UNIVERSITY OF CAMBRIDGE

Cambridge ALTA
Institute for Automated Language Teaching and Assessment

Use of Deep Learning in Free Speaking Non-native English Assessment

Kate Knill

TSD 6th September 2021

Automated Language Teaching & Assessment Institute

UNIVERSITY OF CAMBRIDGE
DEPARTMENT OF
COMPUTER SCIENCE AND TECHNOLOGY

UNIVERSITY OF CAMBRIDGE
DEPARTMENT OF ENGINEERING

Cambridge ALTA
Institute for Automated Language Teaching and Assessment

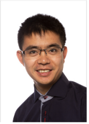
- Virtual Institute for

  cutting-edge research on non-native English assessment

  - Machine Learning and Natural/Spoken Language Processing
  - develop technology to enhance assessment and learning
  - improve learner experience and progress, support teachers

ALTA SLP Technology Team Past and Present


Prof Mark Gales


Dr Xie “Jeff” Chen


Dr Rogier van Dalen

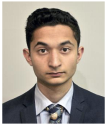

Kostas Kyriakopolous


Yiting “Edie” Lu


Adian Liusie

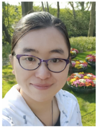

Dr Andrey Malinin

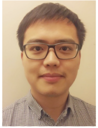

Potsawee Manakul

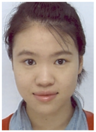

Vatsal Raina


Vyas Raina


Dr Anton Ragni


Dr Linlin Wang


Dr Yu Wang

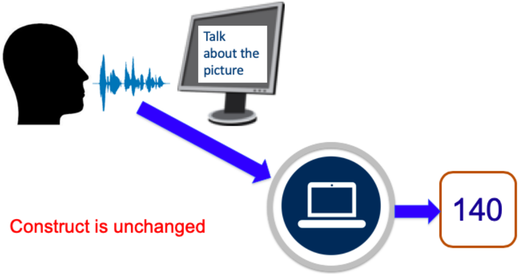

Dr Xizi Wei


Dr Xixin Wu

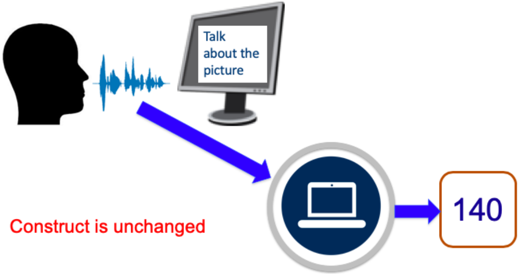
plus undergraduate and masters research project students

Spoken Language Assessment & Learning

Spoken Language Assessment & Learning



Construct is unchanged



- Automate (English) spoken language assessment & learning
  - without simplifying/limiting form of test: “free speaking”
  - possibility for richer, interactive, tests
  - desire to assess communication skills

CEFR - Levels of Foreign Language (L2) Learning

- Internationally agreed standard for assessing level
  - Common European Framework of Reference (CEFR)

- Basic User

  A1 - breakthrough or beginner
  A2 - way-stage or elementary
- Independent User

  B1 - threshold or intermediate
  B2 - vantage or upper intermediate
- Proficient User

  C1 - effective operational proficiency or advanced
  C2 - mastery or proficiency

Linguaskill Speaking Test

Linguaskill

- Cambridge Assessment English computer-based oral English test
  - General and Business (formerly BULATS) English
  - hybrid assessment: auto-marking & human examiners [12]
- Overview of Tasks:
  1. Interview: 8 questions about the candidate
  2. Reading Aloud: read aloud 8 sentences
  3. Presentation: speak on a given topic
  4. Presentation with Visual Info: speak based on graphic info
  5. Communication Activity: opinion on 5 ques. related to a scenario

Overview

- Assessment Framework
- Feature-Based Assessment
- Neural Assessment
- Multi-view Assessment
- Robustness

Assessment Framework

Automatic Assessment Challenges

- Reliability: assessment is consistent with human scores

Automatic Assessment Challenges

- Reliability: assessment is consistent with human scores
- Validity: all aspects associated with a construct are evaluated

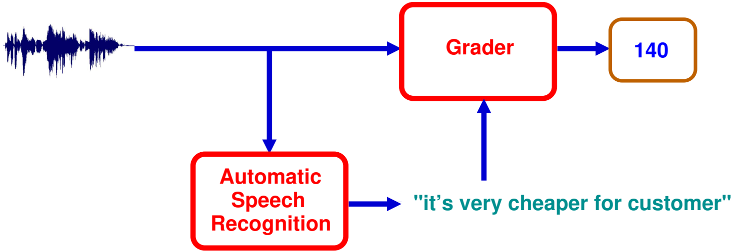
Automatic Assessment Challenges

- Reliability: assessment is consistent with human scores
- Validity: all aspects associated with a construct are evaluated
- Robustness: handles 'gaming' and organised/systemic cheating

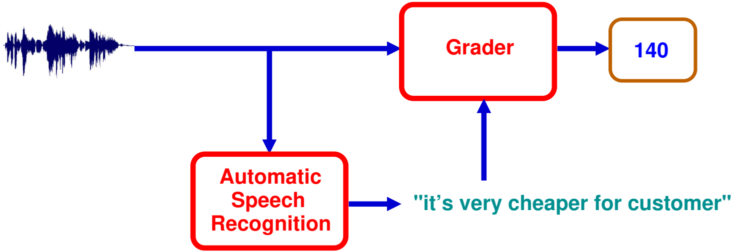
Automatic Assessment Challenges

- Reliability: assessment is consistent with human scores
- Validity: all aspects associated with a construct are evaluated
- Robustness: handles 'gaming' and organised/systemic cheating
- Fairness: the assessment shows no bias for any user group

Assessment Framework [11]



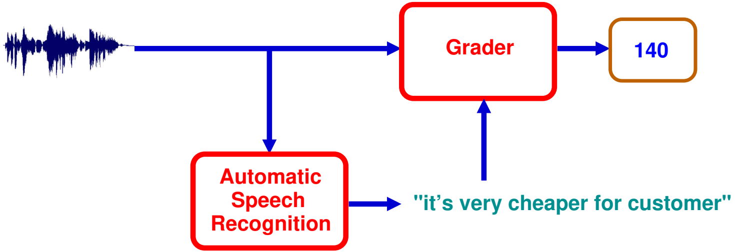
Assessment Framework [11]



Key Issues:

- Input speech variability
  - Speakers: large range of L1s, non-native speech, wide ability
  - Recordings: varying background noises, channel corruptions

Assessment Framework [11]

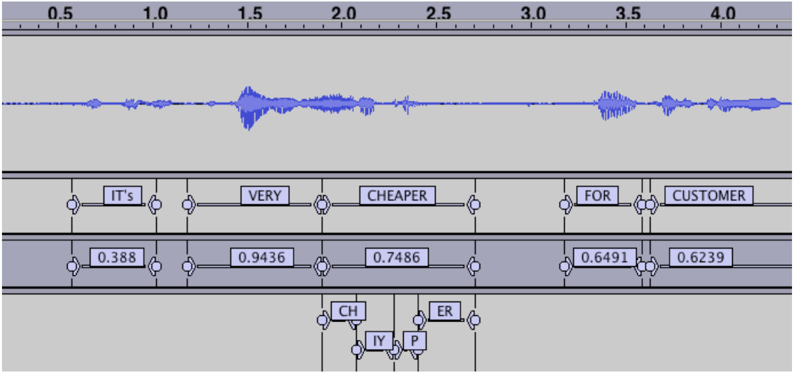


Key Issues:

- Input speech variability
  - Speakers: large range of L1s, non-native speech, wide ability
  - Recordings: varying background noises, channel corruptions

    ⇒ High word error rate (WER): propagates through system

Automatic Speech Recognition [10, 2]



- Baseline Automatic Speech Recognition (ASR) yields:
  - time aligned word/disfluencies/partial-word sequence
  - time aligned phone/grapheme sequence
  - word level confidence scores

Speech Recognition Performance [10, 2]

- Non-Native ASR: real-time decoding (non-RNNLM)
  - "basic users" (A1/A2) highly challenging data

	A1	A2	B1	B2	C	Avg
Baseline ASR	33.8	27.7	21.2	19.9	16.5	21.3
+su-RNNLM	31.8	25.4	19.6	18.0	14.7	19.5

Speech Recognition Performance [10, 2]

- Non-Native ASR: real-time decoding (non-RNNLM)
  - "basic users" (A1/A2) highly challenging data

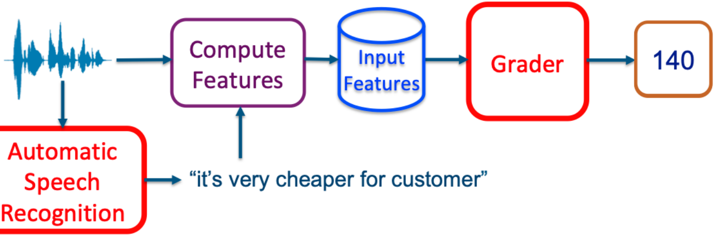
	A1	A2	B1	B2	C	Avg
Baseline ASR	33.8	27.7	21.2	19.9	16.5	21.3
+su-RNNLM	31.8	25.4	19.6	18.0	14.7	19.5

- Need to mitigate for ASR errors in grader

$\Rightarrow$ match train and test i.e. use ASR outputs for both

Feature-Based Assessment

Feature-based Assessment Framework [11]

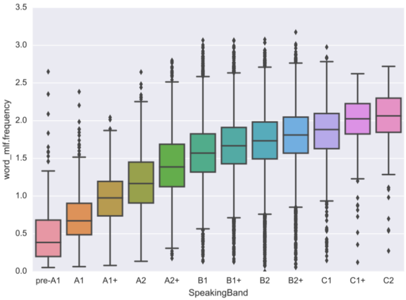


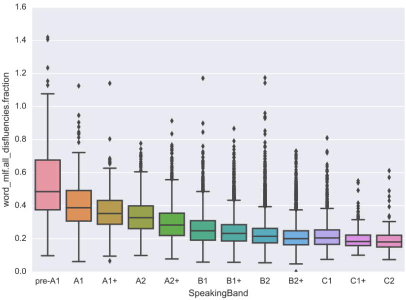
- Hand-craft grader input features to optimise assessment

Baseline Grader Features [11]

- Baseline features mainly fluency based, including:
- Audio Features: statistics about e.g.
  - fundamental frequency (F0)
  - speech energy and duration
- Aligned Text Features: statistics about e.g.
  - silence durations
  - number of disfluencies (`um`, `uh` etc)
  - speaking rate
- Text Features: statistics about e.g.
  - number of repeated words (per word)
  - number of unique word identities

Baseline Features: Correlation with Grades

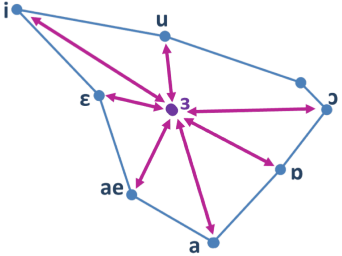




- Examine distribution of extracted features with grade
  - example box-plots for speaking rate and percentage disfluencies

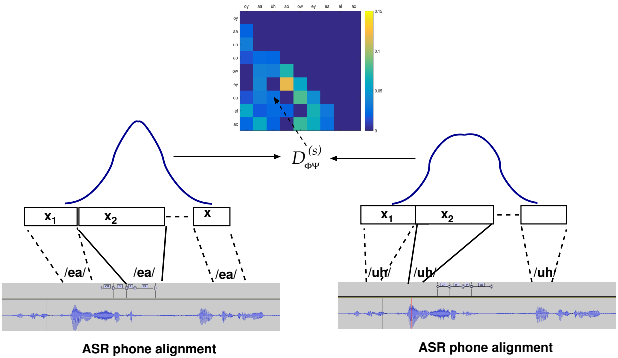
Derived Features: e.g. Phone-Distances [8]

- Pronunciation is an important predictor of proficiency
  - but no reference native speech for free speaking tasks
- Phone distance features are one approach



  - each phone characterised relative to others
  - independent of speaker attributes
  - characterise speaker's pronunciation of each phone

Model-based Pronunciation Features [4]

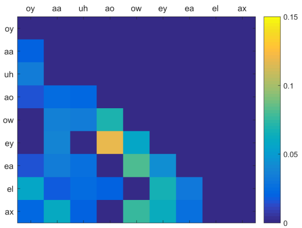


- Train Gaussian model for each phone $\boldsymbol{x}^{(i)}$ and speaker s:

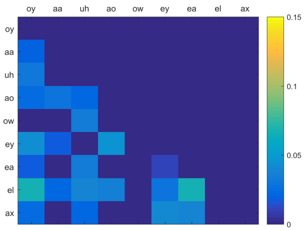
$$p(\boldsymbol{x}^{(i)}|\omega_\phi) = \mathcal{N}(\boldsymbol{x}^{(i)}; \boldsymbol{\mu}_\phi{}^{(s)}, \boldsymbol{\Sigma}_\phi{}^{(s)})$$

- Compute relative entropy between each phone-pair $\mathcal{D}_{\phi,\psi}{}^{(s)}$

Model-based Pronunciation Features



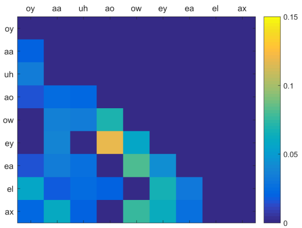
Candidate Grade A1



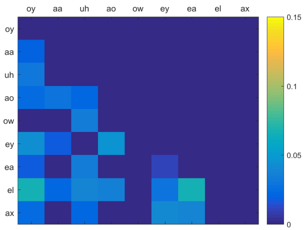
Candidate Grade C1

- Pair-wise entropies used as features in grader
  - yields small gains in assessment performance
  - pattern is first language (L1) dependent

Model-based Pronunciation Features



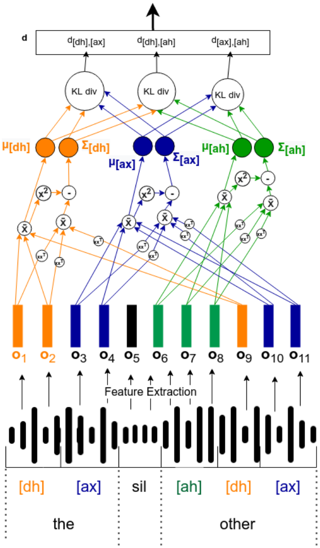
Candidate Grade A1



Candidate Grade C1

- Pair-wise entropies used as features in grader
  - yields small gains in assessment performance
  - pattern is first language (L1) dependent
- General approach $\Rightarrow$ tunable approach based on deep learning

Deep Learning Pronunciation Features [5]



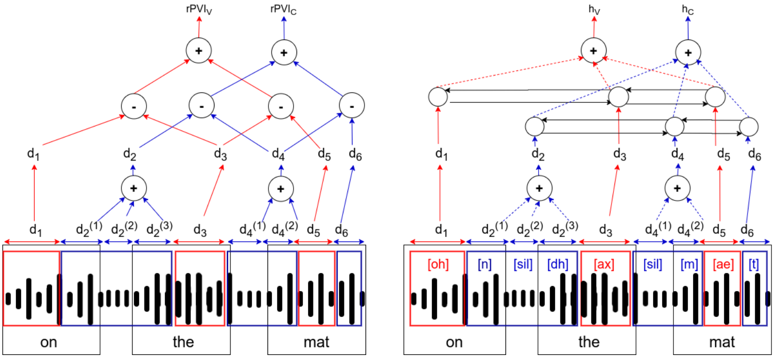
Derived Features: Rhythm

- Standard metrics developed based on durations (d_k)

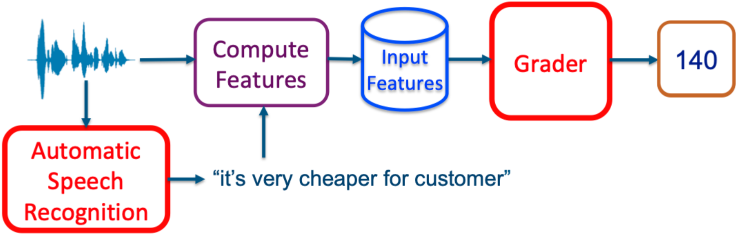
$$\texttt{rPVI} = \frac{1}{m-1}\sum_{k=1}^{m-1}|d_k - d_{k+1}|; \qquad \texttt{nPVI} = \frac{1}{m-1}\sum_{k=1}^{m-1}\frac{|d_k - d_{k+1}|}{(d_k + d_{k+1})/2}$$

  - added as simple features for assessment

Deep Learning Rhythm Features [6]

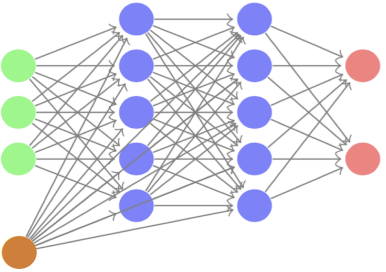


Grader



- Supervision data for assessment is a score
  - assessment run as a regression task: $p(y|\boldsymbol{x}^{\star};\boldsymbol{\theta})$
- For practical use also want to know how trustworthy prediction is

Deep Density Network-based Grader [1, 7]



- Deep Density Networks predict parameters of a distribution

$$p(y|\boldsymbol{x}^{\star};\boldsymbol{\theta}) = \mathcal{N}(y; f_{\mu}(\boldsymbol{x}^{\star};\boldsymbol{\theta}), f_{\sigma}(\boldsymbol{x}^{\star};\boldsymbol{\theta}))$$

  - flexible framework for any form of distribution
  - distribution variance gives measure of confidence in assessment

Ensembles of Models

- Deep learning optimisation is highly complex
  - multiple local minima in cost function
  - not possible to obtain the best model

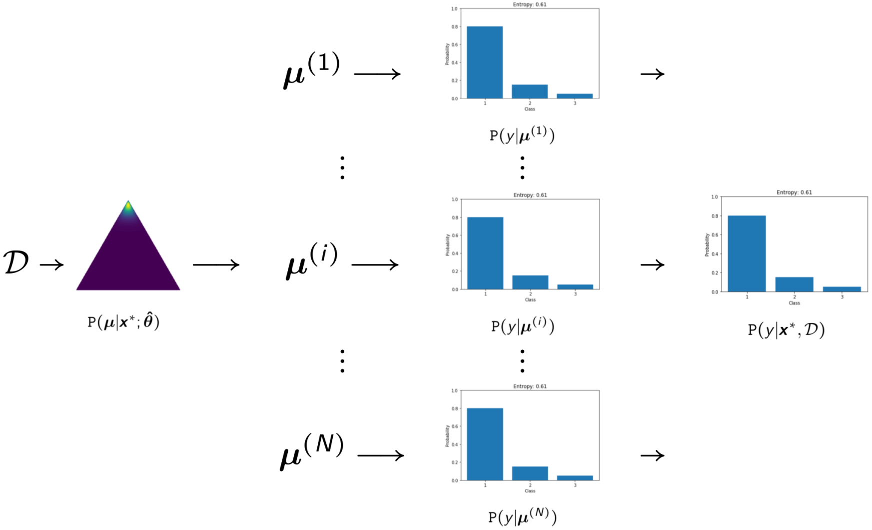
- Simple solution - train multiple models - an ensemble
  - average the prediction from the members of the ensemble
  - also useful for score confidence

DDN Assessment Performance

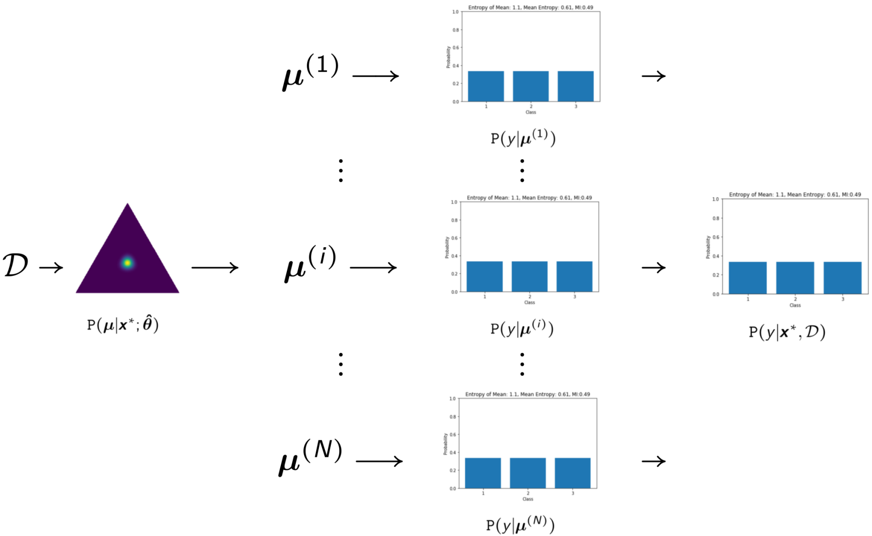
Model	PCC	MSE	MAE	%<0.5	%<1.0
Single	$0.885_{\pm 0.7}$	$0.32_{\pm 0.02}$	$0.43_{\pm 0.01}$	$67.8_{\pm 2.6}$	$93.7_{\pm 1.6}$
Ensemble	0.888	0.31	0.43	68.2	94.2

- BULATS data - “expert” grades, 225 speakers, 6 L1s

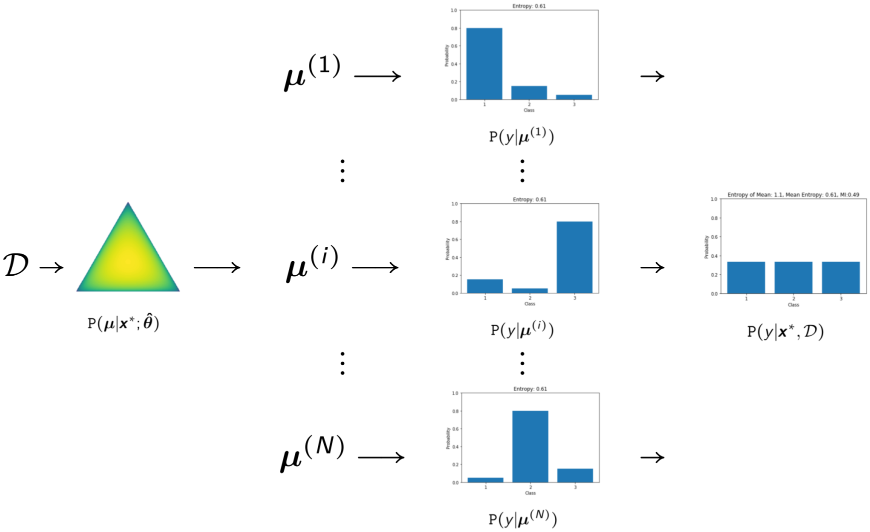
Ensemble Score Confidence



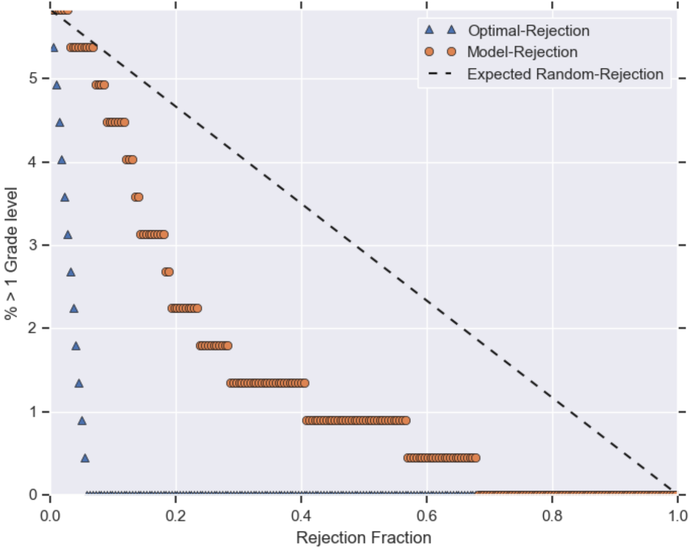
Ensemble Score Confidence



Ensemble Score Confidence

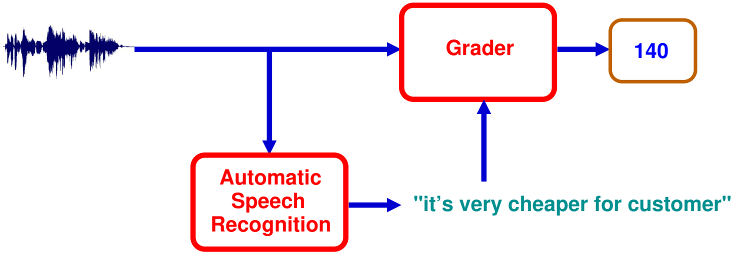


Detecting Outliers (candidates > 1.0 error)



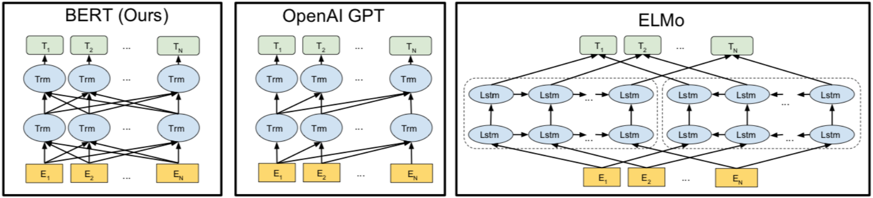
Neural Assessment

Assessment Framework



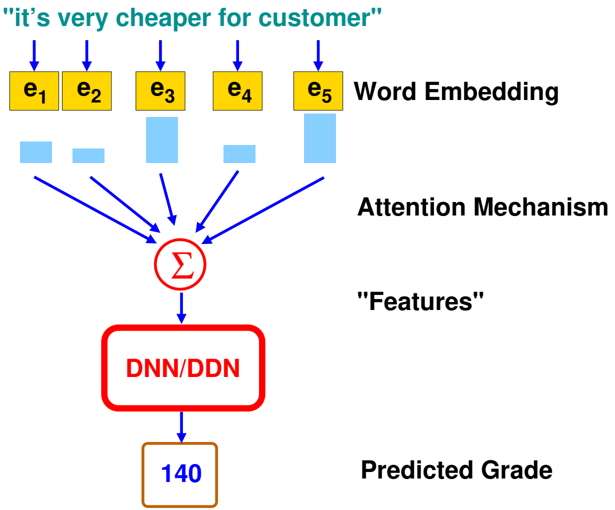
- Expert (handcrafted) features good, but are they optimal?
- Use deep-learning to map from ASR/audio to grade
  - network extracts trainable (optimal?) features from text/audio
  - needs to be able to handle variable length nature of audio/text

Text Processing: Word Embeddings



- First stage is to map from discrete words to continuous vector
  - word-embeddings very popular at the moment
  - use BERT - trained on large amounts of text data

Text: “Vanilla” Neural Assessment



(Text) Neural Assessment Performance

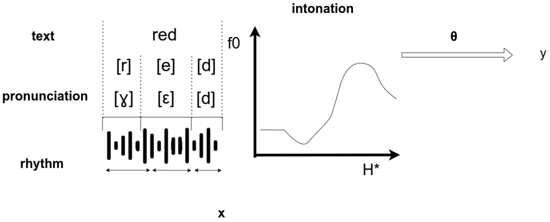
Model	PCC	MSE	MAE	%<0.5	%<1.0
DDN (All)	0.888	0.31	0.43	68.2	94.2
Neural (Text)	0.879	0.34	0.44	68.2	91.4

- Ensemble systems
- Good performance but weak on validity and reliability

Multi-view Assessment

Holistic Grader

- Input $\boldsymbol{x}$ is mapped to holistic score y by model with parameters $\boldsymbol{\theta}$

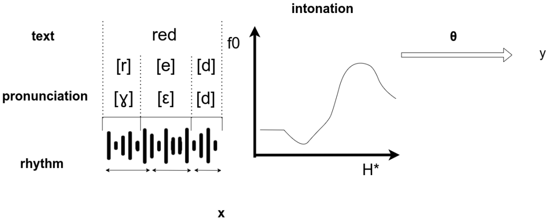


- Holistic proficiency, y, captures overall communicative competence

  e.g. *the candidate can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible (CEFR B2)*

Holistic Grader

- Input $\boldsymbol{x}$ is mapped to holistic score y by model with parameters $\boldsymbol{\theta}$

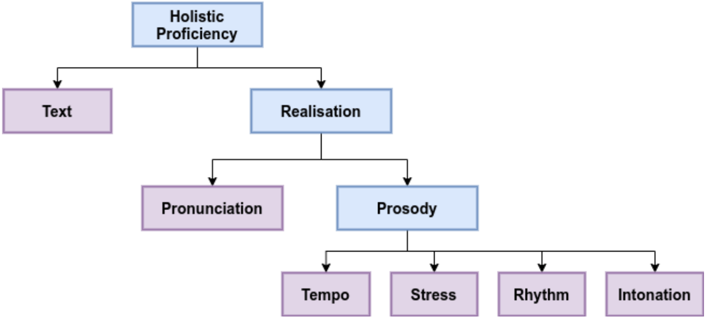


- Holistic proficiency, y, captures overall communicative competence

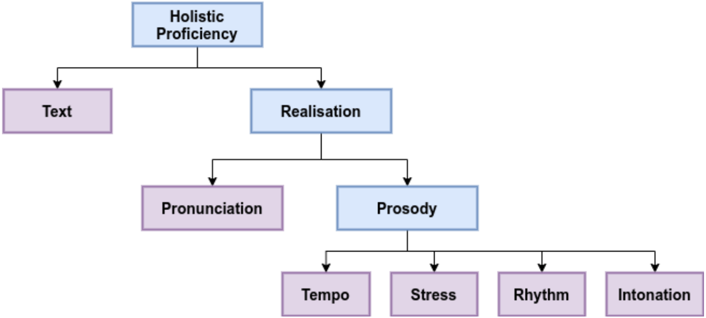
  e.g. *the candidate can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible (CEFR B2)*

- Can we assesss proficiency in a more interpretable way?
  - Give candidate useful feedback to help them improve

Multi-view Assessment and Feedback [3]

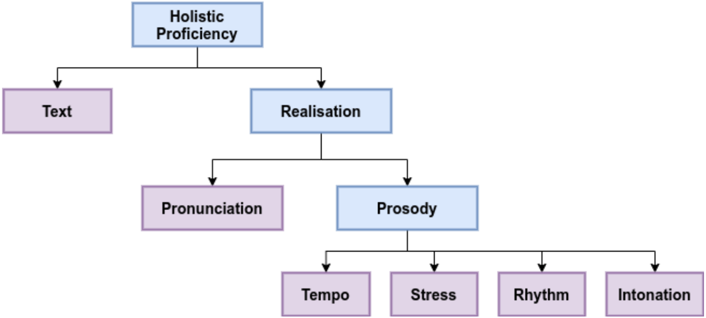


Multi-view Assessment and Feedback [3]



- Single-view proficiency, y_j (e.g. y_{text}, y_{rhythm}), captures one aspect
  - e.g. **Rhythm**: pattern of durations of speaker's words and phones

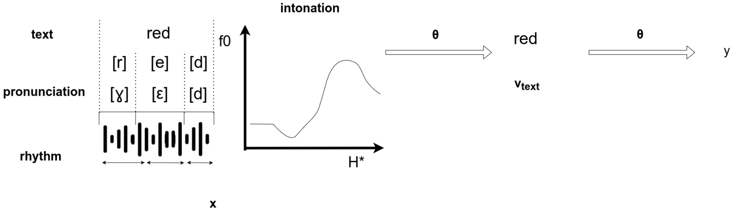
Multi-view Assessment and Feedback [3]



- Single-view proficiency, y_j (e.g. y_{text}, y_{rhythm}), captures one aspect
  - e.g. **Rhythm**: pattern of durations of speaker's words and phones
- Build single view graders: combine for multi-view assessment
  - Challenge: only holistic grades available for training

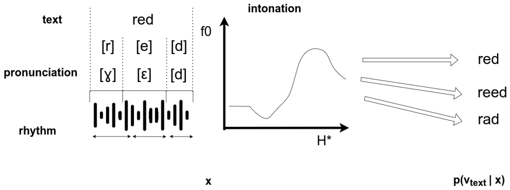
Single-view grading

- To force single-view grading want to limit information to one view
- Add an initial projection $\boldsymbol{x} \rightarrow \boldsymbol{v}_j$
  - to extract information about view j from $\boldsymbol{x}$
  - discard information about other views



Two-stage single-view grader

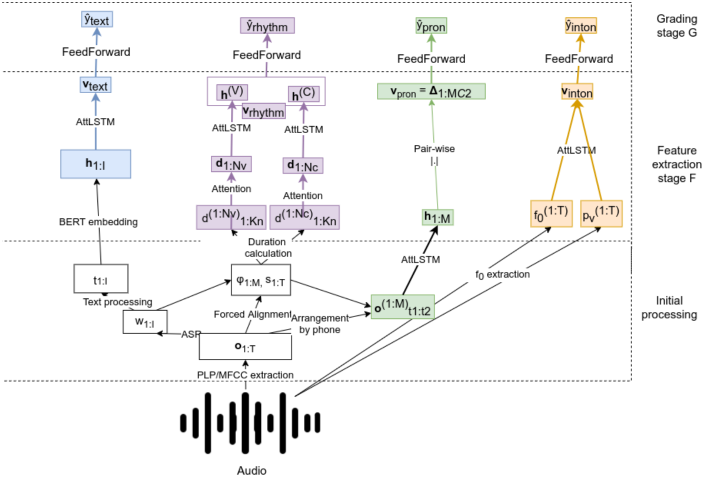
1. Extract a $\boldsymbol{v}_j$ from $\boldsymbol{x}$ according to a distribution $p(\boldsymbol{v}_j|\boldsymbol{x};\boldsymbol{\theta})$:



2. Then map each $\boldsymbol{v}_j$ to a y with $p(y|\boldsymbol{v}_j;\boldsymbol{\theta})$ s.t. for the full grader:

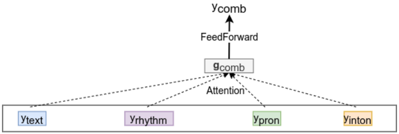
$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \int p(y|\boldsymbol{v}_j;\boldsymbol{\theta})\ p(\boldsymbol{v}_j|\boldsymbol{x};\boldsymbol{\theta})\ d\boldsymbol{v}_j$$

Single-view graders



Multi-view Grader Combination

- Use attention to combine the single-view scores



$$\hat{\hat{y}} = \sum_{j=1}^{J} \alpha_j \hat{y}_j$$

$$\textit{where} \qquad \alpha_j = \frac{\exp(s_j)}{\sum_{n=1}^{I} \exp(s_j)} \qquad s_j = \mathcal{A}(\boldsymbol{v}_j, \boldsymbol{\theta})$$

- Can train on its own or end-to-end with the single-view graders

Single-view Assessment Performance

Grader	PCC	MSE	MAE	%<0.5
holistic	0.888	0.31	0.43	68.2
text	0.820	0.46	0.51	60.7
pron	0.820	0.53	0.57	53.6
rhythm	0.819	0.54	0.58	49.6
intonation	0.826	0.44	0.49	60.7

Similarity of Single-view Graders Predictions

	text	pron	inton	rhythm
text	1.000			
pron	0.638	1.000		
inton	0.588	0.653	1.000	
rhythm	0.613	0.699	0.690	1.000

Kendall's τ between single-view grader predictions

Multi-view Assessment Performance

Grader	PCC	MSE	MAE	%<0.5
holistic	0.888	0.31	0.43	68.2
multi-view	0.881	0.36	0.47	64.2

- Multi-view performance shows single-view graders complementary

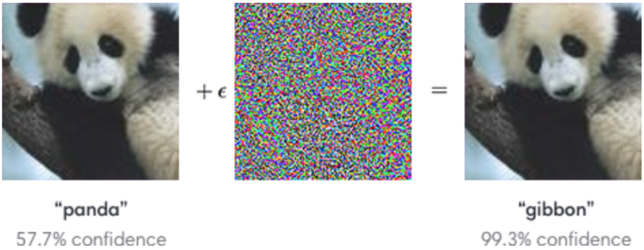
Robustness

Robustness Issues

- L1 Speech Detection
- Speaker Verification
- Off-Topic Response Detection
- Spoken Language Adversarial Attacks and Detection

Adversarial Attacks

- Image adversarial attacks popular/important research area

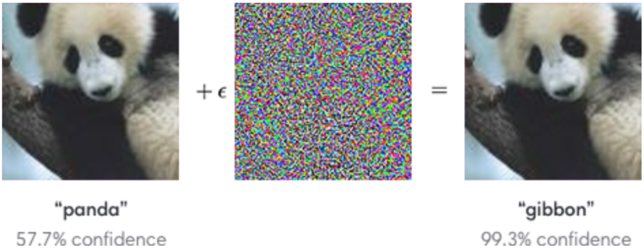


- increasing work for text and ASR attacks as well

What is the equivalent for spoken language assessment?

Adversarial Attacks

- Image adversarial attacks popular/important research area



  - increasing work for text and ASR attacks as well

    What is the equivalent for spoken language assessment?

- Add a phrase to the end of a response that increases score

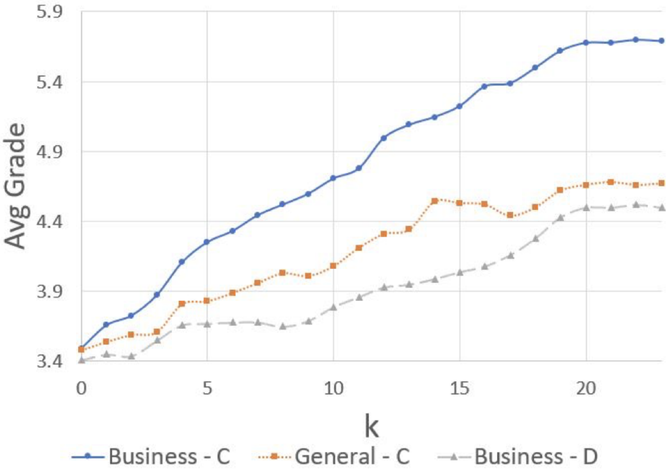
Spoken Language Assessment Attacks [9]

- Add a phrase to a user response (BULATS part 3 used)

  <user response> offensively obese astronauts amazingly ...

- Add a phrase to a user response (BULATS part 3 used)

  <user response> offensively obese astronauts amazingly ...

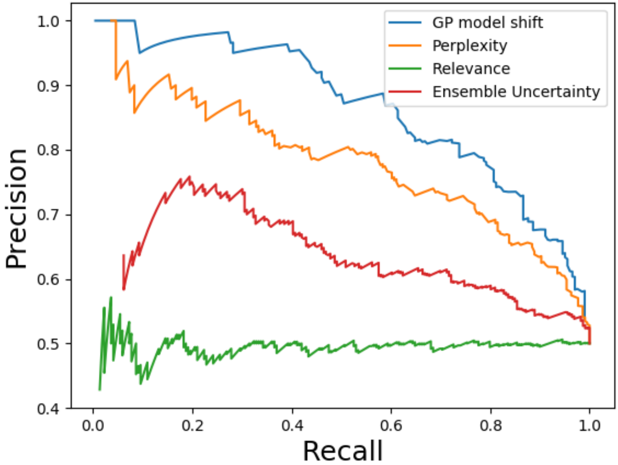


Adversarial Attack Performance (6 words) [9]

Grader (+adv)	Score	PCC	RMSE	%<0.5	%<1.0
Ensemble	3.49	0.749	0.727	59.9	83.2
+ adversarial	4.33	0.700	1.110	27.2	62.9

- Increase average score by 0.9 using 6 words

Adversarial Attack Detection (6 words) [9]



Conclusions

Conclusions

- Spoken language learning and assessment important
  - increasing need for automated (and validated) systems
  - auto-marked free speaking systems now live
- Deep learning is central to current state-of-the-art systems
  - Need to factor in interpretability & robustness to adversarial attacks
- Next steps:
  - Providing more feedback - lack of annotated data a big challenge
  - Assessment of conversational speaking tests

Thanks

- Thanks to Cambridge Assessment, University of Cambridge for supporting this research.
- Thanks to the CUED ALTA Speech Team for their contributions: Prof. Mark Gales, Xie "Jeff" Chen, Rogier van Dalen, Kostas Kyriakopoulos, Adian Liusie, Yiting Lu, Andrey Malinin, Potsawee Manakul, Vatsal Raina, Vyas Raina, Anton Ragni, Linlin Wang, Yu Wang, Xizi Wei, Xixin Wu ...
- http://mi.eng.cam.ac.uk/~mjfg/ALTA/index.html

Speak and Improve: https:speakandimprove.com

- Current beta of free speaking web-application
  - collaboration between ALTA, Cambridge Assessment and Industrial partners

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer Verlag, 2006.

[2] X. Chen, X. Liu, Y. Wang, A. Ragni, J. H. M. Wong, and M. J. F. Gales, "Exploiting future word contexts in neural network language models for speech recognition," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 27, no. 9, pp. 1444–1454, 2019. [Online]. Available: https://doi.org/10.1109/TASLP.2019.2922048

[3] K. Kyriakopoulos, "Deep learning for automatic assessment and feedback of spoken english," Ph.D. dissertation, Cambridge University, 2021.

[4] K. Kyriakopoulos, M. Gales, and K. Knill, "Automatic characterisation of the pronunciation of non-native English speakers using phone distance features," in *Proceedings of Workshop on Speech and Language Technology for Education (SLaTE)*, 2017.

[5] K. Kyriakopoulos, K. Knill, and M. J. F. Gales, "A deep learning approach to assessing non-native pronunciation of english using phone distances," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, 2018, pp. 1626–1630. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-1087

[6] K. Kyriakopoulos, K. M. Knill, and M. J. F. Gales, "A deep learning approach to automatic characterisation of rhythm in non-native english speech," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*. ISCA, 2019, pp. 1836–1840. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-3186

[7] A. Malinin, A. Ragni, M. Gales, and K. Knill, "Incorporating uncertainty into deep learning for spoken language assessment," in *Proc. 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[8] N. Minematsu, S. Asakawa, and K. Hirose, "Structural representation of the pronunciation and its use for call," in *2006 IEEE Spoken Language Technology Workshop*, Dec 2006, pp. 126–129.

[9] V. Raina, M. J. F. Gales, and K. M. Knill, "Universal adversarial attacks on spoken language assessment systems," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*. ISCA, 2020, pp. 3855–3859. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-1890

[10] Y. Wang, J. H. M. Wong, M. J. F. Gales, K. M. Knill, and A. Ragni, "Sequence teacher-student training of acoustic models for automatic free speaking language assessment," in *2018 IEEE Spoken Language Technology Workshop,*

SLT 2018, Athens, Greece, December 18-21, 2018, 2018, pp. 994–1000. [Online]. Available: https://doi.org/10.1109/SLT.2018.8639557

[11] Y. Wang, M. J. F. Gales, K. M. Knill, K. Kyriakopoulos, A. Malinin, R. C. van Dalen, and M. Rashid, "Towards automatic assessment of spontaneous spoken english," *Speech Communication*, vol. 104, pp. 47–56, 2018.

[12] J. Xu, M. Brenchley, E. Jones, A. Pinnington, T. Benjamin, K. Knill, G. Seal-Coon, M. Robinson, and A. Geranpayeh, "Linguaskill - building a validity argument for the Speaking test," 2020.