UNIVERSITY OF CAMBRIDGE

Adapting Whisper for Spoken Language Assessment and Feedback

Rao Ma

Speech Group, ALTA institute, University of Cambridge

October 2023

Team Members

Rao Ma

Dr Stefano Bannò

Dr Mengjie Qian

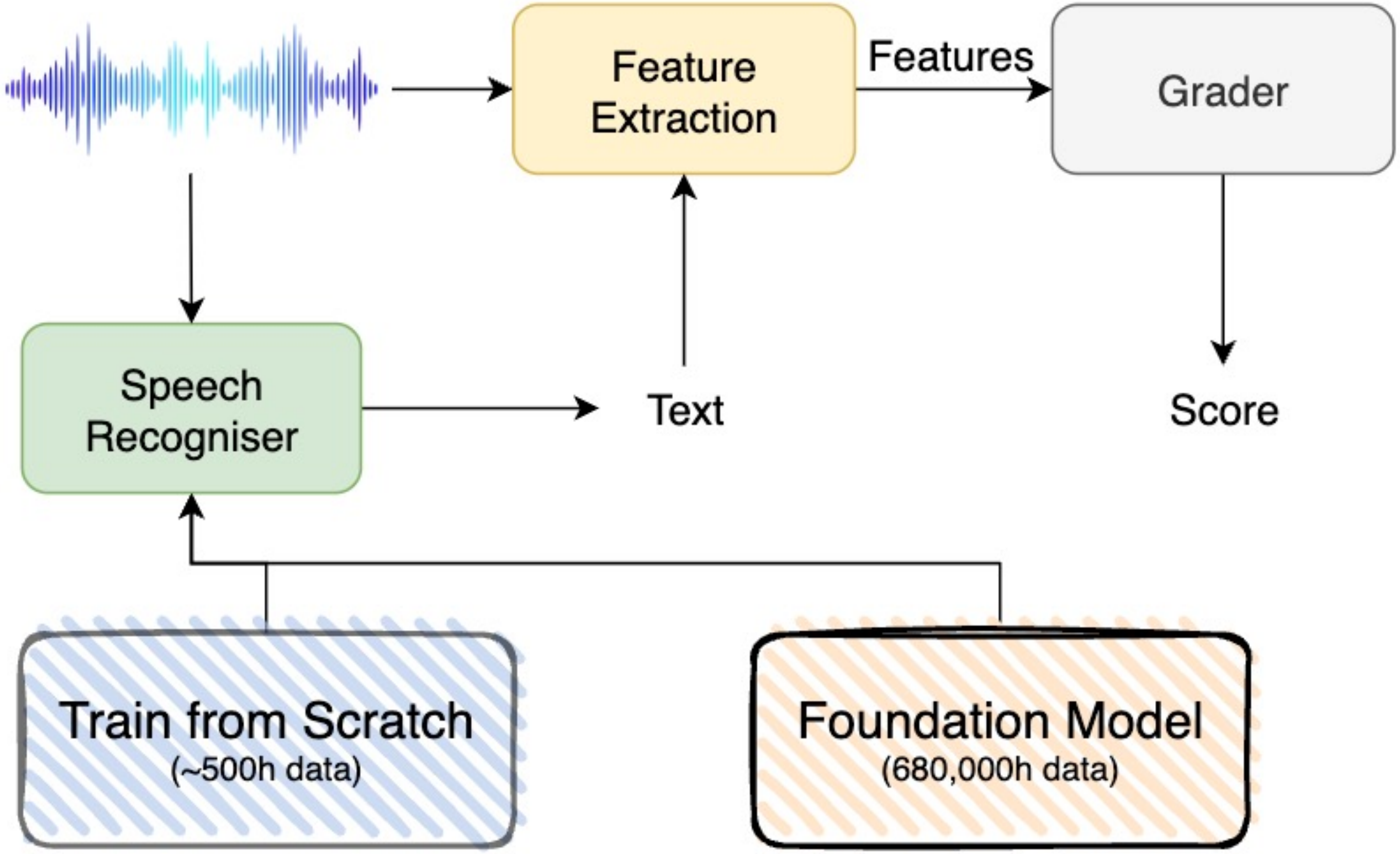
Siyuan Tang

Dr Kate Knill

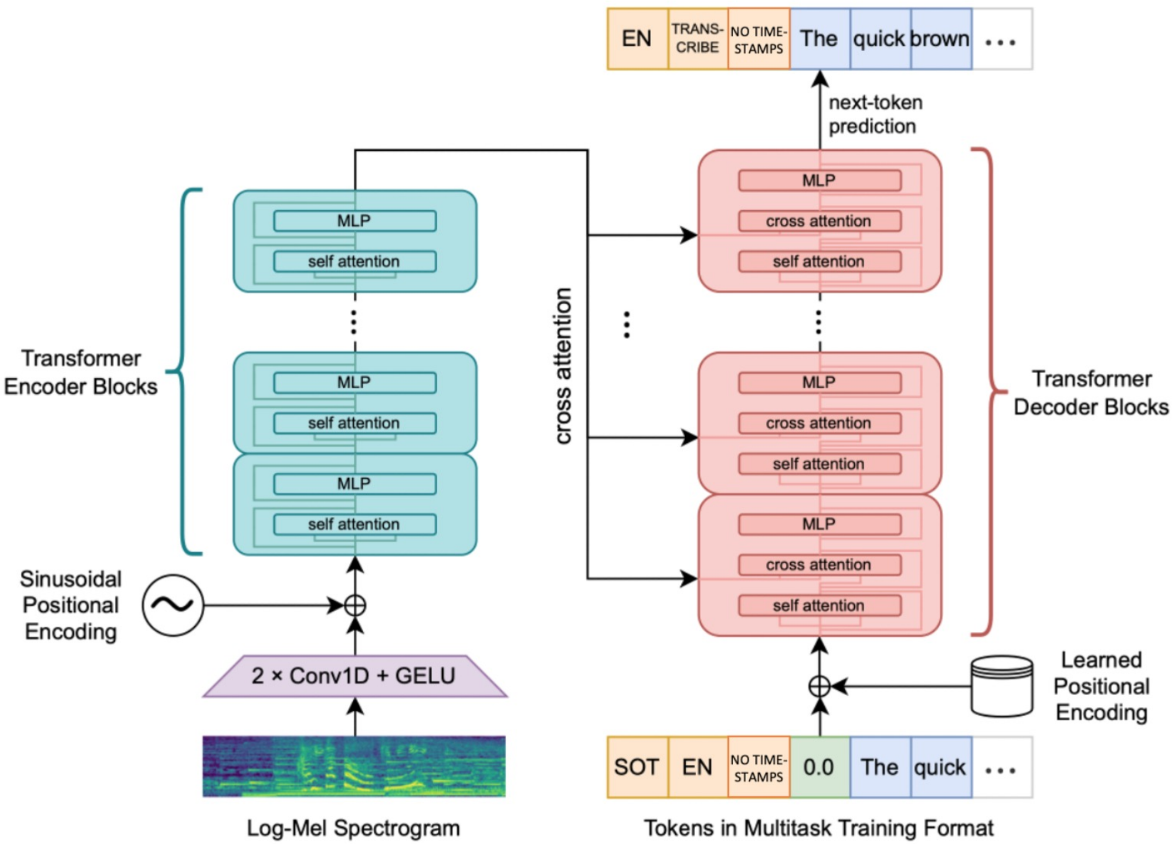
Prof Mark Gales

Spoken Language Assessment

Spoken Language Assessment



Whisper



- 680,000 hours of web data
- Multitask training:
  - multilingual ASR
  - speech translation
  - language identification
  - voice activity detection
- Different sizes:
  - tiny – 39M
  - base – 74M
  - small – 244M
  - medium – 769M
  - large – 1550M

Speech Recognition with Whisper

Dataset	wav2vec 2.0 Large (no LM)	Whisper Large V2	RER (%)
LibriSpeech Clean	**2.7**	**2.7**	0.0
Artie	24.5	**6.2**	74.7
Common Voice	29.9	**9.0**	69.9
Fleurs En	14.6	**4.4**	69.9
Tedlium	10.5	**4.0**	61.9
CHiME6	65.8	**25.5**	61.2
VoxPopuli En	17.9	**7.3**	59.2
CORAAL	35.6	**16.2**	54.5
AMI IHM	37.0	**16.9**	54.3
Switchboard	28.3	**13.8**	51.2
CallHome	34.8	**17.6**	49.4
WSJ	7.7	**3.9**	49.4
AMI SDM1	67.6	**36.4**	46.2
LibriSpeech Other	6.2	**5.2**	16.1
Average	29.3	**12.8**	55.2

WER Results on Linguaskill test set

Hypotheses	LIALTtst02			
	Sub	Del	Ins	WER
Kaldi	8.2	9.4	1.3	18.9
Whisper	7.4	15.9	1.8	25.1

- Kaldi-based system
  - Acoustic model and language model: trained on 400+ hours of Linguaskill data
- Whisper outputs contain more deletion errors
  - L2 English learners have a lot of disfluencies and hesitations/fillers

Problems of Whisper outputs

Type	Sentence
Ref	**mister** lee when you arrive **you could uh** we could take **the most** the most cheap park zone blue zone it costs **um twenty dollar p-** per week
Hyp	**Mr.** Lee, when you arrive, *** ***** ** we could take *** **** the most cheap Park zone, blue zone. It costs ** **$20** ****** ** per week.

- The output is human readable, i.e. punctuation is added, numbers are presented in Arabic numeric format, and disfluencies are skipped.
- Typical ASR error types made by Whisper: 1) abbreviation in red; 2) disfluency (false start and repetition) in blue; 3) hesitation in pink; 4) number in cyan and 5) partial word in orange.

Text Normalisation Rules

- Symbols like currency units and mathematical notations: converted
- Ordinal numbers: converted
- Punctuation: removed or replaced by space
- Abbreviation: mapped
- Combination of numbers and letters: converted case by case
- US/UK spelling difference: converted
- ……

WER Results After Text Normalisation

Hypotheses	LIALTtst02			
	Sub	Del	Ins	WER
Kaldi	8.2	9.4	1.3	18.9
Whisper	7.4	15.9	1.8	25.1
$\text{Whisper}_{\text{std}}$	6.4	14.9	2.2	23.5

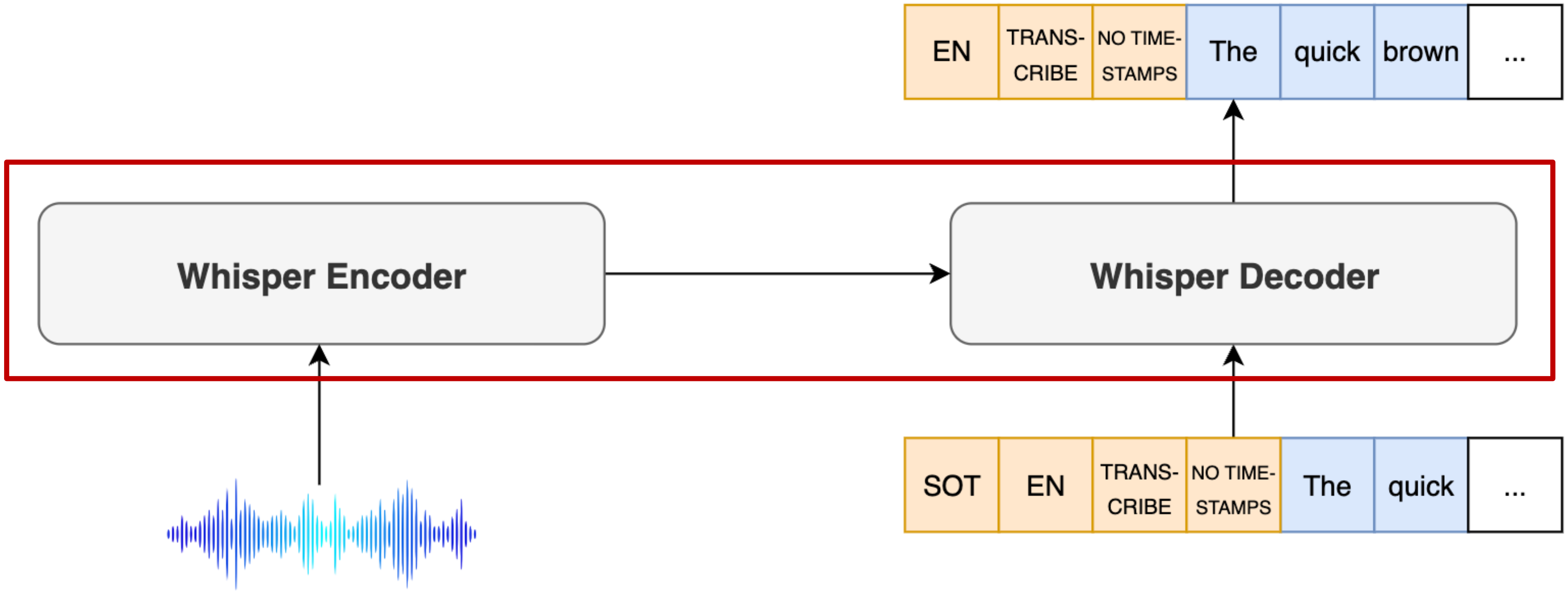
- Cannot recover the skipped hesitations or disfluencies
- Ambiguity: $20 → "twenty dollar" or "twenty dollars" ?

WER Results without Hesitation

Hypotheses	LIALTtst02			
	Sub	Del	Ins	WER
Kaldi	8.6	7.1	1.4	17.1
Whisper	7.4	8.7	2.5	18.5
$\text{Whisper}_{\text{std}}$	6.2	7.7	2.9	16.8

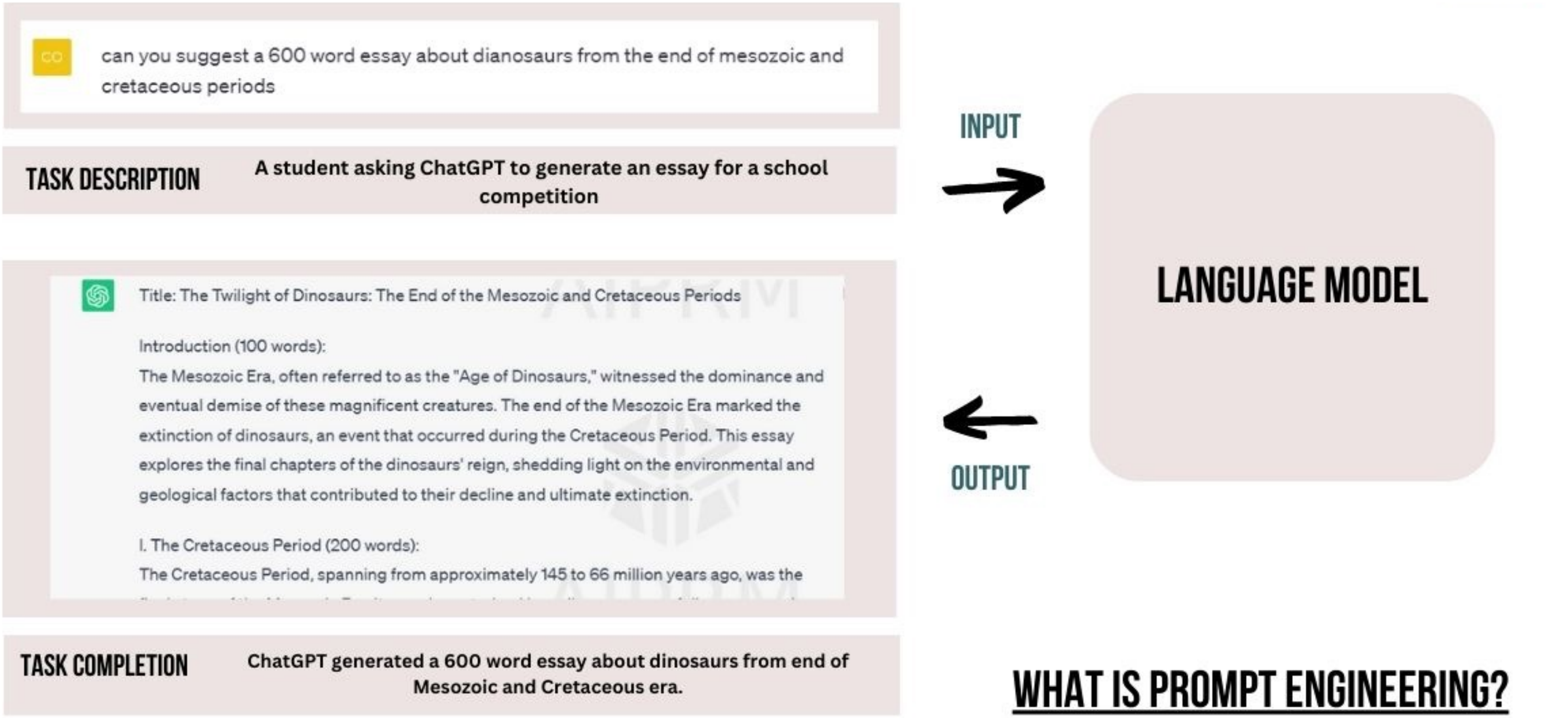
- Better performance than Kaldi-based system in zero-shot evaluation!

Task Adaptation: Fine-tuning



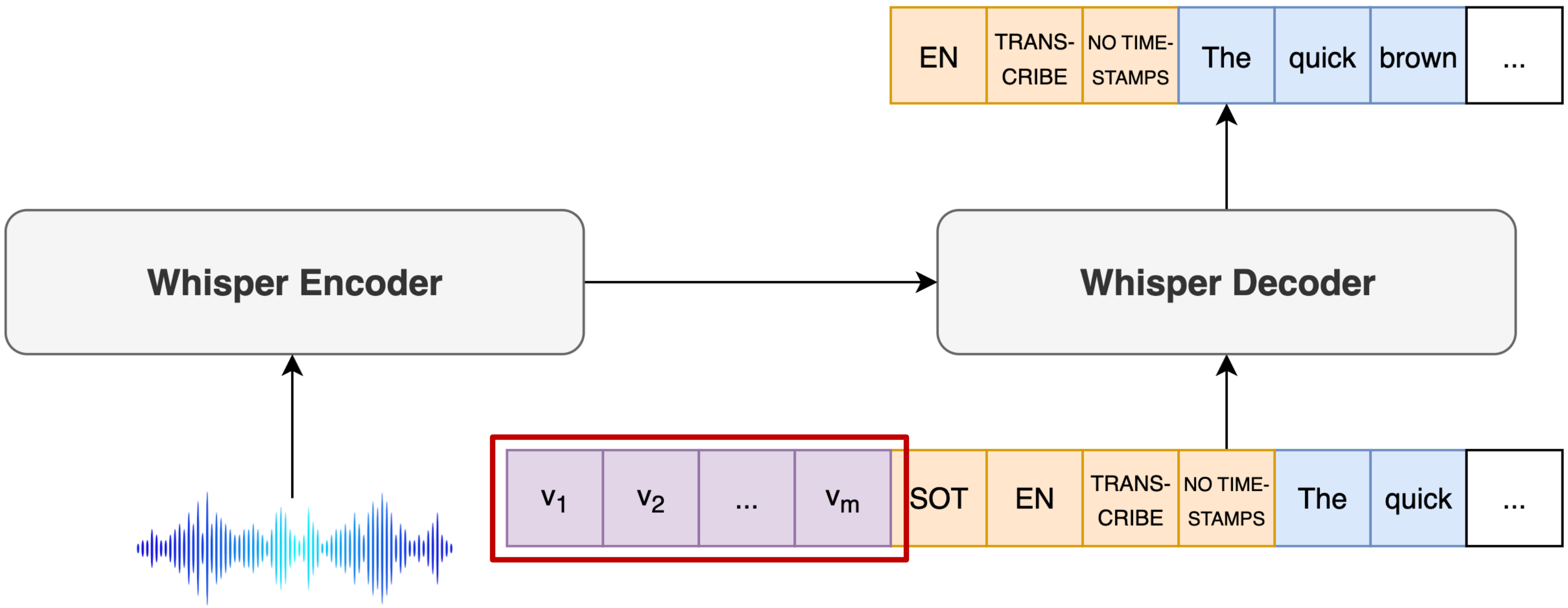
- Update all model parameters based on the task-specific training set

Task Adaptation: Prompting



- Operate in the model input, task specified in natural language
- Human readable, but requires human expertise in prompt designing

Task Adaptation: Soft Prompt Tuning



- Insert 20 trainable vectors in the decoder embedding space
- Optimised via gradient descent
- Parameter efficient compared to fine-tuning (only 0.0006% parameters)

Dataset for Whisper Adaptation

Split	Corpus	Hours	Submissions
train	LNG_subset	17	-
test	LIALTtst02	30	229

- Train/test data annotated by ELiT (LIESTdev01/2/3)
  - Complete responses with no unknown or foreign words
- Train: mix of Linguaskill General and Business Speaking
- Test: Linguaskill Business

ASR Results of Whisper Adaptation

Hypotheses	LIALTtst02			
	Sub	Del	Ins	WER
Kaldi	8.6	7.1	1.4	17.1
$\text{Whisper}_{\text{std}}$	6.2	7.7	2.9	16.8
Whisper-FT	5.6	1.8	2.3	**9.7**
Whisper-SPT	6.0	2.0	2.2	10.3

- Fine-tuned Whisper performs the best (43% WERR to Kaldi)
  - Soft-prompt tuning only slightly worse
- Large reduction in deletions with adapted Whisper
  - suitable to display to learners in Speech and Improve

Case Analysis

Type	Example
Ref	%hes% i think i'm not i'm not really denominal maybe %hes% one hundred because i'm not i'm not like shopping
Baseline	***** i think i'm not i'm not really **the nominal** maybe ***** **a 100** because *** *** i'm not like shopping
SPT	%hes% i think i'm not i'm not really **the nominal** maybe %hes% one hundred because i'm not i'm not like shopping
FT	%hes% i think i'm not i'm not really denominal maybe %hes% one hundred because i'm not i'm not like shopping

Analysis on Word Counts

Word Type	C_{all} Ref	$C_{correct} \uparrow$ Baseline	FT	SPT
Hesitation	2661	5	2213	2267
Number	421	220	388	381
Abbreviation	18	17	17	17
Disfluency	2201	583	1935	1938
Partial Words	358	0	55	51
Recall All	-	15.4%	82.1%	82.9%

Dataset for Spoken Language Assessment

Split	Corpus	Hours	Submissions
train	LIESTtrn04	750	6,809
test	LIALTtst02	30	229

- Both train and test sets are from Linguaskill Business Speaking
- Training data different from that used in adaptation
- Transcriptions from underlying ASR system: Kaldi or Whisper-FT

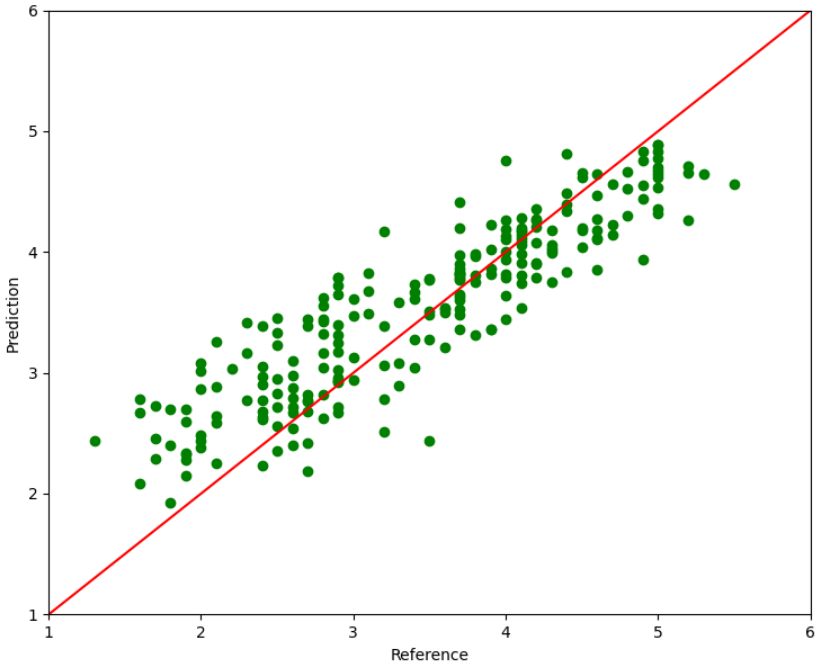
Grader Performance

Model	WER	PCC↑	RMSE↓	% <= 0.5 ↑	% <= 1.0 ↑
Kaldi	17.1	0.896	0.468	72.5	96.1
Whisper-FT	**9.7**	**0.903**	**0.430**	**74.7**	**97.4**

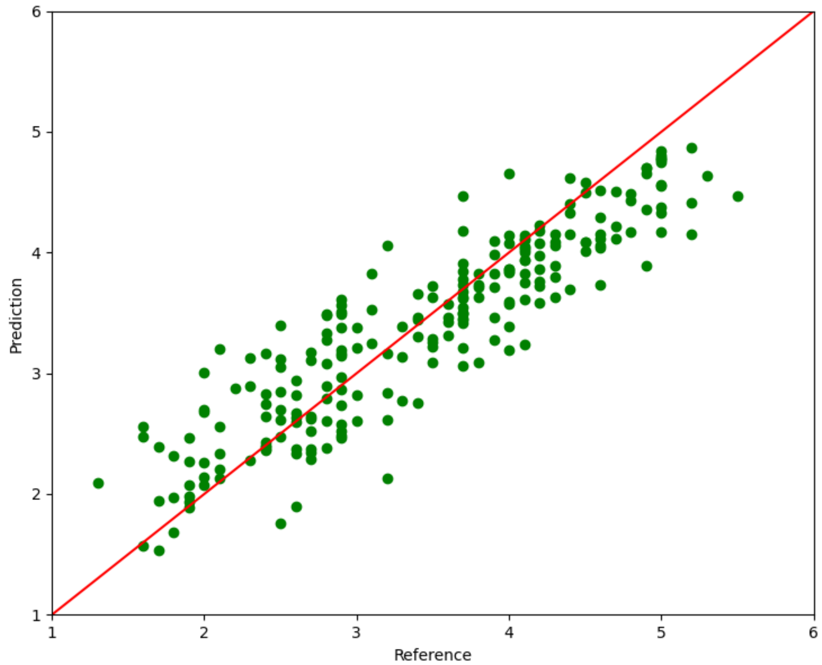
- DDN feature-based grader
  - 24 feature subset used for these preliminary experiments
- Whisper-FT shows improvement on all metrics!
  - Significant reduction in WER
  - Small gains in auto-marking due to approach of mitigating effect of ASR errors by training on ASR transcriptions

Predicted vs Reference Scores

Kaldi



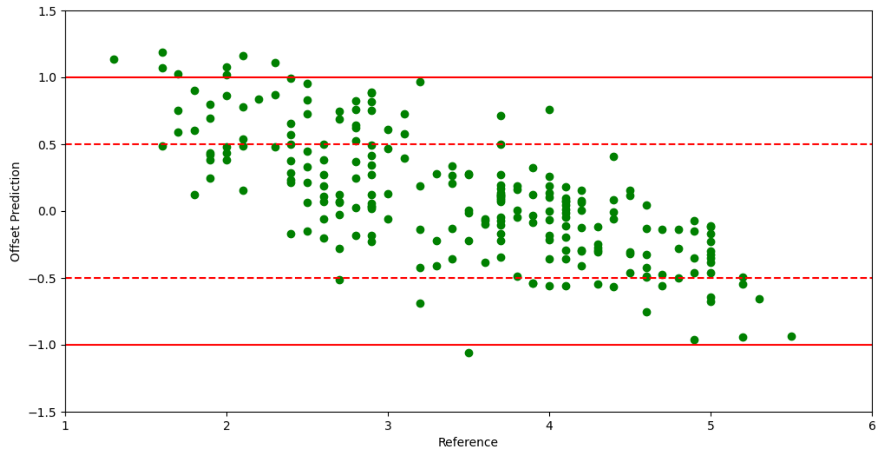
Whisper-FT



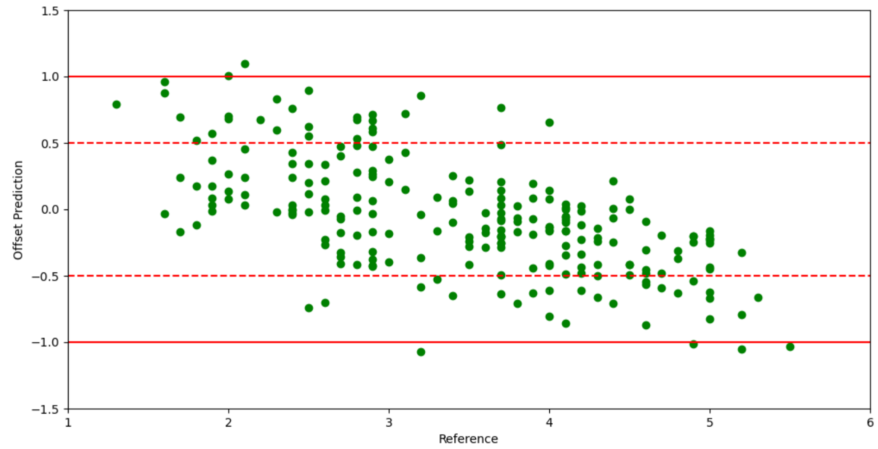
- Whisper-FT a little more consistent

Offset Predicted vs Reference Scores

Kaldi



Whisper-FT



- Whisper-FT more within desired bounds, especially for lowest scores
- Whisper-FT slightly more offset on highest scores

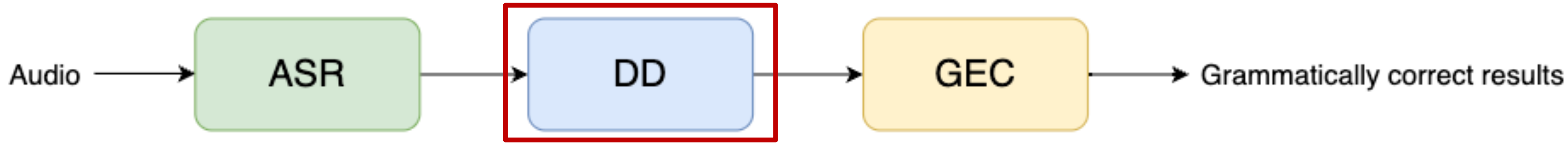
SLA Conclusions

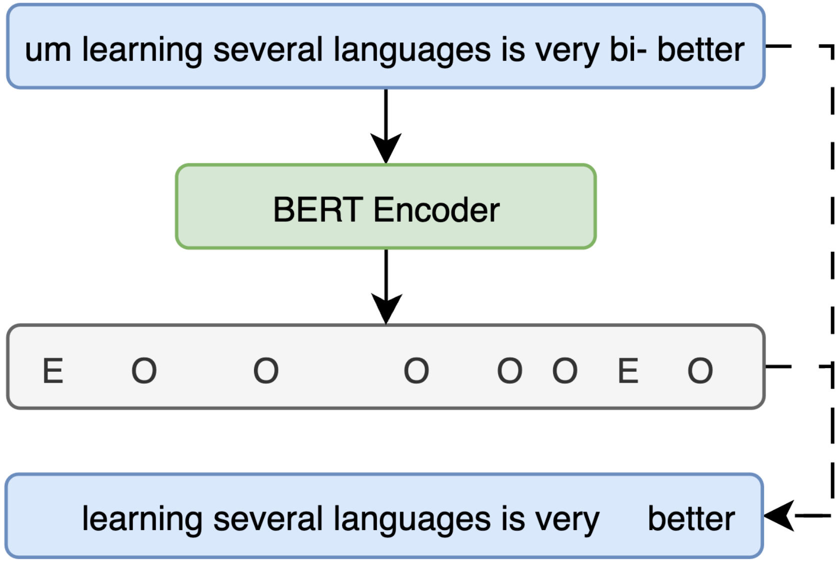
- Standard Whisper deletes parts of L2 learners' speaking transcript
  - Fine-tuning and soft prompt tuning can be used to address the issue
- After fine-tuning on 17h Linguaskill training set, we can achieve 43% WERR compared to a 400h trained Kaldi-based system
- Grader shows performance gain on Linguaskill Business test set

Feedback for Spoken Grammatical Error Correction

Spoken GEC

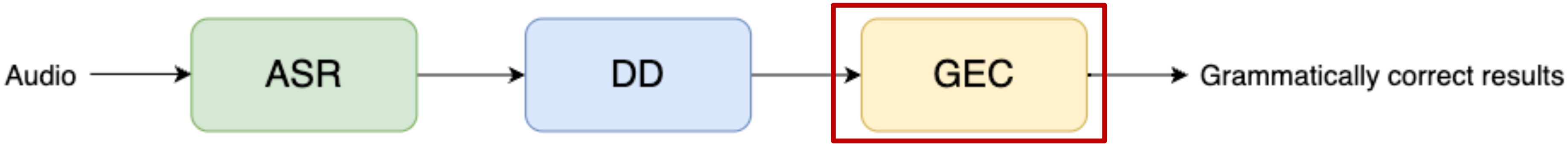
Disfluency Detection (DD)

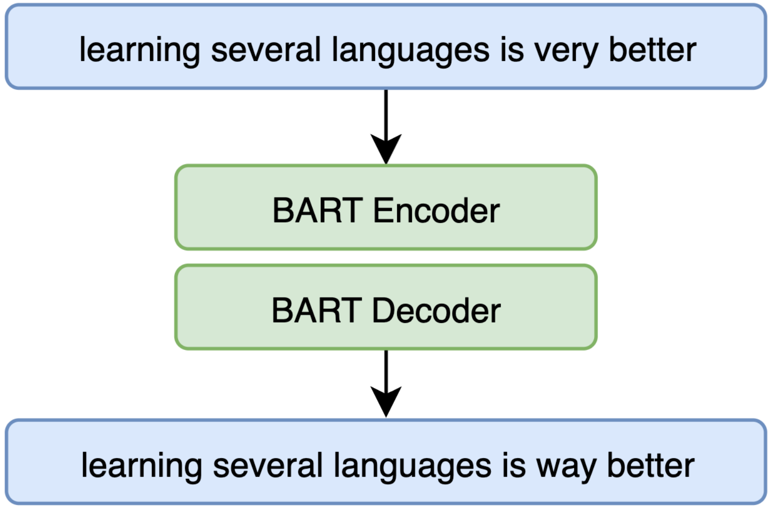




- Pre-trained language model: BERT
  - Capable of high-quality feature representations
  - Fine-tune BERT for DD sequence tagging objective

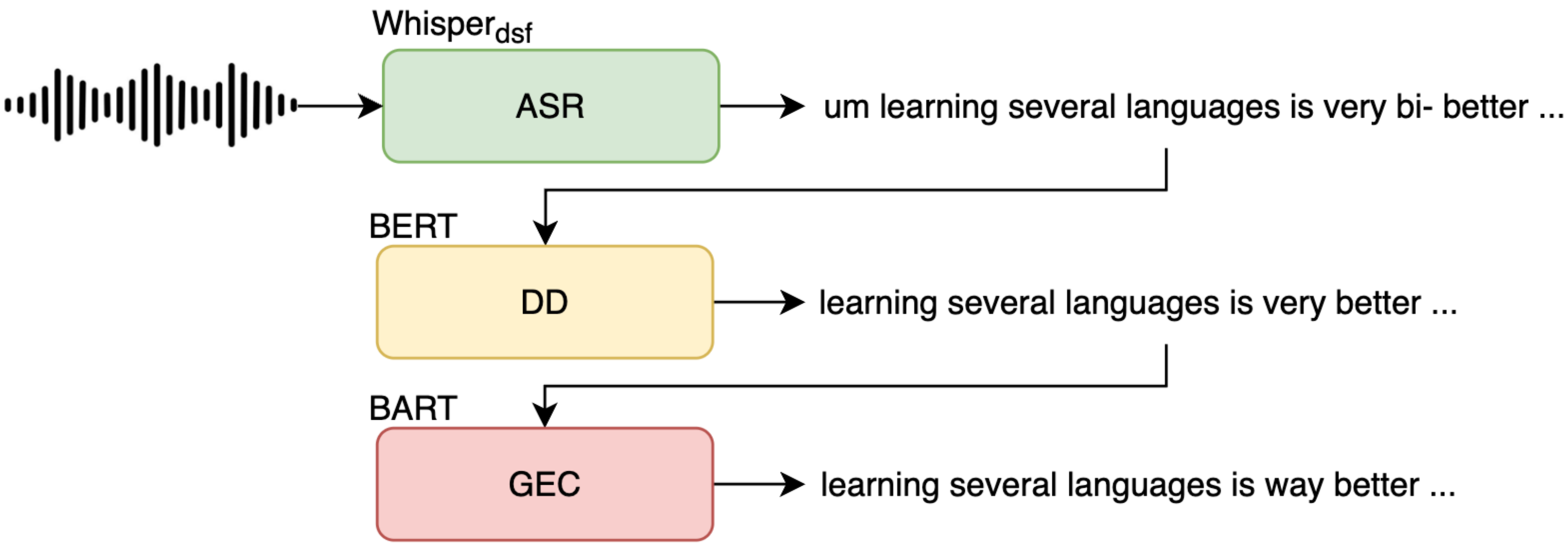
Grammatical Error Correction (GEC)





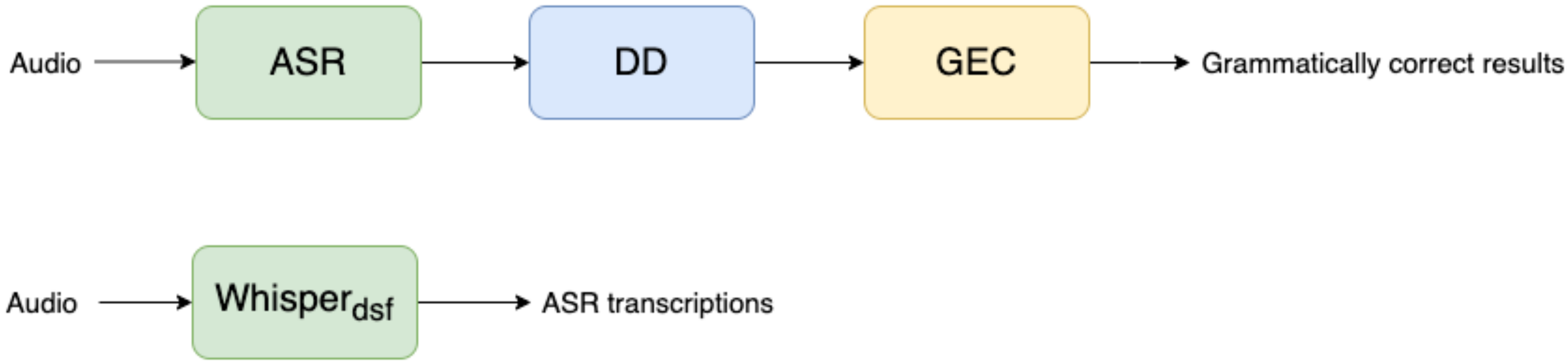
- Pre-trained language model: BART
  - Encoder-decoder architecture
  - Treat spoken GEC as a sequence-to-sequence task

Cascaded System Issues

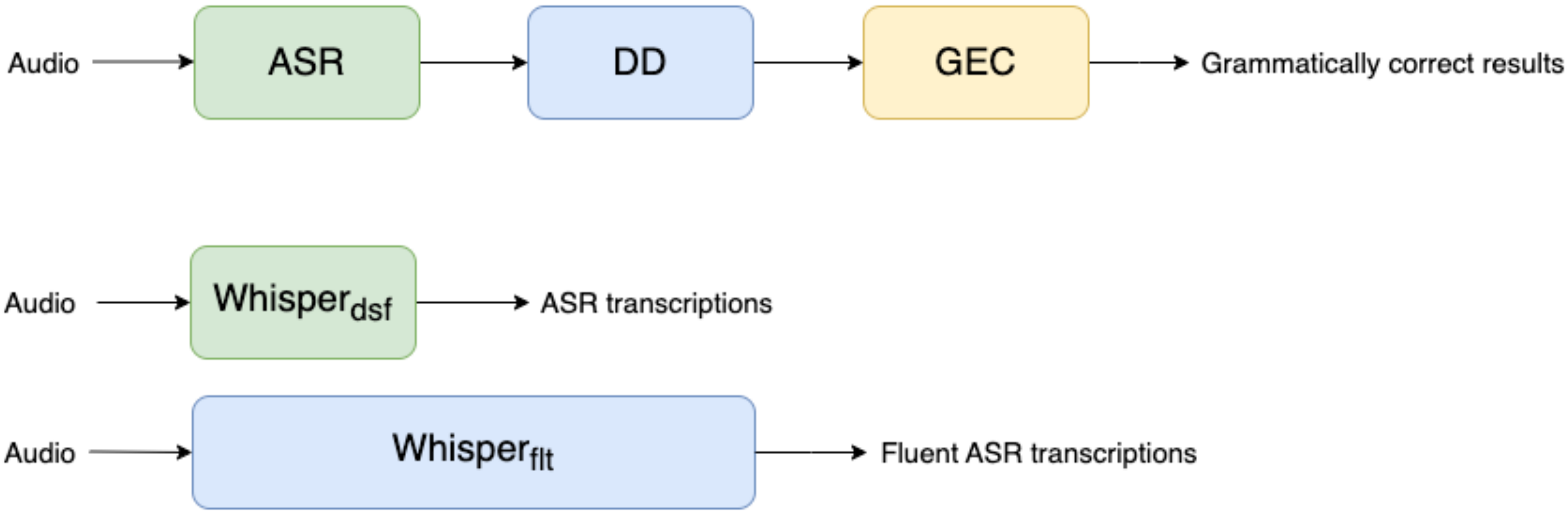


- ASR errors can propagate in the pipeline
- Loss of information (intonation, speaker info, emotion, etc.)
- Training-evaluation mismatch

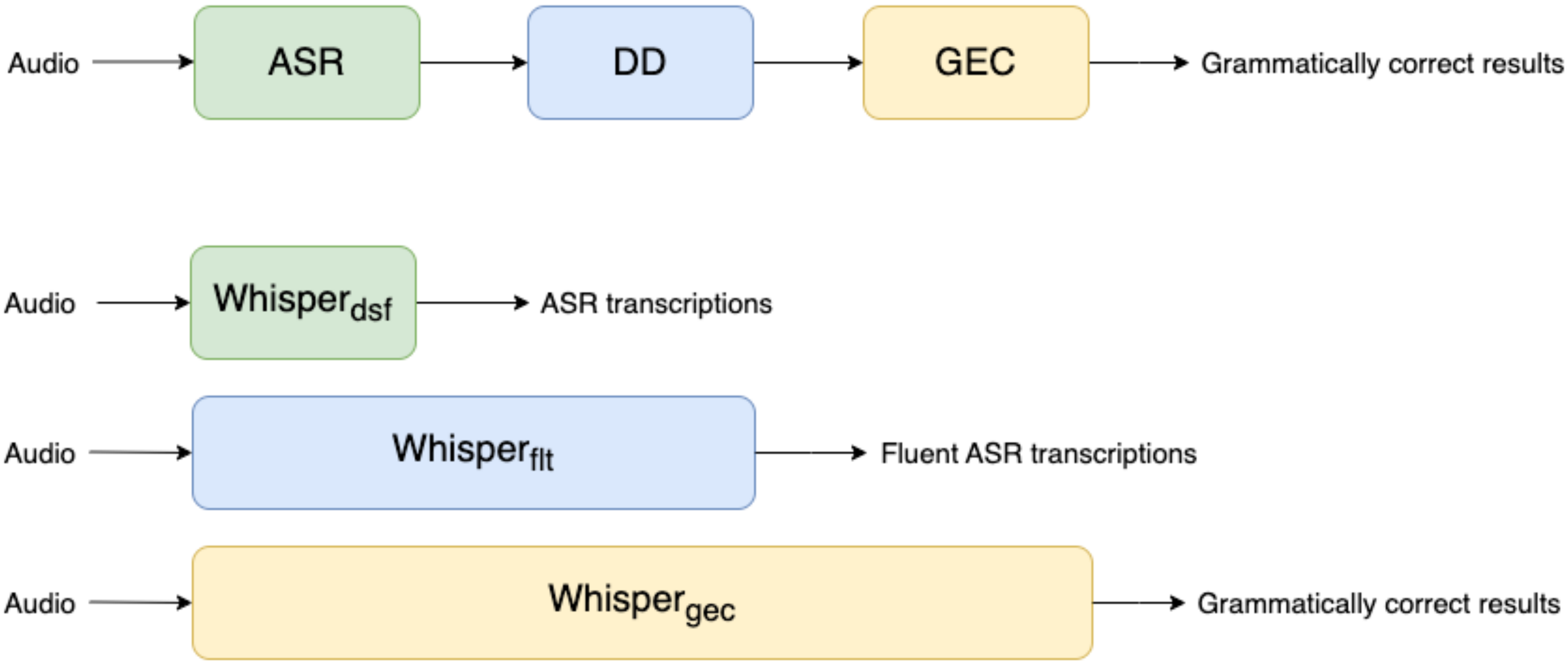
Whisper for Spoken GEC



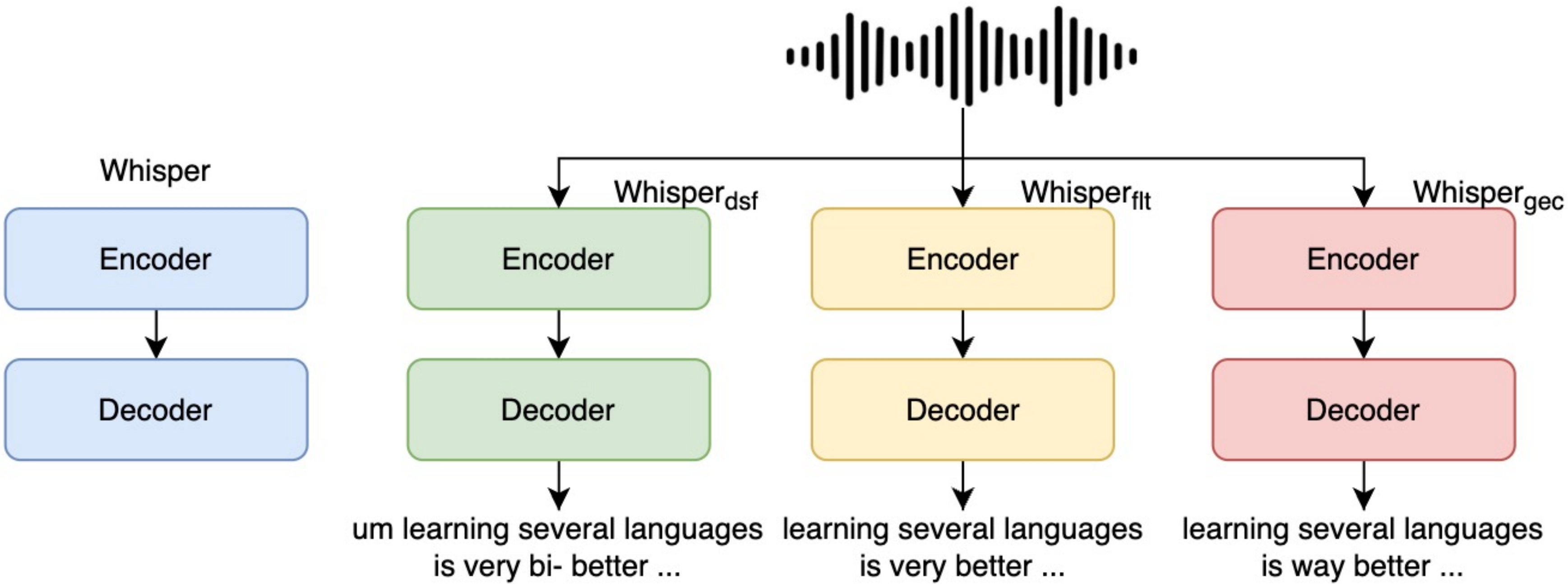
Whisper for Spoken GEC



Whisper for Spoken GEC



Fine-tuning Whisper for Spoken GEC



- **Proposal:** Fine-tune Whisper on three training sets separately to generate ASR transcription in different formats

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

Model Setup

- DD (BERT):
  - Stage 1 fine-tuning: Switchboard NXT
  - Stage 2 fine-tuning: Linguaskill data
- 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

- Standard metrics for DD/GEC challenging for spoken processing
  - ASR errors mean that standard annotation not applicable

- **Disfluency Detection** (DD):
  - Standard Metric: F_1 score on detecting disfluencies
  - **BUT** ASR errors have no disfluency annotation
  - Use WER to assess distance to manual text with disfluencies removed

- **Spoken Grammatical Error Correction** (GEC):
  - Standard Metric: $F_{0.5}$ score on edits to correct manual text
  - **BUT** ASR errors modify edits required to yield correct text
  - Use WER/TER to assess word-level distance from *GEC* manual reference

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**

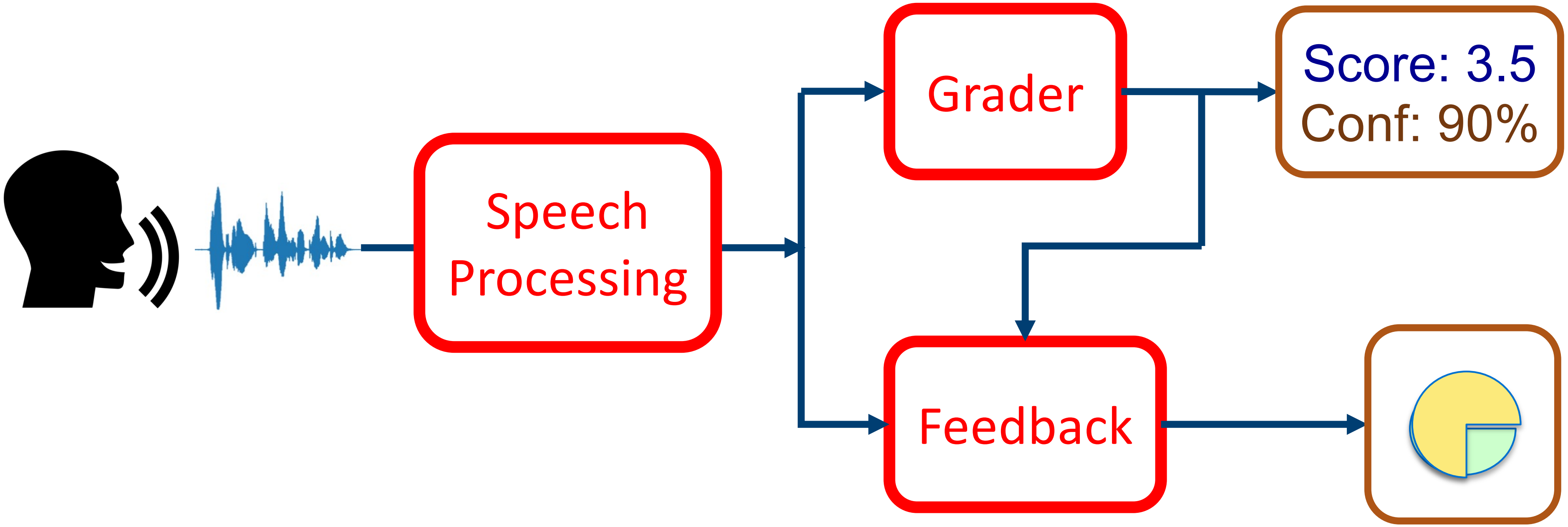
- E2E approach performs better than a cascaded system
- Attention mechanism in Whisper is able to learn to skip words
  - $\text{Whisper}_{\text{flt}}$ has learnt to skip disfluencies

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

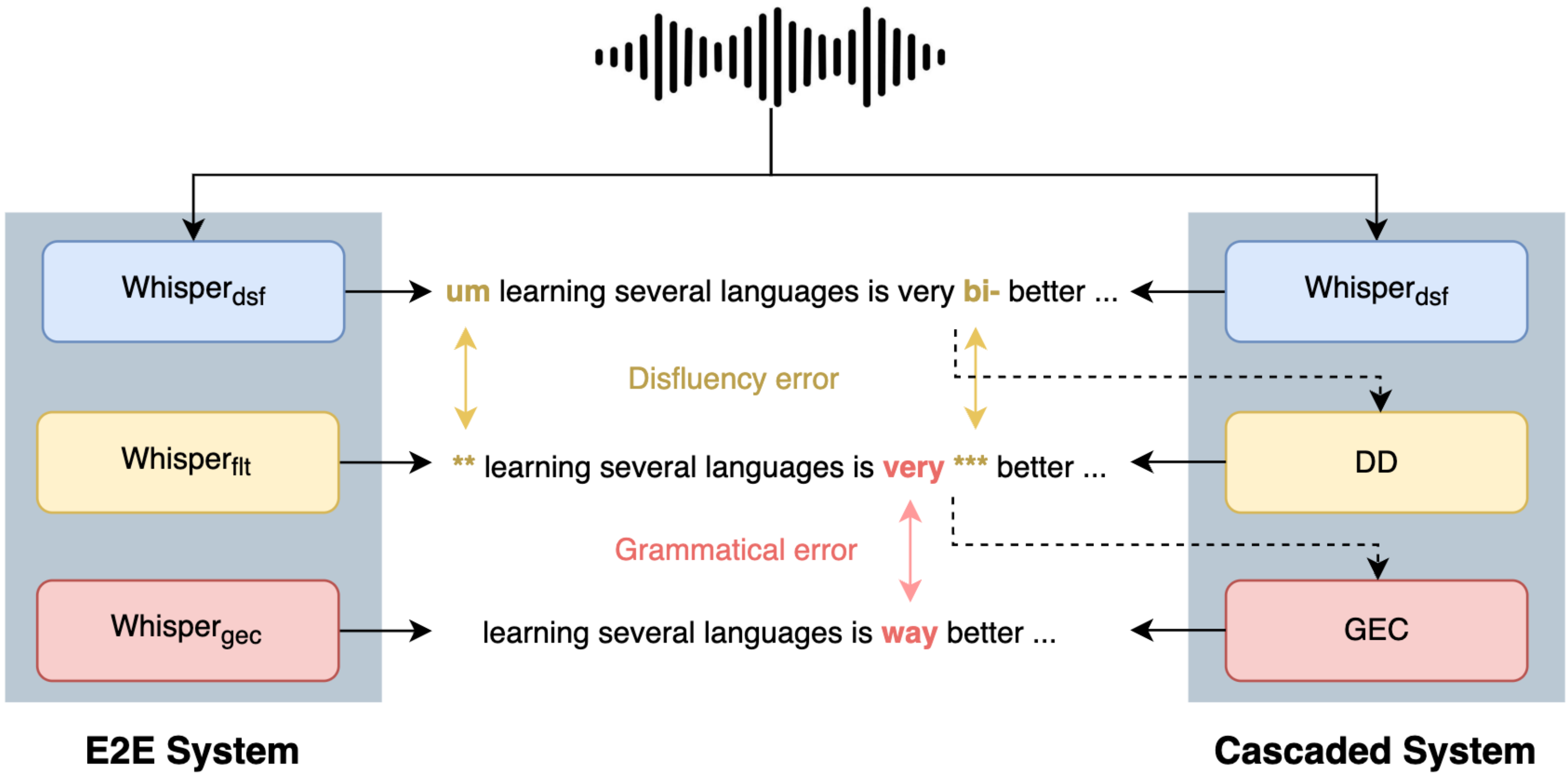
- Comparable performance compared to a fully cascaded system
- $\text{Whisper}_{\text{gec}}$ has learnt to ”translate” to correct text
- Problem: lack of available training data

Feedback for Spoken Grammatical Error Correction



Analytic – holistic feedback across all speech
Fine-grained – feedback on specific errors in words/phrases

Feedback Analysis for Spoken GEC



Feedback Analysis for DD

DD Model	P	R	F_1
$\text{Whisper}_{\text{dsf}}\text{+DD} \xrightarrow{\texttt{del}} \text{Whisper}_{\text{dsf}}$	74.94	75.05	73.35
$\text{Whisper}_{\text{flt}} \xrightarrow{\texttt{del}} \text{Whisper}_{\text{dsf}}$	61.02	68.11	62.30

- Evaluate whether the deletions are accurate
- The cascaded system compares deletions based on a single transcription
- E2E systems compare outputs from two different decoding processes

Feedback Analysis for Spoken GEC

GEC Model	P	R	$F_{0.5}$
$\text{Whisper}_{\text{flt}}\text{+GEC} \xrightarrow{\text{gec}} \text{Whisper}_{\text{flt}}$	38.17	23.52	33.95
$\text{Whisper}_{\text{gec}} \xrightarrow{\text{gec}} \text{Whisper}_{\text{flt}}$	23.54	19.00	22.47

- Evaluate whether the 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

Spoken GEC Conclusions

- For disfluency removal, Whisper outperforms a cascaded system
- For spoken GEC, Whisper shows comparable system performance to a fully cascaded system
- Feedback is more challenging
  - Multiple, possibly inconsistent, decoding runs required to deive edits

Conclusions

- Whisper is a better ASR model than previous Kaldi model
  - adaptation yields performance gains in SLA and more accurate transcriptions
  - want to make it fast and use less computation → distillation

- Whisper can produce fluent spoken GEC output in E2E fashion
  - Feedback more challenging as multiple decoding runs required

Foundation ASR models like Whisper have great potential in building language learning applications!

Questions?

Thank you for listening

Thanks to: Diane Nicholls and the Humannotator team at ELiT for the Linguaskill Speaking annotations. This presentation reports on research supported by Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge.