Characterizing Large LMs for Generative Speech Recognition Error Correction

Huck Yang*

*presenting works mainly done during Georgia Tech
huckiyang@gatech.edu

Nov. 9th 2023
Invited talk at SLS Group MIT CSAIL

Could You Recognize this Speech?

"He has not been dropped."

[Context] at a noisy party

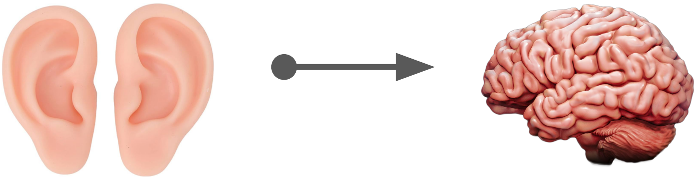
[Context] referring to someone who has not been uninvited or excluded from an event or social circle.

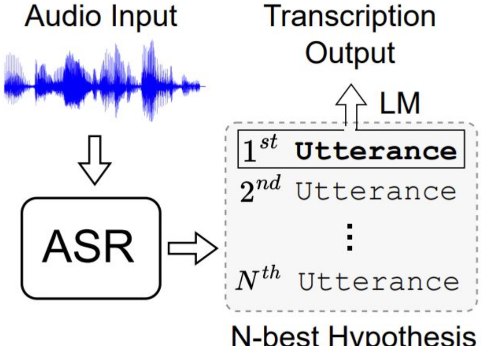
Open Questions (perhaps in mind) During this Talk

Do we need better ears or a better brain to recognize speech?

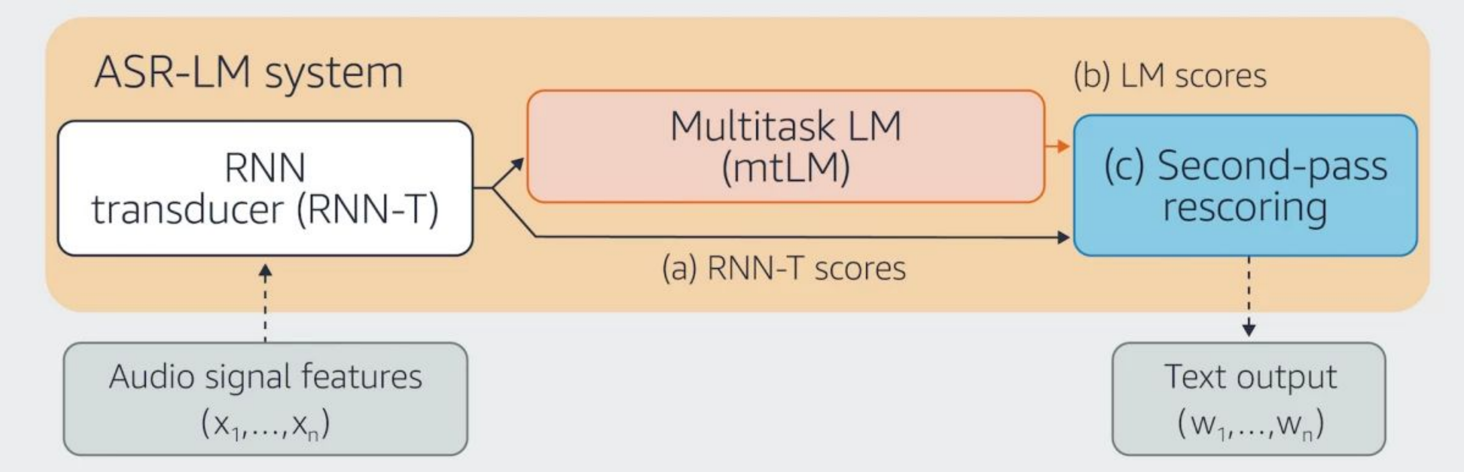
Does word error rate really matter?

What kind of new metrics we need for acoustics and speech understanding?





Background of LM for E2E Speech Recognition



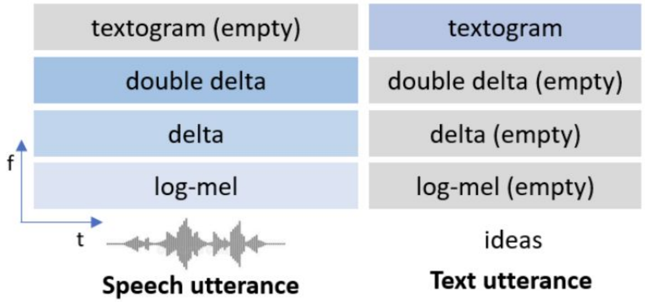
[Using NLU labels to improve an ASR rescoring model - Amazon Science]

[ASRU 21 C.-H. H. Yang et al.]

Challenges and LM Criticisms from 2017 to 2022*

Do we really need a cascade LM in end-to-end ASR model?

- External LM in RNN-T (Microsoft)
- Text Injection; Textogram (IBM; Google)
- Joint Speech and Text Alignment (Google; Meta)



Towards Reducing the Need for Speech Training Data To Build Spoken Language Understanding Systems, ICASSP 22
JOIST: A JOINT SPEECH AND TEXT STREAMING MODEL FOR ASR, SLT 23

Outline

- Frozen Pre-Trained Model Adaptation for Universal Sequence Modeling
- Generative Error Correction for Speech Recognition
- Remaining Challenges
    - What Kind of open data we need?
    - Hallucination
    - How Good Acoustic (Vision) Models we need in a two-pass system?

Theory : A Dual Form of In-Context Learning

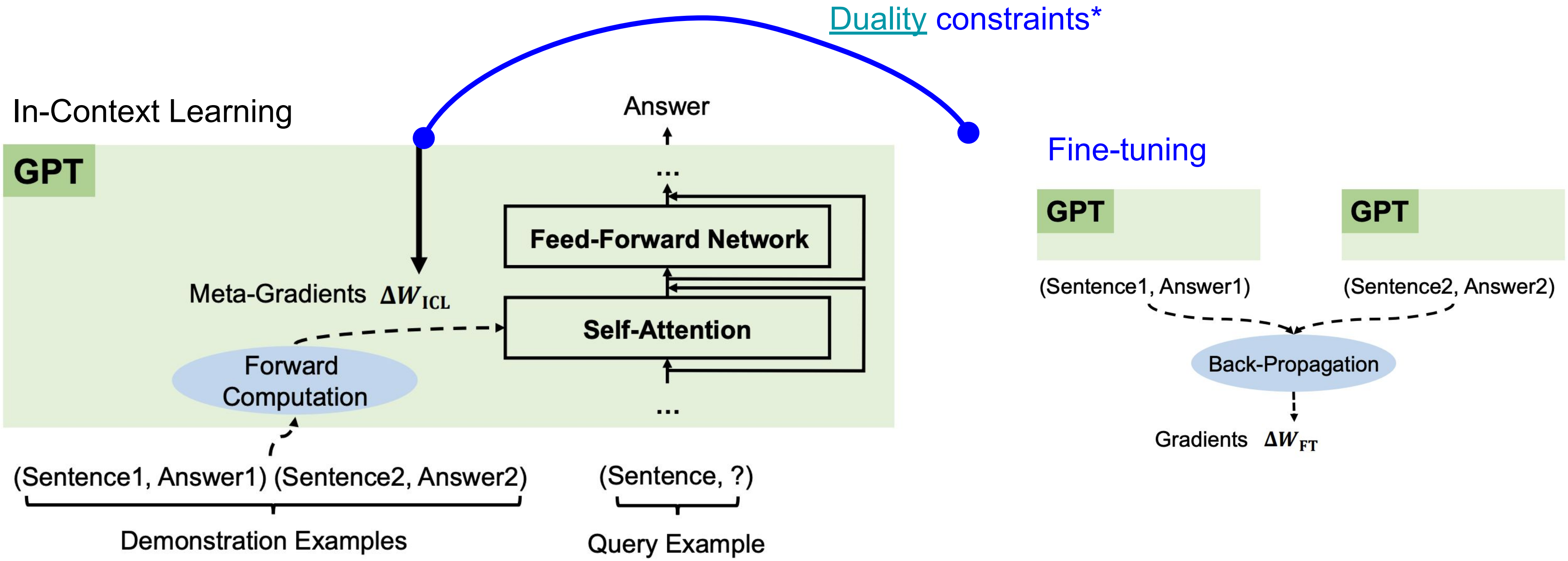


Image Source: Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers, ACL 23 Findings

**The Dual Form of Neural Networks Revisited: Connecting Test Time Predictions to Training Patterns via Spotlights of Attention*, ICML 22*

Takeaways I

- The power of in-context adaptation would be based on the **previous context sequences** (both latent vectors and outputs) in the duality of attention layers

- Can we guide sequences predictions to refunction LLM for zero-shot or few-shot unseen tasks? (e.g., ASR Rescoring)

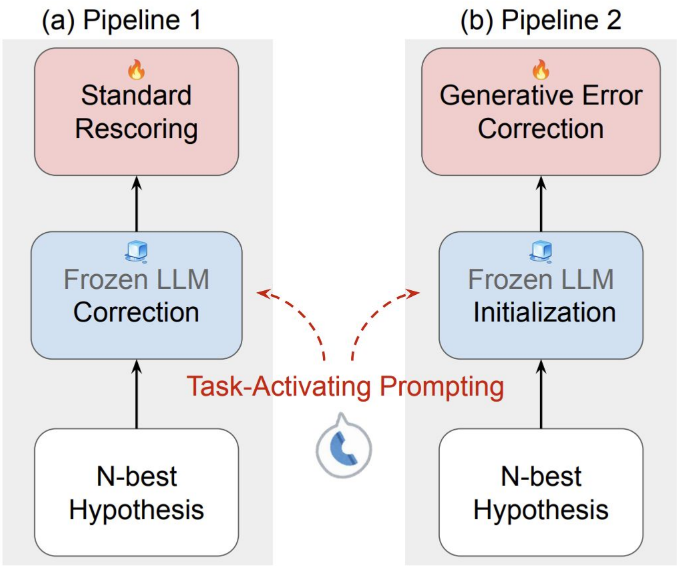
- Other properties on this task-oriented questions
  - Pros: no need to re-deploy a pre-trained model
  - Cons: longer context windows, latency …

Task-Activating Prompting (1/3)

Do you know speech recognition?

Do you know language modeling in speech recognition?

Would you give some examples of n-best rescoring in speech recognition?



Generative Speech Recognition Error Correction with Large Language Models and Task-Activating Prompting, *ASRU 23 C.-H. H. Yang et al.*

Task-Activating Prompting (2/3)

Q: *Do you know speech recognition?*
R: *Yes, I am familiar with speech recognition. Speech recognition, also known as automatic speech recognition (ASR) or speech-to-text, is the process of converting spoken language into text. This technology involves using algorithms and machine learning models to analyze and transcribe the acoustic features of spoken words and phrases. Speech recognition has many applications, including voice-controlled assistants, automated phone systems, and transcription services.*
Q: *Do you know language model for speech recognition?*
R: *Yes, I can provide you with some information about language models used in speech recognition. Language models are an essential component of Automatic Speech Recognition (ASR) systems. They help to predict the likelihood of a sequence of words, given the context. This helps the ASR system to better understand and transcribe spoken language into text. This separate model is typically more complex and powerful than the initial language model used for the transcription, and it is used to re-score the transcription based on the probability of the words occurring in the given context. The rescoring process involves taking the output of the initial language model, which is usually based on statistical methods such as Hidden Markov Models, and then applying a more advanced language model, such as a neural network-based language model, to generate a more accurate transcription. This is accomplished by re-ranking the possible transcriptions based on the probabilities assigned by the more advanced language model. Language model rescoring has been shown to significantly improve the accuracy of speech recognition systems, particularly in noisy or challenging environments where the initial language model may not perform well.*
Q: *Could you give a possible example of language model rescoring with some hypotheses?*
R: *Sure, here is an example of language model rescoring for ASR with 5-best hypotheses:*

1. *recognize speech with artificial intelligence.*
2. *recognized speech with artificial intelligence.*
3. *recognize speech with artificial intelligent.*
4. *reckon eyes speech with artificial intelligence.*
5. *recognize peach with artificial intelligence.*

After rescoring, the model assigns a higher probability to the first hypothesis, which is the correct transcription of the input. The other hypotheses receive lower probabilities, reflecting their decreased likelihood given the improved language model. In this case, the top-ranked hypothesis after rescoring would be chosen as the final output: recognize speech with artificial intelligence
Q: *Nice job, I will provide some examples as a demonstration from [target domain]. The 10-best hypothesis is:[hypotheses list from training set], and I would expect your output is: [corresponding transcription]. Following this example, could you report the true transcription from the following 10-best hypotheses:? [hypotheses list for inference]*

Generative Speech Recognition Error Correction with Large Language Models and Task-Activating Prompting, ***ASRU 23 C.-H. H. Yang et al.***

Task-Activating Prompting for Zero-Shot ASR-LM (2/3)

$\mathcal{P}_2$: zero-shot rescoring setup	WSJ	ATIS
(a) Oracle	9.78	6.43
(b) First pass	11.87	8.82

Table 2. WERs on ATIS and WSJ using prompting variants to enhance the $\mathcal{P}_2$ in-context learning pipeline. We report the results of InstructGPT and BLOOM as LLMs over 100B; GPT-2 and OpenLLaMA do not perform consistently in this setting.

	WSJ		ATIS	
In-context learning variant	InstructGPT	BLOOM	InstructGPT	BLOOM
$\mathcal{P}_1$: LLM-corrected N-best w/ RescoreBERT [31]	10.13	10.46	7.13	8.46
$\mathcal{P}_2$: (c) Zero-shot scoring	10.43	11.23	7.95	8.45
$\mathcal{P}_2$: (c) + zero-shot reasoning [16]	10.20	11.88	7.77	8.53
$\mathcal{P}_2$: (c) + domain-hint prompting [2]	10.98	11.45	7.59	8.49
$\mathcal{P}_2$: (d) Scoring with one example-pair	9.42	9.45	6.34	7.30
$\mathcal{P}_2$: (d) + zero-shot reasoning [16]	9.87	11.46	7.25	8.64
$\mathcal{P}_2$: (d) + domain-hint prompting [2]	9.70	10.99	6.19	7.12
$\mathcal{P}_2$: (d) + task-activating prompting (TAP)	**8.84**	**8.99**	**5.79**	**6.72**

Generative Speech Recognition Error Correction with Large Language Models and Task-Activating Prompting, *ASRU 23 C.-H. H. Yang et al.*

Task-Activating Prompting for Few-Shot ASR-LM (3/3)

Table 3: Cross-domain WER results by task-activated ICL [100] in *zero-shot* and *few-shot* settings. "o_{nb}" and "o_{cp}" respectively denote n-best oracle and compositional oracle that are defined in 5.2.

Domain Shift	Test Set	Baseline	n-shot In-Context Learning (ICL)				*Oracle*	
			$\boldsymbol{n} = 0$	$\boldsymbol{n} = 1$	$\boldsymbol{n} = 5$	$\boldsymbol{n} = 10$	o_{nb}	o_{cp}
Specific Scenario	WSJ-*dev93*	9.0	$8.5_{-5.6\%}$	$7.8_{-13.3\%}$	$7.7_{-14.4\%}$	$7.1_{-21.1\%}$	6.5	5.3
	WSJ-*eval92*	7.6	$7.3_{-3.9\%}$	$6.6_{-13.2\%}$	$6.6_{-13.2\%}$	$6.3_{-17.1\%}$	5.5	4.7
	ATIS	5.8	$5.5_{-5.2\%}$	$5.1_{-12.1\%}$	$5.0_{-13.8\%}$	$4.7_{-19.0\%}$	3.5	2.4
Common Noise	CHiME4-*bus*	18.8	$17.6_{-6.4\%}$	$16.7_{-11.2\%}$	$16.2_{-13.8\%}$	$15.9_{-20.7\%}$	16.8	10.7
	CHiME4-*caf*	16.1	$14.7_{-8.7\%}$	$14.3_{-11.1\%}$	$13.7_{-14.9\%}$	$13.2_{-18.0\%}$	13.3	9.1
	CHiME4-*ped*	11.5	$10.9_{-5.2\%}$	$9.9_{-14.4\%}$	$9.7_{-15.7\%}$	$9.4_{-18.3\%}$	8.5	5.5
	CHiME4-*str*	11.4	$10.9_{-4.4\%}$	$10.0_{-12.3\%}$	$9.7_{-14.9\%}$	$9.2_{-19.3\%}$	9.0	6.0
Speaker Accent	CV-*af*	25.3	$24.9_{-1.6\%}$	$24.2_{-4.3\%}$	$23.6_{-6.7\%}$	$22.6_{-10.7\%}$	23.6	21.7
	CV-*au*	25.8	$25.1_{-2.7\%}$	$24.1_{-6.6\%}$	$24.0_{-7.0\%}$	$23.3_{-9.7\%}$	24.9	21.8
	CV-*in*	28.6	$27.6_{-3.5\%}$	$25.6_{-10.5\%}$	$25.0_{-12.6\%}$	$24.4_{-14.7\%}$	27.1	22.6
	CV-*sg*	26.4	$26.5_{+0.4\%}$	$25.0_{-5.3\%}$	$25.1_{-4.9\%}$	$23.7_{-10.2\%}$	25.5	22.2

HyPoradise: An Open Baseline for Generative Speech Recognition with Large Language Models, NeurIPS 23 Data and Benchmark Track, *Chen & Hu et al.*

Appendix: Zero-shot Acoustic Scores Fusion in ATIS

Nice job, next, can you sum the acoustic model score provided in the fist column with language model score in the provided examples, computing by log likelihood of language model times a coefficient of 0.5?

-4.81 what are the american flights from newark to nashville
-4.71 one of the american floods from newark to nashville
-5.39 what are the american floods from newwork to nashvillevil

Sure! First, let's recap the hypotheses and their language model scores:

1. what are the american flights from newark to nashville (0.90)
2. one of the american floods from newark to nashville (0.05)
3. what are the american floods from newwork to nashvillevil (0.05)

Now, let's compute the log likelihood of the language model scores and multiply them by the coefficient of 0.5:

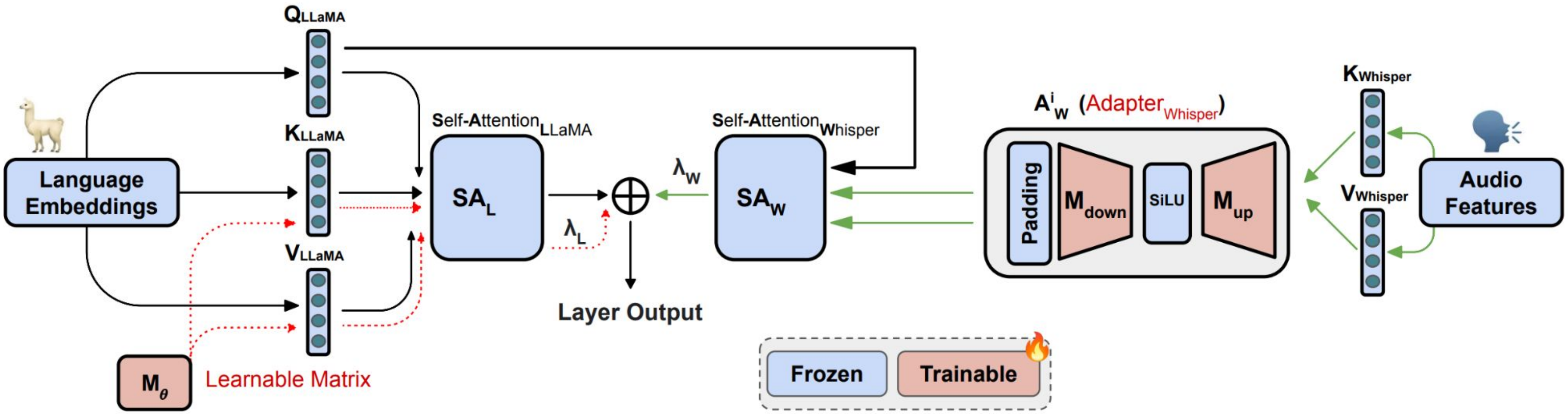
1. log(0.90) * 0.5 = -0.0458 * 0.5 = -0.0229
2. log(0.05) * 0.5 = -1.301 * 0.5 = -0.6505
3. log(0.05) * 0.5 = -1.301 * 0.5 = -0.6505

Finally, let's add the acoustic model scores provided in the first column:

1. (-4.81) + (-0.0229) = -4.8329
2. (-4.71) + (-0.6505) = -5.3605
3. (-5.39) + (-0.6505) = -6.0405

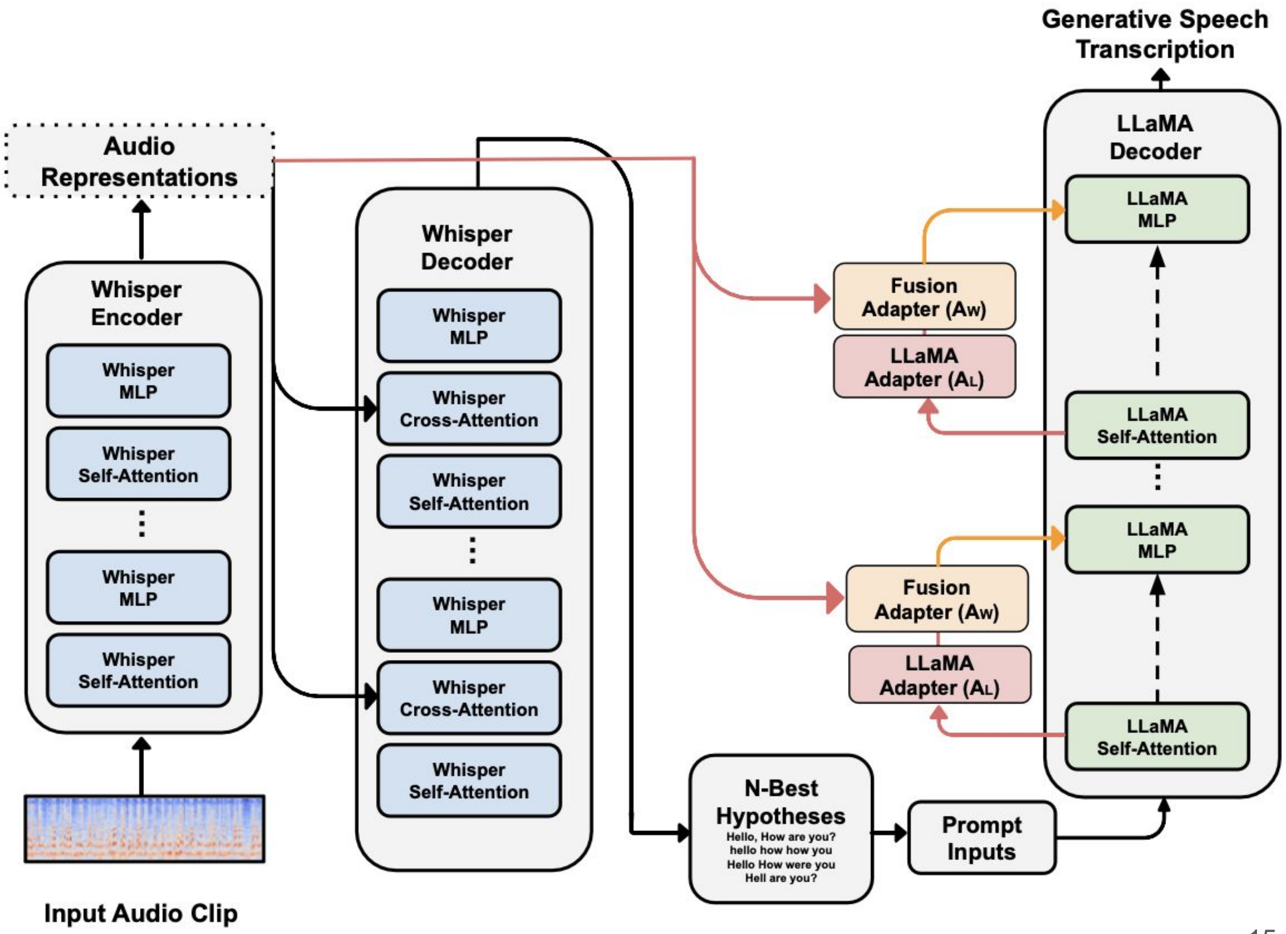
Multimodal Merging (1/2)

[Whispering Llama: A Cross-Modal Generative Error Correction Framework for Speech Recognition]
EMNLP 23, Radhakrishnan et al. [Github]



Multimodal Merging (2/2)

- token-level fusion
- attention heads merging
  - each transformer layer



Prompt Format in Whisper Llama

You

Instruction:
You are an ASR transcript selector. You have a few transcripts generated by an automatic speech recognition model. Your task is to generate the most likely transcript from them. If the generated transcripts have grammatical or logical errors, you will modify them accordingly to produce the most accurate and coherent transcript.
Input:
so that it can carry its momentum to the logic
that he can carry his moment on tour with the logic
that he can carry his momentum through with the logic
that he can carry his momentum to with an logic
that he can carry his momentum true within logic
that he carries momentum through with logic
that it can carry a moment and through with the logic
that it can carry a moment of truth with the logic
that it can carry a moment on tour with the logic
that it can carry a moment on true with the logic
that it can carry a momentum tool within logic
that it can carry as a moment on tour with the logic
that it can carry at moments and through with the logic
that it can carry his moment on tour with a logic
that it can carry his moment on tour with the logic
Response:

ChatGPT

The most likely and coherent transcript from the given options appears to be:

"that he can carry his momentum through with the logic"

This version is grammatically correct and logically coherent, suggesting that 'he' has the ability to maintain his momentum in accordance with logic or logical reasoning.

Whisper Llama (WL) Performance

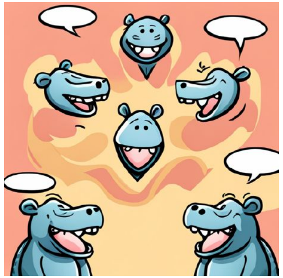
#	Method	#Para.	ATIS	GS_E	GS_P	GS_S	GS_M	$\mathrm{WER}_{\mathrm{Avg}}(\downarrow)$	WERR ($\uparrow$)
1	Oracle (1st-pass)	-	13.76	28.22	22.84	23.93	19.5	21.64	-
2	$\mathcal{WL}_L$	26.40M	2.04	21.76	19.21	20.55	11.6	15.03	30.52
3	$\mathcal{WL}_M$	7.97M	1.77	**21.61**	**16.20**	**18.02**	**9.82**	**13.48**	**37.66**
4	$\mathcal{WL}_S$	4.89M	1.89	22.24	17.23	19.157	10.185	14.144	34.62
5	$\mathcal{WL}_M$ w/o masking	4.89M	3.94	27.56	18.10	21.71	12.79	20.04	22.25
6	$\mathcal{WL}_M$ w/o $H_{\text{acoustics}}$	4.89M	253.20	123.19	203.44	376.81	256.44	242.61	-1020.68
7	$\mathcal{WL}_M$ w/o init.	4.89M	405.83	500.58	414.34	461.63	390.64	434.60	-1907.45
8	$\mathcal{WL}_M$ w/o SA_W	1.22M	1.66	24.99	18.734	20.73	10.86	15.39	28.83
9	Big-scale Adapter	4.91M	**1.45**	23.65	16.59	19.93	10.62	14.45	33.21

Ongoing Development of Hyporadise

If all speech people donate their n-best hypotheses, could we have a general generative error correction model?

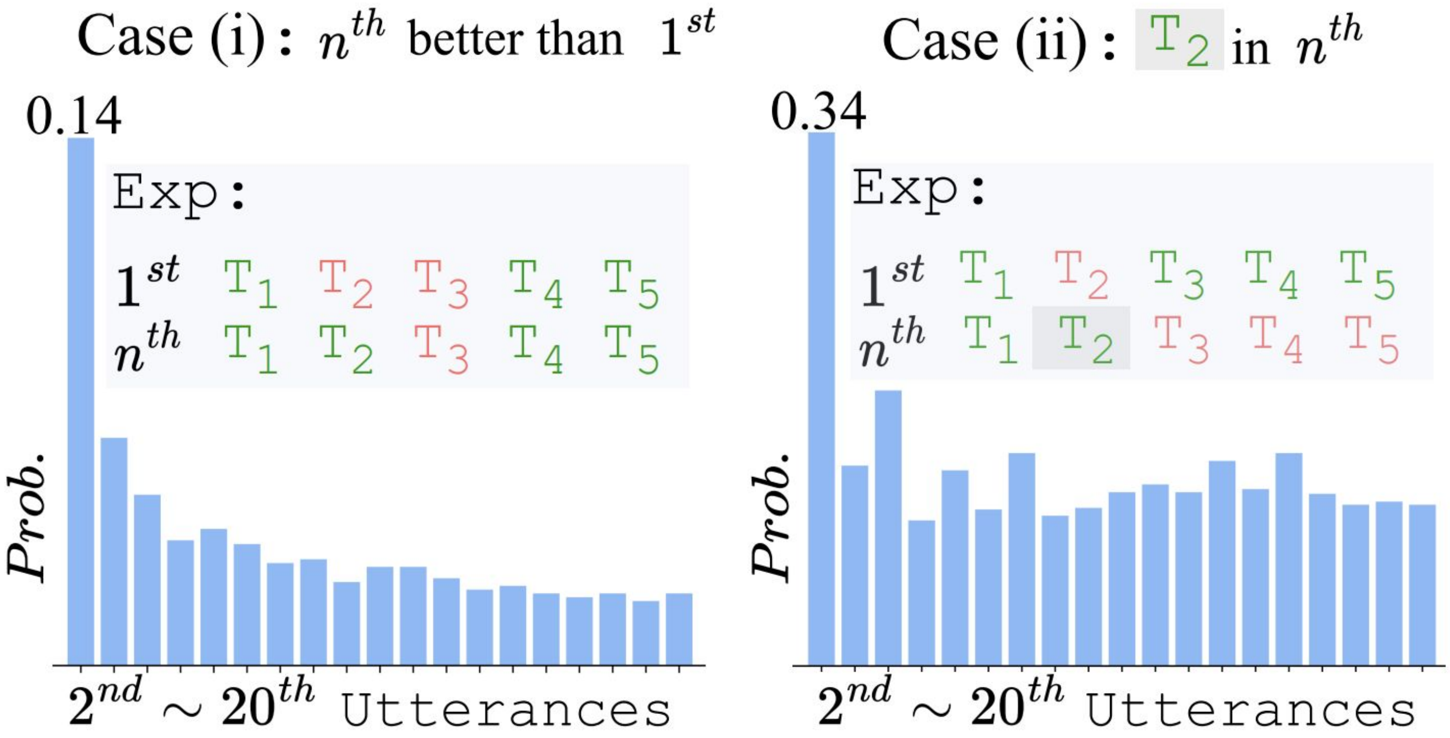
Prior to joining company in Nov 2022, I created a non-profit organization, Peaceful Data to open source data and pre-trained models.

Peaceful Data



HyPoradise

Toward Generative Error Correction for ASR (1/4)



Toward Generative Error Correction for ASR (2/4)

Source	Domain Category	Training Set	# Pairs	Length	Test Set	# Pairs	Length
LibriSpeech	Audiobooks	*train-960*	88,200	33.7	*test-clean* *test-other*	2,620 2,939	20.1 17.8
CHiME4	Noise	*train*	8,738	17.0	*test-real*	1,320	16.4
WSJ	Business news	*train-si284*	37,514	17.5	*dev93* *eval92*	503 333	16.7 17.3
SwitchBoard	Telephone	*train*	36,539	11.8	*eval2000*	2,000	11.8
CommonVoice	Accented English	*train-accent*	49,758	10.5	*test-accent*	2,000	10.5
Tedlium-3	TED talk	*train*	47,500	12.6	*test*	2,500	12.6
LRS2	BBC audio	*train*	42,940	7.6	*test*	2,259	7.6
ATIS	Airline info.	*train*	3,964	12.4	*test*	809	11.3
CORAAL	Interview	*train*	1,728	24.2	*test*	100	24.0
Total		*train*	316,881	18.1	*test*	17,383	14.1

Toward Generative Error Correction for ASR (3/4)

Table 2: WER (%) results of H2T-*ft* and H2T-*LoRA* in *fine-tuning* setting. "o_{nb}" and "o_{cp}" respectively denote n-best oracle and compositional oracle that are defined in 5.2.

Test Set	Baseline	LM_{rank}	H2T-*ft*		H2T-*LoRA*		Oracle	
			T5	LLaMA	T5	LLaMA	o_{nb}	o_{cp}
WSJ	4.5	4.3	4.0	3.8	2.7$_{-40.0\%}$	**2.2**$_{-51.1\%}$	4.1	1.2
ATIS	8.3	6.9	2.7	3.4	**1.7**$_{-79.5\%}$	1.9$_{-77.1\%}$	5.2	1.1
CHiME-4	11.1	11.0	7.9	8.2	7.0$_{-36.9\%}$	**6.6**$_{-40.5\%}$	9.1	2.8
Tedlium-3	8.5	8.0	6.6	5.2	7.4$_{-12.9\%}$	**4.6**$_{-45.9\%}$	3.0	0.7
CV-*accent*	14.8	16.0	12.9	15.5	11.0$_{-25.7\%}$	**11.0**$_{-25.7\%}$	11.4	7.9
SwitchBoard	15.7	15.4	15.9	18.4	14.9$_{-5.1\%}$	**14.1**$_{-10.2\%}$	12.6	4.2
LRS2	10.1	9.6	9.5	10.2	**6.6**$_{-34.7\%}$	8.8$_{-12.9\%}$	6.9	2.6
CORAAL	21.4	21.4	23.1	22.9	20.9$_{-2.3\%}$	**19.2**$_{-10.3\%}$	21.8	10.7

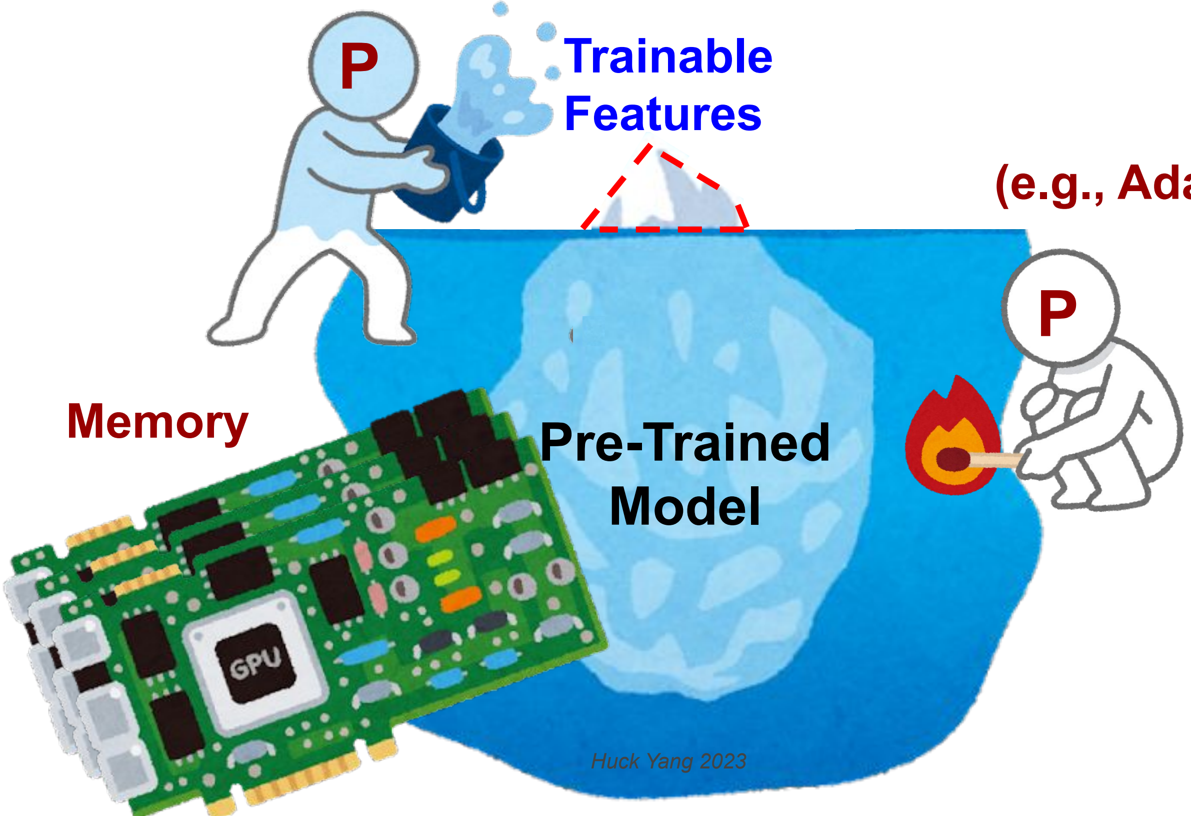
Toward Generative Error Correction for ASR (4/4)

Table 4: Case study of ICL. The utterance is drawn from WSJ-*dev93* dataset.

Type	Utterance	WER
1st Hypo. by AM	Bankers in Hong Kong expect xinnepec to return for more loans as it develops China's petro chemical industry.	16.7
2nd Hypo. by AM	Bankers in Hong Kong expect xinepec to return for more loans as it develops China's petrochemical industry.	8.3
Correction by LLM	Bankers in Hong Kong expect Sinopec to return for more loans as it develops China's petrochemical industry.	0
Ground-truth Transcription	Bankers in Hong Kong expect **Sinopec** to return for more loans as it develops China's **petrochemical** industry.	-

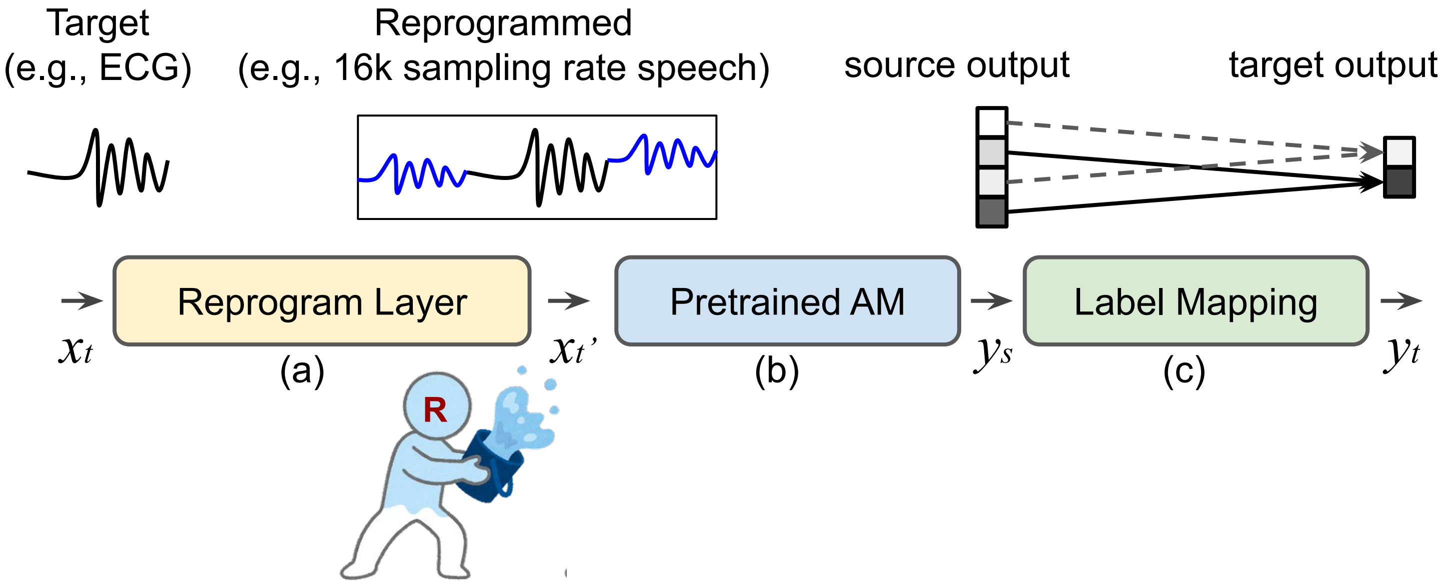
Interspeech 23 Tutorial …

Parameter Efficient Learning (e.g., Reprograms & Prompts)



[Recap] Cross-Modal Reprogram from Speech to Time Series

- Schematic illustration of the proposed Voice2Series Neural Reprogramming (CHH Yang et al. ICML 2021)



Acknowledgments

- **Amazon Alexa:** Dr. Ivan Bulyko, Dr. Andreas Stolcke, Dr. Björn Hoffmeister
- **Georgia Tech:** Prof. Alex Lerch and Prof. Chin-Hui Lee
- **Industry:** Dr. Pin-Yu Chen (IBM); Dr. Bo Li, Dr. Yu Zhang, Dr. Nanxin Chen, Dr. Heiga Zen and Dr. Tara N. Sainath (Google)
- **Academics:** Prof. Sabato Marco Siniscalchi (NTNU), Prof. Chao Zhang (THU), Prof. Jesper Tegner (KAUST), Prof. Hung-yi Lee (NTU), Prof. Yu Tsao (Sinica), Prof. Jen-Tzung Chien (NYCU), and Prof. Eng Siong Chng (Nanyang Tech)

Thank you

Let’s open source this science!

huckiyang@gatech.edu