BACK TO BASICS: REVISITING ASR IN THE AGE OF VOICE AGENTS

Geeyang Tay[†], Wentao Ma[†*], Jaewon Lee, Yuzhi Tang, Daniel Lee, Weisu Yin, Dongming Shen, Silin Meng, Yi Zhu, Mu Li, Alex Smola

Abstract

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

- 大背景：ASR（Automatic speech recognition）在测试集上表现良好，真实场景下存在易识别崩盘，语义脑补，跨语言失效问题

- 贡献：WildASR

环境退化(environmental degradation)
人群偏移(demographic shift)
语言多样性(linguistic diversity)

Background

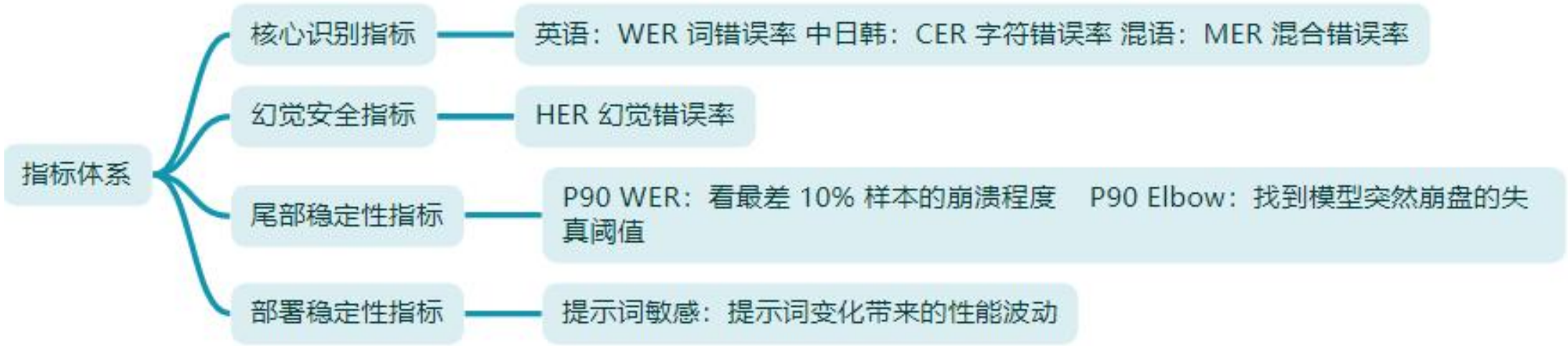
- 真实场景：电话压缩、远场、儿童语音、语码转换等场景 表现拉胯
- 评测体系失效：现有基准只测域内干净数据，用平均 WER 掩盖具体失效因素
- 无诊断工具：开发者无法知道哪种环境、哪类人群、哪种语言现象会导致模型崩溃；
- 幻觉风险被忽视:残缺音频下模型自动补全语义，直接导致语音智能体执行错误指令

Construction of dataset

Table 1: **Overview of the proposed WildASR.** Each OOD dimension is decomposed into explicitly defined subcategories. For each subcategory, we report the covered languages, the number of samples per language, the average utterance duration, and the curation steps applied (defined in §3.1). Detailed data sources are listed in Appendix D.

Categories	Languages	#Samples	Avg Duration (s)	Curation Steps
Environmental degradation				
Reverberation	EN/ZH/JA/KO	2841/3735/2850/2046	10.0/12.9/12.5/10.4	DC→QF→NR→AA→MV
Far-field	EN/ZH/JA/KO	2841/3735/2850/2046	10.0/13.0/12.5/10.4	DC→QF→NR→AA→MV
Phone codec	EN/ZH/JA/KO	1894/2490/1900/1364	7.5/10.4/10.0/7.9	DC→QF→NR→AA→MV
Noise gap	EN/ZH/JA/KO	3788/4980/3800/2728	8.6/11.6/11.2/9.1	DC→QF→NR→AA→MV
Clipping	EN/ZH/JA/KO	947/1245/950/682	7.5/10.4/10.0/7.9	DC→QF→NR→AA→MV
Demographic shift				
Children	EN/ZH	300/1000	4.06/3.02	DC→SF→QF→NR→MV
Older adults	EN/ZH	300/1000	5.93/1.95	DC→SF→QF→NR→MV
Accent	EN/ZH	1000/1000	3.48/5.69	DC→SF→QF→NR→MV
Linguistic diversity				
Short utterances	EN/ZH/JA/KO	318/367/467/255	1.2/0.7/1.1/1.0	DC→QF→NR→MV
Incomplete audio	EN/ZH/JA/KO	2345/2517/195/396	3.9/2.1/1.9/2.6	DC→QF→NR→MT→MV
Code-switching	(EN)+ZH/JA/KO	700/700/700	8.6/11.7/11.5	DC→QF→NR→MV

Indicator


指标体系
核心识别指标
英语：WER 词错误率 中日韩：CER 字符错误率 混语：MER 混合错误率
幻觉安全指标
HER 幻觉错误率
尾部稳定性指标
P90 WER：看最差 10% 样本的崩溃程度　P90 Elbow：找到模型突然崩盘的失真阈值
部署稳定性指标
提示词敏感：提示词变化带来的性能波动

Experiments

Figure 1: **Multilingual ASR robustness under real-world distribution shifts in WildASR.** We evaluate seven ASR systems across four languages and aggregate performance over three OOD dimensions. The horizontal line denotes the in-distribution clean-set model-average reference (5.7%), defined as the average error rate on the FLEURS test set across all models and languages. The sharp and uneven degradation across OOD conditions shows that human-parity performance on in-distribution data does not reliably transfer to real-world settings.

实验使用的模型:

Whisper Large V3（开源标杆）

GPT-4o Transcribe（OpenAI）

Gemini 2.5 Pro / Gemini 3 Pro（Google）

Qwen2-Audio（通义千问）

Nova 2（Deepgram）

Scribe V1（ElevenLabs）

Experiments

Table 2: **Impact of environmental degradations on multilingual ASR performance.** Average error rates across seven ASR models under controlled acoustic perturbations. Results are reported as MagicData / FLEURS. Δ denotes the absolute increase in error rates relative to the clean condition. Bold highlights the largest degradation magnitude per language and dataset.

Perturbations	EN		ZH		JA		KO	
	WER (%)	Δ	CER (%)	Δ	CER (%)	Δ	CER (%)	Δ
Original	*19.9/4.1*	–/–	*14.6/7.8*	–/–	*19.7/5.1*	–/–	*19.5/5.9*	–/–
Reverberation	31.9/9.5	+12.0/+5.3	25.7/13.0	+11.1/+5.2	45.3/15.5	+25.5/+**10.4**	46.6/15.5	+27.0/+9.6
Far-field	26.0/15.8	+6.1/+**11.7**	23.1/12.3	+8.5/+4.5	33.7/13.5	+13.9/+8.4	40.1/19.1	+20.6/+**13.2**
Phone (G.711)	20.5/10.4	+0.6/+6.3	16.9/8.6	+2.3/+0.8	29.1/6.7	+9.4/+1.6	24.8/8.8	+5.3/+2.9
Phone (GSM)	22.9/5.2	+3.0/+1.1	25.0/9.4	+10.4/+1.6	33.8/10.5	+14.1/+5.4	47.6/7.9	+28.0/+2.0
Noise gap	87.6/6.7	**+67.7**/+2.5	24.9/13.2	+10.3/+5.4	138.7/10.0	**+118.9**/+5.0	140.5/12.8	**+121.0**/+6.8
Clipping	30.6/15.6	+10.7/+11.5	37.3/17.9	**+22.7/+10.1**	52.0/13.6	+32.3/+8.5	46.6/18.5	+27.0/+12.5

Experiments

Table 3: **ASR performance under demographic shift.** English remains relatively robust, while Chinese and child speech exhibit substantially higher error rates.

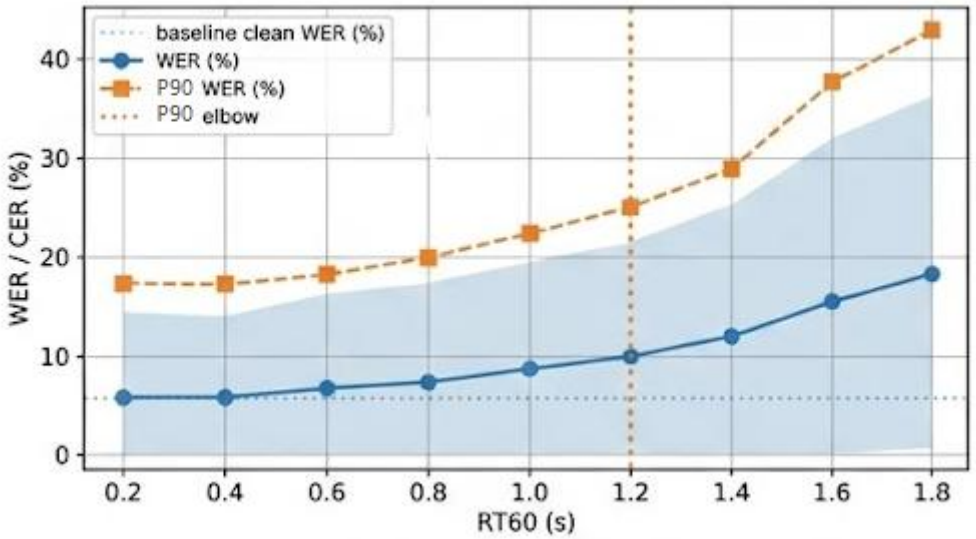
	Accent		Children		Older	
Model	ZH	EN	ZH	EN	ZH	EN
Nova 2	59.2	6.6	54.4	27.4	51.6	2.9
GPT-4o Transcribe	40.7	2.6	39.9	29.4	36.0	1.1
Gemini 2.5 Pro	49.9	5.0	58.6	25.1	52.6	1.8
Gemini 3 Pro	62.5	3.0	55.3	**18.2**	41.4	0.7
Qwen2-Audio	**7.5**	6.8	**23.4**	26.7	**18.6**	1.5
Scribe V1	37.9	**2.2**	65.1	29.3	42.3	0.8
Whisper Large V3	51.0	4.1	52.0	21.7	34.0	**0.2**

Experiments

Table 4: **ASR performance and hallucination behavior under linguistic diversity.** We can see that short and truncated inputs induce high error and frequent hallucinations, revealing semantic failures not captured by lexical metrics alone (EN not applicable for code-switching).

Model	Category	WER/CER/MER (%)				HER (%)			
		ZH	EN	JA	KO	ZH	EN	JA	KO
Nova 2	code-switch	33.7	-	32.0	56.4	68.4	-	58.1	71.9
	short	57.6	43.2	56.8	65.3	52.6	36.3	46.9	59.6
	incomplete	35.0	13.5	37.2	61.0	38.1	7.8	37.4	56.8
Gemini 2.5 Pro	code-switch	20.7	-	9.8	18.2	7.0	-	9.4	19.1
	short	40.6	64.4	48.6	55.6	30.5	35.4	28.1	34.1
	incomplete	31.9	15.3	37.1	23.5	31.6	10.5	32.8	11.1
Gemini 3 Pro	code-switch	7.2	-	9.0	9.4	3.7	-	6.3	11.9
	short	33.9	73.9	55.4	47.8	15.5	27.3	31.2	23.5
	incomplete	21.7	10.6	36.7	18.5	16.9	6.7	25.6	14.1
GPT-4o Transcribe	code-switch	21.9	-	24.4	29.9	12.0	-	17.9	36.9
	short	26.9	38.7	37.3	26.9	21.5	20.5	21.9	21.2
	incomplete	25.4	38.1	26.6	22.9	22.3	12.4	26.1	12.6
Qwen2-Audio	code-switch	12.3	-	80.5	211.7	8.9	-	35.6	85.7
	short	21.4	40.7	59.2	102.6	14.7	23.3	40.6	73.3
	incomplete	20.5	13.0	224.4	34.2	14.7	6.1	21.5	37.6
Scribe V1	code-switch	10.2	-	22.8	23.9	7.6	-	20.1	31.7
	short	38.3	57.2	94.9	58.3	30.0	38.5	50.0	32.6
	incomplete	25.5	12.9	36.4	18.8	30.5	10.9	38.5	15.7
Whisper Large V3	code-switch	12.0	-	22.8	29.6	10.9	-	23.6	38.0
	short	41.6	39.8	154.1	92.0	31.6	21.4	9.4	22.0
	incomplete	24.0	12.2	26.7	17.7	21.0	7.7	19.5	12.9

Experiments



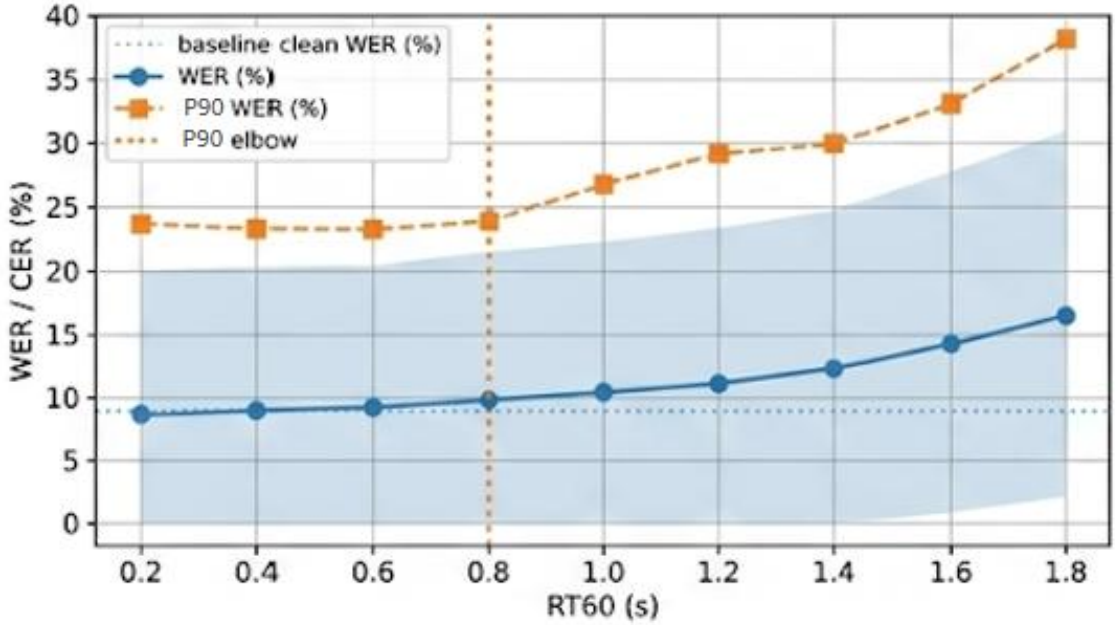


Figure 3: **ASR error dynamics under increasing reverberation** for Qwen2-Audio on FLEURS (top: English, bottom: Chinese).

Thanks