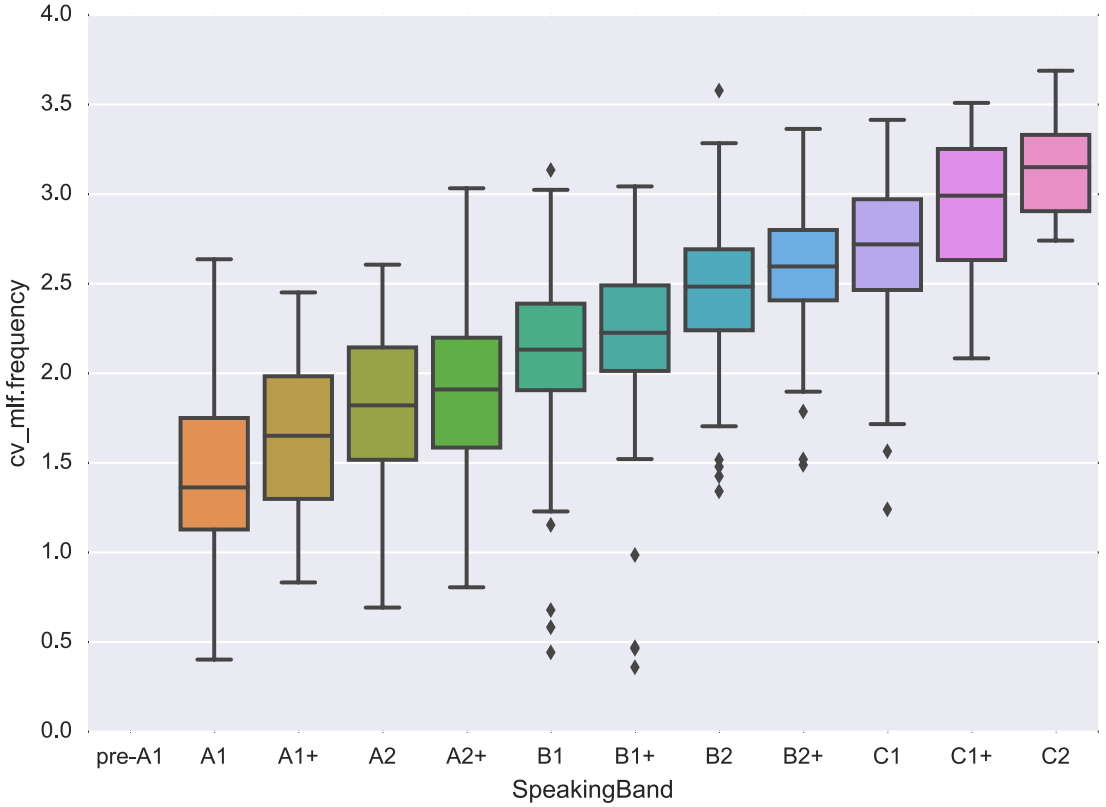
- 400 hour BULATS training set

Auto-marker: Grader Features

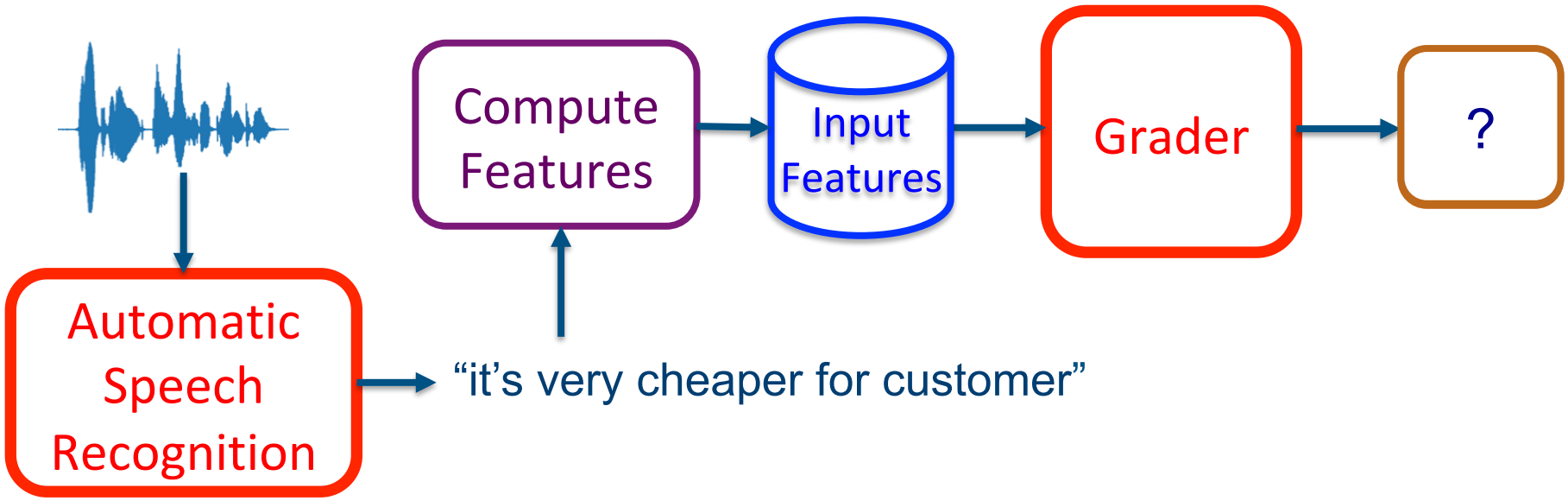


Example Auto-marker Feature

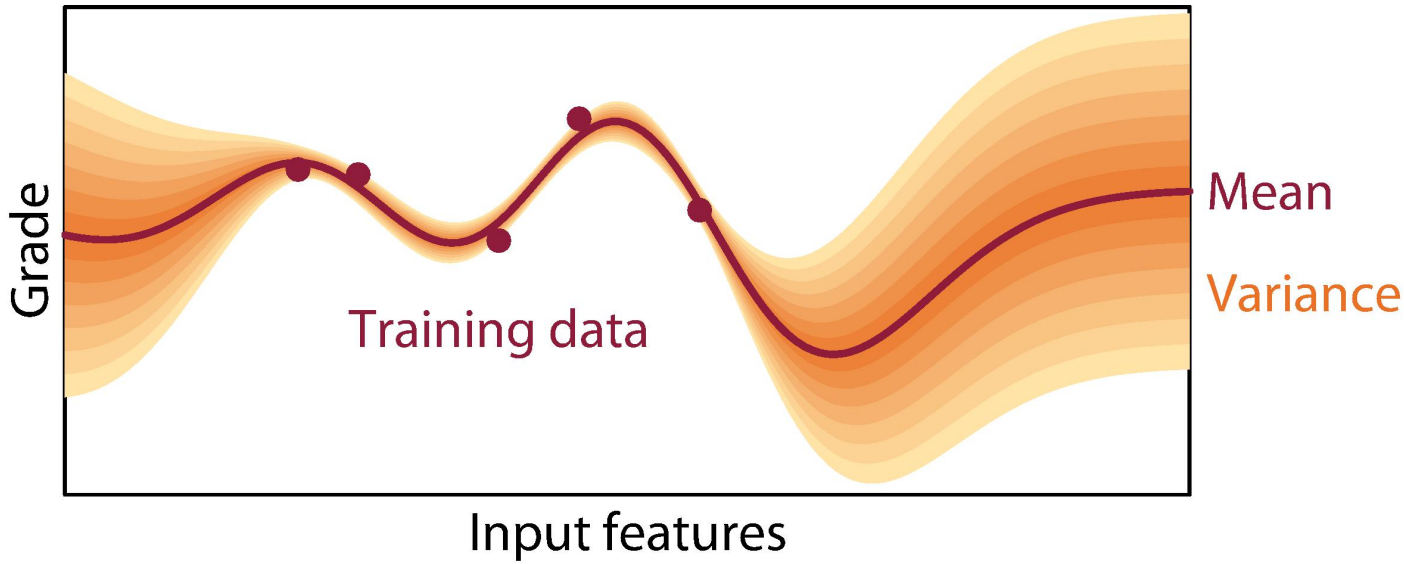
Speaking Rate vs Candidate Speaking Band



Auto-marker: Grader

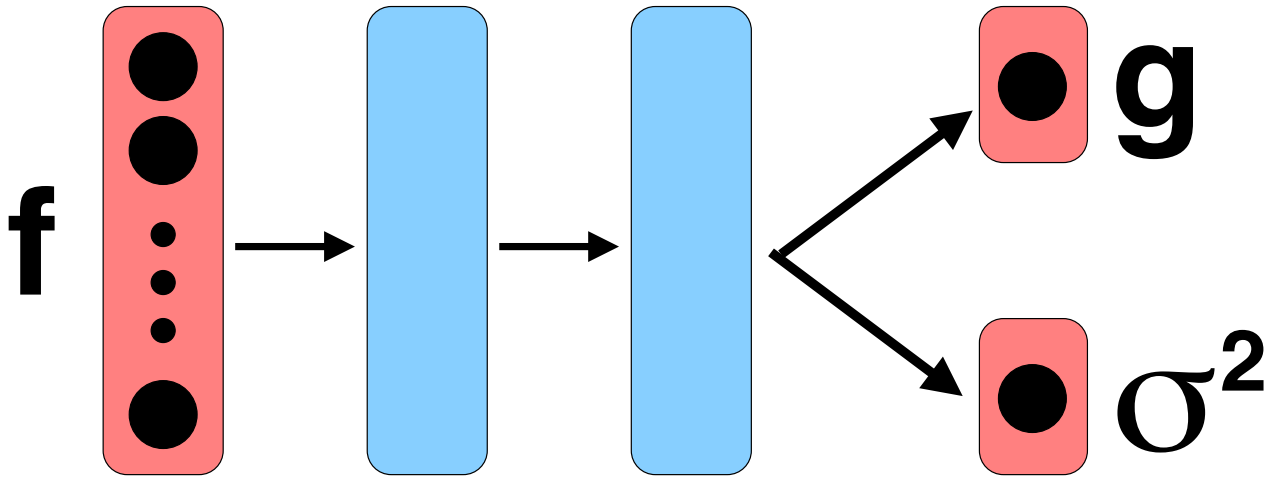


Gaussian Process Grader



- Powerful non-parametric Bayesian model
- Use variance to provide confidence in prediction
- Limitation − $O(n^2)$ memory usage, $O(n^3)$ computational load

Deep Density Network (DDN) Grader



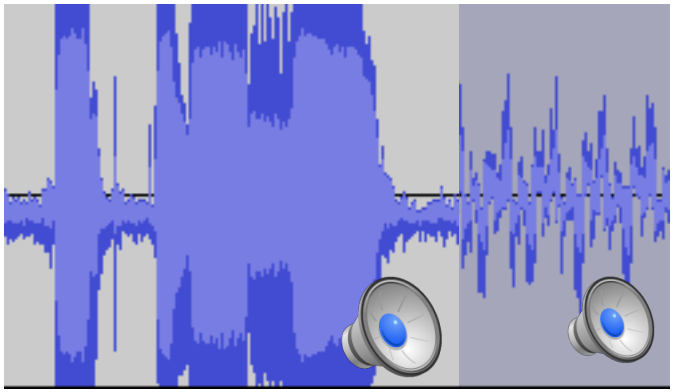
- Predict score and score variance
- Use predicted variance to provide confidence in prediction
- Model size and computational independent of training data set size

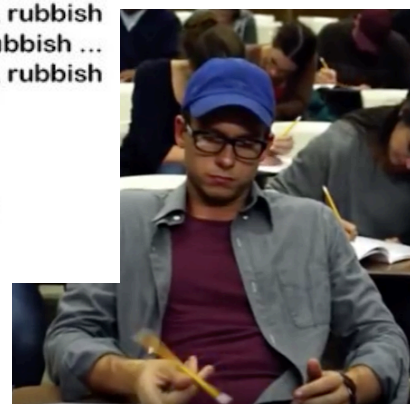
Grader Performance

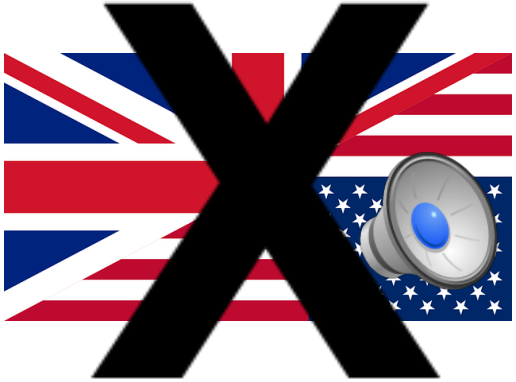
Grader	PCC	MSE
Exam	8.48	0.864
GP	8.56	0.449
DDN	0.86	0.420

- Automatic graders match human non-expert examiners
- Combine human and auto grades to boost accuracy
- Regions with limited data a challenge: low A1/C2

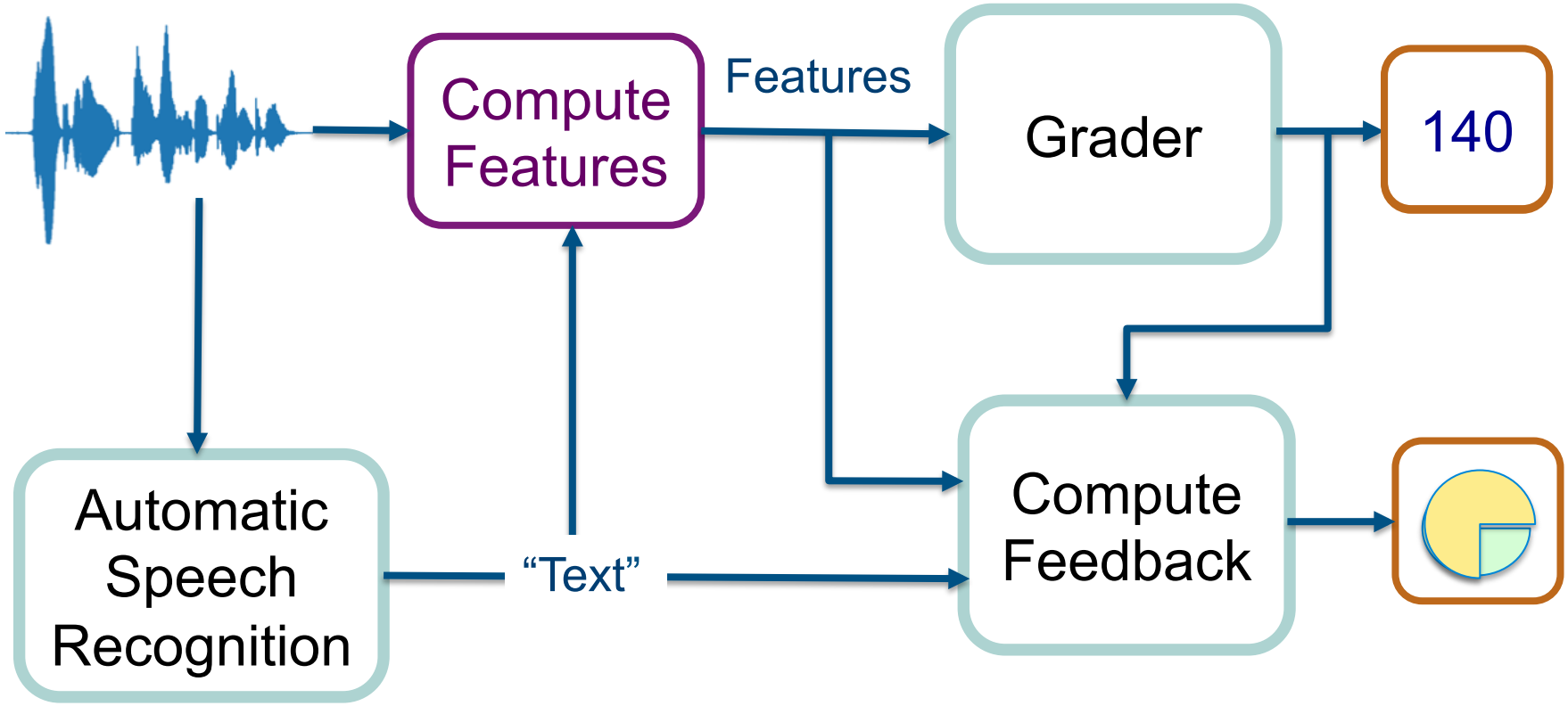
Practical Considerations







Automatic Assessment with Feedback



Pronunciation Error Detection

- Substitution errors
  - e.g.

Correct	Actual
/**sh**/ /aa/ /p/ /ih/ /ng**/**	**/**s/ /aa/ /p/ /ih/ /ng/

- Insertion errors
  - Word final e.g.

/k/ /ah/ /m**/** ➔ /k/ /ah/ /m/ /ah/

  - Inter-consonant e.g.

/r/ /iy/ /s/ /ah/ /n/ /t/ ➔ /r/ /iy/ /s/ /ah/ /n/ /**eh** / /t/

- Deletion errors (final phone, 1st/2nd consonant in consonant pair)

Common Substitutions by L1

L1	Substitution		% correct pronounced as incorrect
	Correct	Incorrect	
French	ch	sh	4.76
	dh	z	3.40
Dutch	th	t	2.11
	dh	d	1.99
Thai	dh	d	7.24
	oh	aa	5.21

- Top 2 recurrent substitution errors for speakers in each L1

Pronunciation Error Detection and Feedback

- Use Siamese networks and attention networks to detect bad phones

Grammatical Error Detection: Challenge

- Spoken language consists of

  Text + Pronunciation + Prosody + Delivery

- Challenge for feedback on “grammatical” errors in spoken language

  Spoken Text ≠ Written Text

  - We don’t speak in sentences, we repeat ourselves, hesitate, mumble etc
  - There is no defined spoken grammar standard
- Advantages of speech
  - There are no spelling or punctuation mistakes
  - We provide additional information within the audio signal

Example of Learner Speech (Manual)

{F %HES%}THE VISITOR CAN {F %HES%} GO TO THE RESTAURANT AND HAVE {F %HES%} {REP BUFFET + BUFFET} {F %HES%} AFTER THAT THEY CAN {F %HES%} {F %HES%} HAVE THE {F %HES%} KARAOKE ACTIVITIES TO {F %HES%} {REP GET AC- + GET ACQUAINTED} WITH THE COMPANY

THE VISITOR CAN GO TO THE RESTAURANT AND HAVE BUFFET AFTER THAT THEY CAN HAVE THE KARAOKE ACTIVITIES TO GET ACQUAINTED WITH THE COMPANY

THE VISITOR CAN GO TO THE RESTAURANT AND HAVE BUFFET AFTER THAT THEY CAN HAVE THE KARAOKE ACTIVITIES TO GET ACQUAINTED WITH THE COMPANY

Example of Learner Speech (ASR)

{F %HES%}THE VISITOR CAN {F %HES%} GO TO THE RESTAURANT AND HAVE {F %HES%} WITH A BUDGET {F %HES%} AFTER THAT THEY CAN {F %HES%} HAVE THE {F %HES%} CAN OKAY ACTIVITIES TO {F %HES%} GET A QUITE GET ACQUAINTED WITH THE COMPANY

THE VISITOR CAN GO TO THE RESTAURANT AND HAVE WITH A BUDGET AFTER THAT THEY CAN HAVE THE CAN OKAY ACTIVITIES TO GET A QUITE GET ACQUAINTED WITH THE COMPANY

THE VISITOR CAN GO TO THE RESTAURANT AND WITH A BUDGET AFTER THAT THEY CAN HAVE THE CAN OKAY ACTIVITIES TO GET A QUITE GET ACQUAINTED WITH THE COMPANY

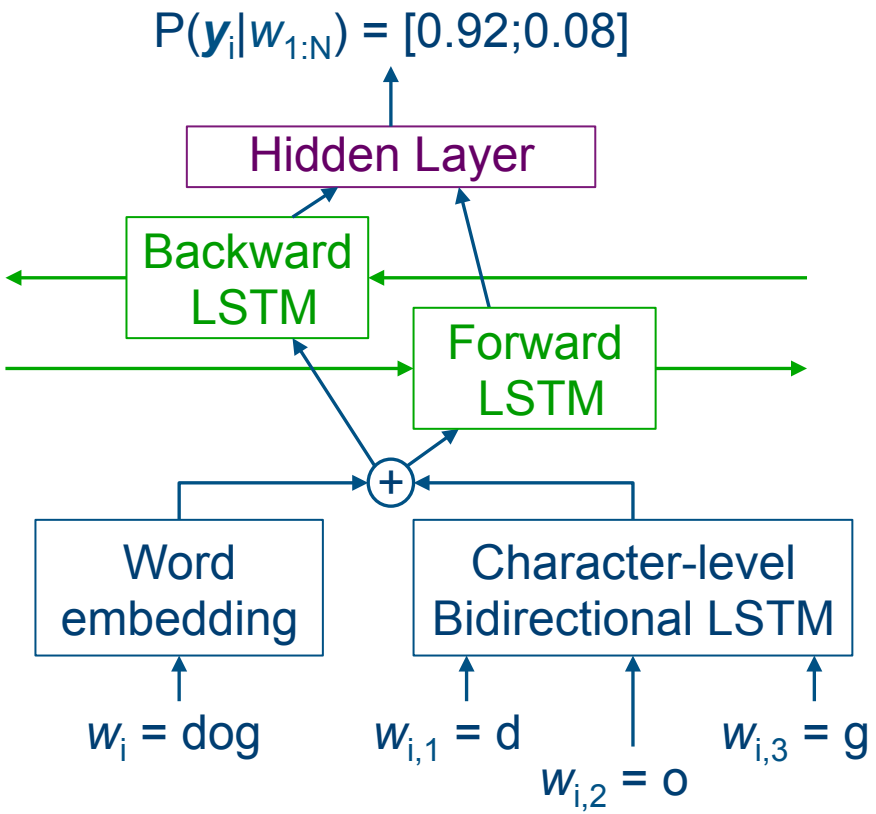
Grammatical Error Detection Sequence Labelling

- Task: given a sentence automatically label each word with
  - P_{word} (grammar is correct) and P_{word} (grammar is incorrect)

- Example sentence

	Internet	was	something	amazing	for	me	.
P(c)	0.02	0.96	0.97	0.97	0.95	0.98	0.99
P(i)	0.98	0.04	0.03	0.03	0.05	0.02	0.01

Sequence Labeler



Marek Rei

Approach for Spoken Grammatical Error Detection

1. Match data processing in training and testing
   a. Test: Speech data - convert speech transcriptions to be “like” text

```
// flor company is an engineering company in the poland
// we do business the refinery business and the chemical business
// the job we can offer is a engineering job
// basically this is the job in the office
```

   b. Train: Text data - correct spelling errors, remove punctuation and casing

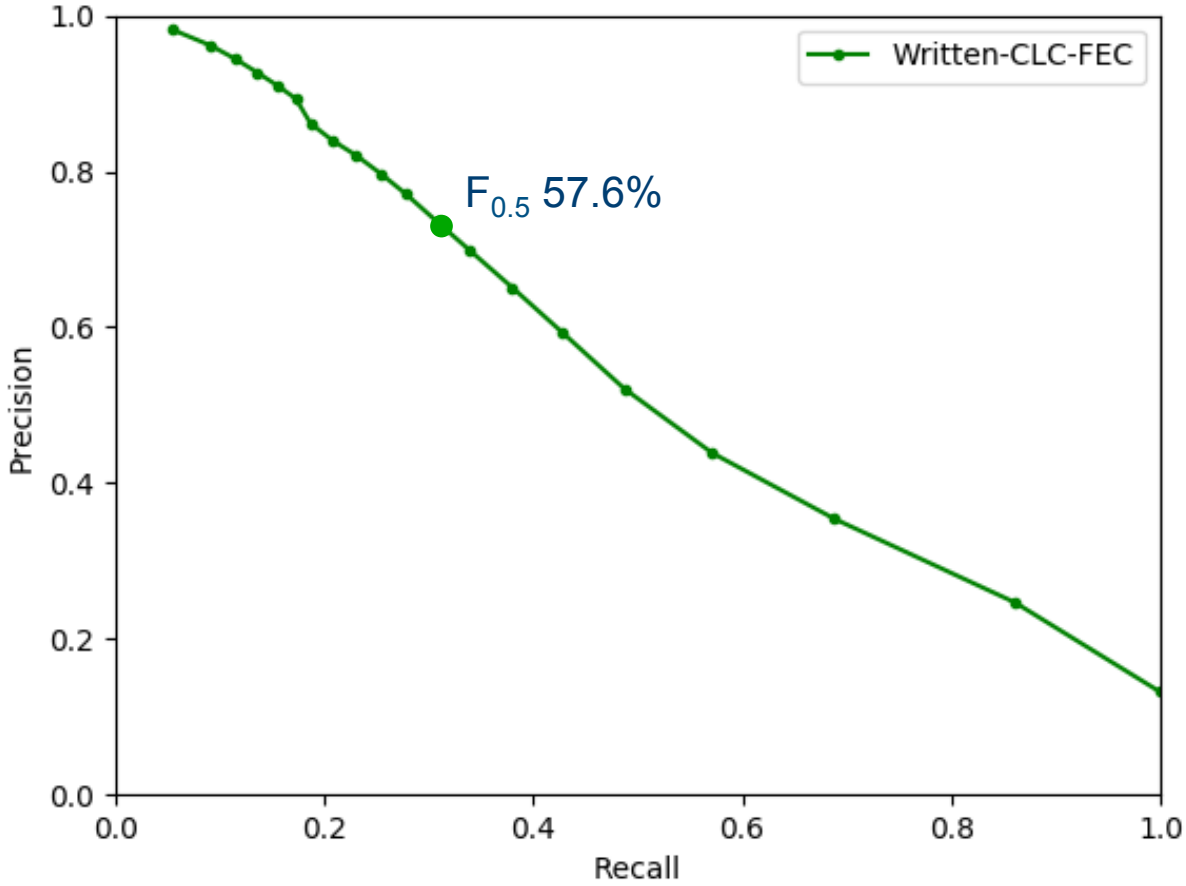
2. Train GED sequence labeler
   a. Each word is labelled as correct or incorrect

3. Apply GED model to test data
   a. Predict P_{word} (grammar is correct) and P_{word} (grammar is incorrect)

GED Experimental Corpora

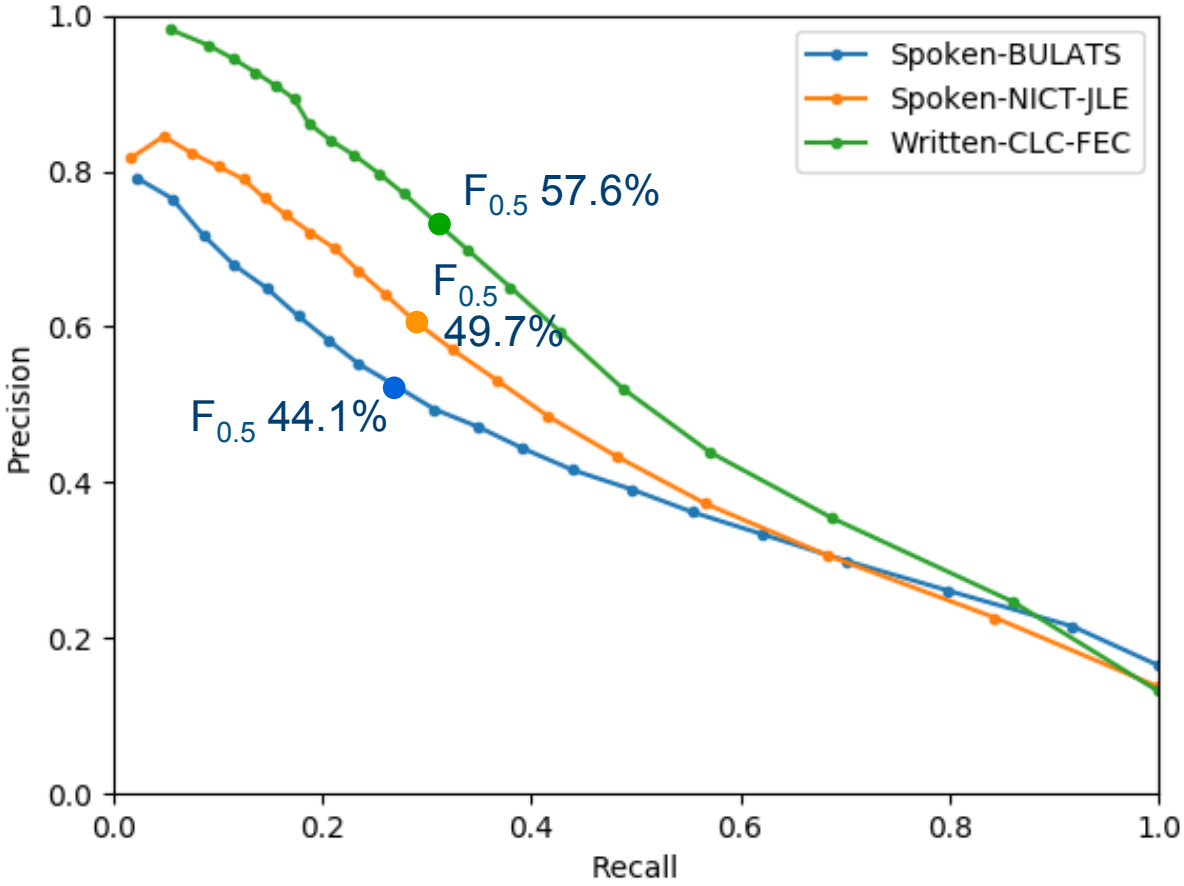
- Cambridge Learner Corpus (CLC)
  - Learner written texts
  - Training: FCE, BULATS, IELTS, CPE, CAE; Test: FCE public evaluation set
- NICT Japanese Learner English (JLE)
  - Manual transcription of interviews from English oral proficiency test
  - 167 GE marked interviews – test on interviewees; grades A1-B2
- CA BULATS Spoken Corpus
  - Multi-level test with single speaker free speaking responses (sections C,D,E)
  - Manual and ASR transcriptions of 226 speakers, 1438 responses
  - Speakers evenly distributed across CEFR grades A1-C

GED Using CLC Trained Model



- Lower $F_{0.5}$ than usual as spelling mistakes and punctuation removed

GED Using CLC Trained Model



(Small Scale) System Error Analysis

- True precision higher for Spoken BULATS than scores suggest
  - System error (~27%)

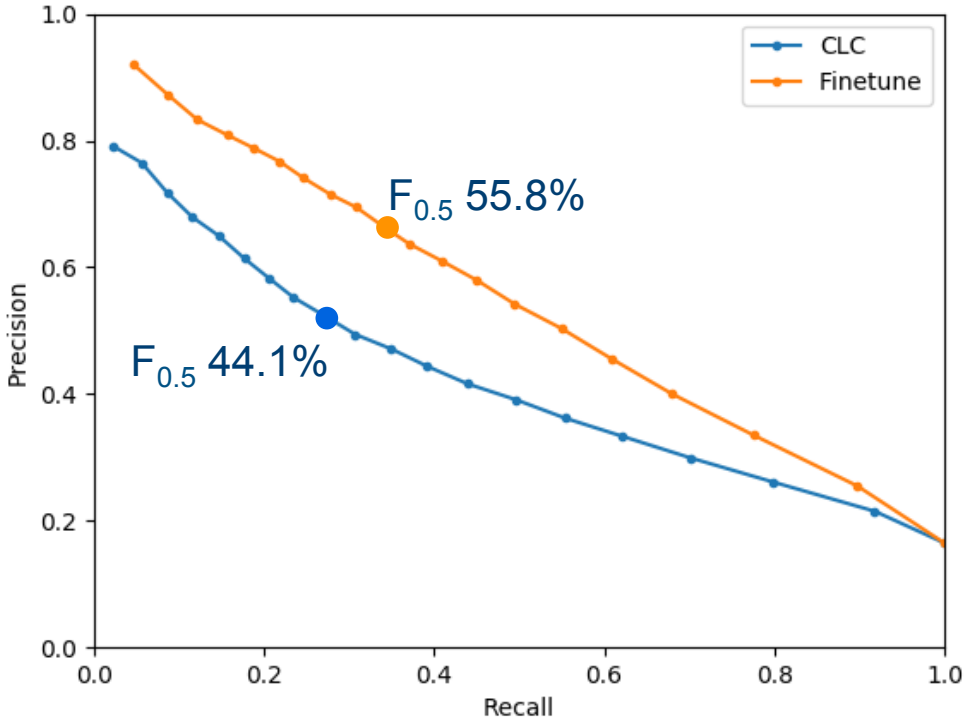
    .. and i have to practice more because I have ..
  - Unmarked error (~40%)

    .. so I think you need taxi
  - Next to error tagged word(s) (~27%)

    .. and continue to inform with customer when we have ..
- To provide feedback we need to boost recall of high precision items
  - Issue: lack of labelled learner speech corpora
    - Corrupt native speaker transcriptions to make learner errors
    - Adapt/“fine-tune” CLC trained system to subset of target speech data

Boosting GED Performance on Spoken BULATS

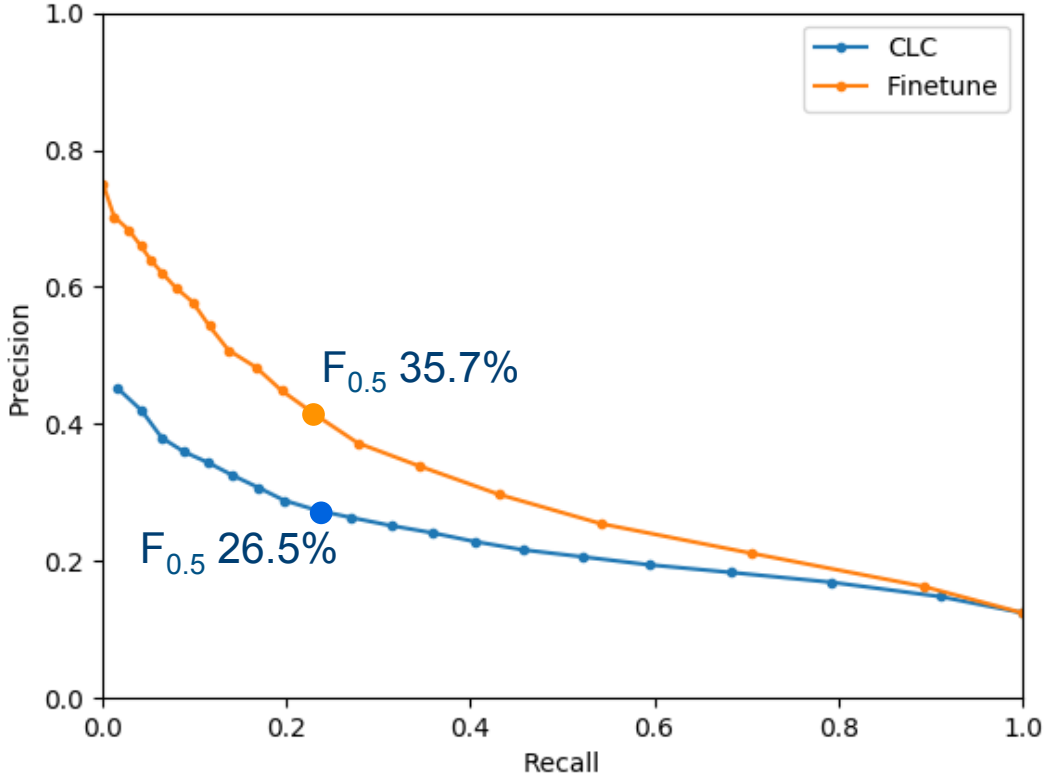
- Fine-tune CLC system with 80% data, dev 10%, test 10% x10



- Fine-tuning produces significant boost in performance
  - has also learnt some annotator bias e.g. “two thousand eight”

GED on BULATS ASR Transcriptions

- Manual transcriptions used for GE marking and meta-data extraction

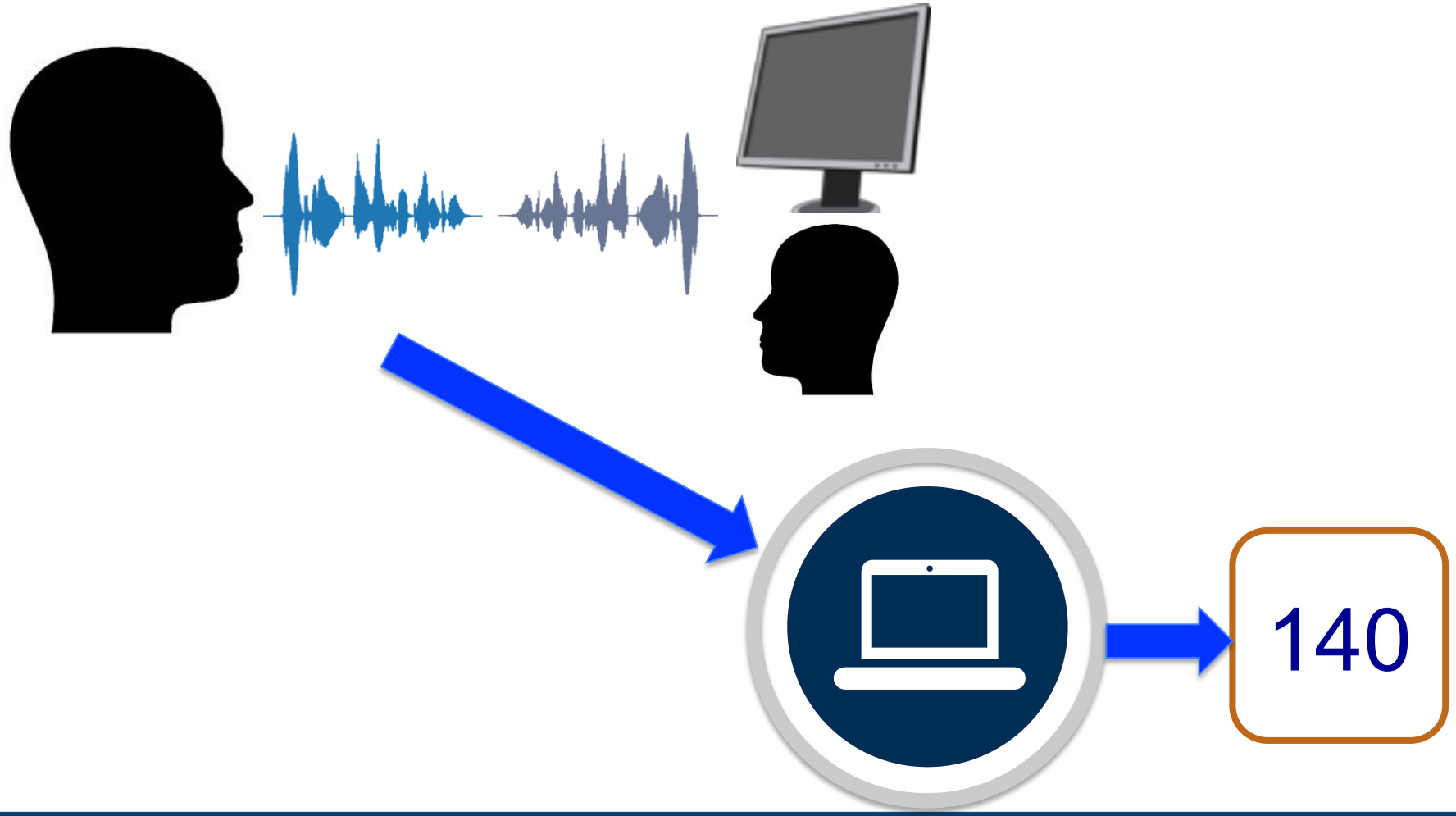


- Significantly lower performance than manual transcriptions
  - Focus on boosting high precision region

Conclusions

- Learning Orientated Assessment
  - Assessment, feedback, learning activities synced
- Non-native spoken learner English
  - Large variation due to proficiency, L1 etc
  - Assessment for prompt-response on a par with human examiners
  - Research into feedback on errors progressing
    - Pronunciation
    - Use of English/grammar

Future: Conversational Auto-marking



Questions?

- Thanks to ALTA members especially CUED team, Andrew Caines , Marek Rei, Helen Yannakoudakis
- Email: kate.knill@eng.cam.ac.uk

- Try out Speak&Improve https://speakandimprove.com