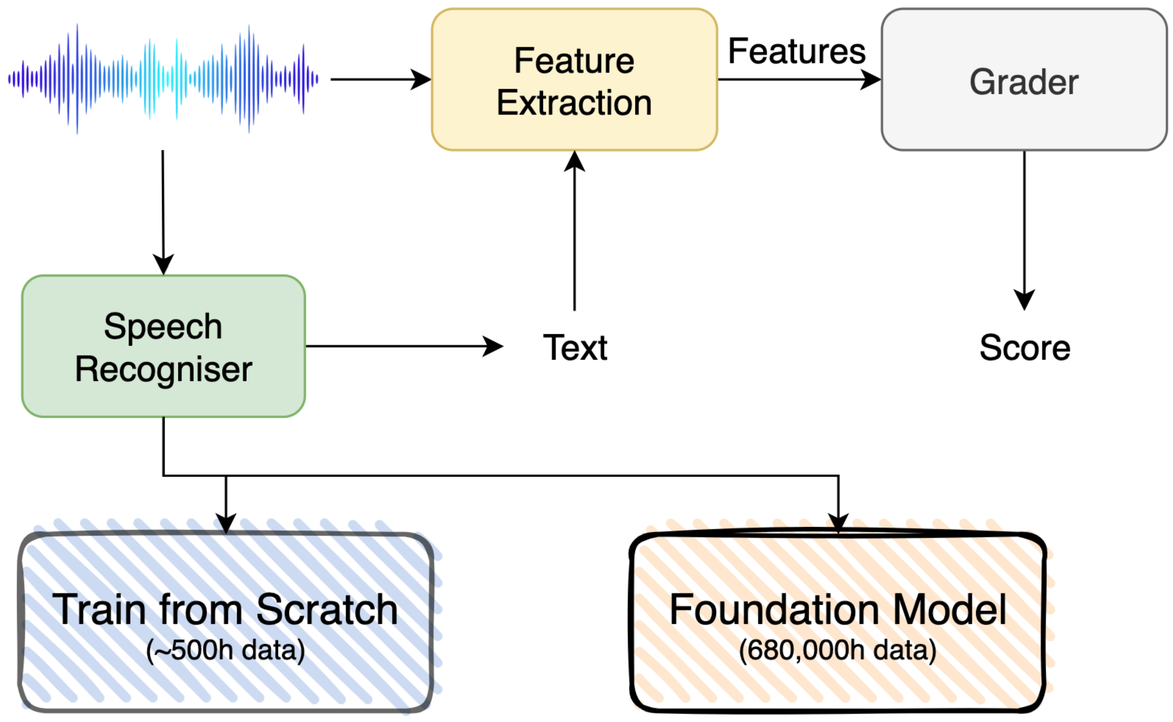
Adapting an ASR Foundation Model for Spoken Language Assessment

Rao Ma, Mengjie Qian, Mark J. F. Gales, Kate M. Knill
{rm2114,mq227,mjfg100,kmk1001}@cam.ac.uk

ALTA Institute / Department of Engineering, University of Cambridge

Spoken Language Assessment



- The ASR module is essential for auto-marking and providing feedback.

Motivation

- For assessment we want to know exactly what was said.
- Problems of Whisper output:
  - The output is human-readable, i.e. punctuation is added, and numbers are presented in Arabic numeric format.
  - The model has a tendency to skip disfluencies and hesitations.

Table: 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 purple and 5) partial word in orange.

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.

Proposed Method: Fine-Tuning (FT)

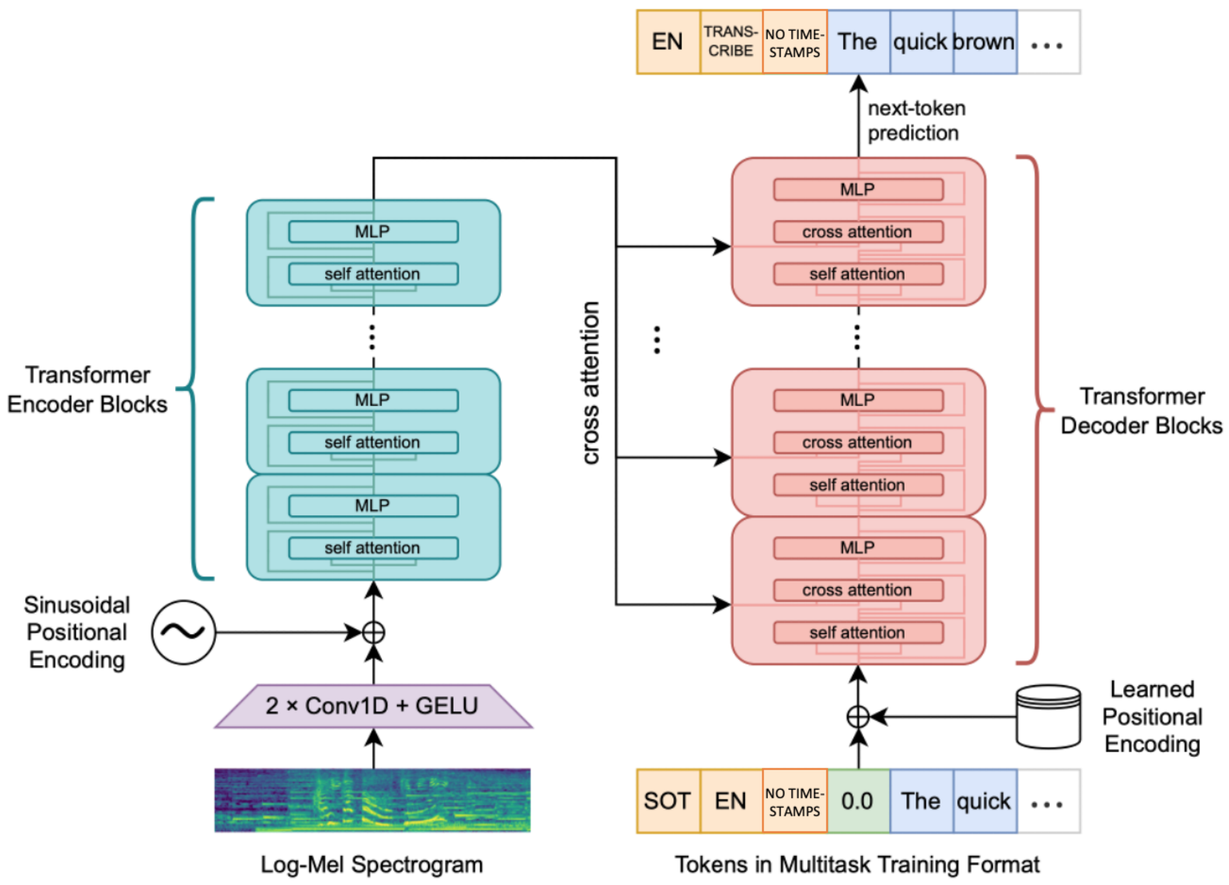
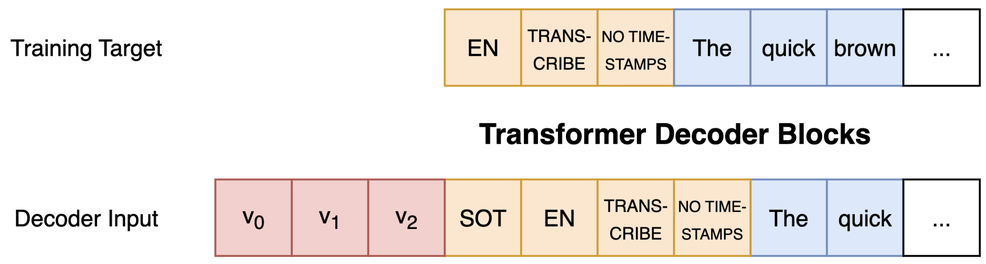


Figure: Whisper model architecture from the original paper.

This paper reports on research supported by Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge. Mengjie Qian is supported by EPSRC Project EP/V006223/1 (Multimodal Video Search by Examples). Thanks to EST Ltd for initial analysis of Whisper on Linguaskill.

Proposed Method: Soft Prompt Tuning (SPT)



- Discrete prompts:
  - Operate in the model input, task specified in natural language.
  - Human readable but requires human expertise in prompt designing.
- Soft prompts:
  - Operate in the embedding space, continuous vectors are concatenated with the token embeddings and optimised via gradient descent.
  - The original model parameters are fixed and only a small number of prompt parameters are learned (15KB vs 244MB).

Linguaskill Dataset

- A submission includes various tasks including reading aloud, describing pictures, and speaking freely on a given topic.
- Linguaskill General focuses more on everyday conversation while Linguaskill Business puts emphasis on business situations.
- Each of the test sets contains around 8h speech data. We randomly sample 17h data as the training set.

Case Analysis

Table: Case analysis on Ling_general (**Errors in red**).

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
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
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

Experimental Results

Table: Word counts and overall recall on Ling_general.

Word Type	C_{all} Ref	$C_{correct}$ ↑ 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%

Table: Overall Speech WER results and breakdown of different error types.

Model	Ling_general Sub	Del	Ins	WER	Ling_business Sub	Del	Ins	WER
Baseline	4.5	10.7	1.3	16.4	5.8	16.6	1.4	23.9
FT	4.3	1.7	2.1	8.1	5.0	2.4	2.3	9.7
SPT	4.4	1.7	2.8	8.9	5.3	2.5	3.2	11.0

Conclusions

- The output of Whisper is designed to be human-readable, which is not helpful for building a spoken language assessment system.
- We propose two solutions: fine-tuning and soft prompt tuning.
- Results on Linguaskill show we can effectively alter the decoding behaviour of Whisper to generate the exact spoken words.