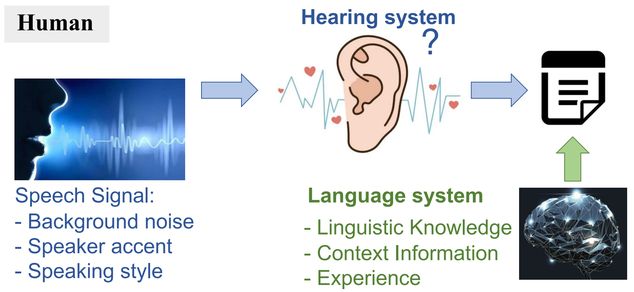
It's Never Too Late: "Fusing" Acoustic Information into Large Language Models (LLMs) for Automatic Speech Recognition

Chen Chen, Ruizhe Li, Yuchen Hu, Ruizhe Li, Sabato Marco Siniscalchi, Pin-Yu Chen,
Eng Siong Chng, and Huck Yang

CHEN1436@e.ntu.edu.sg, hucky@nvidia.com

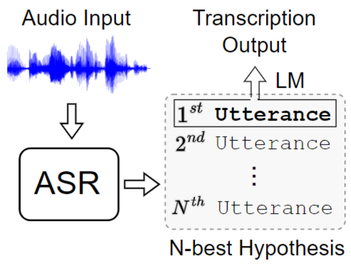
ICLR International Conference On Learning Representations

Research Motivation



Language system contributes to the robustness of our auditory system.

LM in ASR: Rescoring



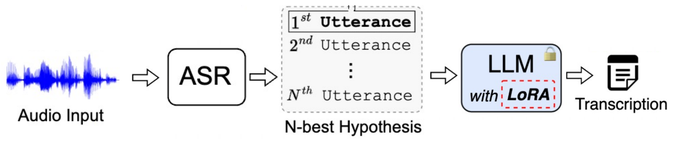
Traditional LM Rescoring:

(i) Beam search ➔N-best list.

(ii) In-domain LM ➔ Top1 candidate

(iii) Discarding others.

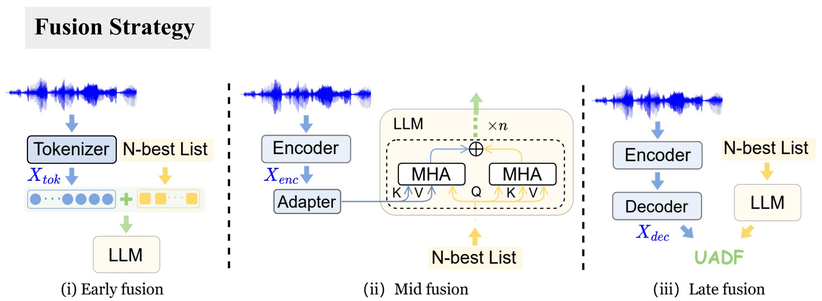
LLM in ASR: Generative Error Correction [1,2,3,4]



- GER makes full use of N-best hypos and LLM to predict GT.
- **Can we integrate acoustic information in GER process?**

[1] Chen, et al. Hyporadise: An open baseline for generative speech recognition with large language models, NeurIPS 2023.
[2] Hu et al. LLMs are Efficient Learners of Noise-Robust Speech Recognition, ICLR 24
[3] Yang et al. Generative ASR Error Correction with LLMs and Task-Activating Prompting, ASRU 2023
[4] Radharishnan et al. Whispering LLaMA: A Cross-Modal Generative Error Correction Framework for Speech Recognition, EMNLP 2023

"3 Body Problem" of N-Best, ASR, LLM



(i) Early Fusion: Concatenate speech representation and word embedding.

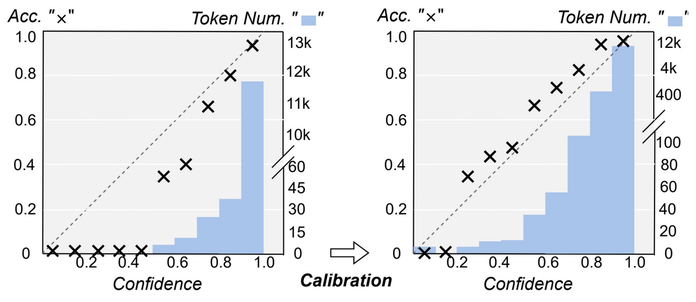
(ii) Mid Fusion: Leverage cross-attention mechanism for text-speech alignment.

(iii) Late Fusion: Combine the token-level logits in auto-regressive decoding.

UADF: Uncertainty-Aware Dynamic Fusion

- Calibration: align the model confidence with true accuracy

$$\mathrm{Conf}_{llm}(f^{llm}, \tau_1) \approx 1 - \mathrm{TER}_{llm}(f^{llm})$$



- Fusion: dynamically assign token-level weight according to uncertainty

$$P(Y_T) = \prod_{t=0}^{T} \mathrm{softmax}(\frac{f_t^{llm}}{\tau_1}) + (\mathrm{sigmoid}(\mathcal{U}_t^{llm}) - \beta)\ \mathrm{softmax}(\frac{f_t^{asr}}{\tau_2})$$

Fusing-LLM Experimental Result

WER on WSJ and ATIS datasets

Table 1: WER (%) and WERR results of early, mid, and late fusion on ATIS and WSJ dataset. "*W2v.*", "*Hub.*" and "*Whis.*" indicate Wav2vec2-large, HuBERT and Whisper model, respectively. "*Conc.*", "*Atten.*", and "*Stat.*" indicate concatenation, cross-attention and static fusion strategies introduced in [3]. "GER" denotes the H2T results of LLM that is consistent across the three fusion

Acoustic Info.	Fusion *where*	Fusion *how*	GER *ATIS*	GER *WSJ*	ASR-*only* *ATIS*	ASR-*only* *WSJ*	*WER* ↓ *ATIS*	*WER* ↓ *WSJ*	*WERR* ↑ *ATIS*	*WERR* ↑ *WSJ*
X_{tok} *by* W2v.	*early*	Conc.	1.61	2.83	-	-	2.16	3.21	-34.2%	-13.4%
by Hub.					-	-	2.02	3.11	-25.5%	-9.9%
X_{enc} *by* Whis.	*mid*	*Atten.*	1.61	2.83	-	-	1.75	2.59	-8.7%	8.5%
X_{dec} *by* ASR	*late*	*Stat.*	1.61	2.83	4.67	9.21	1.36	2.55	15.5%	9.9%
		UADF					**1.24**	**2.47**	**23.0%**	**12.7%**

ASR-LLM Ablation Study

Table 2: Ablation study of WER (%) and WERR results on the ATIS dataset based on UADF using late fusion. The difference between the system ID-1 to ID-3 is the different performance of ASR-*only* model (X_{dec}), and the system ID-4 to ID-5 varies based on the different combination of "*Cali.*" and "*Dyn.*". "Static" does not utilize either "*Cali.*" and "*Dyn.*".

System ID	GER	ASR-*only* (X_{dec})	ID-C *WER*	ID-C *WERR*	Static *WER*	Static *WERR*	UADF *Cali.*	UADF *Dyn.*	UADF *WER*	UADF *WERR*
1		12.16	2.41	-49.7%	1.51	6.2%	✓	✓	1.52	5.6%
2	1.61	8.22	1.96	-21.7%	1.45	9.9%	✓	✓	1.39	13.7%
3		4.67	1.57	2.5%	1.36	15.5%	✓	✓	**1.24**	**23.0%**
4	1.61	4.67	1.57	2.5%	1.36	15.5%	✓	✗	1.33	17.4%
5							✗	✓	1.39	13.7%

Noise-robustness on CHiME

Noise Type	ASR-*only*	GER	Static *WER*	Static *WERR*	UADF *WER*	UADF *WERR*
bus	12.45	8.67	8.05	7.2%	**7.98**	**8.0%**
caf	11.48	6.96	6.37	8.5%	**6.22**	**10.6%**
ped	11.36	5.49	4.96	9.8%	**4.82**	**12.2%**
str	12.28	5.86	5.28	9.9%	5.28	9.9%
Avg.	11.89	6.75	6.17	8.6%	**6.08**	**9.9%**

Github

Conclusion

- We focus on fusing acoustic information into LLM-based GER.
- We present a simple yet effective solution UADF that performs late fusion in the auto-regressive decoding process.
- UADF dynamically assimilates information from the audio modality, leading to more reasonable token-level decisions.
- UADF seamlessly adapts to noise-robust ASR as well as AVSR.