UNIVERSITY OF CAMBRIDGE

Cambridge ALTA
Institute for Automated Language Teaching and Assessment

End-to-end systems for L2 spoken English assessment and feedback

Stefano Bannò

ALTA Institute, Department of Engineering, University of Cambridge

14th March 2025

ALTA SLP Project Team

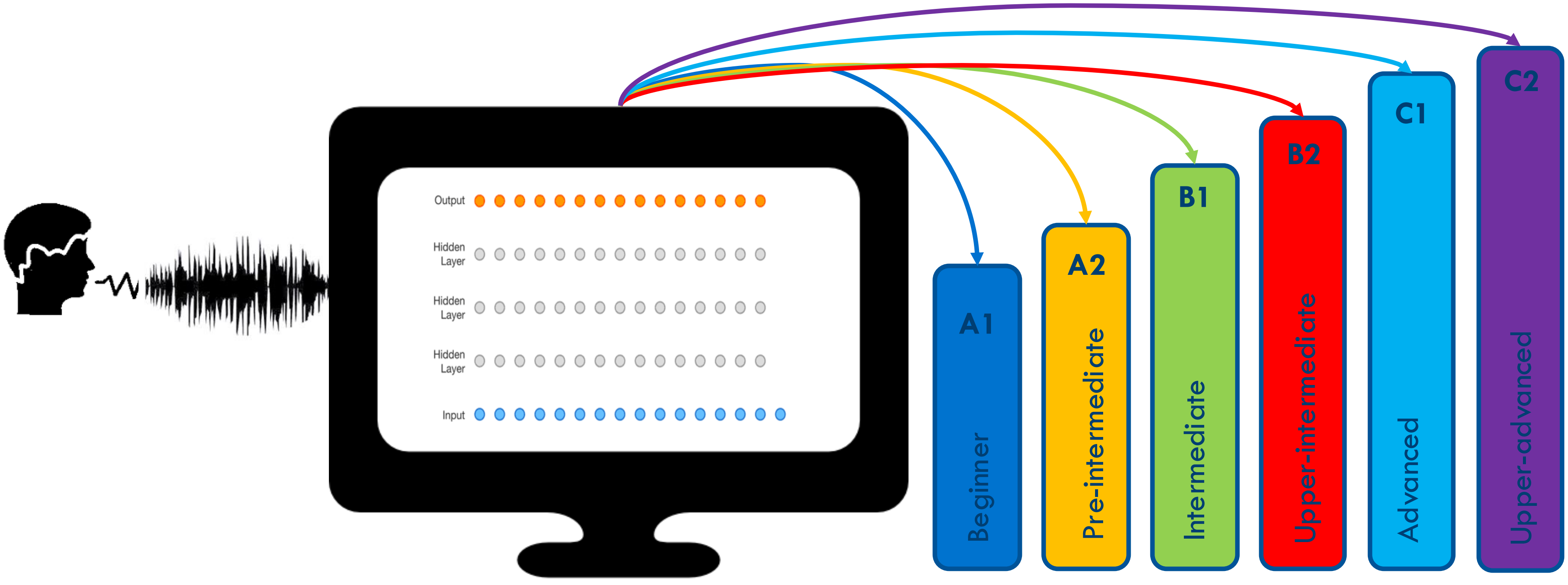
- Principal Investigators: Dr Kate Knill, Prof Mark Gales
- Postdocs: Dr Stefano Bannò, Dr Penny Karanasou, Dr Mengjie Qian
- Research Assistant: Siyuan Tang
- PhD students: Yassir Fathullah, Adian Liusie, Rao Ma, Charles McGhee, Vatsal Raina, Vyas Raina
- 4th year Engineering students

- Webpage: http://mi.eng.cam.ac.uk/~mjfg/ALTA/index.html

Overview

- **Spoken Language Assessment and Feedback**
- **End-to-End Spoken Language Assessment**
    - Why End-to-End Spoken Language Assessment?
    - Proposed Methods
    - Data and Evaluation Metrics
    - Experimental Results
    - Conclusions and Future Work
- **End-to-End Spoken Grammatical Error Correction**
    - What is Spoken Grammatical Error Correction (GEC)?
    - Proposed Methods
    - Data and Evaluation Metrics
    - Experimental Results
    - Feedback Analysis
    - Conclusions and Future Work
- **Discussion and Future Work**

Spoken Language Assessment and Feedback

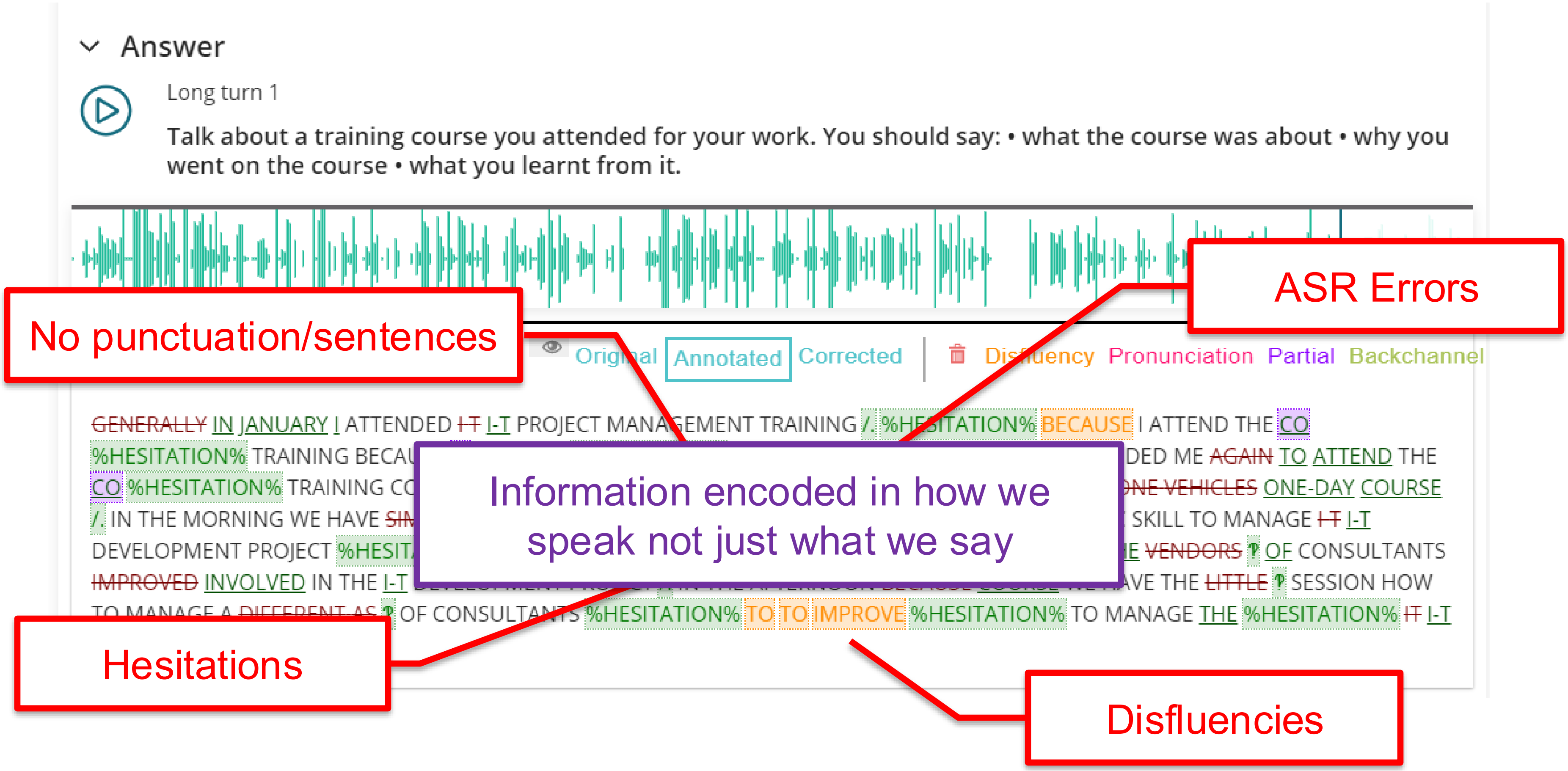


Spoken Language Assessment and Feedback

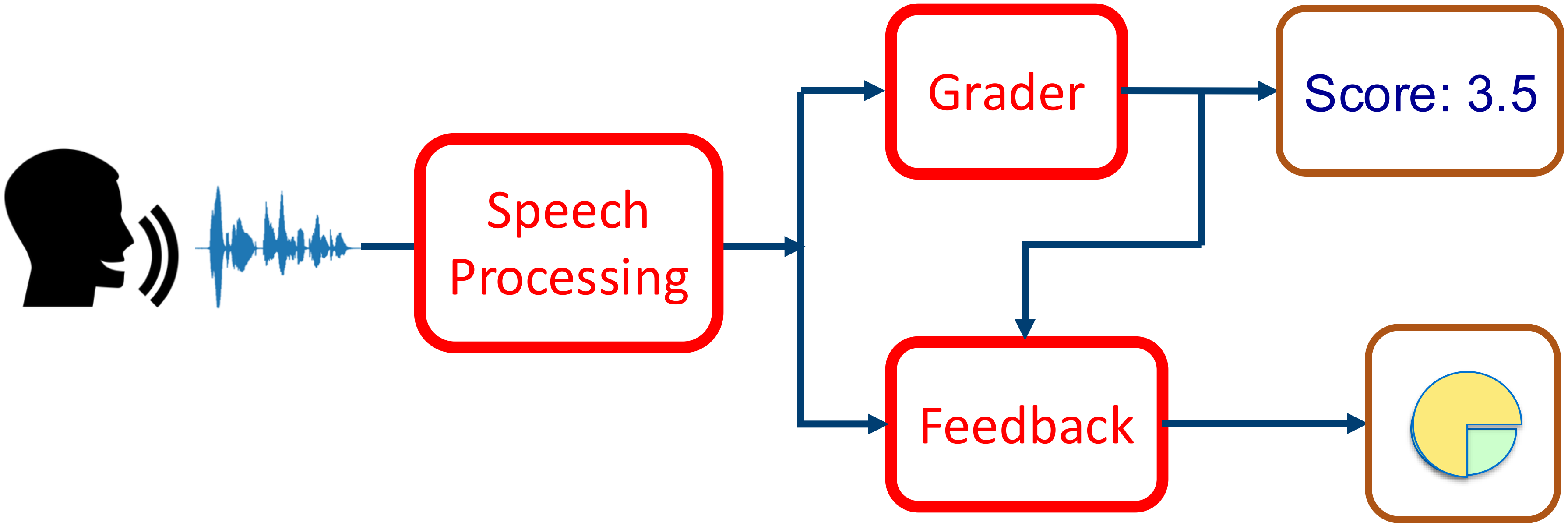
- Almost 2 billion people worldwide use and/or are learning English as a second language
  - Not enough teachers or examiners
  - Automated assessment and Computer-Assisted Language Learning (CALL) systems play an important role

- Speaking is key skill for communication
  - Many systems ignore or heavily restrict speech input – not testing communication

L2 learner speech is challenging!



Spoken Language Assessment and Feedback



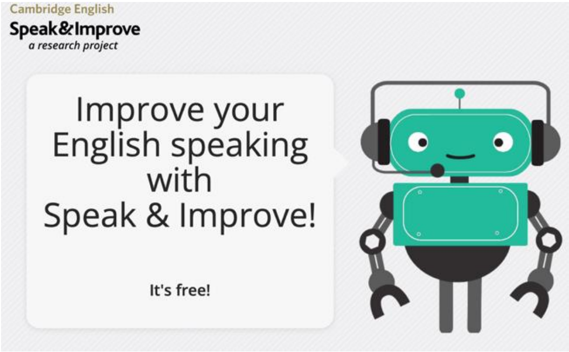
- Holistic – overall feedback across all speech
- Analytic – fine-grained feedback on specific elements in words/phrases (grammar, fluency, pronunciation, etc.)

Spoken Language Assessment and Feedback

Linguaskill

Upskill from Cambridge

>300k SUBMISSIONS
April 2023



https://speakandimprove.com

> 150 COUNTRIES

> 400k CANDIDATES / VISITORS

>9M SUBMISSIONS
June 2022

- Achieved through medium to long-term research at ALTA SLPTP
  - with technology transfer and collaboration with CUP&A and technology partners

End-to-End Spoken Language Assessment

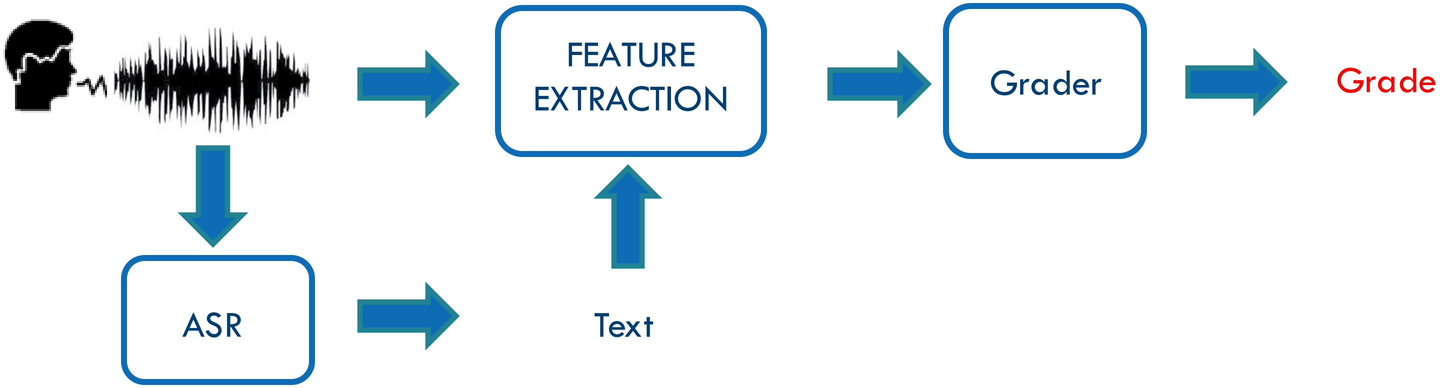
SLaTE 2023

Assessment of L2 Oral Proficiency Using Self-Supervised Speech Representation Learning

Stefano Bannò, Katherine M Knill, Marco Matassoni, Vyas Raina, Mark Gales

A standard pipeline for automated spoken language assessment is to start with an automatic speech recognition (ASR) system and derive features that exploit transcriptions and audio. Although efficient, these approaches require ASR systems that can be used for second language (L2) speakers and preferably tuned to the specific form of test being deployed. Recently, a self-supervised speech representation-based scheme requiring no ASR was proposed. This work extends the initial analysis to a large-scale proficiency test, Linguaskill. The performance of a self-supervised, wav2vec 2.0, system is compared to a high-performance hand-crafted assessment system and a BERT-based system, both of which use ASR transcriptions. Though the wav2vec 2.0 based system is found to be sensitive to the nature of the response, it can be configured to yield comparable performance to systems requiring transcriptions and shows significant gains when appropriately combined with standard approaches.

Why End-to-End Spoken Language Assessment?



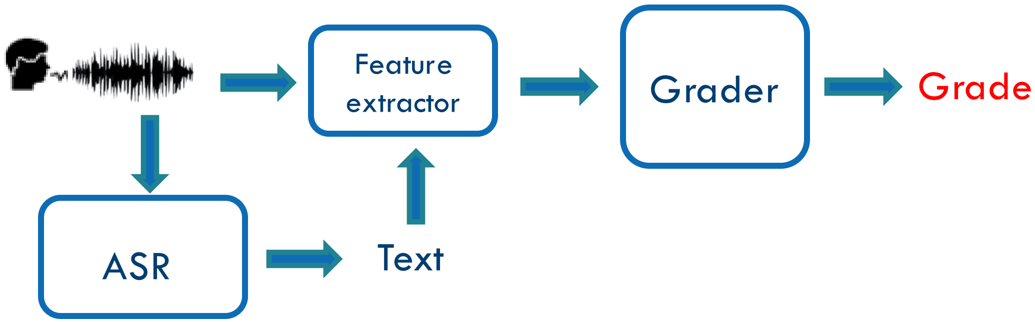
- Efficacy of **handcrafted features** relies on their particular underlying assumptions and they **risk discarding potentially salient information about proficiency**

- **ASR transcriptions** may not faithfully render the contents of learners' performances nor yield **any information about intonation, rhythm, fluency, and prosody**

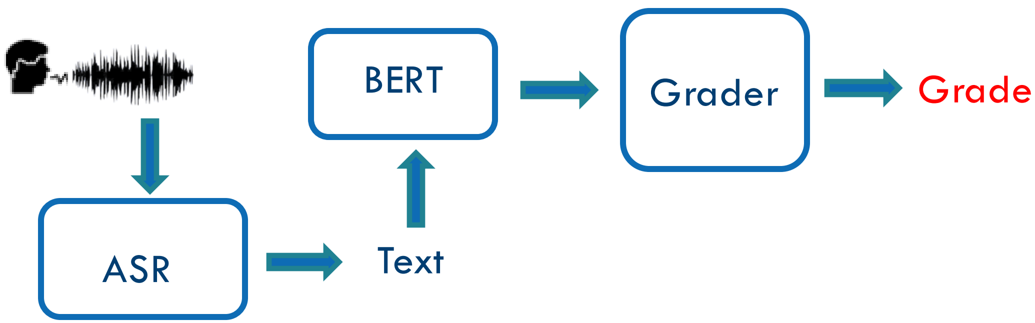
Proposed methods

- Following our preliminary work (Bannò & Matassoni, 2023), we compared three different systems:

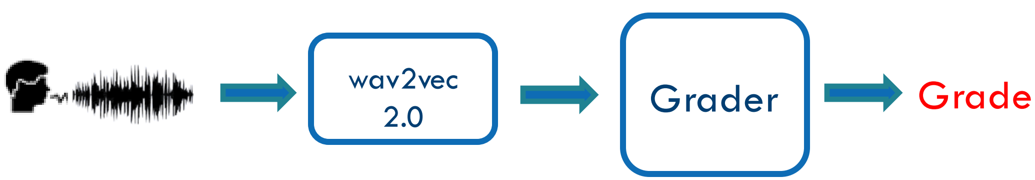
- feature-based



- BERT-based



- wav2vec2-based

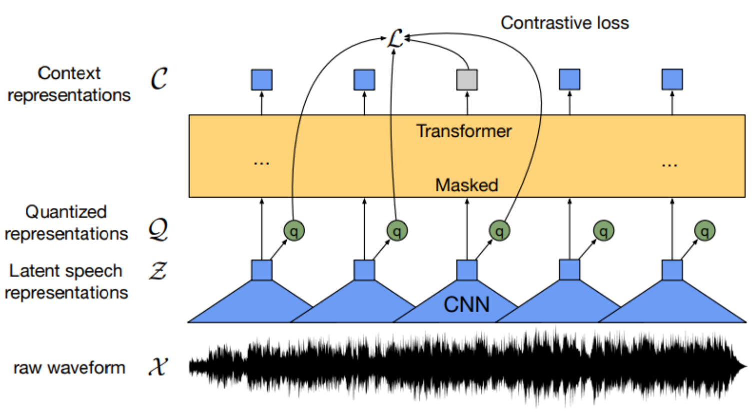


Foundation models for assessment (text)

- **BERT and similar models** have been massively applied to speech transcriptions for assessment (Craighead et al., 2020; Raina et al., 2020; Wang et al., 2021)
  - **Suitable** for assessing content-related, lexical, and – to a certain extent – grammatical elements of learners' productions.
  - **Not suitable** for assessing acoustic-related information, e.g., fluency and pronunciation.

Foundation models for assessment (speech)

- **Speech foundation models such as wav2vec 2.0 and HuBERT** were initially investigated for mispronunciation detection and diagnosis (Peng et al., 2021; Wu et al., 2021; Xu et al., 2021) and pronunciation assessment only (Kim et al., 2022)
  - **Not suitable** (?) for assessing content-related, lexical, and grammatical elements of learners' productions
  - **Suitable** for assessing acoustic-related information, e.g., fluency and pronunciation.



Data

- **Linguaskill data** obtained from Cambridge University Press & Assessment
  - **Training set:** 31475 speakers
  - **Dev set** (also used as calibration set)**:** 1033 speakers
  - **Two test sets**, **LinGen** (General English) and **LinBus** (Business English): 1049 and 712 speakers, respectively.
  - Sets feature around **30 L1s** and are balanced for gender and proficiency level from **1 to 6** (CEFR ~A1 to C)
  - Exam is divided into **5 parts**. Parts 1 and 5 include short answers (10-20 seconds), Part 2 contains read speech, and Parts 3 and 4 include long turns (around 1 minute)

Linguaskill from Cambridge

Evaluation metrics

To measure the average magnitude of prediction errors:

- Root-mean-square error (**RMSE**)

To evaluate the linear relationship between predicted and actual scores:

- Pearson's correlation coefficient (**PCC**)

To evaluate the strength and direction of the monotonic relationship:

- Spearman's rank coefficient (**SRC**)

To check the model's ability to make precise predictions:

- Percentage of the predicted scores that are equal to or lie within 0.5 (**% ≤ 0.5**) of the actual score.
- Percentage of the predicted scores that are equal to or lie within 1.0 (**% ≤ 1.0**) of the actual score.

Experimental results

LinGen

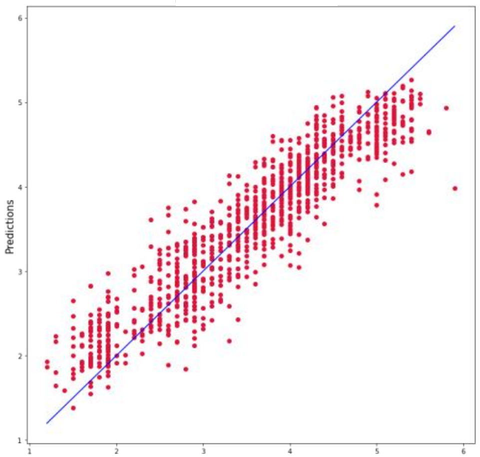
Model	PCC	SRC	RMSE	% ≤ 0.5	% ≤ 1.0
F.-based	0.932	0.937	0.383	**81.5**	98.6
BERT	0.929	0.934	0.395	80.3	98.5
w2v2	**0.934**	**0.938**	**0.383**	80.9	**99.0**
F+B+w	0.943	0.947	0.353	85.0	99.2

- The results for wav2vec 2.0 are different from the ones in the paper, where we used a mean pooling mechanism which was replaced by an attention mechanism afterwards.
- **F+B+w** consists of a linear regression model trained on the predictions of the dev data obtained from the three systems.
- The results on LinBus show very similar trends.

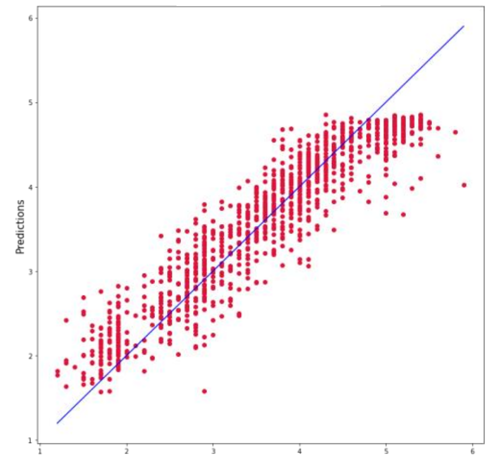
Experimental results

LinGen

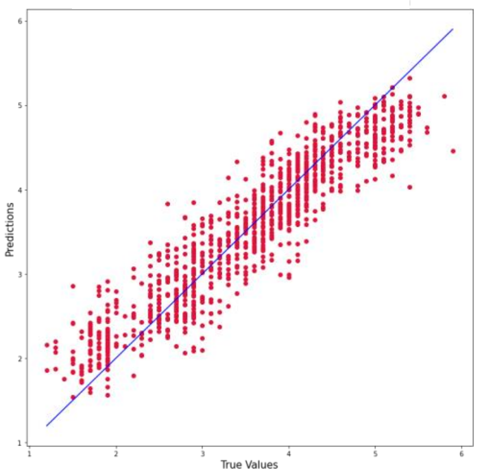
Feature-based



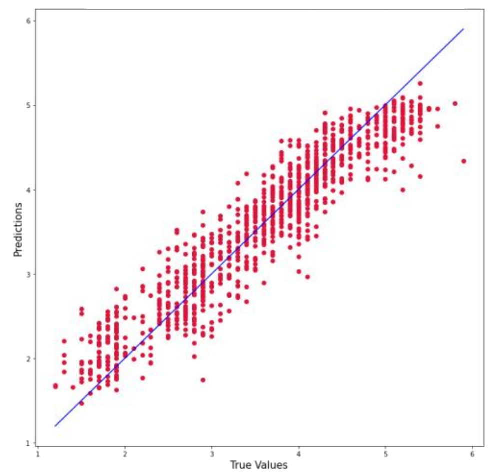
BERT



wav2vec 2.0



Feature-based+wav2vec 2.0+BERT



End-to-End Spoken Language Assessment – Conclusions and Future Work

- **Recap**
  - Compared three different speaking assessment systems: feature-based, BERT-based, and wav2vec2-based.
    - Wav2vec 2.0 achieves slightly better results than the other systems (**no need for transcriptions**!);
    - Combination of the three systems **boosts performance** and **enhances validity and explainability** of results as the feature-based grader can rely on explainable features.
    - Since holistic assessment also encompasses content-related aspects, **does this mean that wav2vec 2.0 is able to grasp information about them in addition to acoustic-related aspects**?
- **Future work**
  - We have recently used Whisper in a similar fashion and obtained promising results.
  - Use of **multi-modal (audio+text) LLMs** for holistic (and analytic) assessment

End-to-End Spoken Grammatical Error Correction

Conferences > ICASSP 2024 - 2024 IEEE Inter...

Towards End-to-End Spoken Grammatical Error Correction

Publisher: IEEE Cite This PDF

Stefano Bannò ; Rao Ma ; Mengjie Qian ; Kate M. Knill ; Mark J. F. Gales **All Authors**

Abstract

Document Sections

Authors

Abstract:
Grammatical feedback is crucial for L2 learners, teachers, and testers. Spoken grammatical error correction (GEC) aims to supply feedback to L2 learners on their use of grammar when speaking. This process usually relies on a cascaded pipeline comprising an ASR system, disfluency removal, and GEC, with the associated concern of propagating errors between these individual modules. In this paper, we introduce an alternative "end-to-end" approach to spoken GEC, exploiting a speech recognition foundation model, Whisper. This foundation model can be used to replace the whole framework or part of it, e.g., ASR and disfluency removal. These end-to-end approaches are compared to more standard cascaded approaches on the data obtained from a free-speaking spoken language assessment test, Linguaskill. Results demonstrate that end-to-end spoken GEC is possible within this architecture, but the lack of available data limits current performance compared to a system using large quantities of text-based GEC data. Conversely, end-to-end disfluency detection and removal, which is easier for the attention-based Whisper to learn, does outperform cascaded approaches. Additionally, the paper discusses the challenges of providing feedback to candidates when using end-to-end systems for spoken GEC.

What is Spoken GEC?

- Mastering grammar is a key aspect for L2 speakers
  - Grammatical errors are highly correlated with holistic proficiency
  - A poor grammatical proficiency impacts intelligibility, e.g., a typical error by Italian speakers:

Please translate: *Mi piace la pizza.*

What is Spoken GEC?

- Mastering grammar is a key aspect for L2 speakers
  - Grammatical errors are highly correlated with holistic proficiency
  - A poor grammatical proficiency impacts intelligibility, e.g., a typical error by Italian speakers:

Please translate: *Mi piace la pizza.*

Literally: The pizza appeals to me.

What is Spoken GEC?

- Mastering grammar is a key aspect for L2 speakers
    - Grammatical errors are highly correlated with holistic proficiency
    - A poor grammatical proficiency impacts intelligibility, e.g., a typical error by Italian speakers:

Please translate: *Mi piace la pizza.*

Literally: The pizza appeals to me.

Learner: The pizza likes me.

What is Spoken GEC?

- Mastering grammar is a key aspect for L2 speakers
  - Grammatical errors are highly correlated with holistic proficiency
  - A poor grammatical proficiency impacts intelligibility, e.g., a typical error by Italian speakers:

Please translate: *Mi piace la pizza.*

Literally: The pizza appeals to me.

Learner: The pizza likes me.

Correct: I like pizza.

What is Spoken GEC?

- Mastering grammar is a key aspect for L2 speakers
  - Grammatical errors are highly correlated with holistic proficiency
  - A poor grammatical proficiency impacts intelligibility, e.g., a typical error by Italian speakers:

Please translate: *Mi piace la pizza.*

Literally: The pizza appeals to me.

Learner: The pizza likes me.

Correct: I like pizza.

- Two errors: word order and unnecessary determiner.

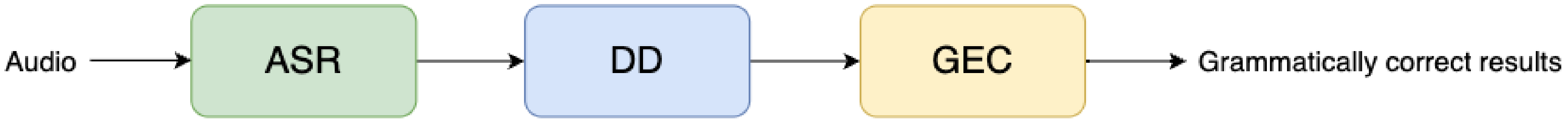
What is Spoken GEC?

- Grammatical error correction (**GEC**) is an established area of study, with several shared tasks organised in the last 15 years;
- Spoken GEC tackles the complex challenge of **correcting errors within spoken language**;
- Spoken language features **disfluencies**, such as hesitations, repetitions and false starts, which **make spoken GEC more difficult** than written GEC.

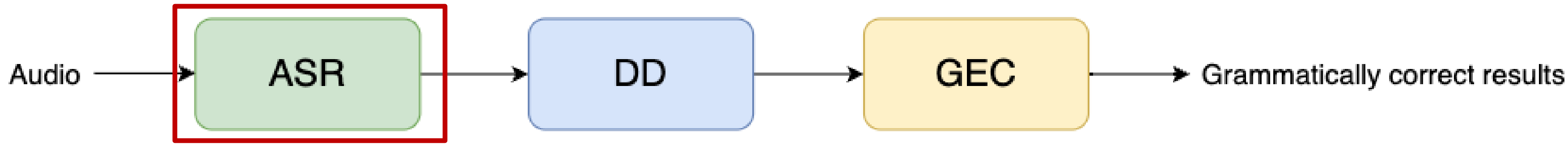
What is Spoken GEC?

- Aim of GEC is to produce grammatically correct sentences:
  - **Original:** Learning several languages is very better.
  - **Corrected:** Learning several languages is way better.
- Speech makes it more challenging:
  - **Original:** um learning several languages is very bi- better
  - **Corrected:** learning several languages is way better

Spoken GEC

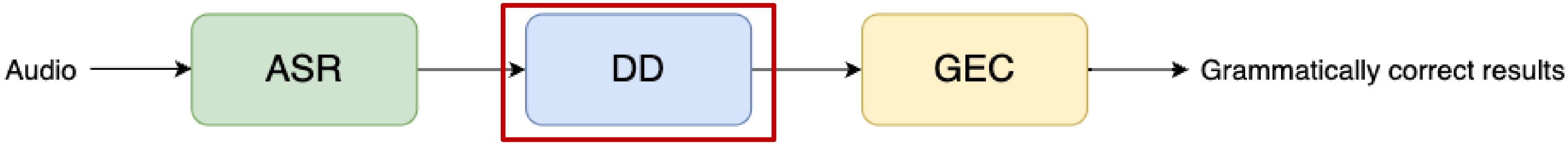


Spoken GEC



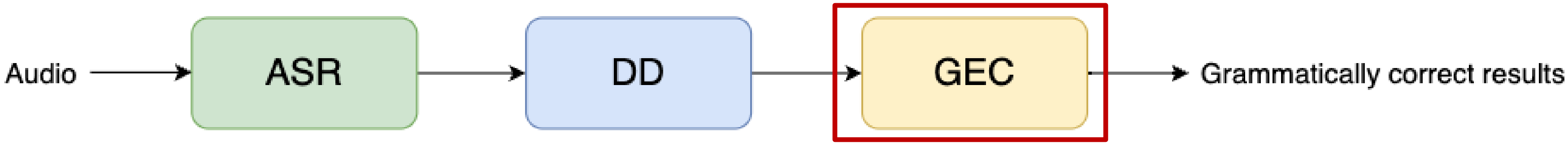
um learning several languages is very bi- better

Spoken GEC



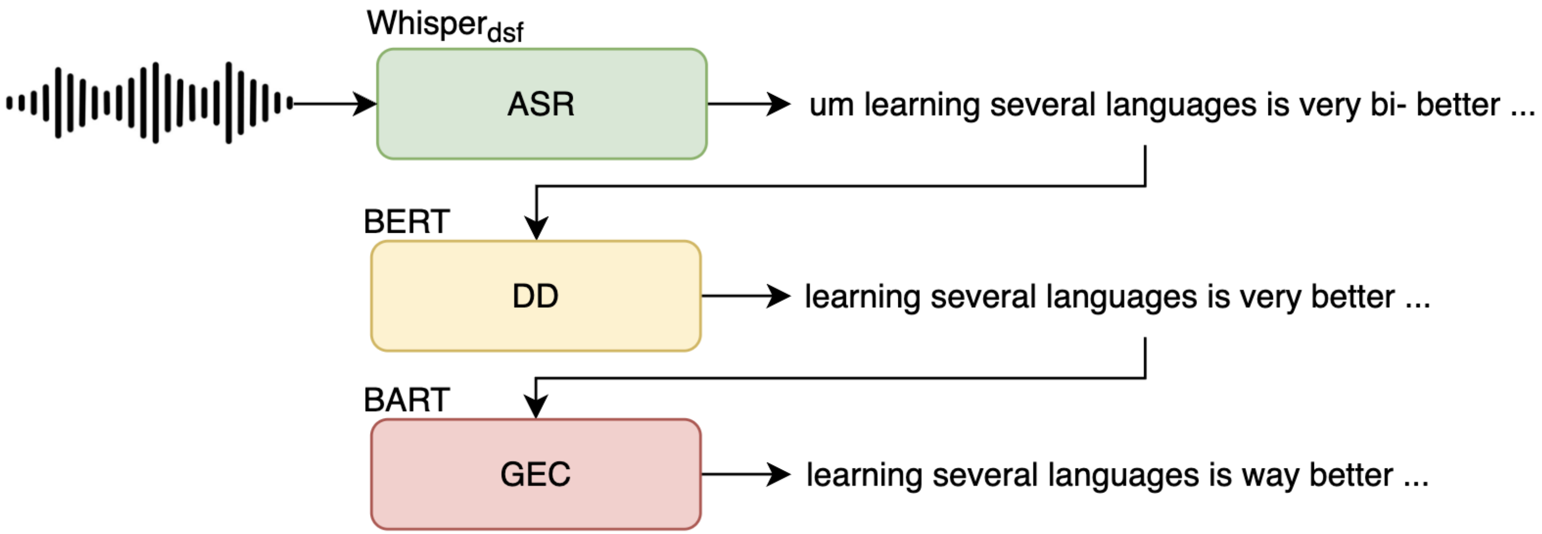
learning several languages is very better

Spoken GEC



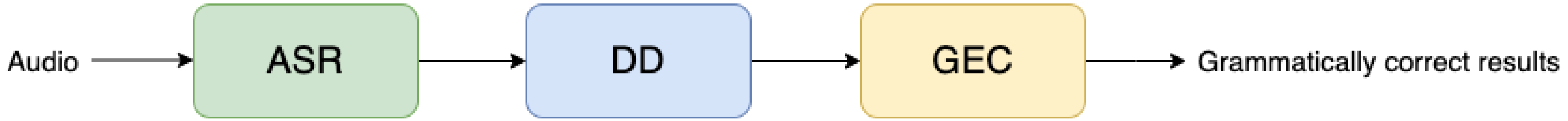
learning several languages is way better

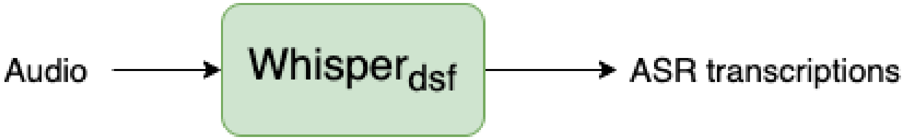
Cascaded System Issues



- ASR errors might propagate through the pipeline
- Loss of information (intonation, speaker info, emotion, etc.)
- Training-evaluation mismatch

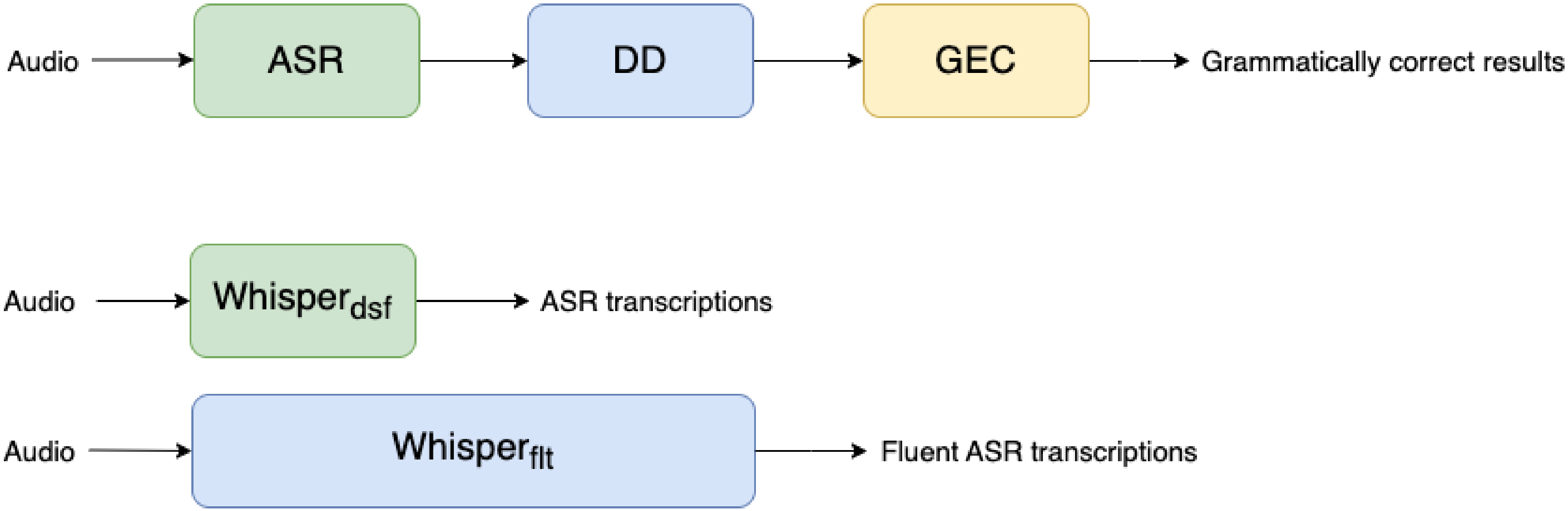
Whisper for Spoken GEC





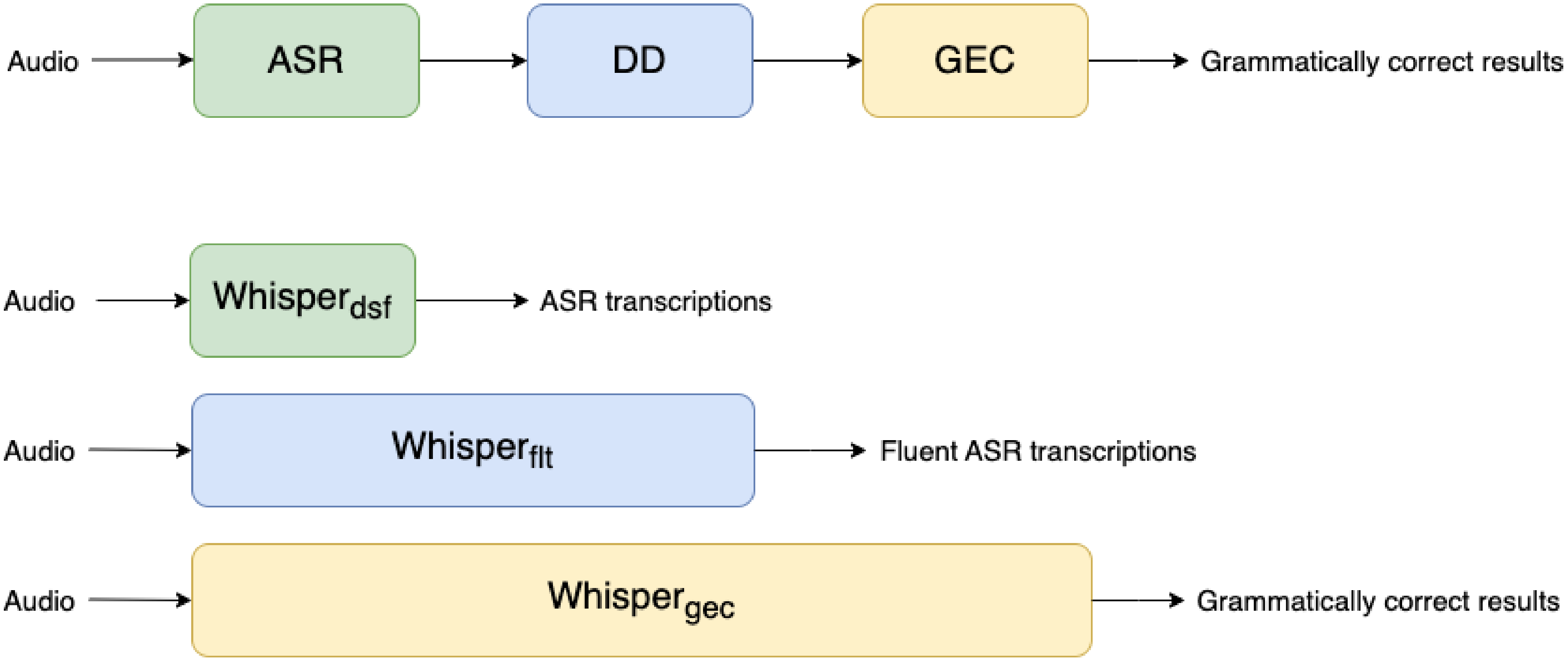
um learning several languages is very bi- better

Whisper for Spoken GEC



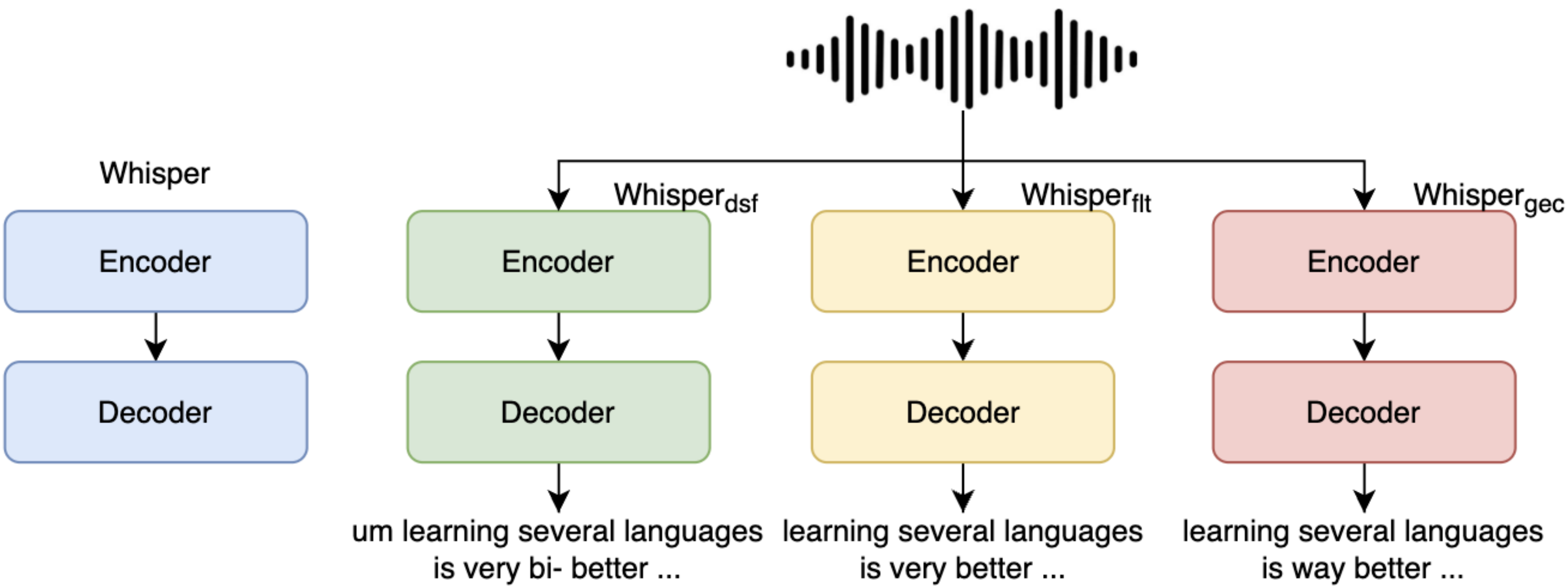
learning several languages is very better

Whisper for Spoken GEC



learning several languages is way better

Fine-tuning Whisper for Spoken GEC



- **Proposal:** Fine-tuning Whisper on three different sets of transcriptions separately to generate ASR transcriptions in different formats

Data

	Corpus	Split	Hours	Speakers	Utts/Sents	Words
Spoken	Switchboard	train	50.8	980	81,812	626K
		dev	3.8	102	5,093	46K
		test	3.7	100	5,067	45K
	Linguaskill	train	77.6	1,908	34,790	502K
		dev	7.8	176	3,347	49K
		test	11.0	271	4,565	69K
Written	EFCAMDAT +BEA-2019	train	-	-	2.5M	28.9M
		dev	-	-	25,529	293K

Linguaskill

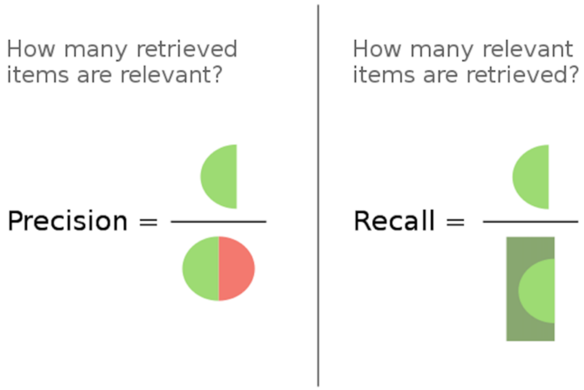
- Data obtained from Linguaskill examinations for L2 learners of English, provided by Cambridge University Press & Assessment
- Each speaker is graded on a scale from 2-6 based on CEFR (A2 to C)
- Each set balanced for gender, proficiency and L1s (around 30)
- Data have been: a) manually transcribed; b) annotated with disfluencies; c) annotated with grammatical error corrections

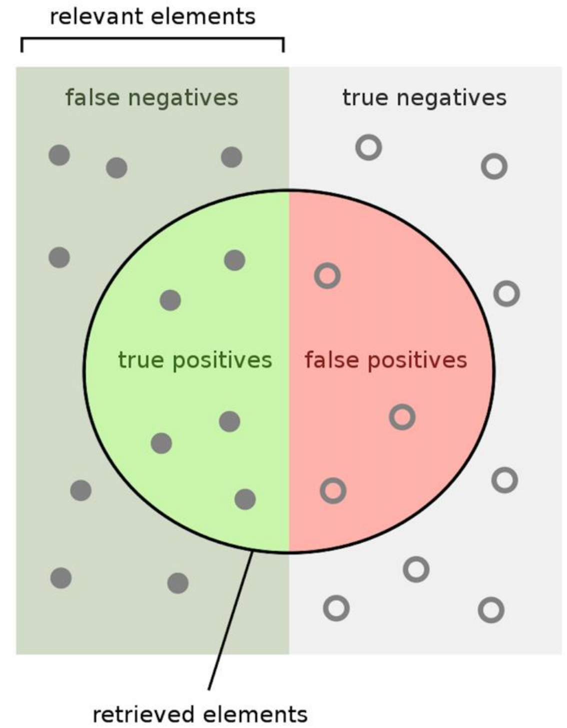
Model Setup

- DD (BERT):
  - Stage 1 fine-tuning: Switchboard
  - Stage 2 fine-tuning: Linguaskill
- GEC (BART):
  - Stage 1 fine-tuning: EFCAMDAT+BEA-2019
  - Stage 2 fine-tuning: Linguaskill
- Whisper_{dsf}, Whisper_{flt}, Whisper_{gec}:
  - Fine-tuning: Linguaskill

Evaluation Metrics

- Typically, ASR is evaluated using WER, while DD and GEC using Precision, Recall, and F scores:
  - Disfluency detection: F_1
  - Grammatical error correction: $F_{0.5}$





$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

N.B.: recall is **beta** times as important as precision!

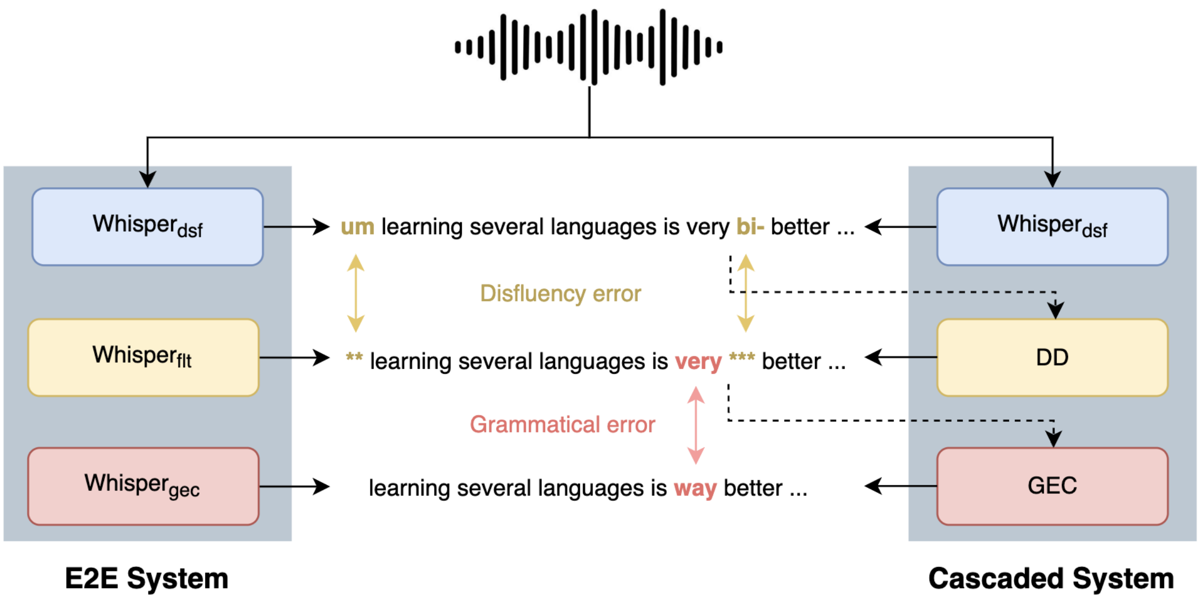
Evaluation Metrics

- Standard metrics for DD/GEC are challenging for spoken processing

- **Disfluency Detection** (DD):
    - ASR transcriptions do not have manual disfluency annotations
    - Use **WER**
- **Spoken Grammatical Error Correction** (GEC):
    - ASR errors might modify edits required to provide correct text
    - Use **WER** and **TER (translation edit rate)**

Evaluation Metrics

- However, standard metrics for DD/GEC are still useful (although still challenging!) **for feedback analysis**
- We don't want to give learners the corrected text only, but informative feedback as well!



WER of E2E Models based on Whisper

Model	dsf	flt	gec
$\text{Whisper}_{\text{dsf}}$	**5.92**	9.97	19.17
$\text{Whisper}_{\text{flt}}$	9.22	**5.77**	14.89
$\text{Whisper}_{\text{gec}}$	13.73	10.37	**13.49**

- Whisper models are trained on three tasks separately
  - Matching training to task achieves best performance

Disfluency Detection Performance

System	Model	flt
Cascaded	$\text{Whisper}_{\text{dsf}}$+DD	6.31
E2E	$\text{Whisper}_{\text{flt}}$	**5.77**

- E2E approach performs better than a cascaded system
- Attention mechanism in Whisper is able to learn to skip words
  - $\text{Whisper}_{\text{flt}}$ has learnt to skip disfluencies

Spoken GEC Performance

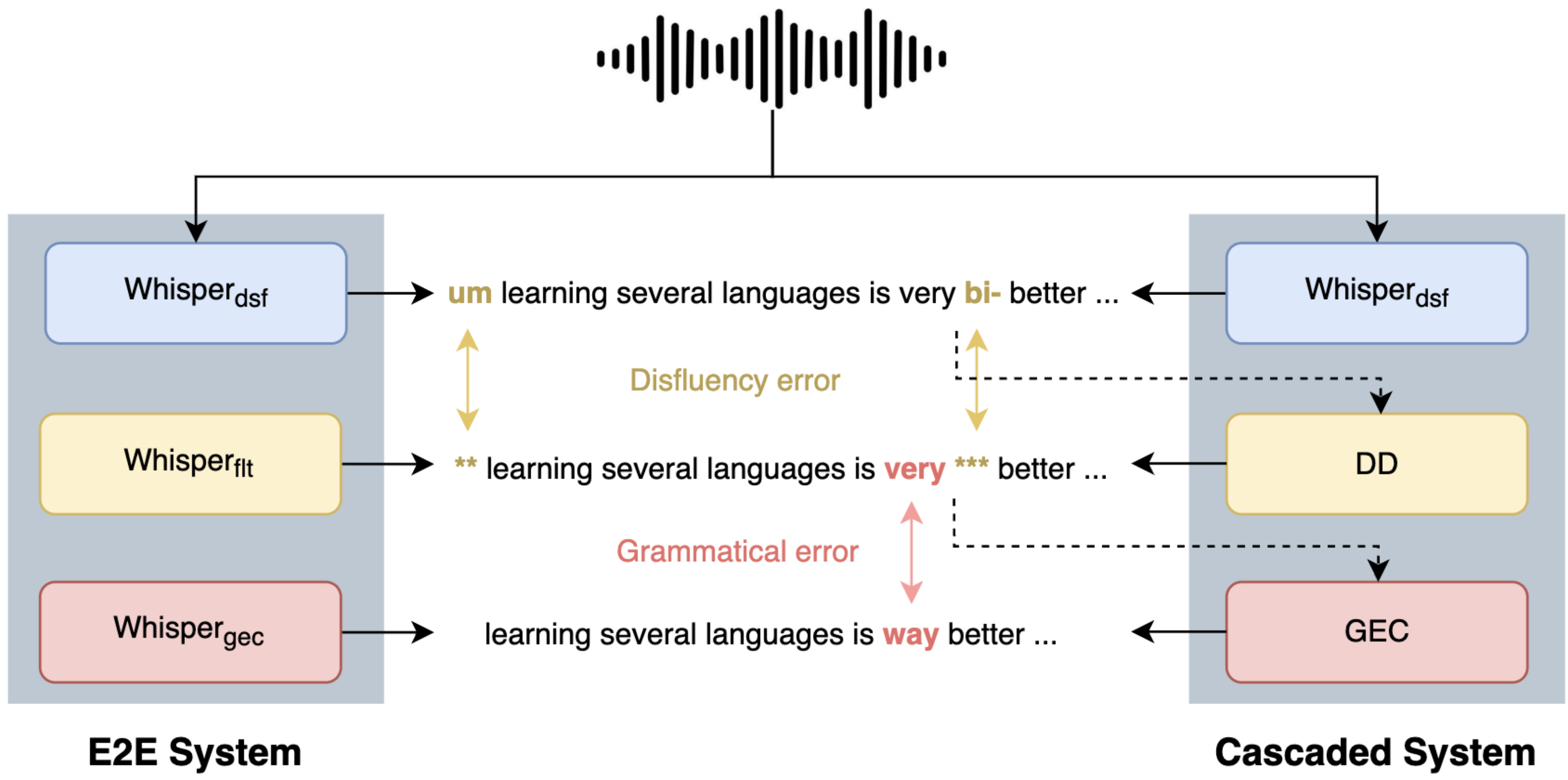
System	Model	gec WER	gec TER
Cascaded	$\text{Whisper}_{\text{dsf}}$+DD+GEC	13.34	12.96
	$\text{Whisper}_{\text{flt}}$+GEC	**12.96**	**12.54**
E2E	$\text{Whisper}_{\text{gec}}$	13.49	13.08

- Comparable performance compared to a fully cascaded system
- $\text{Whisper}_{\text{gec}}$ has learnt to 'translate' to correct text
- Problem: lack of available training data

Data for Spoken GEC

	Corpus	Split	Hours	Speakers	Utts/Sents	Words
Spoken	Switchboard	train	50.8	980	81,812	626K
		dev	3.8	102	5,093	46K
		test	3.7	100	5,067	45K
	Linguaskill	train	77.6	1,908	34,790	502K
		dev	7.8	176	3,347	49K
		test	11.0	271	4,565	69K
Written	EFCAMDAT +BEA-2019	train	-	-	2.5M	28.9M
		dev	-	-	25,529	293K

Feedback Analysis



Feedback Analysis for Spoken GEC

- We extract GEC edits using the ERRor ANnotation Toolkit (ERRANT)
  - Automatically extracts edits from parallel original and corrected sentences
  - Classifies them according to a dataset-agnostic rule-based framework
  - Facilitates error type evaluation at different levels of granularity

Auto:	the	cat	sit	on		mat
Ref:	the	cat	sat	on	the	mat
Edit:			R:VERB:TENSE		M:DET	

Feedback Analysis for Spoken GEC

End-to-end

GEC Model	P	R	$F_{0.5}$
$\text{Whisper}_{\text{gec}} \xrightarrow{\text{gec}} \text{Whisper}_{\text{flt}}$	27.77	22.31	26.40
$\text{Whisper}_{\text{flt}}\text{+GEC} \xrightarrow{\text{gec}} \text{Whisper}_{\text{flt}}$	44.70	27.53	39.74
$\text{Manual}_{\text{flt}}\text{+GEC} \xrightarrow{\text{gec}} \text{Manual}_{\text{flt}}$	58.64	35.84	52.02

Partially cascaded

- Evaluate whether the **ERRANT edits** are accurate
- Outputs from the cascaded system are conditioned on the transcription generated by $\text{Whisper}_{\text{flt}}$
- E2E systems generate outputs only based on the audio input

End-to-End Spoken Grammatical Error Correction - Conclusions

- Grammatical proficiency is an important part of overall language proficiency
- Spoken grammar is different (and more complex) than written grammar
- In addition to correcting learners, we should be able to give informative feedback about their grammar

End-to-End Spoken Grammatical Error Correction - Conclusions

- For DD, **the end-to-end outperforms the cascaded system**
- For spoken GEC, the end-to-end shows **comparable system performance** to a fully cascaded system.
  - The partially cascaded system is the best-performing system, most likely because it uses **a much higher amount of GEC training data**
- **Feedback is more challenging** using end-to-end systems as we do not have ‘full access’ to intermediate steps

End-to-End Spoken Grammatical Error Correction - Future Work

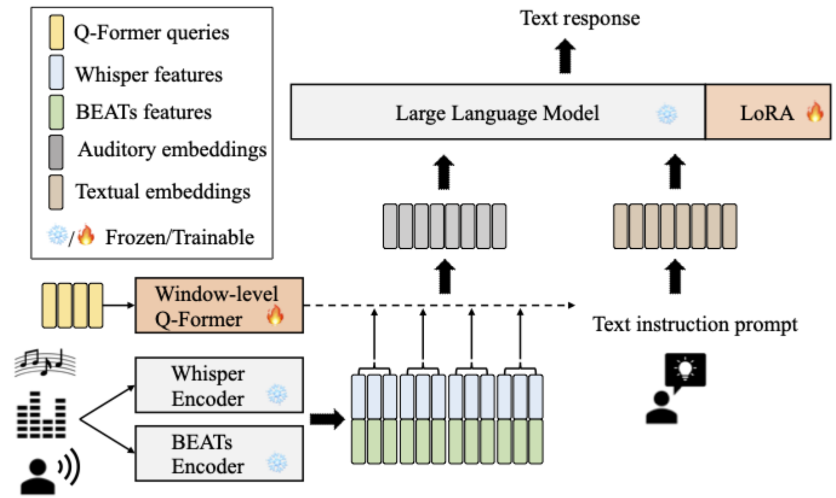
- **Extend the analysis of feedback**
- **Data augmentation:** we are currently investigating the use of text-to-speech and voice cloning algorithms to augment the training data
- **Use of multi-modal (audio+text) LLMs for DD and GEC**

Discussion and Future Work

Discussion and Future Work

- For both assessment and spoken GEC, recently we have started experimenting with multimodal LLMs such as **SALMONN** (Tang et al., 2024) and **QwenAudio** (Chu et al., 2023) in a zero-shot fashion.

SALMONN architecture



Discussion and Future Work

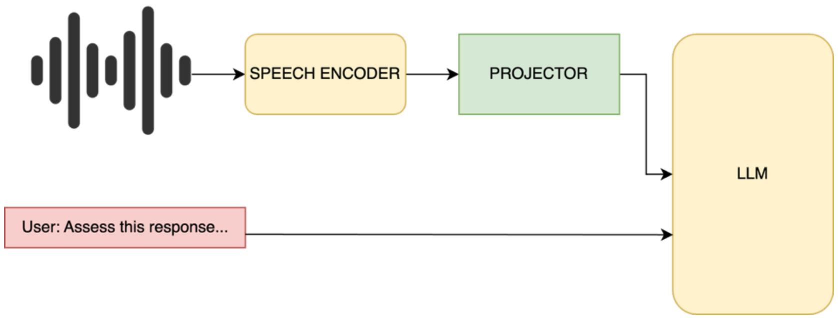
- Based on the results shown by Yancey et al. (2023) on writing assessment, **zero-shot LLMs are good but do not outperform previous systems when we have a decent amount of training data**.
- Our preliminary results on spoken assessment (paper submitted to Interspeech 2025) show **interesting but moderate improvements** when fine-tuning an audio LLM.
- A similar conclusion can be drawn about GEC, as **zero-shot LLMs tend to overcorrect**, while **previous systems still achieve competitive results when training data are available**.

In such situations, using a **bespoke model** seems to be a better solution than using an off-the-shelf general-purpose LLM.

On the other hand, **LLMs could be very efficient for more challenging tasks**, such as analytic assessment.

Discussion and Future Work

- Recently, Bellver-Soler et al. (2024) proposed an approach based on a **speech encoder in combination with an LLM** for emotion recognition.
- A similar approach has been investigated by Fu et al. (2024) for **pronunciation assessment** showing promising performances.



Discussion and Future Work

- For spoken GEC, we explained that, despite an acceptable WER, feedback poses very challenging problems;
  - To tackle them, we have recently investigated **pseudo-labelling and prompting techniques** using Whisper, which bring remarkable improvements, especially for feedback (paper submitted to Interspeech 2025).
  - Data augmentation techniques using voice editing and TTS systems are also ongoing.

Bonus: The S&I Challenge 2025

- In December 2024, we distributed the training and dev data obtained from Speak & Improve for a challenge that includes 4 shared tasks:
  - ASR of L2 speech
  - L2 assessment
  - Spoken GEC
  - Spoken GEC feedback

The full S&I corpus will be released in April.

Webpage: https://mi.eng.cam.ac.uk/~mq227/sandi2025.html

CAMBRIDGE

UNIVERSITY OF CAMBRIDGE
DEPARTMENT OF ENGINEERING

Speak & Improve Challenge 2025: Spoken Language Assessment and Feedback

Questions?

Thanks for your attention

This presentation reports on research supported by Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge.