Whisper ASR 幻觉问题研究

从问题发现到解决方案

ASR 论文阅读报告

2025 年 12 月 31 日

汇报人：王宇东

目录

背景介绍

Whisper 幻觉问题

幻觉现象深入调查

Calm-Whisper 解决方案

总结与展望

背景介绍

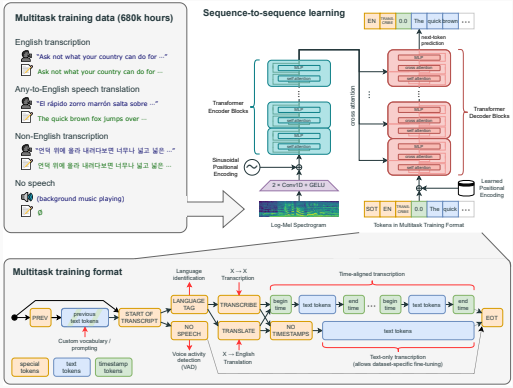
Whisper **模型概述**

OpenAI Whisper (2022)

- **训练数据**: 680,000 小时多语言弱监督数据
- **架构**: Encoder-Decoder Transformer
- **多任务**: 语音识别、翻译、语言识别、VAD
- **多语言**: 支持 96+ 种语言

核心优势

- 零样本迁移能力强
- 无需针对特定数据集微调
- 接近人类水平的准确率

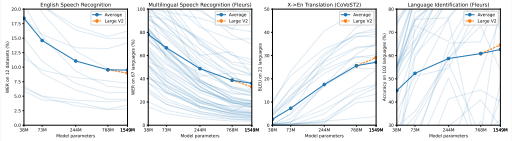


图源: Radford et al., 2022

Whisper 的卓越性能

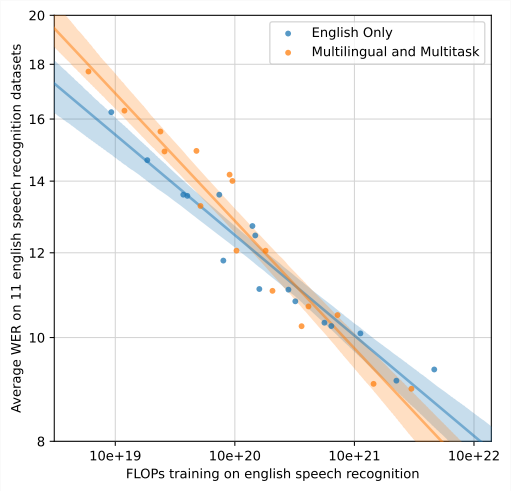
零样本鲁棒性

- LibriSpeech test-clean: **2.5% WER**
- 相比 wav2vec 2.0 平均错误率降低 **55.2%**
- 在噪声环境下表现优异



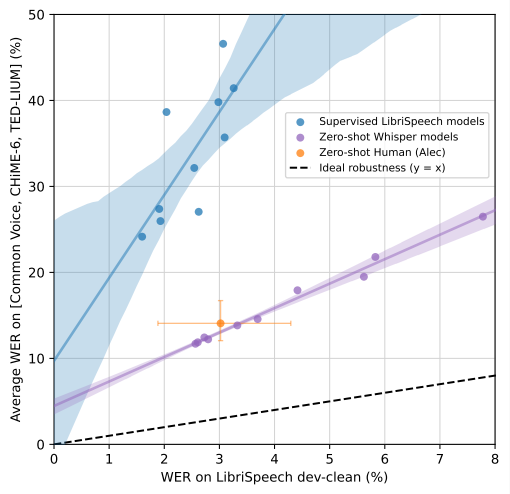
模型规模与性能关系

多任务格式

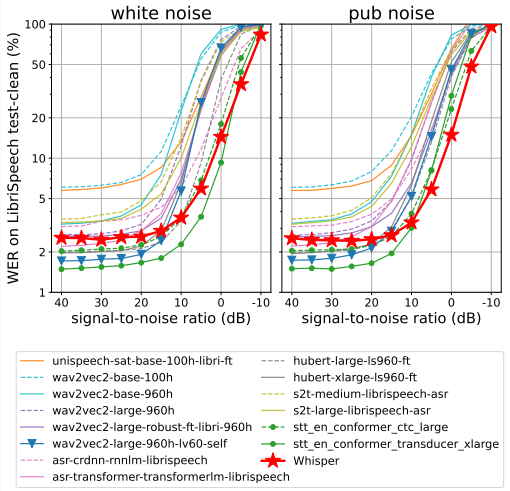


特殊 token 指定：

Whisper 的鲁棒性与人类水平



零样本 Whisper 接近人类鲁棒性



噪声环境下 Whisper 表现更优

Whisper **幻觉问题**

什么是 ASR 幻觉?

定义

ASR 幻觉是指模型生成与原始音频**没有语音或语义联系**的文本输出。

幻觉类型

- 非语音幻觉: 对纯噪声生成文字
- 循环幻觉 (Looping): 重复之前的文本
- 内容幻觉: 生成与音频无关的内容

产生原因

- 自回归解码的累积误差
- 训练数据中的噪声和偏差
- 模型对模糊输入的过度自信

危害

部分幻觉可能包含**暴力、色情**等有害内容，在实际应用中造成严重问题!

幻觉示例

非语音幻觉案例

输入音频	Whisper 输出
▷ 狗叫声	*"Thank you for watching!"*
▷ 汽车引擎	*"Please subscribe..."*
▷ 纯静音	*"I'm going to show you..."*

这些输出与音频内容完全无关

循环幻觉案例

实际语音: "Hello"

"Hello hello hello hello hello hello hello hello..."

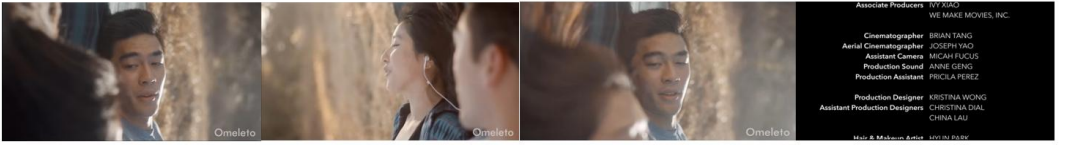
有害内容幻觉

研究发现部分幻觉包含:

- 暴力描述
- 色情内容
- 种族歧视言论

来源: Koenecke et al., PNAS 2020

whisper-large-v3 vs parakeet-tdt-0.6b-v3



woman : Hey, do you like the soulmates?

woman : I wonder if you have more than one, or they're like different kinds.

man:Yeah, I think so.

09:01.040 --> 09:04.040
Hey, do you like the soulmates?

09:05.540 --> 09:10.040
I wonder if you have more than one, or if there are different kinds?

09:13.540 --> 09:15.540
Yeah, I think so.

09:20.040 --> 09:25.040
Translated by Hua Chenyu English Subs

09:50.040 --> 09:55.040
Subtitles brought to you by the Figaro Cuisine Squad at Viki

Parakeet TDT

00:09:00.799 --> 00:09:03.440
Hey, do you like the soulmates?

00:09:05.360 --> 00:09:09.519
I wonder if you have more than one, or if they're like different kinds.

00:09:13.440 --> 00:09:14.960
Yeah, I think so.

Whisper tends to produce hallucinations in segments without speech, especially in background music segments, whereas Parakeet performs much better.

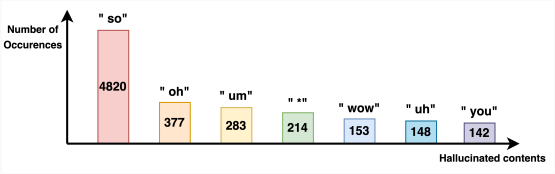
非语音幻觉的严重性

UrbanSound8K 测试结果

(城市环境声音数据集：警笛、狗叫、引擎声等)

模型	幻觉率
Whisper-large-v3	99.97%
Conformer-CTC-large	13.52%

几乎所有非语音音频都会触发幻觉



图源: Wang et al., Calm-Whisper, 2025

幻觉现象深入调查

评估指标介绍

语音识别性能指标

WER (Word Error Rate)

$$\text{WER} = \frac{S + D + I}{N} \times 100\%$$

- S = 替换词数
- D = 删除词数
- I = 插入词数
- N = 参考文本总词数

WER 越低越好，Whisper large: ∼2-4%

幻觉评估指标

幻觉率 (Hallucination Rate)

$$\text{HR} = \frac{\text{产生幻觉的样本数}}{\text{总样本数}} \times 100\%$$

循环率 (Looping Rate)
检测输出中重复片段的比例

其他指标

- CER: 字符错误率
- Top-K **率**: 最常见 K 种幻觉的占比

大规模幻觉实验 (ICASSP 2025)

实验设置

- 数据集: AudioSet + Musan + UrbanSound8K + FSD50K
- 总计 301,317 个非语音音频文件
- 移除所有可能包含语音的音频

实验结果

- 幻觉发生率: 40.3%
- 其中 9.1% 涉及循环 (looping)
- 生成 41,231 种不同输出
- 67% 的幻觉来自 1,270 个重复短语

关键发现

- 超过 35% 的幻觉是两个短语
- Top 10 占所有幻觉的一半以上
- 与 YouTube 字幕训练数据相关

音频长度对幻觉的影响

长度 [s]	幻觉率	循环率	Top30 率
原始	70.5%	18.5%	76.7%
1	52.1%	0.7%	84.2%
10	**11.6%**	3.4%	47.1%
20	27.4%	4.8%	47.5%
30	62.3%	9.6%	84.6%

* Top30 = 最常见的 30 种幻觉

观察结论

- 10 秒左右幻觉率**最低**
- 极短 (1s) 和较长 (30s) 音频幻觉率高
- **30 秒**是 Whisper 的解码窗口边界
- 音频内容与幻觉内容关联性弱

幻觉词袋 (Bag of Hallucinations)

构建方法

1. 收集所有非语音音频的转录结果
2. 使用 n-gram 语言模型计算概率
3. 过滤条件:
   - log 概率 < -10
   - 出现次数 > 4

典型 BoH 内容

- ”thanks for watching”
- ”thank you for watching”
- ”subtitles by the amara org community”
- ”i’m not sure what i’m doing here”

幻觉内容	占比	n-gram
thank you	24.76%	-9.22
thanks for watching	10.32%	-13.32
so	3.80%	-7.76
thank you for watching	2.58%	-12.42
the	2.50%	-6.67

高亮行属于 BoH（低概率但高频）

后处理缓解方案

方法流程

1. **去循环** (Delooping): 检测并移除重复文本
2. **BoH 移除**: 使用 Aho-Corasick 算法匹配并删除
3. (可选) 强制音素对齐验证

对比方法

- Whisper 参数调整 (beam size, threshold)
- VAD 预处理 (WebRTC, SileroVAD)
- 去噪 (DCCRN) —效果不佳

方法	WER
无处理	104.8%
beam size 1	107.2%
hall. threshold 20	39.8%
WebRTC VAD	68.3%
SileroVAD	8.0%
Deloop + BoH	17.1%
SileroVAD + Deloop + BoH	**6.5%**

组合方法效果最佳

Calm-Whisper **解决方案**

核心假设：Crazy Heads

假设

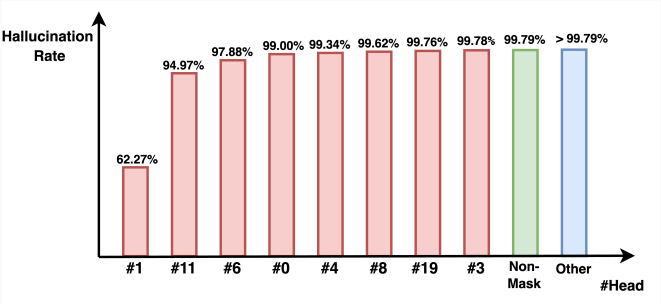
Whisper 解码器中的某些自注意力头对噪声特别敏感，是导致幻觉的主要原因。

实验方法

1. 逐个掩码 20 个注意力头
2. 测量幻觉率变化
3. 识别”crazy heads”

发现

- 8 **个头**导致幻觉增加
- 掩码 #1 头：幻觉率降 30%+
- 其余 12 个头抑制幻觉



图源: Wang et al., Calm-Whisper, 2025

多头掩码实验

掩码头	幻觉率	WER test-clean	WER test-other
无掩码	99.97%	**2.12%**	**4.07%**
#1, #6	50.16%	5.70%	5.48%
#1, #11	70.03%	3.80%	6.07%
#6, #11	57.08%	2.37%	4.66%
#1, #6, #11	**24.10%**	3.57%	5.98%
#0, #1, #6, #11	28.91%	15.87%	21.39%

关键发现

- 三头 (#1, #6, #11) 贡献 75%+ 的幻觉
- 掩码四头会严重损害识别性能

权衡

- 掩码三头是最佳平衡点
- 但仍有 WER 损失
- 需要更好的解决方案

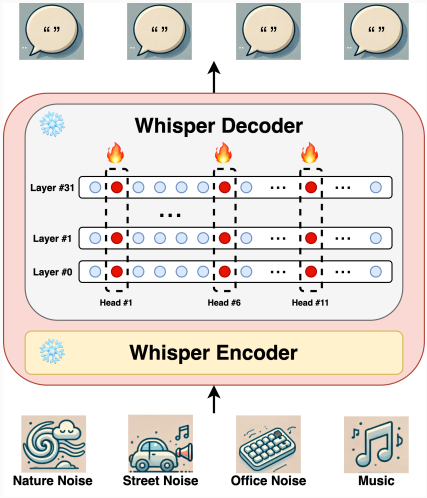
Calm-Down 微调策略

核心思路

保留所有头进行推理，但只对 crazy heads 进行微调

训练设置

- **冻结**: 除 #1, #6, #11 外的所有参数
- **数据**: 105 小时非语音音频
  - AudioSet (无语音部分)
  - DEMAND
  - Musan (音乐 + 噪声)
- **标签**: 空字符串
- **超参**: batch=128, lr=10^{-6}



图源: Wang et al., Calm-Whisper, 2025

Calm-Whisper 实验结果

模型	幻觉率	WER test-clean	WER test-other
原始 (Non-Mask)	99.97%	**2.12%**	**4.07%**
掩码三头 (Mask)	24.10%	3.57%	5.98%
ft-decoder-3epochs	0.01%	100%	100%
ft-3heads-3epochs	69.79%	2.16%	4.08%
Calm-Whisper (5 epochs)	**15.51%**	2.19%	4.13%

关键结论

- 幻觉率下降 84%
- WER 仅增加 ~0.07%
- 长幻觉大幅减少

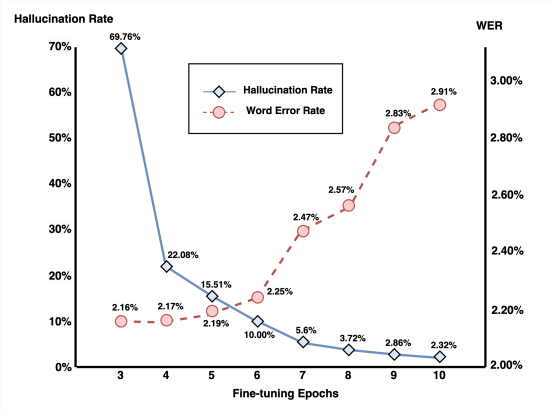
对比：全解码器微调

- 幻觉率降至 0.01%
- WER 飙升至 100%
- 完全失效

局限性

- 仍有 15% 幻觉率
- 仅验证 large-v3
- 需要额外训练资源

微调深度的影响



图源: Wang et al., Calm-Whisper, 2025

观察

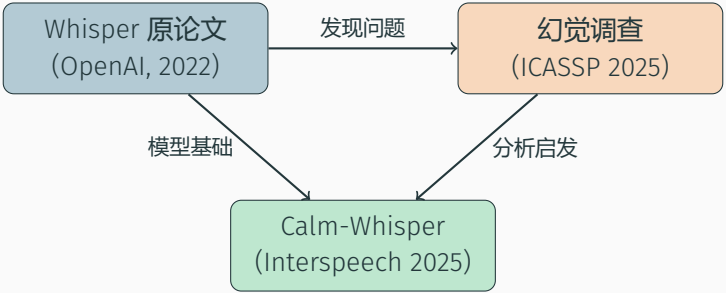
- 幻觉率：初期快速下降，后平缓
- WER：初期缓慢上升，后加速
- 5 epochs 是最佳平衡点

为什么有效?

- 非微调头保留原始能力
- 提供必要的冗余和鲁棒性

总结与展望

三篇论文的关系



Whisper **原论文**

- 展示弱监督的潜力
- 暴露幻觉问题

ICASSP **调查**

- 量化幻觉现象
- 提出后处理方案

Calm-Whisper

- 定位问题根源
- 从模型内部解决

方法对比

方法	类型	幻觉降低	WER 影响	复杂度
VAD 预处理	前处理	中等	低	低
参数调整	推理时	低	低	低
BoH 后处理	后处理	高	可能误删	中
Calm-Whisper	模型微调	**很高**	**极低**	中
组合方案	多阶段	最高	低	高

推荐策略

- 对准确性要求高: Calm-Whisper + VAD + BoH
- 快速部署: VAD 预处理 + BoH 后处理
- 研究探索: 注意力头分析 + 针对性微调

未来研究方向

模型层面

- 更深入的注意力头分析
- 其他架构的幻觉研究
- 多语言幻觉特性
- 强化学习优化解码

数据层面

- 更好的训练数据过滤
- 对抗性训练
- 多样化非语音数据

应用层面

- 实时幻觉检测
- 置信度校准
- 工业场景适配

开放问题

- 为什么特定头容易幻觉?
- 如何预测幻觉发生?
- 幻觉与模型能力的关系?

总结

核心要点

1. Whisper 在非语音音频上存在严重幻觉问题（接近 100%）
2. 幻觉主要由解码器中的少数注意力头引起
3. Calm-Whisper 通过定向微调将幻觉率降低 84%，同时保持识别性能
4. 后处理方法可作为补充安全措施

实践建议

- 工业部署需要幻觉防护
- 多种方法组合效果最佳
- 关注有害内容过滤

研究启示

- 大模型的可解释性研究
- 局部微调的有效性
- 数据质量的重要性

谢谢！

Q & A

参考文献 i

1. Radford, A., et al. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*. ICML 2022.
2. Wang, Y., et al. (2025). *Calm-Whisper: Reduce Whisper Hallucination On Non-Speech By Calming Crazy Heads Down*. Interspeech 2025.
3. Barański, M., et al. (2025). *Investigation of Whisper ASR Hallucinations Induced by Non-Speech Audio*. ICASSP 2025.
4. Koenecke, A., et al. (2020). *Racial disparities in automated speech recognition*. PNAS.
5. Peng, Y., et al. (2024). *Owsm v3. 1: Better and faster open whisper-style speech models based on e-branchformer*. arXiv.