arXiv:2512.17562v1 [cs.SD] 19 Dec 2025

当降噪适得其反：语音增强对现代医疗自动语音识别系统影响的系统性研究

技术报告

Sujal Chondhekar **Vasanth Murukuri** **Rushabh Vasani** **Sanika Goyal**

Rajshree Badami **Anushree Rana** **Sanjana SN** **Karthik Pandia** **Sulabh Katiyar**

Neha Jagadeesh **Sankalp Gulati**

EkaCare（Orbi Health Private Limited），班加罗尔，印度（sankalp@eka.care）

摘要

人们通常认为，语音增强方法能够提升噪声环境下自动语音识别（ASR）的性能。然而，对于在多样化含噪数据上训练的现代大规模自动语音识别模型而言，这些技术的有效性并不能想当然。我们使用九种噪声条件下的500段医疗语音录音，系统评估了MetricGAN-plus-voicebank降噪对四种最先进自动语音识别系统——OpenAI Whisper、NVIDIA Parakeet、Google Gemini Flash 2.0和Parrotlet-a——的影响。自动语音识别性能采用语义词错误率（semWER）进行衡量，这是一种纳入领域特定规范化的归一化词错误率（WER）指标。结果揭示了一个违反直觉的发现：在所有噪声条件和模型下，语音增强预处理都会降低自动语音识别性能。在测试的全部40种配置（4个模型 $\times$ 10 条件）中，原始含噪音频的semWER均低于增强音频，绝对semWER增幅为1.1%至46.6%。这些发现表明，现代自动语音识别模型具备足够的内部噪声鲁棒性，而传统语音增强可能会移除对自动语音识别至关重要的声学特征。对于在噪声临床环境中部署医疗抄录系统的实践者而言，我们的结果表明，使用降噪技术对音频进行预处理不仅可能造成计算资源浪费，还可能损害转录准确率。

关键词 Medi cal A SR · 语音增强 · 去噪 · 噪声鲁棒性 · 临床文档

1 引言

在印度，将自动语音识别（自动语音识别）用于临床文档面临着重大挑战，原因是环境噪声水平较高。医院经常受到医疗设备、交谈和基础设施产生的高噪声影响 [Joseph et al., 2020], ，噪声水平远超世界卫生组织为医疗设施建议的 40 dB 限值 [1]。传统观点认为，语音增强预处理应通过去除噪声并改善信号质量来提高自动语音识别准确率 [Zhu and O'Shaughnessy,2003]。

然而，这一假设 ption 源自自动语音识别系统使用高斯混合模型 (GMMs 的时代) 和隐马尔可夫 在干净语音上训练的模型（隐马尔可夫模型） [Hinton 等人， 2012]。现代端到端神经 l

[1] ht tps://cpcb.nic.in/who-guidelines-for-noise-qual ity/

自动语音识别模型（如 Whisper） [Radford 等人， 2022], 在数十万小时的多样化真实世界音频（包括含噪录音）上进行训练，可能具有截然不同的噪声鲁棒性特征。如果这些模型已经学会了能够处理噪声的鲁棒的内部表示，那么外部预处理可能会移除有用的声学信息。

本研究的动机源于 EkaScribe[2], 在真实场景中部署时的实际观察：其中大多数部署于非移动设备上，且这些设备通常只有一个麦克风。音频通过操作系统（OS）管理的音频处理管线采集，该管线通常会默认应用噪声抑制或回声消除等信号处理技术。这些语音增强机制旨在提升人类的听觉体验，但它们是否总能改善自动语音识别性能则尚无保证。这引发了一个重要的实际问题：如果现代自动语音识别模型本身已具备较强的抗噪能力，那么此类预处理可能没有必要，甚至会适得其反。因此，了解显式去噪与现代自动语音识别之间的相互作用，对于实际部署至关重要。

为弥补这一空白，我们系统评估了 MetricGAN-plus-voicebank 去噪 [Fu 等人， 2021]这一先进的语音增强模型在四种领先的自动语音识别系统上的表现：OpenAI Whisper、NVIDIA Para- keet、Google Gemini Flash 2.0, 以及 Parrotlet-a——一种专门针对印度医疗保健领域英语语音识别训练的模型。我们在九种代表印度典型临床环境的噪声条件下测试了 500 段医疗录音。我们公开了评估代码、数据集和详细结果，以便复现和扩展这项研究。这些资源可在 GitHub[3]上公开获取。

2 相关工作

2.1 现代自动语音识别系统

近期自动语音识别领域的进展源于在大规模数据集上进行自监督或弱监督学习。Whisper [Radford 等人， 2022] 使用 68万小时的多语言网络音频进行训练，其中明确包含多样的录音条件和背景噪声。这种对真实世界声学变化的接触，可能使模型天生具有噪声鲁棒性。同样，NVIDIA 的 Parakeet 模型 [Rekesh 等人，2023，Xu 等人， 2023] 采用大规模训练，并进行包括噪声注入在内的数据增强。Google 的 Gemini 2.0 [Gemini 团队， 2024]通过在数百万个示例上预训练的多模态编码器处理音频。
这些模型与传统自动语音识别系统之间存在一个关键区别：现代模型能够学习从原始音频到文本的端到端映射，而无需显式的声学建模假设。这可能使它们无需外部预处理，便能从含噪信号中提取鲁棒特征。

2.2 语音增强技术

谱减法和维纳滤波等传统语音增强技术旨在提高人类语音的可懂度，并服务于主要使用干净数据训练的传统自动语音识别系统 [Loizou，2007，Benesty 等人， 2006]。这些方法通常在频域中运行，通过估计并去除噪声来恢复干净语音信号。

包括深度神经网络（DNN）和生成对抗网络（GAN）在内的深度学习方法已在很大程度上取代了传统方法。MetricGAN [Fu 等人， 2019] 旨在直接优化感知质量指标（感知语音质量评估、短时客观可懂度），而非重建误差。改进版 MetricGAN+变体 [Fu 等人， 2021] 在标准基准上实现了最先进性能。

然而，仍存在一个关键缺口：大多数语音增强评估都侧重于语音质量 指标或传统 自动语音识别系统的性能。对现代大规模神经自动语音识别进行系统评估，仍是一个有待研究的课题。

2.3 语音增强对自动语音识别的影响

随着自动语音识别架构不断演进，语音增强与自动语音识别准确率之间的关系一再受到重新审视，但始终未能达成共识。

经典自动语音识别时代：在干净语音上训练的传统 GMM-HMM 系统可从预处理中获益良多。Flynn 和 Jones [2008] 证明，结合语音增强与听觉建

[2]https://ekascribe.ai/ [3]
https://github.com/eka-care/when-denoising-hurts

模能够提升噪声条件下的分布式语音识别性能。这些发现确立了一种范式：增强应先于识别。这一假设至今仍存在于许多部署流程中。

深度学习转型： 深度神经网络在声学建模中的出现，引发了关于预处理必要性的根本性问题。Delcroix 等人 [2013] 在 2013年提出了一个富有前瞻性的问题："在使用深度神经网络进行声学建模时，语音增强预处理是否仍然重要？"他们对基于深度神经网络的声学模型的研究表明，随着模型日益复杂，增强带来的益处正在减弱。Donahue 等人 [2018] 探索了用于鲁棒语音识别的基于生成对抗网络的增强，发现效果不一，且因噪声类型和训练条件而异。

现代自动语音识别时代： 近期研究揭示了日益复杂的规律。Prasad 等人 [2021] 研究了用于鲁棒语音识别的端到端模型，发现即使不进行增强，现代架构也表现出令人意外的噪声鲁棒性。Braun 和 Gamper [2022] 分析了噪声抑制损失，并发现一个关键的权衡：过度降低噪声可能会引入语音失真，对自动语音识别性能造成的损害甚至超过原始噪声。这表明，为优化感知质量而设计的增强方法，未必符合自动语音识别需求。

近期的比较研究揭示了自动语音识别性能和增强效果中的重要细节。Agarwal 和 Misra [2025] 证明，Whisper 对背景噪声和领域偏移表现出显著敏感性，而 Gemini 的多模态架构则通过上下文适应展现出更强的鲁棒性。Trabelsi 等人 [2024]直接评估了噪声降低是否能够提升开源自动语音识别引擎的性能，发现增强效果会因模型和条件不同而显著变化。Rai 等人 [2024] 推出了 DENOASR，揭示去噪可能通过选择性声学滤波，无意间在自动语音识别系统中引入或放大人口统计偏差。最近，Nasretdinov 等人 [2025]提出了专为自动语音识别设计的、基于薛定谔桥的增强方法，表明传统增强方法可能与神经识别目标存在根本性的不匹配。

当前研究空白： 尽管近期研究已开始探讨现代自动语音识别系统的噪声鲁棒性，但在这一背景下，常用语音增强流水线的有效性仍缺乏充分刻画。以往研究往往聚焦于有限的模型规模、特定噪声类型或感知质量指标，所得结论不一，或因模型而异。尤其是，目前缺少系统评估，尚未厘清广泛采用的增强方法如何与大规模自动语音识别模型（如 Whisper）相互作用；这些模型是在大量含噪数据上训练的，在医疗语音等特定领域场景下更是如此。鉴于在嘈杂且资源受限的临床环境中部署此类系统具有重要的现实意义，有必要在受控噪声条件下开展审慎的实证研究。

在本研究中，我们通过使用代表真实临床环境的医疗语音和噪声条件，在四种现代自动语音识别系统上评估一种广泛使用的增强方法（MetricGAN+），以弥补这一空白。

3 方法论

3.1 研究问题

我们研究：*MetricGAN-plus-voicebank* 语音增强能否提高现代模型处理含噪医疗录音时的自动语音识别准确率？

我们的零假设是，与转录原始含噪音频相比，应用 MetricGAN-plus-voicebank 等语音增强技术不会使自动语音识别准确率产生统计显著变化。

3.2 数据集与噪声条件

我们从模拟临床咨询场景中汇集了 500 段经过标注的医疗音频录音，并由专业人员转录，以确立真实值。这些模拟咨询由 EkaCare 的员工进行。所选样本主要为英语，以避免引入可能降低自动语音识别性能的另一项变量。为模拟真实的医院声学环境，我们使用 Audiomentations library，以不同强度将三种噪声类型合成添加到原始音频文件中：

背景噪声： 持续的环境声音（暖通空调、设备嗡鸣、远处交谈，这些噪声取自 hospital-ambient-noise-dataset），信噪比为 10、30 和 50 dB。

短时噪声： 信噪比为 10、30 和 50 dB 的瞬时噪声爆发（设备提示音、摔门声、附近谈话，均取自 Pixabay）。

高斯噪声： 振幅为 0.001、0.009 和 0.01 的加性白高斯噪声 7.
信噪比定义如下：

$$\mathrm{SNR}_{\mathrm{dB}} = 10 \log_{10} \left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}} \right) \tag{1}$$

其中， P_{signal} 和 P_{noise} 分别表示非静音区域内的语音功率和噪声功率。
信噪比为 10 dB 对应于繁忙临床区域常见的严重噪声，30 dB 对应于中等噪声，50 dB 对应于轻微背景噪声。由此得到 4,500 条含噪测试语句（500 条录音， $\times$ 9 条件）。该数据集包含原始录音、噪声增强数据及其去噪版本，可在 HuggingFace[4] 上公开获取。

3.3 自动语音识别系统

OpenAI Whisper Large-v3 [Radford 等人， 2022]：一种在 68万小时多语言数据上训练的编码器-解码器Transformer。处理 80 通道的对数梅尔频谱图。我们使用的是 Large-v3 变体（15.5 亿个参数），该版本可在 HuggingFace[5],获取，并部署在本地。

NVIDIA Parakeet-TDT-1.1B [Rekesh 等人，2023；Xu 等人， 2023]：采用快速 Conformer编码器和 Token- and-Duration Transducer解码器，在 64,000+ 小时数据上训练。使用 8 倍深度可分离卷积进行下采样。我们使用可在 HuggingFace[6], 获取并通过 NeMo 部署在本地的模型版本。

Google Gemini Flash 2.0 [Gemini 团队， 2024]：具备原生音频理解能力的多模态模型，通过 Gemini API 调用。

Parrotlet-a：这是一款专门为印度医疗环境中的英语语音而训练、并针对医疗语音转录进行优化的专用自动语音识别（ASR）模型。该模型通过精简的投影层，将 Whisper V3 large编码器与 MedGemma-3-4B解码器结合起来，以高效完成语音转文本转换。模型说明和发布详情见此处[7]。

3.4 降噪方法

降噪时，我们使用了 SpeechBrain 的 [Ravanelli 等人， 2021] MetricGAN-plus-voicebank 模型 [Fu 等人，2021]。这种基于 GAN 的方法优化的是感知质量（感知语音质量评估、短时客观可懂度），而非重建误差：

1. 计算短时傅里叶变换： $X[n,k] = \mathrm{STFT}\{x[n]\}$2. 使用训练好的生成器生成频谱掩码：
$M[n,k] = G_\theta(|X[n,k]|)$ ˆ 3. 将掩码应用于带噪幅度谱： $S[n,k] = M[n,k] \cdot |X[n,k]|$4.
使用原始相位通过逆短时傅里叶变换进行重建：$\hat{s}[n] = \mathrm{ISTFT}\{\hat{S}[n,k] \cdot e^{j\angle X[n,k]}\}$

处理前，所有音频均重采样为 16 千赫兹单声道。

3.5 评估方法

我们采用全面的评估流程，评估自动语音识别系统在含噪和增强条件下的性能，具体如下：

1. **基线建立**

[4]https://huggingface.co/datasets/ekacare/denoising-impact-evaluation-dataset [5] https://huggingface.co/openai/whisper-large-v3 [6]https://huggingface.co/nvidia/parakeet-tdt-1.1b [7] https://info.eka.care/services/parrotlet-a-en-5b-releasing-our-purpose-built-llm-for-english-asr-in-indian-healthcare

- 转录了 500 个原始音频文件（未添加噪声）
- 对 500 个原始音频文件应用 MetricGAN+ 增强，并对其进行转录

2. **噪声污染与双重转录**
   - 使用 9 种噪声条件（3 种类型、 $\times$ 3种配置）污染 500 个原始音频文件中的每一个
   - 生成了 4,500 个含噪音频文件
   - 每个含噪文件均在两种条件下进行转录：– **含噪条件**：不进行增强，直接转录– **增强条件**：经过 MetricGAN+ 增强后进行转录

3. **每个自动语音识别系统的转录量**
   - 基线转录： $500 + 500 = 1{,}000$
   - 含噪条件下的转录：4,500
   - 增强条件下的转录：4,500
   - **每个自动语音识别系统的转录总数**: $1,000 + 4,500 + 4,500 = 10{,}000$

全部 4 个自动语音识别系统的总计: $4 \times 10,000 = 40,000$ 条转录

这个规模庞大的数据集能够全面分析噪声和增强对每个自动语音识别系统的影响。词错误率（WER）计算如下：

$$\mathrm{WER} = \frac{S + D + I}{N} \times 100\% \tag{2}$$

其中， S 表示替换， D 表示删除， I 表示插入， N 表示参考词总数。文本规范化处理（转为小写、移除标点符号）采用统一标准。

语义词错误率（semWER）[Vasanth, 2025]：印度医疗咨询中经常出现语码混用——在对话过程中交替使用英语医学术语和地区语言（印地语、泰米尔语、泰卢固语等）。标准词错误率会对语义等价的表达予以惩罚。我们通过应用额外的规范化处理计算语义词错误率：

1. **单位规范化**
   - 将单位缩写展开为对应的词形或完整表示
   - 示例： `5mg` → `five mg`，五毫克， `five milligrams`
   - 示例： `2°C` → `2 degree C`，2 डगि॒री सेल॒सयिस
2. **符号规范化**
   - 将数学符号和特殊符号转换为对应词
   - 示例： `3/4` → `3 by 4`、 `3 per 4`、 `three per four`
3. **缩约形式展开**
   - 将缩约形式展开为完整词形
   - 示例： `I'm` → `I am`, `It's` → `it is`
4. **数字转词** [AI4Bhārat, 2025]
   - 将数字转换为不同语言和文字系统中的词语表示
   - 示例： `150` → `one hundred and fifty`, एक सौ पचास, १५०
5. **标点符号移除** [Kunchukuttan, 2020]
   - 移除文本中的所有标点符号，包括逗号、句号等。
6. **Unicode规范化**
   - 将所有文本转换为规范化Unicode形式（NFC/NFKC），并合并等价的码位
   - 示例：合并印度文字系统中元音附标的组合形式与分解形式
7. **字符规范化** [Kunchukuttan, 2020]
   - 将视觉上表示同一字符的多个码位规范为单一规范形式
   - 示例：修正组合标记的规范顺序（ +ा）
8. **Nukta变体移除** [Kunchukuttan, 2020]

- 将努克塔符号（下加点）形式替换为对应的基本字符
- 示例：क़ → क、फ़ → फ

9. **变音符号规范化** [Kunchukuttan, 2020]
   - 规范组合变音符号，以确保顺序一致，并将变体形式转换为规范表示形式
   - 示例：规范化随韵符（कं）、月点符（कँ）和送气符（कः）
10. **转写** [Google, 2025]
    - 在保留语音发音的同时，将文本从一种文字转换为另一种文字
    - 示例： `computer` → कंप्यूटर

虽然 semWER 旨在处理印度医疗咨询中常见的多语言和代码混用场景，但由于所考虑的数据集为英语，因此上文提到的规范化处理中，只有部分处理方式（单位扩展、数字转词、缩略形式以及英语中的符号规范化）具有相关性。

这为印度医疗保健领域的自动语音识别生成了更具临床意义的错误度量指标。除非另有说明，所有报告的结果均采用 semWER。

4 结果

4.1 自动语音识别准确率与噪声的影响

图1展示了四种最先进模型在原始录音及其含噪版本上的自动语音识别性能。图中给出了各模型的基线语义词错误率（%），结果显示，Parrotlet-a 在所有条件下的错误率始终最低。为便于分析噪声对性能的影响，图2展示了这些模型在各种噪声条件下相较于原始条件的性能差异（Δ语义词错误率）。一个值得注意的初步观察是，添加高信噪比噪声（即噪声水平较低）甚至会使部分模型的自动语音识别性能略有提升。然而，随着噪声水平升高，自动语音识别性能会显著下降，且不同模型的下降幅度各异。Whisper 的性能下降幅度最大，尤其是在高斯噪声条件下（振幅为 0.017 时， Δ语义词错误率约为 9.72%）；其次是 Parakeet，表现出中等程度的性能下降。相比之下，Gemini 和 Parrotlet-a 在所有噪声条件下都表现出最强的鲁棒性，即使在噪声极大的环境中也能保持相对稳定的性能。这表明，与其他受测模型相比，Gemini 和 Parrotlet-a 的架构更能适应声学变化和受噪声干扰的输入。

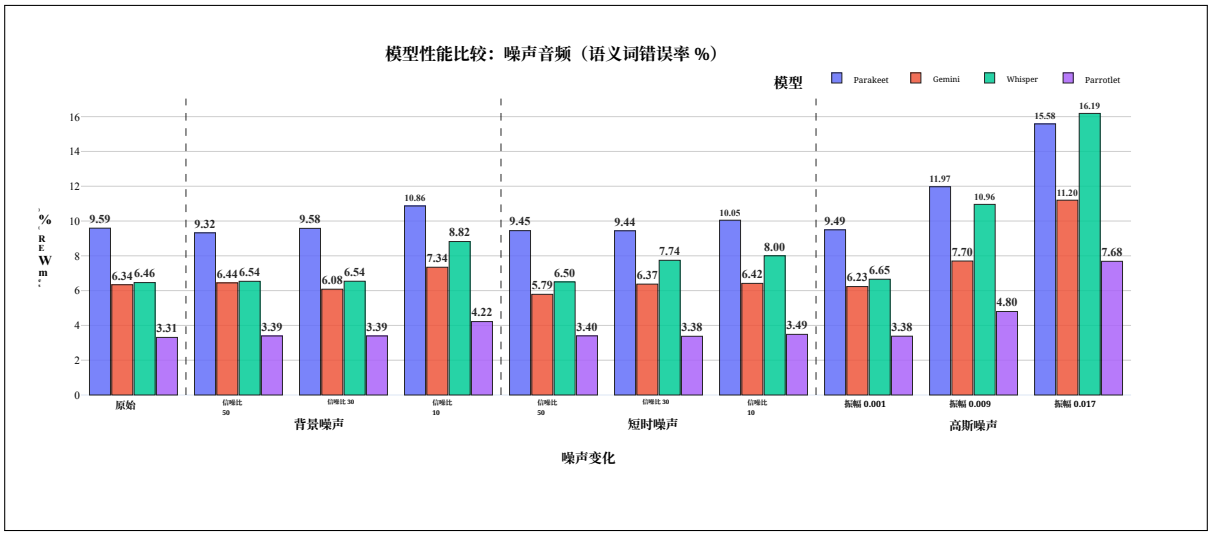

图 1：所有评估的自动语音识别模型在不同噪声条件下的语义词错误率（%）（越低越好）。Parrotlet-a 在所有条件下的错误率均为最低。

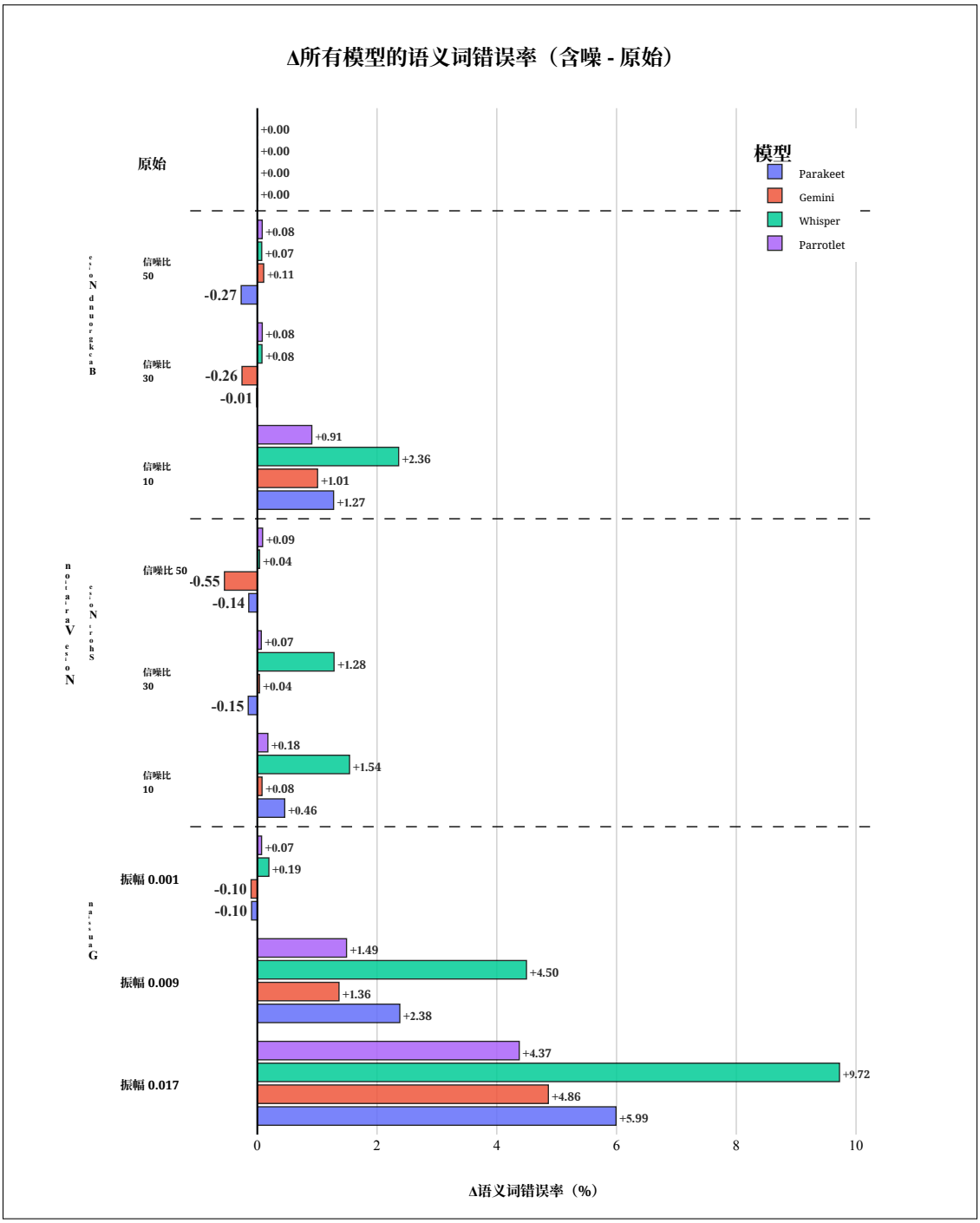

图2：四种自动语音识别模型在不同条件下的性能。我们观察到，这些模型的性能在超过一定信噪比水平后出现下降。

4.2 主要发现：去噪会降低性能

使用 SpeechBrain 去噪器处理过去噪录音后，各自动语音识别系统的性能如图3所示。图中展示了所有模型经过增强处理后的语义词错误率（%）。值得注意的是，在某些噪声条件下去噪后，Gemini 和 Whisper 的性能遭受灾难性下降，语义词错误率飙升至极高水平（例如，Gemini 的语义词错误率超过 50%），使它们从竞争力强、表现稳健的模型沦为表现最差的模型。总体而言，Parrotlet-a 仍是最稳健的模型，在所有过去噪条件下，其绝对错误率增幅最小，最终性能也最佳

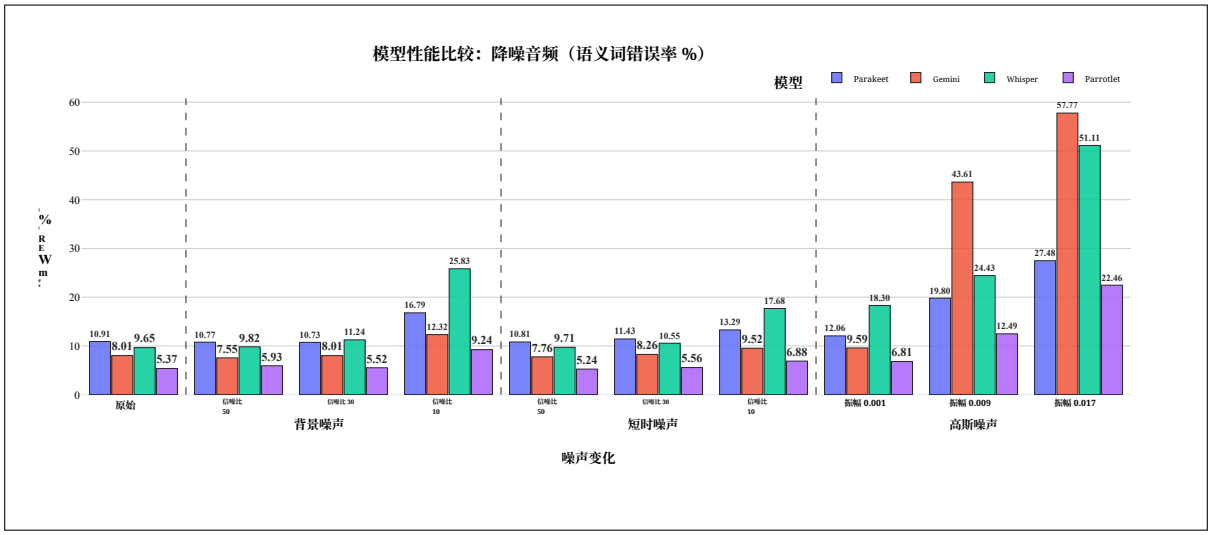

图3：语义词错误率（ %) 使用 SpeechBrain 去噪后所有自动语音识别模型的性能 r.

图4展示了去噪过程对自动语音识别性能造成的变化。图中呈现了所有模型的Δ语义词错误率（去噪后 – 含噪），直接比较噪声音频与同一音频经增强后的错误率。正的 Δ语义词错误率值表明，与直接处理噪声音频相比，处理去噪后的音频会导致更高的语义词错误率。性能下降幅度因噪声类型和模型而异，差异显著。在中等噪声条件下（背景噪声信噪比 10-50 和短时噪声信噪比 10-50），性能下降相对有限。然而，在高斯噪声条件下，影响会明显加剧。当高斯噪声振幅较高（振幅 0.009 和 0.017）时，Gemini 的性能出现灾难性崩溃， Δ语义词错误率分别高达 35.91% 和 46.57%。Whisper 在这些噪声水平下也出现显著退化（13.48% 和 34.93%）；Parakeet 的性能下降相对温和，但仍然显著（7.83% 和 11.9%）。Parrotlet-a 也呈现类似的退化趋势（7.69% 和 14.78%）。值得注意的是，即使在原始条件下，去噪也会使所有模型的性能出现可测量的损失（1.32% 至 3.19%），这表明去噪过程本身会引入伪影或造成信息损失，从而对自动语音识别准确率产生不利影响。这一发现挑战了语音增强预处理普遍有益于下游自动语音识别任务这一传统假设，并表明现代自动语音识别模型或许更擅长直接处理含噪输入，而非依赖去噪后的表征。

4.3 按噪声类型进行详细分析

背景噪声： 图5展示了 Whisper 的表现。在信噪比为 10dB（严重噪声）时，含噪语义词错误率为 8.82%，而增强后的语义词错误率为 25.83%——出现了灾难性的 17.0 个百分点下降。即使在信噪比为 50 dB（轻度噪声）时，增强也会使语义词错误率从 6.54% 上升至 9.82%。

Parakeet 表现出类似的模式（图6），但退化程度较轻。在信噪比为 10dB 时，语义词错误率从 10.86% 上升至 16.79%。Gemini（图 7）表现出中等程度的退化：在信噪比为 10dB 时，语义词错误率从 7.34% 上升至 12.32%；而 Parrotlet-a（图8）的退化相对较小，在信噪比为 10dB 时仅从 4.22% 上升至 9.24%。

短时噪声： 瞬态噪声突发呈现出类似的模式。在信噪比为 10dB 时，Whisper 的语义词错误率从 8.0% 上升至 17.68%。Parakeet：从 10.05% 上升至 13.29%。Gemini：从 6.42% 上升至 9.52%。Parrotlet-a：从 3.49% 上升至 6.88%。对于短时噪声，在任何信噪比水平下，增强都没有带来益处。

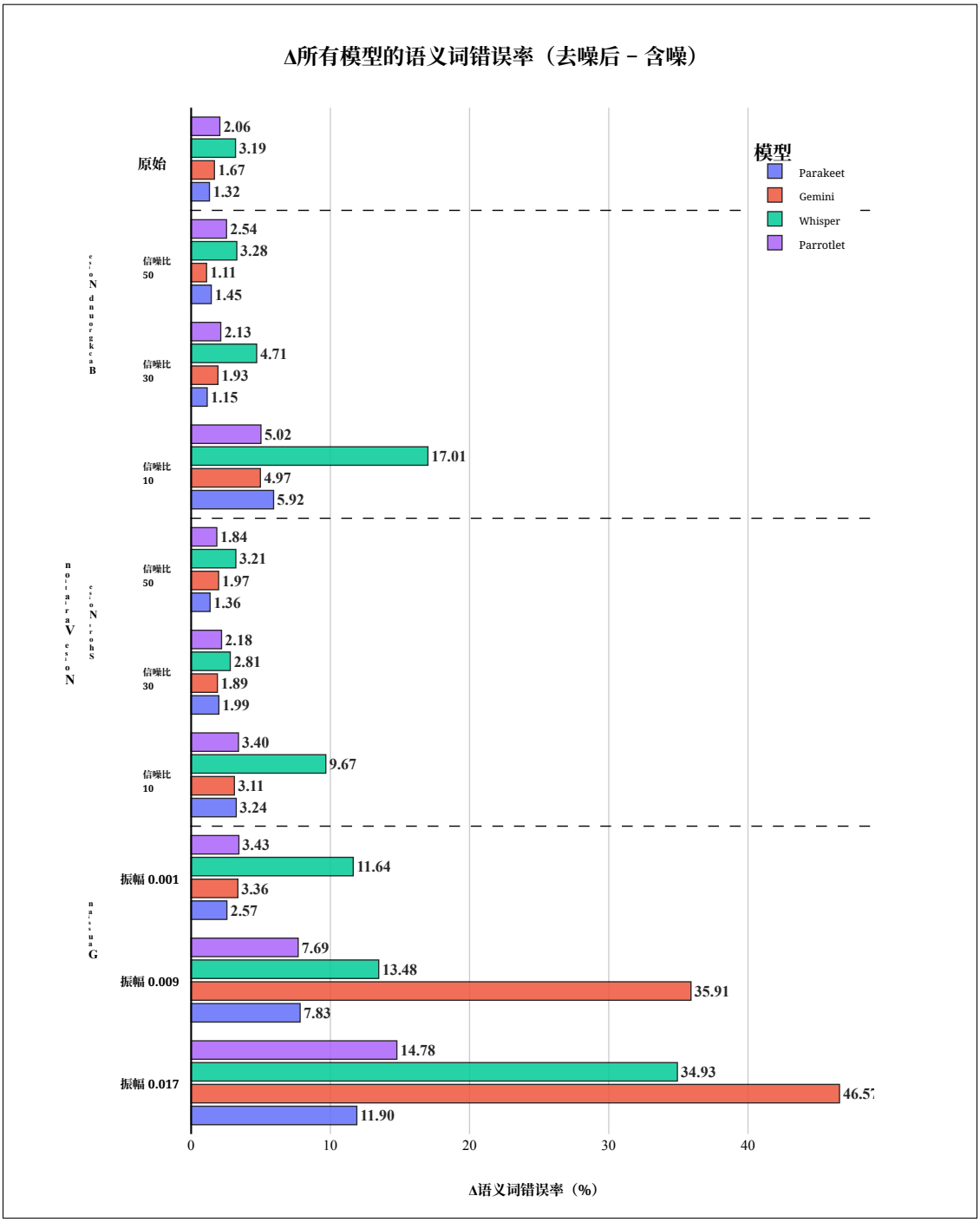

图4：去噪后，各自动语音识别模型在不同噪声条件下语义词错误率（Δ语义词错误率）的变化。我们观察到，去噪会导致性能下降。结果表明，去噪效果在不同模型间差异显著，其中 Whisper 对增强处理最为敏感。

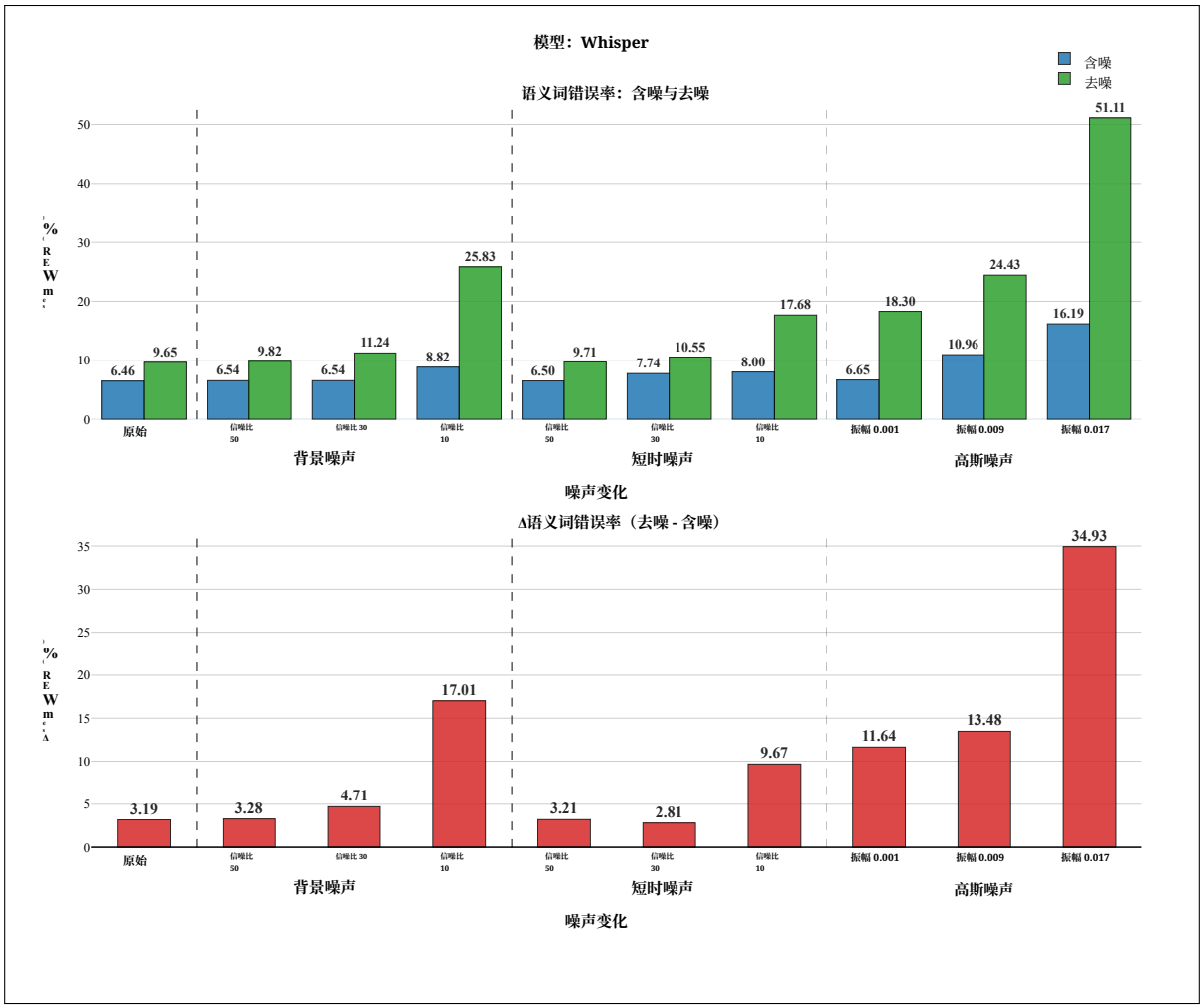

图5：Whisper 的语义词错误率（%）和 Δ语义词错误率。在所有条件下，增强都会持续提高语义词错误率，并且在高斯噪声下出现极其严重的性能下降。

高斯噪声： 该条件会造成最显著的性能下降。在振幅为 0.017 时，Whisper 的语义词错误率从 16.19% 上升至 51.11%，而 Gemini 的性能下降更为严重，从 11.2% 上升至 57.77%。Parakeet 也出现了显著上升，从 15.58% 增至 27.48%。Parrotlet-a 同样呈现出类似的性能下降模式，从 7.69% 上升至 22.46%。即使在较低振幅（0.001）下，增强仍会使所有模型的性能下降，而非带来改善。

4.4 定量总结

表 1 展示了在全部 40 种测试配置下计算得出的 Δ语义词错误率统计数据。这些配置包括在 10 种不同噪声条件下评估的 4 个不同自动语音识别模型：一种原始（干净）条件，以及涵盖三种噪声类型、每种均包含三个强度级别的九种含噪条件。 Δ语义词错误率指标用于量化应用语音增强后的语义词错误率变化：正值表示增强降低了模型的性能。

在**所有配置中，没有任何一种**语音增强能降低语义词错误率。这一负面结果具有普遍性，表明 MetricGAN+ 增强与现代自动语音识别模型之间存在根本性的不兼容。

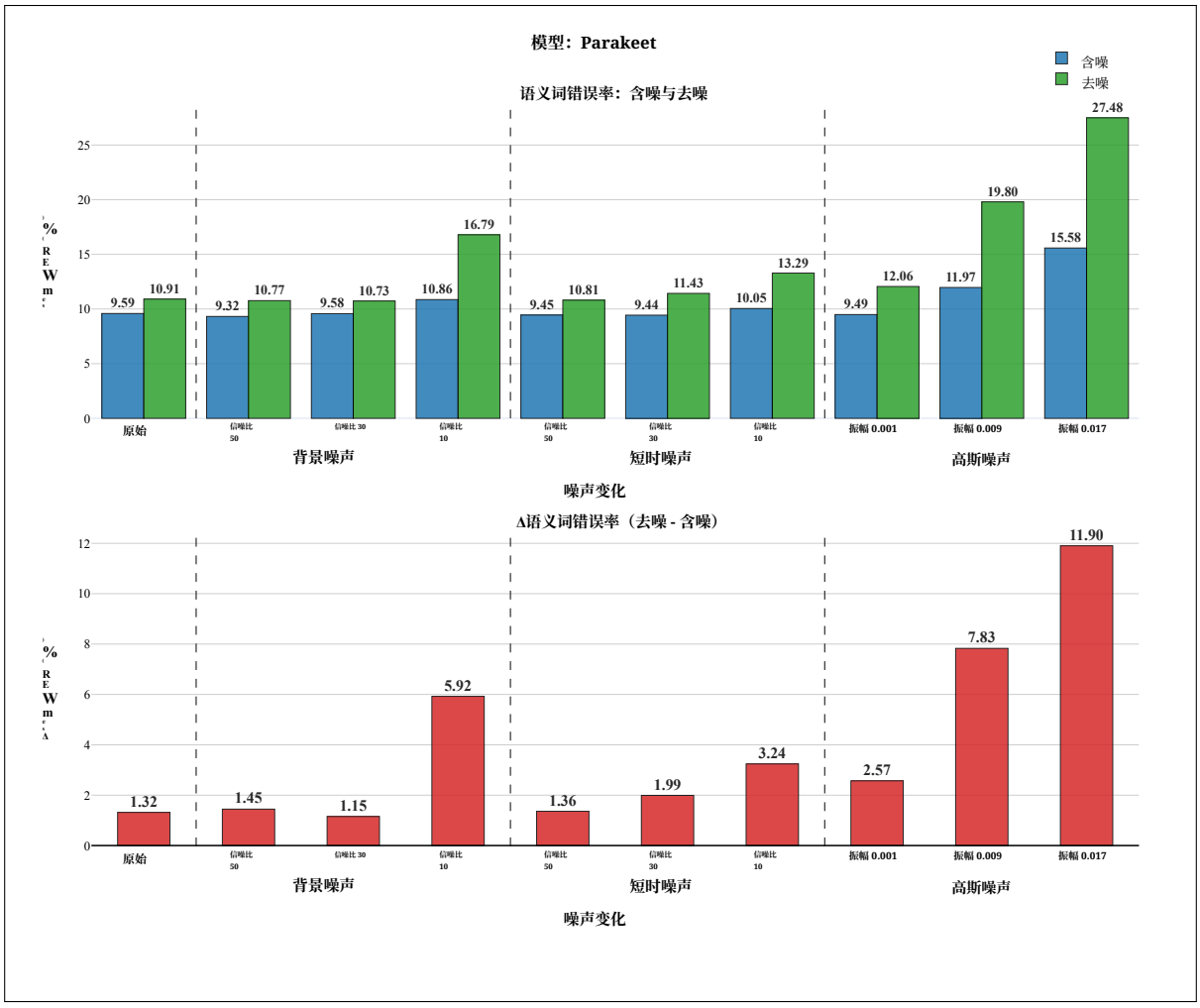

图6：Parakeet 的语义词错误率（%）和 Δ语义词错误率。语音增强会降低所有噪声条件下的性能，但降幅小于 Whisper 或 Gemini。Parakeet 的性能相对稳定，但预处理仍会导致普遍的性能下降。

表1： Δ语义词错误率统计。正值表示语音增强使语义词错误率上升（性能下降）。全部 40 种配置均出现性能下降。

统计量	所有配置的原始背景			短	高斯噪声
包含 Δ语义词错误率 > 0 的配置	40 / 40	4 / 4	12 / 12	12 / 12	12 / 12
平均 Δ语义词错误率（%）	+7.83	+2.06	+4.26	+3.05	+16.17
中位数 Δ语义词错误率 (%)	+3.32	+1.86	+2.91	+2.49	+11.77
最大退化（%）	+46.57	+3.19	+17.0	+9.67	+46.57
最小退化（%）	+1.11	+1.32	+1.11	+1.36	+2.57

5 讨论

我们的结果对“语音增强能够改善自动语音识别”这一最常见的假设提出了挑战。我们提出了两个主要假设：

假设1：现代自动语音识别已具备内部噪声鲁棒性。Whisper 的训练数据明确包含 68万小时的多样化真实世界音频，其中包括含噪录音 [Radford 等人， 2022]。该模型可能已经学会了鲁棒的内部表示，能够处理典型的声学噪声，无需进行外部预处理——

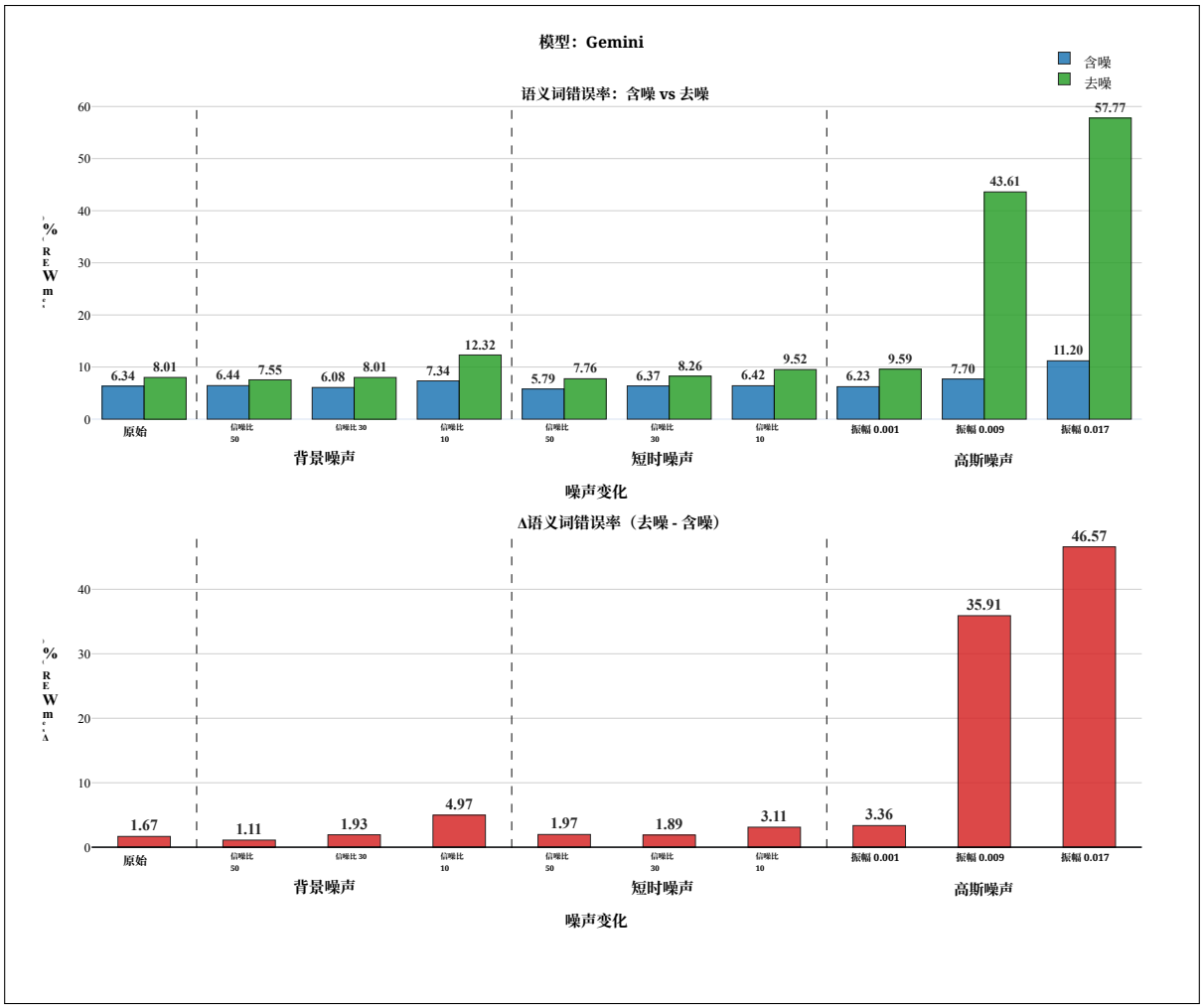

图 7：Gemini 的语义词错误率（%）和 Δ语义词错误率。高斯噪声下的增强会显著降低性能（振幅为 0.017 时，语义词错误率从 11% 上升至 58%）。其他噪声类型也会导致性能持续下降，但程度较轻。

增强实际上可能会去除有助于识别的细微声学线索（韵律、精细的频谱结构）。

假设 2：增强伪影。语音增强可能会引入处理伪影，例如频谱模糊、时间不连续和不自然的共振峰过渡。虽然这些伪影可能不会被人类听众察觉（因此在感知语音质量评估/短时客观可懂度指标上得分较高），但使用自然语音训练的神经网络自动语音识别模型可能会对这类伪影较为敏感。

5.1 泛化性与局限性

泛化性： 我们的结果可能适用于其他在多样化数据上训练的现代大规模自动语音识别模型（例如 wav2vec 2.0），但仍需验证。该发现可能无法推广至：

- 仅在干净语音上训练的传统自动语音识别系统（高斯混合模型-隐马尔可夫模型）
- 在有限数据上训练的小规模模型
- 超出我们测试范围的极端噪声条件（信噪比 < 0 dB）

局限性： 我们的研究存在一些限制：

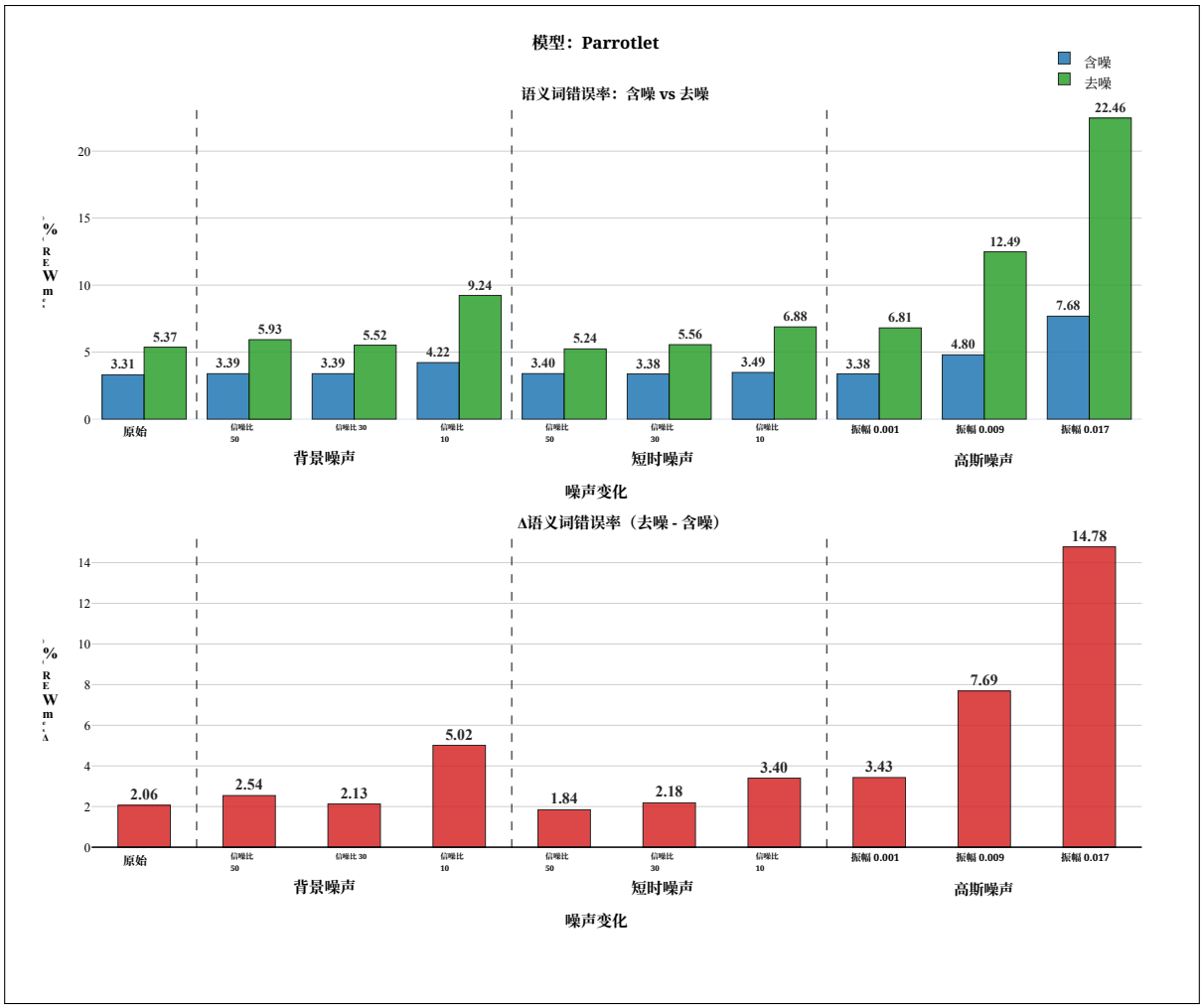

图8：Parrotlet-a 的语义词错误率（%）和 Δ语义词错误率。与其他模型相比，Parrotlet-a 在含噪条件下的语义词错误率最低，但去噪仍会降低其性能。

单一增强方法： 我们仅评估了 MetricGAN-plus-voicebank 模型。其他方法（基于 Transformer 的增强方法、扩散模型）的表现可能有所不同。
合成噪声： 我们采用了合成噪声添加方法。虽然这使我们能够进行受控比较，但真实的临床声学环境包含混响和复杂的多源干扰。仍需进行实地验证。
数据集规模： 我们的样本包含 500 段录音，对于在这一主题上得出具有普遍性的结论而言，规模仍不够理想。实验数量的组合性质使我们无法进一步增加实验数量。
医疗领域： 在其他领域（对话、广播）中，结果可能有所不同。包含专业术语的医疗语音可能呈现出不同的噪声鲁棒性特征。
印度英语： 我们的数据集由印度英语医疗咨询录音构成。是否适用于其他口音和语言，仍需验证。

5.2 未来方向

替代增强方法： 评估现代方法（例如，基于 Transformer 的增强方法、扩散模型），以确定性能下降是否仅发生于基于 GAN 的方法。
端到端联合训练： 开发可联合优化增强与识别的架构，以提升自动语音识别准确率，而非感知质量。

失效模式分析： 详细调查哪些声学现象会导致增强失效——共振峰失真、韵律去除、伪影引入——有助于指导改进。

真实世界部署研究： 在印度医院开展试点部署，使用原始噪声音频和增强音频，在生产环境中验证结果。

多麦克风方法： 评估波束形成和空间滤波，作为单通道增强的替代方案。

6 结论

本研究系统地评估了广泛使用的语音增强方法 MetricGAN-plus-voicebank 在含噪临床环境中对四种现代自动语音识别系统的影响。在所有测试的噪声条件和模型下，我们观察到，与直接转录噪声音频相比，这种特定的增强方法始终会提高词错误率，导致性能持续下降。

我们的结果表明，旨在优化人类感知指标的增强方法，可能无法很好地适应在含噪真实世界数据上训练的大规模自动语音识别模型所学习的表征。需要特别指出的是，这并不意味着语音增强本身会对自动语音识别产生不利影响。相反，这凸显出增强的有效性高度取决于所采用的具体技术、噪声特征以及自动语音识别模型的训练方法。其他方法——例如面向自动语音识别的增强、联合优化或领域特定微调——可能会产生不同的结果，仍需进一步研究。

从实际角度来看，我们的结果表明，在医疗自动语音识别流程中，不应默认采用 MetricGAN+ 预处理等去噪技术；应针对每项具体任务和模型评估其影响。更广泛地说，这些发现表明，有必要重新审视关于音频预处理的常见假设，尤其是在现代自动语音识别系统已经过训练、能够在噪声条件下运行的背景下。

参考文献

Manav Agarwal 和 Anurag Misra。《自动语音识别中的鲁棒性与幻觉：Whisper 和 Gemini 的比较评估》。*IOSR 计算机工程期刊*，27(3):65–69，2025年。

AI4Bhārat。indic-numtowords：一个用于将数字转换为印度语言和英语单词的 Python 库。Python 包索引（PyPI），2025年。网址：`https://pypi.org/project/indic-numtowords/`。

Jacob Benesty、Shoji Makino 和 Jingdong Chen。《语音增强》。Springer，2006年。

Sebastian Braun 和 Hannes Gamper。噪声抑制损失对语音失真和自动语音识别性能的影响。发表于 *IEEE I CASSP*，第 996–1000 页，2022 年。

Marc Delcroix、Yusuke Kubo、Tomohiro Nakatani 和 Atsushi Nakamura。使用深度神经网络进行声学建模时，语音增强预处理是否仍然相关？发表于 *INTERSPEECH*，第2992–2996页，2013年。

Chris Donahue、Bo Li 和 Rohit Prabhavalkar。探索利用生成对抗网络进行语音增强，以实现鲁棒语音识别。发表于*IEEE ICASSP*，第 5024–5028 页，2018年。Robert Flynn 和 Edward Jones。结合语音增强与听觉建模，实现鲁棒的分布式语音识别。*Speech Communication*，50:797–809，2008年。Szu-Wei Fu、Chien-Feng Liao、Yu Tsao 和 Shou-De Lin。MetricGAN：基于生成对抗网络的黑盒度量分数优化，用于语音增强。发表于国际机器学习会议 *(ICML)*，第 2031–2041 页，2019年。Szu-Wei Fu、Cheng Yu、Tsung-An Hsieh、Peter Plantinga、Mirco Ravanelli、Xugang Lu 和 Yu Tsao。MetricGAN+：用于语音增强的 MetricGAN 改进版本。*arXiv* 预印本 *arXiv:2104.03538*，2021年。Gemini 团队。Gemini 1.5：解锁跨数百万词元上下文的多模态理解。*arXiv* 预印本 *arXiv:2403.05530*，2024年。Google。google-transliteration-api：用于 Google 转写服务的 Python 接口。Python 包索引 (PyPI)，2025年。URL `https://pypi.org/project/google-transliteration-api/`。

Geoffrey Hinton 等。用于语音识别声学建模的深度神经网络。*IEEE*《信号处理杂志》，29(6):82–97，2012年。B. E. Joseph、H. Mehazabeen 和 M. U. 医院噪声污染——一项关于公众认知的研究。《噪声与健康》，22(104):28–33，2020年。doi:10.4103/nah.NAH_13_20。Anoop Kunchukuttan。IndicNLP 库。GitHub 仓库，2020年。URL https://github.com/anoopkunchukuttan/indic_nlp_library。Philipos C. Loizou。《语音增强：理论与实践》。CRC 出版社，2007年。Rustam Nasretdinov、Roman Korostik 和 Ante Jukić。利用薛定谔桥式语音增强实现鲁棒语音识别。发表于*IEEE ICASSP*，第 1–5 页，2025年。Anusha Prasad、Preethi Jyothi 和 Raghuraman Velmurugan。端到端模型在鲁棒语音识别中的应用研究。发表于*IEEE ICASSP*，第 6893–6897 页，2021年。Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Casey McLeavey 和 Ilya Sutskever。通过大规模弱监督实现鲁棒语音识别。*arXiv* 预印本 *arXiv:2212.04356*，2022年。A. K. Rai、S. D. Jaiswal、S. Prakash、B. P. Sree 和 A. Mukherjee。DENOASR：通过选择性降噪对自动语音识别系统进行去偏。发表于*IEEE* 国际知识图谱会议 *(ICKG)*，第 283–290 页，2024年。Mirco Ravanelli 等。SpeechBrain：通用语音工具包。*arXiv* 预印本 *arXiv:2106.04624*，2021年。Dinesh Rekesh、Nithin Rao Koluguri、Samuel Kriman、Somshubra Majumdar、Vahid Noroozi、Haicheng Huang、Oleksii Hrinchuk、Kishore Puvvada、Ankit Kumar、Jagadeesh Balam 和 Boris Ginsburg。采用线性可扩展注意力的快速 Conformer，实现高效语音识别。发表于*IEEE* 自动语音识别与理解研讨会 *(ASRU)*，2023年。Ahmed Trabelsi、Loic Werey、Sylvain Warichet 和 Eric Helbert。噪声降低是否能提升开源自动语音识别转录引擎的质量？发表于*ICAART*，第 3 卷，第 1221–1228 页，2024年。Vasanth。karma-medeval：医疗人工智能模型评估框架。Python 包索引 (PyPI)，2025年。URL https://pypi.org/project/karma-medeval/。Hainan Xu、Feng Jia、Somshubra Majumdar、Haicheng Huang、Shinji Watanabe 和 Boris Ginsburg。通过联合预测词元与时长实现高效序列转导。发表于国际机器学习会议 *(ICML)*，第 38491–38510 页，2023年。Wei Zhu 和 Douglas O'Shaughnessy。利用噪声降低和频谱强调技术提升噪声条件下的自动语音识别性能。发表于*IEEE* 自动语音识别与理解研讨会 *(ASRU)*，第 357–362 页，2003年。