arXiv:2611.32455v1 [cs.CL] 24 Nov 2026

音素身份深度剖面无法可靠预测构音障碍自动语音识别中的 LoRA 放置

Bernard Muller[1,2] LaVonne Roberts[1,2]

[1]Scott-Morgan基金会，英国[2]SMF Labs，法国巴黎
Colonel Pierre Avia 街 28 号，邮编 75015

通信作者： lavonne@scottmorganfoundation.org

摘要

对语音基础模型进行参数高效适配以处理构音障碍语音，需要做出一个鲜有充分依据的选择：应当适配哪些编码器层。一种近期提出的策略是探测编码器，找出语音信息最容易被解码的位置，并在那里放置适配器。我们在两个自动语音识别模型（Qwen3-ASR、Whisper-small）和两个构音障碍语料库（普通话CDSD、英语UASpeech）上检验这一推断，在整个过程中保持实验方案和随机种子规范不变，并在每个语料库内保持数据划分和门控不变。

这一推断并不成立。 在同一流程下进行测量，而不是沿用已发表图表中的结果后，四条音素错误曲线的最小值均出现在相对深度的83%至92%之间；对于每个模型而言，其达到最小值的层在不同语料库之间的移动幅度，并不超过不同探测随机种子之间的移动幅度。最佳适配层段则并非同样恒定。由曲线映射到层段的三种自然方法——最大改进幅度、最低频带均值错误率或拐点位置——都无法找出测试配置中表现最佳的配置。我们并不声称曲线中的任何统计量都不可能奏效，只是这三种方法无法奏效。

位置仍然重要，但因模型和语料库而异。 在 CDSD 上，模型的偏好并不一致——Qwen3-ASR 偏好上三分之一区间（字符错误率为 29.88%，三随机种子均值），而 Whisper-small 偏好全块 LoRA（53.56%）；在 UASpeech 上，两者都不偏好上层带。将 CDSD 扫描实验以三个随机种子重新训练后，上层与全部的比较仍已解决（-1.31 个百分点，置信区间 $[-2.16, -0.25]$），但上层与中部的比较则未解决（-0.83 个百分点，置信区间 $[-2.06, +1.86]$）——训练随机种子造成的变化幅度足以涵盖这一差距。两个模型的偏好差异通过直接检验交互作用得以确认（-4.71 个百分点，置信区间 $[-6.17, -2.89]$）。我们根据其作用将其指定为主要主张——这是我们的第一项贡献所要论证的内容，也是论文试图回答的问题——并注明，这一指定是在早期版本将另一种对比列为主要对比之后作出的（§6.1.1）。无论采用微平均字符错误率、说话人等权加权，还是精确配对检验，这一结果均已解决（未调整的精确 p 值 $= 0.031$）；但 Qwen3-ASR 自身的偏好仅在微平均下成立。预算匹配对照未发现额外容量带来统计学上已确认的收益

对于任一位置，这都不能证明容量无关紧要：在匹配预算下，放置对比在当前样本量下本身无法得到可靠判定，而已解决的结果是等秩结果。在 Qwen3-ASR $\times$ UASpeech 上，在输入附近进行适配的效果可靠地差于完全不进行适配（CER 相差 $+9.90$ 个百分点）。因此，我们建议根据手头的模型和语料库进行简短的适配扫描来选择频带，而不是根据探针读数来决定。

1. 引言

自动语音识别已能用于典型语音，但对许多构音障碍人士而言仍然无法使用。主流解决方案是适配：取一个大型预训练模型，在现有的言语障碍语音数据上对其进行全量或部分微调。由于这类数据稀缺，而且说话人级适配日益成为部署模式，参数高效方法——尤其是低秩适配器（LoRA； [8]）——已被广泛采用。构音障碍语音识别文献印证了这一趋势：Jiang等人 [11] 在 Whisper 上使用 LoRA 进行微调

，然后为中文言语障碍语音加入说话人提示；Tan 等人 [26] 使用课程调度 AdaLoRA 进行低资源构音障碍识别；Wagner 等人 [27] 报告称，AdaLoRA 在说话人个性化方面优于全量微调和标准 LoRA；Hu 等人 [9] 则在 UASpeech 上通过混合专家说话人适配器进行路由。**但这些研究都没有解决适配器位置的问题**。Hu 等人 [9] 的研究与我们的问题最接近，但两者仍各自独立：它选择的是适配某个说话人的哪个专家，而非适配器在编码器的哪个位置。这些研究都沿用或搜索适配器配置，却没有测量相关信息位于何处，因此一个设计问题无法回避：**适配器应该放在编码器的什么位置?**

这个问题促使我们从机制角度进行思考：如果能够测量与目标领域相关的语音信息是在编码器深度的哪个位置形成的，那么理论上就应该能够将适配器放置在那里。Muthu 等人 [18] 正是以此为依据研究构音障碍自动语音识别的。他们在普通话构音障碍语音上对一个包含 12 个模块的 Whisper-small 编码器进行探测，发现音素错误率在早期至中部层中大幅下降，并在第 6 层之后趋于平稳，由此推断“这些层应当是构音障碍语音适应最有效的目标”，并对此进行了验证：在第 7 层放置单个适配器，以及使用第5–8层频带，其效果都接近全编码器适配，而其上层带 9–12 则是效果较弱的连续分组。（他们的 v2 报告显示，全部模块的字符错误率为 64.35，而前两种设置分别为 68.64 和 66.21，即相对值分别为 6.67% 和 2.89%。他们的 v1 摘要给出的数值为 3.5% 和 2.48%，但这无法由其自身的表格推导出来；v2 摘要对此进行了更正。**我们全文引用 v2** ——它还将未适配基线从 105.08 字符错误率修订为 89.98，这一点对第 §5 节很重要。）这是一个简洁明了的结果，但它容易引出一种我们认为并不稳妥的解读：构音障碍语音适应属于编码器中部。

混杂因素在于，层索引在不同编码器中代表的含义不同。深度、时间分辨率、训练目标以及下游连接的组件，在不同自动语音识别模型系列之间都各不相同。向 Whisper 风格文本解码器提供输入的编码器，其顶层所承受的压力不同于向投影器提供输入、再连接至大语言模型的编码器。如果语音信息成熟的深度不同，那么适配能够产生效果的频带也应随之不同——而沿用文献中的层放置建议反而会造成严重误导。

这就是我们的假设，而事实证明，其中后半部分是错误的。在普通话语料库上，位置建议是因模型而异的：在其他条件完全不变的情况下更换编码器，会改变表现最佳的频带。在所有情况下，它并非都因模型而异——在模型 $\times$ 语料库设计的四个单元格中进行相同的对比显示，两个模型在英语孤立词上达成了一致（§6.8、§6.9），因此这种差异是某个语料库的属性，而不是两种架构的属性。但所提出的机制——不同的成熟深度——在两种情况下都无法经受测量检验。根据我们自己的流程而非引用的数据进行测量，Whisper-small 的音素错误率最低点位于相对深度的 83% 处，而不是已发表曲线所暗示的 50%，基本上与 Qwen3-ASR 的位置相同，因此深度剖面不可能是决定频带的因素。

我们沿两个相互正交的轴进行测试。**首先，保持数据不变，改变编码器**。我们使用与 Muthu 等人相同的语料库——中文构音障碍语音数据库（CDSD； [28]）——并探测 Qwen3-ASR-1.7B [25], 的 24 个块组成的音频编码器；其解码器是语言模型，而不是 Whisper 风格文本解码器。因此，深度剖面的差异应归因于编码器，而非语料库、语言或任务——但不能归因于某个单一的架构因素，因为这两个系统在深度、帧步幅、预训练数据和目标函数、规模以及解码器类型方面同时存在差异。**其次，保持编码器不变，改变语料库**。我们在 **UASpeech** [12], 上重复完全相同的适配扫描；该语料库在我们能够考察的每个维度上都与 CDSD 不同：使用英语而非普通话，包含 449 个孤立词而非新闻朗读句子，并采用临床严重程度标签而非基于字符错误率的伪标签。两个深度探测都在**两个**语料库上运行（§5.5、§5.6），因此第二个轴检验的是用于解释排序的机制，而不仅仅是预测结果本身——结果发现，探针曲线几乎没有变化，但放置结果发生了变化。

本文有两点并非如此。本文并不是声称某个基础模型在构音障碍语音方面优于另一个基础模型：这里没有进行这种比较，而据我们所知，唯一一项已发表的正面对比研究 [10] 在单个说话人上得出了相反的适配性结论（§2）。本文也不是声称适配前探测是一个新想法——Pasad 等人 [21]、Beheshti 等人 [2] 和 Muthu 等人 [18] 都提出过某种形式的这一观点。我们的贡献更为有限：**受控地证明，在一个构音障碍语音识别模型上获得的 LoRA 放置结果不能安全地迁移到另一个模型上，并提供证据表明，三个自然的音素 PER 启发式方法无法解释我们所观察到的放置差异**。

贡献。

1. **一项在同一语料库、采用相同实验方案的跨架构适配比较表明，在所测试的两个语料库之一上，这两个模型的最佳适配器放置位置有所不同**。 在固定语料库、数据划分、完整性门控、LoRA 配置方案、每个模型的参数预算和随机种子规范，并更换模型后，CDSD 中的 $\texttt{upper} - \texttt{all}$ 对比方向相反；在我们采用的所有估计量下，二者的交互作用均已解决（§6.9.1）——不过，Qwen3-ASR 本身的主效应在微平均字符错误率下成立，在说话人等权加权下则不成立（§6.10）：Qwen3-ASR 最适合采用其上层带，其效果优于以三分之一的参数适配全部编码器块；而 Whisper-small 最适合采用全块 LoRA，在四种方案中排名 upper 最差（§6.1、§6.6）。先前的文献无法进行这种比较，因为相关研究分散在不同论文中，而非在同一项研究内开展。 **在某个模型的全部四个单元格中进行相同的对比 $\times$ 表明，语料库设计的反转现象仅见于 CDSD：**在 UASpeech 上，两个模型都没有偏好 upper （§6.8、§6.9）。位置与 $\times$ 容量的分析 2×2 未发现总预算带来已解决的收益（§6.7）：将任一位置的秩增加三倍，其字符错误率的变化幅度仍未达到已解决的程度。

2. **一个负面结果：三种音素身份 PER 启发式方法均无法预测两个自动语音识别模型和两个语料库中最佳的 LoRA 区间**。这里只将音素身份 PER 作为位置选择信号进行测试；冻结的 CTC 读出探究的是在某一层是否能够通过 *CTC* 读出恢复音素身份，而 LoRA 适配探究的是哪一处扰动最能提升生成式序列模型的表现，因此这一结果最好解读为：**通过 CTC 读出恢复音素身份的能力，并不等同于局部适配效用**。根据我们自己的流程测得（并非引用他人结果），Whisper-small 的音素错误率最低点位于相对深度的 83% 处，Qwen3-ASR 的则位于 83–92% 之间——两者都处于较深层，但两个模型偏好的区间不同。在 *PER* 改善最陡峭处进行适配这一策略在 Qwen3-ASR 上不成立：最陡峭的区间反而是最差的目标；在区段平均 *PER* 最低处进行适配仅在四个单元格中的一个选中了最佳目标；而在拐点处进行适配——这是我们在测量第二种架构之前自行提出的一种调和方案——对 Whisper-small 的预测结果为upper ，在四种方案中排名最差。**根据我们自己的测量结果，我们撤回这一规则**。在 UASpeech 上重复两种探针（§5.6），排除了曲线依赖语料库这一替代解释：四个最低点均落在相对深度的**83% 到 92%**之间，四者对三个比例区间的排序完全相同；对于每个模型而言，其最小值位置在不同语料库间的变化幅度，不超过同一语料库内不同探测随机种子间的变化幅度。

3. **第二语料库的复现将可泛化与不可泛化的结果区分开来**。在两个语料库上，有两项对比的区间均不包含零—— upper 优于lower，以及 middle 优于 lower。第三项对比 upper 优于 middle，在UASpeech上已解决，但在CDSD上未解决；在对说话人进行重采样的同时也对训练随机种子进行重采样后，结果依然如此（§6.1.1），但每次运行的方向都一致；因此我们将其报告为方向性结果。两个语料库上没有任何对比呈现相反方向的已解决结果。 upper与all 的对比是唯一无法复现的：UASpeech上该对比无论哪个方向都未解决，因此该对比及以其为基础构建的参数匹配 2×2 均报告为仅在CDSD上成立。第二个语料库还解答了第一个语料库无法解答的问题：

在CDSD上，不进行适配与基线之间的对比均未达到显著性；而在UASpeech上，四项对比均达到显著性，表明适配有所帮助（CER最高改善 -9.20个百分点），并且在该模型和语料库上，下部频带适配**确实比完全不适配更差**（CER高出$+9.90$个百分点）。每个频带使用三个匹配的随机种子，种子方差为0.12–0.38个百分点，远低于说话人抽样不确定性。

4.**关于层探测本身的两个方法学结果，解释了被舍弃的论断是如何产生的。探针读出对训练数据量的敏感性取决于深度，并具有架构特异性**（在固定训练轮次预算下，因此数据量与优化器步数同步变化——§5.5）：增加探测训练数据 $5.4\times$ (6,200 $\rightarrow$ 33,617 个话语) 后，Whisper-small 的最深层模块提升了 9.15 个百分点，即其 $1.87\times$ 堆栈均值；相比之下，Qwen3-ASR 的提升为 2.25 个百分点和 $0.74\times$ 。因此，低数据量探测会人为制造出两个架构之间表面上的后段堆叠层差异——最小值后上升分别为 6.56 和 1.85 个百分点——而在匹配数据量下，这一差异缩小为 2.07 和 1.82。并且，**我们无法复现先前报告的 Whisper-small 中层平台**：探针配置与先前研究逐项一致，而无论采用哪种填充处理以及匹配训练数据量，该差异都依然存在。我们已无法排除未适配基线能力是影响因素之一：先前研究的 v2 报告字符错误率为 89.98，而我们修正后为 77.57（§5）。我们将这一差异报告为尚未解释的现象，并将基线能力列为仍在考虑的候选因素之一。

5.**作为该方法的架构前提条件，这一点被报告为一项发现，而非研究空白**。Qwen3-ASR 音频编码器每 ~77 毫秒输出一帧，而 Whisper- small 的帧长为 20 毫秒；相较于该语料库的强制对齐结果（音素中位时长为 80 毫秒），前者有 49.4% 的音素短于单个编码器帧，而在 20 毫秒帧长下这一比例为 1.3%。因此，我们不比较两个编码器的帧级边界F1：这样会将表征信息与输出网格施加的量化混为一谈。我们**不**声称该表征无法编码子帧时序；受限的是帧级探针能够放置边界的位置。

6. **浅层适配的一种已记录失效模式**。在输入附近进行适配所造成的损害集中于声学特征异常的说话者，其错误**与**语言先验接管相一致——这是一种机制性假设，而非已得到证明的因果路径，因为解码器被冻结，适配改变的是其条件表示，而不是任何语言模型权重：已适配模型输出流畅、符合语料库典型特征且与音频无关的文本，而未适配模型则能正确转写。我们还报告了一项负面结果，以界定这一现象的泛化范围。

---

2. 相关工作

2.1 语音表示的逐层分析

Pasad 等人 [20] 建立了对自监督语音编码器的逐层分析，展示了 wav2vec 2.0 中类似自编码器的行为：中部层承载最多的语言内容，而顶层则回归输入。其后续研究 [21]提出了本文将在新场景中检验的两点内容：其一，“逐层属性会因模型不同而以不同方式跨层演变”；其二，逐层分析“提供[s]可靠指导，以选择下游任务所关注的层”。**尚未经过检验的是，这些结论是否适用于适配病理性语音条件下监督式ASR基础编码器的深度选择**。Ashihara等人 [1] 进行了相应的跨模型家族比较

针对自监督语音模型和说话人模型，并为在保持数据不变的同时改变编码器提供了方法论先例。

近期有两项研究探讨了非典型语音中的逐层结构。Yue 等人 [33] 对 Whisper-Medium 进行了构音障碍检测和严重程度评估探针分析，定位出第 13–15 层附近信息量最大的表征，并发现微调仅引起了轻微的逐层变化；Mojarad 和 Tang [17] 则探究了 wav2vec 2.0 和 Whisper 在非裔美国人英语中的辅音丛简化现象。两者都没有将探针分析与适配器放置联系起来。

关于自动语音识别的机制可解释性研究从不同方向得出了模型同一区域的结论。Glazer 等人 [4] 发现，上下文偏置在编码器内部 出现，并覆盖了声学证据——这与我们在§7.2中记录的失效模式密切相关；Pluth 等人 [22] 报告称，对 Whisper 编码器特征进行因果引导时，“高层特征比下部特征更可靠”；Ma 等人 [15] 则描述了经过 LoRA 适配的 Whisper 中存在“延迟专门化”。这三项研究都与我们的适配结果方向一致，但单独来看，都不能作为该结果的证据。

2.2 与适配器放置相关的逐层探测

Muthu 等人 [18] 是本文的直接前身，也是我们部分提出异议的研究。Muthu 等人针对 CDSD 上的 Whisper-small 编码器，在三种转录文本匹配条件下进行探测，报告了贯穿深度的任务相关层级，并据此提出中间层适配器放置；随后，他们对单层以及第 1–4 / 5–8 / 9–12 层和全部层进行扫描，进一步证实了这一点。他们所处的是一种失效基线情形：未经适配的 Whisper-small 在构音障碍 CDSD 上达到 89.98% 的字符错误率，而采用全编码器低秩适配后降至 64.35%（其 v2 结果；v1 报告的相应数值分别为 105.08 和 68.85）。我们完整沿用了他们的探测设计、适配器配置和层段结构，因此两项研究之间的差异在于编码器而非协议——而且我们**重新运行了他们的架构，而不只是引用它**：我们在自己的语料库、数据划分、门控、实验方案和评分设置下，于 Whisper-small 上重复进行了深度探测（§5.5）和四区间扫描（§6.6、§6.8），从而将论文的核心比较从跨论文比较转变为内部受控比较。由于我们尚未找到原因（§5.5），他们的探针结果未能在我们的流程中重现；但他们的层段结果部分得到了重现。

2.3 在哪里 适配，构音障碍之外——一个由来已久且备受争议的问题 n

“中部并非自动就是适配的正确位置”这一一般性主张**并不新颖**，而且文献中的观点**并不一致**。

- **一种 U 形规律**。 DAMA [30] 研究了低资源语言多语言自动语音识别的深度感知适应，并报告称，早期和后期层具有语言特异性，需要进行适配，而中间层则保留共享结构；在 18 种语言上，训练参数减少了 80%，准确率持平或更高。
- **自动语音识别中的层选择性 LoRA 已有先例**。 Han 等人 [6] 将 LoRA 应用于大型自动语音识别模型中自动选出的一部分层，以适配低资源多语言场景，由此确立了一般做法——并非每一层都要适配——且这一做法已具体应用于自动语音识别。该研究既未探讨哪一深度适合病理性语音，也未将选择与实测的语音成熟度概况联系起来；但这意味着我们不能再宣称“在自动语音识别中选择性地设定位置”是新颖之处。
- **探测引导的选择**。 Beheshti 等人 [2] 使用廉价探针对任何特定任务模型训练之前的层进行排序，并指出信息丰富的层“聚集在深度区域中，而非位于某个唯一的最优位置”，其结果是“骨干网络特定选择，而不是固定的层实验方案”。他们选择层是为了读取 而非适配，但这是现有文献中与我们所检验原则最接近的表述。
- **相反的结果**。Lin 等人 [14] 研究了自监督语音编码器中逐层的参数高效微调放置，发现这种方法**并未**优于在全部层中均匀插入。我们的发现

即 `upper` 在 CDSD × Qwen3-ASR 上优于 `all` ——参数量仅为其三分之一（§6.1）——是来自不同模型家族的一个反例，而这种分歧本身就是支持本文论点的证据。

- **跨模态迁移**。 Ogawa等人 [19] 依据CKA相似度选择LoRA层，在编码器独占、解码器独占和多模态模型中，以相当的性能将可训练参数减半；Xu等人 [31] 提出一种由理论驱动的"层卡片"诊断指标，用于确定大语言模型中的位置。

- **学习层选择，而非从探针中读取**。 Gumbel-BEARD[29] 以针对Whisper编码器层的端到端硬Gumbel-Softmax选择器取代手动搜索层，并通过BEST-RQ自监督目标进行训练；研究报告称，选出的深度因领域而异。这是与本文实践建议最接近的当代替代方案：我们认为，由于音素探针无法迁移，应进行简短扫描；该研究则主张直接从适应目标中学习选择。两者都认同本文所捍卫的前提——有用的深度取决于模型和任务，而不是可以从通用诊断指标中读出的东西。

- **自动语音识别中的编码器深度与 LoRA**。 Kolluri 等人 [13] 对 Whisper 编码器层进行剪枝，并将剪枝与 LoRA 结合，发现 LoRA 借助语言模型先验，能够在一定程度上弥补被移除的编码器容量，但在低资源场景下效果较差——这与 §7.2 中的失效模式密切相关。

- **调整编码器，而非解码器**。 SALSA [32]报告称，调整语音编码器"尤其是后几层，比调整大语言模型骨干网络更有效"。

2.4 构音障碍语音的基础模型选择

Huber 等人 [10] 比较了 Whisper-large-v3 与 Qwen3-ASR-1.7B，评估它们作为构音障碍语音适应基线的效果，并得出结论认为 Whisper 更优。他们报告称，使用全部可用适配数据时，词错误率为 9.7% ；而 Qwen3-ASR 微调和 LoRA 的词错误率为 14–16%，并指出 LoRA 的表现不及全量微调。他们提出的假设是，构音障碍语音与基础分布差异过大，低秩容量难以弥补。这项研究针对的是**n = 1 说话人**，使用了 92 小时朗读语音和 8.8 小时来自单个用户的野外纠正数据；它衡量的是在每位说话人拥有充足数据时的适应能力，而非表征结构或适配器放置。

我们之所以记录这一点，是因为读者会看到它，也因为它具有双重影响。这说明我们不应将本文的任何部分表述为关于哪个基线模型更优的论断——我们并未作出这样的论断。但这并不削弱我们实际提出的主张：不同频带之间的对比是在同一个编码器内部进行的，无论该编码器的绝对表现如何，这些对比都成立；而且，我们发现，在 CDSD测试集 上，任何已适配变体与未适配模型之间的差异都不显著（§6.1），这一结果若要说有什么意义，倒是与他们的怀疑一致。

2.5 语料库与工具

CDSD [28] 是本文以及 Muthu 等人所使用的普通话构音障碍语料库。对齐由 Montreal Forced Aligner [16]生成；音素身份探针是 CTCprobe [5]；适配器采用 LoRA [8]。

---

3. 编码器几何结构：前提条件与发现

在定义任何探针之前，必须根据待探测的现象测量编码器的时间分辨率。通常这一步不会被特别说明，而在这种架构中，它决定了哪些探针是可行的。

Qwen3-ASR 技术报告 [25]指出，128 维滤波器组特征经过 $8\times$ 下采样后，名义令牌速率为 12.5 Hz。直接测量时长为 0.40–9.40 s 的五段 CDSD 话语后发现，发布的编码器每秒输出 **12.5–13.2 帧，即 ~77 每帧毫秒数**；本文始终采用实测值，而名义上的 80 ms 会让下文的每个数值都显得略为极端。Muthu 等人探测的编码器 Whisper-small 的运行速率为**20 ms**[23], ，因此 Qwen3-ASR 编码器的时间分辨率大约 **4× 更粗**，且 $2\times$ 更深（24 个块，而非 12 个），模型维度 d_为 1024，而非 768。编码器返回 25 个隐藏状态：索引 0 是卷积后的编码器输入，索引 1–24 是区块输出，与先前研究中第 0 层为输入的约定一致。隐藏状态 0 已经经过卷积，因此不存在分辨率更高的更早层可供提取。

根据语料库的强制对齐结果（21,251 个音素、2,640 个话语、44 位说话人）：**表 1.** CDSD 中的音素时长分布及其与编码器帧周期的关系。

统计量	ms
p5 / p10 / p25	30 / 30 / 30
中位数	**80**
p75 / p90 / p95	140 / 230 / 300
mean	110.4

帧周期	音素 <1 帧	音素 <2帧	每帧均值，每个音素
20毫秒（Whisper-small）	**1.3%**	28.6%	5.52
77毫秒（Qwen3-ASR, 实测）	**49.4%**	77.8%	**1.43**

后果。 帧级音素边界检测——“帧 t 是否为边界？”——无法在此编码器上进行测量。全部音素中有一半短于采样间隔，因此，无论得到什么 F1 值，反映的都会是帧网格，而非表征。这与先前研究发现构音障碍语音的边界信息在每一层都较弱，是不同的说法；该发现针对的是表征，而我们无法在此检验。图中的数值对 77 毫秒还是 80 毫秒这一问题也不敏感，因为该语料库的音素时长中位数本身就是 80 毫秒。

我们报告这一限制，是因为它最清楚地体现了论文的总体观点：**探测协议并非架构可移植；编码器一旦改变，其低层探测就可能在不知不觉中变得不再适定**。任何将层探测研究移植到新编码器的人，都应先测量帧周期，并与目标现象的时长分布进行比较。这不仅关乎方法论：构音障碍会扭曲亚音素级时序，因此，分辨率处于这一水平的编码器在结构上无法表示临床医生关注的某些现象。其余探测不受影响：通过CTC探针探测音素身份是可行的（99.9%的话语满足 $T \geq L$，中位数T/L为 $= 3.47$）；话语级难度不需要时间分辨率；跨条件分析和适应性分析则采用合并数据。

4. 实验设置

4.1 语料库与完整性门控

§5.1–§5.5的探测实验和§6.1–§6.4的适配扫描使用CDSD。§6.5和§6.8的复现实验扫描，以及§5.6的复现深度探测，使用UASpeech [12]，其他实验方案均相同；其语料库、数据划分和评分细节见§6.5，§5.6中针对UASpeech探测而不得不采用的两项协议变更亦见该节。

CDSD [28]：44 位构音障碍普通话说话人朗读取自新闻文本的句子材料。该语料库包含两个子集目录：**1 小时** 子集包含来自全部 44 位说话人的 55,851 条话语（35.06 小时），**10 小时** 子集包含来自 8 位说话人的 92,730 条话语（64.32 小时）。**6,103 个（说话人，话语 ID）组合在每个子集中各对应一个文件，且在所有这些情况下，两个文件的转录文本均不相同**。先前研究报告了 ~55,850 条话语，这表明其使用的是 1 小时子集；我们出于同样的理由采用该子集，因为它本身也是合适的选择：保留了全部 44 位说话人。**关于语料库规模**，两个子集合计为 99.38 小时，而语料库论文报告的时长为 133 小时；我们无法解释这一差异，本文所有结果均仅基于**1 小时子集**。

话语标识符全程以 `speaker/subset/utt_id` 为键，并且每一条保留记录的对齐转录文本都会与其清单转录文本进行核验，不匹配的记录将被删除。强制对齐由 Montreal Forced Aligner [16]生成；音素层中不含真实音素的话语会被排除，完全没有 TextGrid 的话语也同样排除。在 1 小时子集中，55,851 条话语中**有 8,164 条（14.6%）被删除**——其中 7,720 条因不含真实音素，444 条因缺少 TextGrid——最终剩余 47,687 条。

门控删除了什么，以及它是否使评估变得更容易。强制对齐失败在病理性语音中并非随机发生，因此完整性门控可能会在不经意间筛选出一个更容易的评估人群。我们对此进行了测量，而不是想当然地忽略这一点。

层	排除	Rate
mild	323 / 3,211	10.1%
中度	3,256 / 22,778	14.3%
重度	4,585 / 29,862	**15.4%**
保留语句	—	平均时长 **2.37 秒**（中位数 2.06）
排除语句	—	平均时长 **1.59 秒**（中位数 1.28）

确实存在严重程度梯度，而且其方向与我们担忧的预测一致——轻度材料中有 10.1% 被剔除，重度材料中则有 15.4%——因此保留集略易于原始子集。整个严重程度范围内的梯度为 5.3 个百分点；我们认为，相较于第6节中的效应，这一幅度较小，但不会对此掉以轻心。**主要因素是时长，而非严重程度：**排除语句的平均时长比保留语句短 0.78 秒，且 94.6% 的排除项属于“无真实音素”，这是极短或实际上为空的录音的典型特征。逐说话人的排除率介于 3.5% 至 32.5% 之间。

由于门控对每个频带以及未适配基线都采用相同方式，因此它不会按条件差异化分配话语，也就无法凭空制造频带间差异。**不过，它确实改变了用于估计各项对比的总体；如果位置效应在所选特征上存在异质性，还可能改变其幅度**——这并非无关紧要的假设，因为§7指出，浅层适配尤其会损害声学特征异常的说话者，而门控过滤掉短话语和重度程度稍高的话语的比例也更高。门控必然会影响绝对 CER 水平，以及与使用不同门控的研究进行的任何比较——包括跨论文 Whisper-small 比较，后者并未应用等效过滤器。LoRA 训练和 CER 评分本身不需要对齐，因此，重新运行

在完整有效音频子集上进行关键 CDSD 条件的测试即可解决这一问题；我们尚未进行此项测试，并将其列为一项待处理的敏感性检查（§9）。

数据划分按说话人互斥：训练集有 31 位说话人／验证集有 7 位／测试集有 6 位，探针和适配扫描使用相同的数据划分，以便进行比较。探针数据集包含 8,800 个话语（每位说话人 200 个）；适配清单包含 47,687 个话语（训练集 33,617 个／验证集 8,682 个／测试集 5,388 个），比先前研究少 14.6%，因为它沿用了门控。

声调单元由转录文本推导，而非由对齐推导。对齐中的音素标签保留了声调字母，但在 43,137 个音素词元中，T1（高平调）一次也未出现——这是数据缺陷，并非普通话的实际情况，因为 T1 大约占普通话音节的五分之一。基于相同转录文本推导的拼音呈现了预期分布（T1 18.3%／T2 19.3%／T3 19.4%／T4 33.2%／轻声 9.9%），因此我们完全沿用先前研究的方法：从文本中提取拼音声母／韵母／声调单元（`tong2` → {t, ong, T2}），对齐仅用于确定片段时间。

4.2 模型与层索引

探测的编码器是从已发布的 Apache-2.0 检查点中读取的 **Qwen3-ASR-1.7B** [25],AuT 音频塔：包含 24 个 Transformer 块，d_model 为 1024，16 个注意力头，前馈网络宽度为 4096，并投影至 2048 以供语言模型解码器使用。遵循先前研究的做法，探针层 0 为编码器输入，层 1–24 为块输出；适应带按**块**索引命名，索引从 1 开始，因此 `upper` = 为块 17–24，探针抽取点 17 对应块 17。

在当前版本下加载此检查点，需要显式进行键重映射并重建子配置；直接采用简单方式会导致键重叠数为零，并得到一个随机初始化的模型（见警告 §10）。

4.3 探针

我们沿用先前研究的探测设计，不作任何改动，从而使比较聚焦于编码器，而非探针容量。

- **探针 B — 音素身份**。一个 CTC 探针 [5]：层归一化 → 128维瓶颈 → 残差时序卷积 → 线性分类器；丢弃率为 0.3，空白符偏置为 0.0，CTC损失，Adam 学习率 1e-3，批量大小 256，训练 10 轮，梯度裁剪为 1.0；在帧级状态上逐层独立训练。指标为音素错误率，分别采用声调敏感和声调不敏感设置，并依据声调敏感音素错误率进行模型选择。训练使用 6,200 个话语（200 × 31 训练说话人），评估使用 6 名未见测试说话人的 1,200 个话语，因此绝对音素错误率不能与先前研究的 ~39,000 个话语相比较；只有随深度变化的趋势可以比较。同一探针也用于 §5.6 中的 UASpeech，评估来自发音词典的 ARPAbet音素。

- **探针 C——基础模型难度预测**。名称反映了它的测量对象：标签依据被探测模型自身的字符错误率得出，因此探针探查的是，在该模型的编码器中，最终会导致识别失败的部分，何处能够通过训练好的 CTC 读出得到恢复——这种读出包含一个瓶颈和一个时序卷积，因此并非线性探针；它并不探查独立定义的难度概念或临床严重程度所在的位置。在线性探针中，使用均值池化编码器状态训练线性 softmax 分类器（交叉熵、Adam，学习率 1e-3、批量大小 4096、训练 10 轮），并采用**说话人不重叠的**5 折交叉验证——若随机划分数据折，探针可能会记住说话人身份。指标为三个难度层级上的宏平均 F1；随机猜测水平为 0.333。必须说明两处偏差。第一，难度标签来自**基线模型本身**的解码：由于没有更强且未受污染的普通话解码器，我们将难度定义为“通用自动语音识别模型难以识别”。第二，层级划分是被迫采用的：字符错误率分布存在两个较大的质量点（2,267 条话语的字符错误率恰为 0.0，另有

2,853 条恰为 1.0），因此普通百分位数阈值会使层级数量变成 2,943 / 5,503 / 354。在自然分界点（高 $=$ 字符错误率 ≥ 1.0，低 $=$ 字符错误率 < 0.2）处分箱后，数量为 2,813 / 2,780 / 3,207，层级均值（0.024 / 0.531 / 1.061）之间差异明显。

- **探针A — 音素边界**。 未运行：在此输出步长下，帧级表述与 20 ms 的情况不可比较，除非重新定义任务（§3）。这限定了探针 在帧分辨率下能够测量的内容；这并不意味着表征中不存在子帧时序，池化表述或基于容差的表述仍可对其进行检验。

4.4 跨条件差异与一项明确的设计弱点

先前研究比较了每个话语的三种转录文本匹配条件：(d) 原始构音障碍语音，(r) 对同一转录文本进行说话人条件零样本TTS重合成，以及 (u) 无条件TTS。我们复现了这些条件，但**未进行逐话语匹配**。我们的合成数据池是为严重程度匹配的数据增强而生成的，因此没有理由保留谁说了什么这一对应关系：在基于 CDSD 说话人生成的 13,627 条构音障碍条件合成记录中，93.0% 使用了某段 CDSD 语料中出现过的文本，但只有**3.0%（411 条记录）**使用了条件说话人实际说过的文本。这是配对问题，而非数据量问题。

因此，我们采用**分布匹配** 设计——37 名经过构音障碍条件合成的说话人用于 (r)，39 名健康人条件说话人用于 (u)，全部 44 名真实 CDSD 说话人用于 (d)，每位说话人对 ~400 个话语进行均值池化——并通过**411 条严格配对的行** 验证，将其作为精确匹配一致性检查。这不如逐话语匹配严格；后者在设计上而非仅在期望上消除了词汇混杂因素。各层条件均值之间的差异以余弦距离（$1 - \cos$）衡量。另一个不足是：条件 (u) 采用了来自不同语料库的健康对照说话人，因此 d–u 对比混合了说话人集合差异与条件差异。**d–r 是干净的对比**。

4.5 适配扫描

LoRA [8] 应用于选定模块的编码器自注意力投影 `q_proj`, `k_proj`, `v_proj`, `out_proj`，**其余部分全部冻结，包括整个解码器** ——与先前研究的配置完全相同：$\mathrm{r} = 64$, $\alpha = 128$，丢弃率 0.1，AdamW，学习率 3e-5，有效批量大小 128，训练 6 轮，余弦调度，权重衰减 0.03。适配从**未适配的基础模型**开始；若从已有的普通话构音障碍微调模型开始，就无法测量任何效果。全文中， `all` **均指全块 LoRA** ——在每个编码器模块上使用低秩适配器——**而非**全权重微调；本文所有实验均未直接更新基础权重。

表2. 适配变体。三个八块带的可训练参数预算完全相同； `all` 的参数预算是它们的三倍。训练前通过试运行核实了模块数量，训练后检查合并差异，确认每个适配器只作用于自身频带，未触及其他部分——这排除了适配器挂载悄然失效、导出基础模型副本，导致各频带因虚假原因而得分相同的失效模式。

变体	编码器模块	模块	可训练参数	模型的%	训练时长	合并差异（张量）
`lower`	1–8	32	4,194,304	0.205%	3.25	32
`middle`	9–16	32	4,194,304	0.205%	3.16	32
`upper`	**17–24**	**32**	**4,194,304**	**0.205%**	**3.12**	**32**
`all`	1–24	96	12,582,912	0.614%	3.54	96
base	—	0	0	0%	—	—

`middle` （第9–16块）是先前研究中在12块编码器上表现最佳的第5–8层频带的比例对应项； `upper` （第17–24块）对应其表现较弱的第9–12层频带，而 `lower` （第1–8块）对应第1–4层频带。四次运行中的每一次都成功读取了全部33,617条训练语句。

4.6 评估与统计

适配器通过模型自身的封装推理路径（§10）在6位未见测试说话人的5,388条留出语句上进行合并和解码。**普通话没有正字法词边界，因此评分器按字符进行标记化，这些标记上的编辑率即为字符错误率；因此，我们仅针对普通话报告字符错误率，不将按字符标记化的编辑率称为"词错误率"**。由于6位测试说话人的样本量较小，我们报告每位说话人的字符错误率、微观和宏观平均值、**每项对比的配对说话人自助法95%置信区间（B = 10,000）**以及留一说话人敏感性表。**每种CDSD适配配置均使用三个随机种子（42/43/44）进行训练**，且所有CDSD区间均按说话人 × 随机种子计算（§6.1.1）；§5.1中的CDSD 探针曲线仍为单次运行结果（§9）。

---

5. 探测结果

5.1 音素身份在后期成熟，而非中层堆栈

该频带根据验证说话人选取，而非测试说话人。 探针在31名训练说话人上进行训练；用于选择自适应层段的深度剖面取自**7名验证说话人**，而6名测试说话人仅在第6节的适配扫描中被使用一次。

表 3。 基础模型按编码器层划分的声调敏感词错误率（%）。**VAL** （7名未见说话人）作为选择依据；**TEST**（6名未见说话人）仅为描述而报告。所选层；完整的25层曲线见发布的 CSV。

层	0	4	8	12	16	20	**21**	24
音素错误率 — **VAL**（选择）	71.17	62.66	54.20	47.04	39.29	34.06	**33.01**	36.56
音素错误率 — TEST（描述性）	75.14	66.96	57.64	49.46	41.07	34.39	**33.58**	35.43
PER	78.07	67.08	55.54	45.92	37.85	33.41	33.27	32.95
声调-不可知的（TEST）								

两个数据划分在所有用于选择的特征上都一致。在验证说话人上，音素错误率从第 0 层到第 21 层单调下降，随后趋于平缓；两个数据划分的最低点都位于**第 21 层**。从中点（第 12 层，47.04）到该最低点，曲线在 VAL 上还提升了**14.03 个百分点**，而在 TEST 上提升了 15.88 个百分点。第 9–16 个模块的平均音素错误率在 VAL 上为**46.00** ； 第 17–24 个模块的平均音素错误率为**34.88**。

表4。 验证集上的基线模型声调敏感词错误率曲线分频段摘要（用于选择的依据；括号内为 TEST 测试值）。请注意，改善最大的频带并非 PER最低 的频带，而第6节表明，这两种排序给出了相反的预测。

Band	模块	平均 PER	频带上的 PER 条目 → 退出	下降幅度 band
`lower`	1–8	61.08 (64.66)	71.17 → 54.20	**16.97 百分点** **（最陡）**
`middle`	9–16	46.00 (48.17)	54.20 → 39.29	14.91 百分点
`upper`	17–24	**34.88 (35.20) — 最低**	39.29 → 36.56	2.72 百分点（最平坦）

曲线的拐点——即音素错误率不再改善的深度——位于第 21 层，在 `upper` 频带内。Muthu 等人报告称，音素错误率在其第 12 层中的第 6 层之后趋于平稳，即位于**相对深度的 50%**处；而这里的改善持续到 ≈**85–88%**。

这种对比存在于我们的测量结果与他们已发表的图表之间，而在由我们自行测量两者后，这种对比并不成立。§5.5 表明，按照本论文自身的流程，Whisper-small 的最小值位于 12 层中的第 10 层——**相对深度的 83%**，而不是 50%——与此同时，在探针训练量匹配的情况下，Qwen3-ASR 的最小值在第 20–23 层之间呈现平坦的平台（跨度为 0.27 个百分点）。**两个探针的最小值都位于堆叠层的后部**，因此，已发表比较中的表面深度剖面差异并不是这两种架构的固有属性。由于最小值所在层位于单随机种子估计的噪声区间内，因此对于任一架构，能够提出的结论只有这个区间。

这一许可的预测是在适配扫描运行之前记录的。先前研究以编码器中部的音素成熟为依据，论证了中间区间定位；但在 Qwen3-ASR 上，该前提并不成立，因此有效区间应当是上层带，而不是按比例对应的第 9–16 层。第6节检验了这一预测，并且**在 Qwen3-ASR 上得到了确认**——但其依据并不成立，因为 Whisper-small 自身的拐点也位于其上层带中，而在此处同样的推理是错误的（§6.6）。

一项次要观察，但存在限制其意义的注意事项。同一基础模型的普通话构音障碍微调模型，在第 0–12 层的表现与基线相差不超过 ~1 pp，之后则持续偏离（第 16 层：41.07 对 37.80；第 20 层：34.39 对 25.67）——微调重塑了编码器上部三分之一的层，而下部三分之二未受影响。不过，该微调模型在自身训练期间接触过 81.9% 的探针话语和全部六位测试说话人。**层定位具有参考价值；但这一差异幅度不能解读为泛化能力**。所有主要结论均基于从未接触过该语料库的基础模型。

声调。 声调敏感与声调无关差距在第 0–4 层为负值（第 0 层为 −2.93 个百分点），在第 5 层反转，在第 10–16 层达到 +3.5 至 +3.8 个百分点的峰值，随后在顶层收窄。从第 5 层起，这重现了先前研究的发现：要求识别声调会降低准确率。早期层的反转是归一化伪影：在 75–78% 的音素错误率下，探针的表现接近随机水平；去除声调标记会使参考文本缩短约三分之一，因此相近的错误数会除以更小的分母。

5.2 跨条件差异在编码器中部达到峰值，随后下降

图1，上部面板绘制了本节讨论的曲线；其下部面板属于§5.3。

替代文本（图1）。两幅上下堆叠的折线图共享一个横轴，表示Qwen3-ASR编码器层索引，范围为0至24。上部面板：跨条件表征距离，即音节障碍语音与说话人条件化重合成语音之间的1减去余弦相似度。该曲线上升

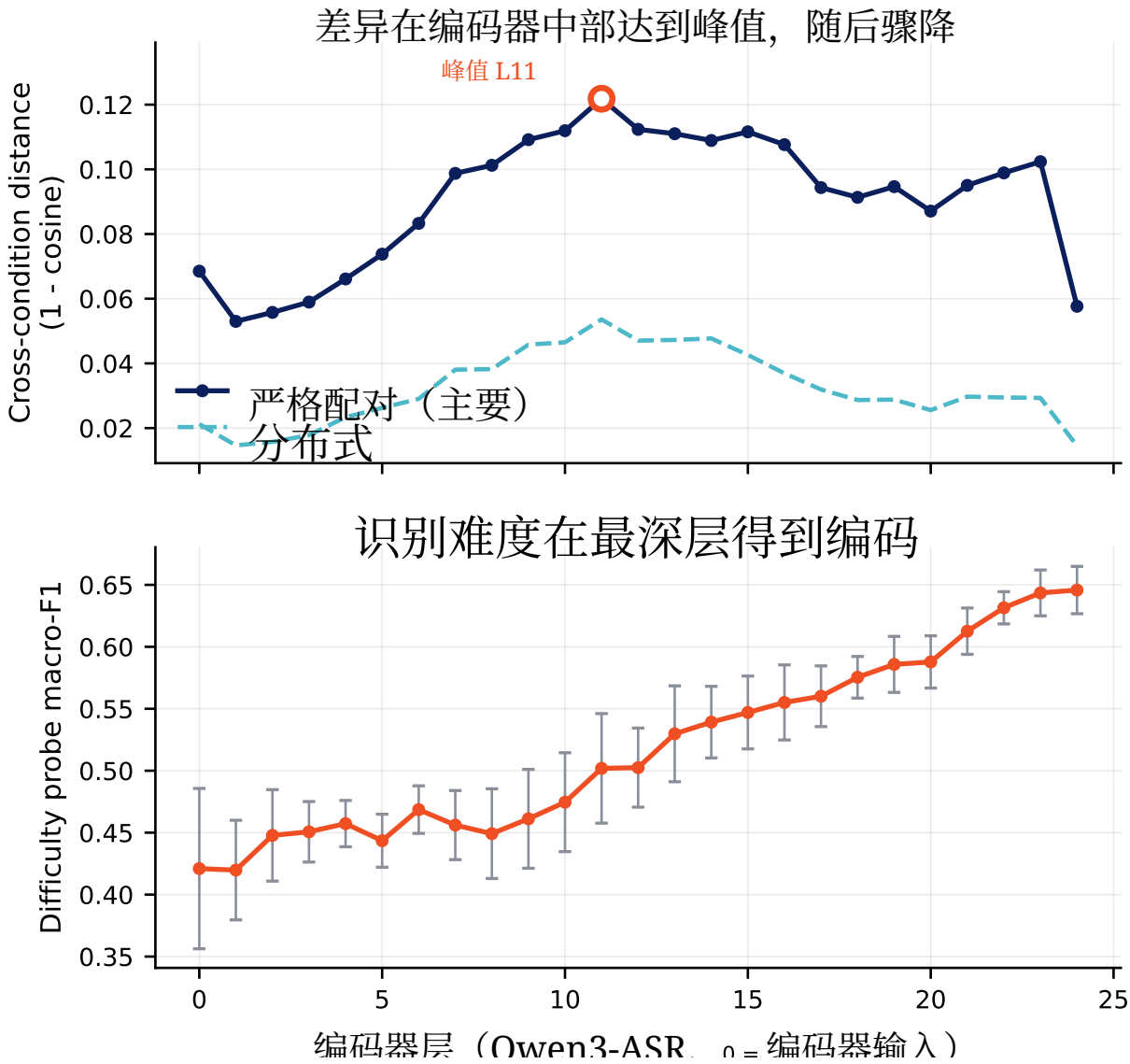


图1： 上部：按 Qwen3-ASR 编码器层展示构音障碍语音与重新合成语音之间的跨条件表征距离（1 − 余弦相似度），以严格配对结果为主，并展示分布变体以供比较。该距离在 24 层中的第 11 层达到峰值，随后向输出端骤降；而先前研究报告的 Whisper-small 结果则随深度增加而上升。下部：按层展示识别难度探针的宏平均 F1，采用说话人互斥五折交叉验证，误差线 ±1 表示各折之间的标准差。难度信息编码得最深——这是先前研究中唯一在此得到复现的发现。

从输入端接近零开始，在堆叠层中部达到峰值，随后在上三分之一区间持续下降，因此两种条件在编码器顶部再次最为相似。图中绘制了两条曲线——以带标记的实线表示严格配对的主要分析，以虚线表示分布变体——二者都呈现先升后降的形态，并在第11层达到峰值；表格5以数值形式报告了另外两种敏感性变体，而未将其绘入图中。下图面板：难度探测覆盖全部25个层位，并带有各折的误差线；其数值随深度增加，在第24层达到宏平均 F1 最大值0.646。

表格5。 跨条件余弦距离：主要分析及三种敏感性变体中的峰值位置和向输出端的衰减情况。（d）真实构音障碍语音，（r）说话人条件重合成，（u）无条件合成。完整的逐层曲线见图1和已发布的CSV；此表格报告了图中无法体现的信息——四种分析在峰值位置及其幅度上均一致。

分析	Pair	n	峰值层	Peak 距离	在 L24	峰值/L24
主要，成对	**d–r**	**411**	**11**	**0.1218**	0.0577	**2.11×**
敏感性 — 分布性，每位说话人	d–r	37	11	0.0764	0.0232	3.29×
敏感性 — 分布式，全局	d–r	8,800	11	0.0536	0.0145	3.69×
敏感性 — 分布式的，全局	d–u	8,800	11	0.0567	0.0242	2.35×

主要分析采用严格配对的 411 行（d、r）子集，其中每一对都是同一位说话人在两种条件下说出同一句子，因此词汇内容和说话人身份在设计上保持匹配。分布式行属于敏感性分析；d–u 行将条件与源语料库混杂在一起，因此应赋予其最低权重。

四条曲线的峰值都出现在**第 24 层中的第 11 层**（r–u 对的峰值出现在第 15 层），随后在**朝输出端下降 2.1–3**。7× Muthu 等人报告，发散程度随深度增加；这一上升趋势正是其核心主张的证据，即言语障碍对高层表示的影响大于对低层声学表示的影响。**在这个编码器上，这一主张并不成立：**发散程度在相对深度的 ≈46% 处达到最大值，而编码器顶端三分之一的部分使三种条件重新趋于一致。三项分布式分析重现了主要子集的曲线形状及其第 11 层峰值；只有在这种情况下，匹配分布设计才适用。数值幅度较小（余弦距离 ≤ 0.12），且**本表中所有曲线均未计算置信区间**——这是本节的一项实际局限（§9）。在每一层，r–u 的发散程度都比 d–r 或 d–u 小一个数量级，其峰值为 0.0116；因此，两种合成条件彼此之间的相似程度远高于它们各自与真实构音障碍语音的相似程度。

结合 §5.1 阅读。 音素错误率（PER）在上部各层仍持续改善，而跨条件差异在此处下降，因此该编码器上三分之一区间的表示会淡化说话人身份和合成伪影，同时继续强化语言内容——这正是为连接至语言模型解码器的投影器提供输入的层应有的表现；而层数只有一半的12块编码器要么不具备这种能力，要么无法达到这一点。

5.3 识别难度编码最深——这一点得到复现 s

图1下图面板展示了难度探针在全部25个层上的结果，并标出了各折的误差线。宏平均 F1 随深度单调上升，从第0层的**0.421**升至第24层的最高值**0.646** （随机水平 $= 0.333$；说话人互斥五折交叉验证）；折标准差也随深度增加而收敛（第0层为 ± 0.065，第24层为 ± 0.019）。这一**重现了**先前研究的发现：识别难度主要编码在最深层的编码器表示中——本研究使用了不同的编码器系列、骨干网络规模和标签来源。微调模型呈现出几乎相同的曲线（最终层为0.642），因此这一信号是架构深度分布的属性，而非适配的结果。

5.4 评分卡

表6。 先前研究针对该编码器、同一语料库提出的可检验主张。

Muthu 等人的发现，关于 Whisper-small	Qwen3-ASR 结果	判定
音素边界 信息在每一 层	无法测量 — ~77 毫秒 帧，占音素的 49.4% 子帧（§3）	**在此架构上无法测试 架构**
音素身份趋于平稳 在编码器中部之后	单调提升至 ≈85%深度；过去 −14.03个百分点 选择划分的中点 数据划分，在TEST上为 −15.88个百分点（§5.1）	**被否定**
跨条件差异 随深度增加而扩大	在第 11 层达到峰值的倒U形 （共 24 层），随后下降 2.1–3.7× （§5.2）	**与此相矛盾**
识别难度编码在 最深层中	单调上升至宏平均 F1 第 24 层达到 0.646（§5.3）	**已复现**
普通话声调是一个持续存在的 错误来源	声调敏感词错误率更差 从第 5 层起，峰值差距 +3.8 个百分点（ §5.1）	**已复现**

有两项需要加以说明。各行均以先前研究已发表的 Whisper- small 结果为基准进行评分；对于我们后来自行测量了相同指标的项目（第 2 行），我们自己的 Whisper-small 探针并未复现其平台期（§5.5），因此该行记录的是与已发表曲线不一致，而非已证实的架构差异。第 4 行和第 5 行是复现结果，不受此影响。

5.5 Whisper-small 配置：在此测量，而非引用

以上内容均将我们对 Qwen3-ASR 的测量结果与其他文献报告的 Whisper-small 结果进行比较。因此，我们使用相同的 CDSD 清单、说话人划分、完整性门控、探针架构、优化器（Adam，学习率 1e-3，批量大小 256，训练 10 轮，梯度裁剪 1.0，丢弃率 0.3）、单元清单和模型选择准则，自行运行了 Whisper-small 探针。Whisper-small 提供 13 个层（嵌入输出加上 12 个模块），相当于 Qwen3-ASR 的 25 个层。

最小值出现在 12 层中的第 10 层——即相对深度的 83% 处——而且第 6 层之后没有出现平台期。图2在同一坐标轴上绘制了两种架构相对于相对 深度的变化，这是对本文负面结果最清晰的表述。

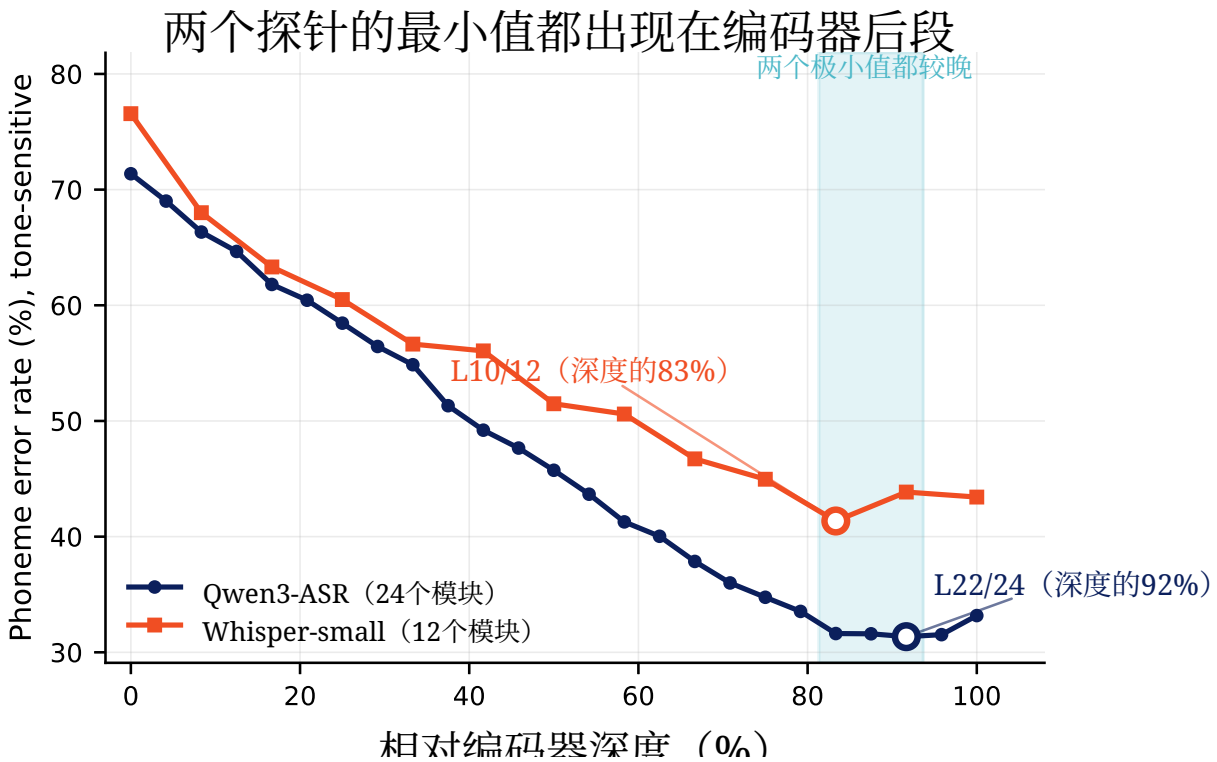


图2： 声调敏感音素错误率随相对编码器深度的变化；两种架构的探针训练量相同（33,617条话语），且使用相同的语料库、数据划分、探针和完整性门控。空心标记表示各条曲线的最小值。Whisper-small 的最小值出现在第 10 层（共 12 层，深度的 83%），Qwen3-ASR 的最小值出现在第 22 层（共 24 层，深度的 92%）；阴影区间覆盖这两个位置。两个最小值都位于编码器后段，但两种模型偏好的适应带不同（图4），因此无法根据这条曲线确定适应带。

替代文本（图2）。一张折线图，纵轴为声调敏感音素错误率，横轴为相对编码器深度，从输入端的 0 到输出端的 1，每种架构各有一条曲线。两条曲线在堆叠层的前半段都急剧下降。Qwen3-ASR 曲线持续改善至约 92% 的深度处，并在第 22 层（共 24 层）达到最低点；Whisper-small 曲线也在堆叠层后段继续改善，并在第 10 层（共 12 层）达到最低点，约为深度的 83%；在第 6 层到第 10 层之间，它进一步提升了 10.14 个百分点，而非在中层堆栈趋于平缓。采用相对而非绝对深度作图，才能直接比较模块数量不同的两种架构。

在匹配数据量下，曲线在第 6 层到第 10 层之间进一步提升了 10.14 个百分点：在这一流程下，Whisper-small 的成熟过程与 Qwen3-ASR 完全一样，都发生在后段。

我们检查了与已发表曲线不一致的四种可能解释，但没有一种能解释这一分歧。

1. 探针配置。 我们逐项对照先前研究的描述进行了核查：探针架构、优化器、批量大小、训练轮次、单元清单，以及按说话人不重叠的 70/15/15 数据划分。

2. 基线能力——现在不能再排除这一因素，原因也颇具启发性。 本节的早期版本曾报告，我们的未适配基线与先前研究的结果相差仅 1.5 个百分点。这一比较采用了他们预印本**v1**中的数据；我们于 2026年8月13日读到该版本。**他们在 8月15日发布的 v2 修订了结果表格**，基线字符错误率从 105.08 降至**89.98** （全块 LoRA 也从 68.85 降至 64.35）。我们全文引用 v2；与该版本相比，双方结果并不一致：我们修正后的基线为 77.57，比他们好**12.4 个百分点**。**因此，我们无法排除基线能力是探针曲线分歧的影响因素，也不再声称可以排除**。基线能力应列入仍待考察的候选因素，而不是已排除的因素。

> 有两点观察使这一结果具有信息价值，而不只是令人泄气。首先，**他们的修订方向与我们的相同，且看起来出于合理的相同原因**：未适配的 Whisper 基线字符错误率超过 100%，经校正后下降了 ~15 个百分点，正如我们的结果在应用重复控制后从 103.57 降至 77.57（§6.6.1）。两个团队在同一语料库上独立撤回一个被夸大的 Whisper 基线，这比单独的任何一次校正都更能证明该伪影确实存在且具有普遍性。其次，**这一检查无论如何都与探针结果无关**：该探针是基于编码器激活训练的 CTC读出，从不运行自回归解码器，因此<style id='6'>本文中的任何探针结果都不依赖于解码配置 ——无论是我们的还是他们的。

3. 填充。 Whisper 会将每个输入填充至固定的 30 秒／1500 帧窗口，而先前研究会截断至 1500 帧；相比之下，我们按每个话语的 ~133 实际帧数进行切片。匹配对提取在<style id='4'>相同的 1,188 个话语和<style id='6'>相同的 7 个层上分别生成了一个填充分片集和一个切片分片集，并通过数值验证确保填充是唯一变量。**两种设置都在第 10 层达到最小值**。

4. 探针训练数据量。 先前研究使用 ~39,000 个话语进行训练；我们的首次运行使用了 6,200 个。在 33,617 个话语的情况下，曲线平均下移 ~4.9 个百分点，且**后期最小值的位置未发生变化**，但最小值后上升的幅度发生了显著变化（表7）。

因此，第10层最小值在四种配置中均保持不变，这些配置涵盖了样本量范围以及相反的填充处理（8,800 个切片；1,188 个切片；1,188 个填充；33,617 个切片）。**我们将这一差异报告为无法解释**。

关于填充的一个意外结果应记录在案。填充并未稀释探针，反而显著增强了它的效果（匹配对实验中最低为87.89 → 53.54音素错误率）。Whisper的编码器不遵循注意力掩码，而是关注整个

30秒窗口，因此填充位置携带了语音的上下文化、模糊化副本，使CTC探针能够对齐数百个信息丰富的位置，而不是仅对齐 ~15个单元。基于这一解读，**Whisper的填充帧是信息通道，而非噪声**——这与掩码Qwen3-ASR分支形成了真正的不对称（§9）。我们将其作为一个**假设**提出：实验表明，填充会显著改变探针的读出结果及其变化方向，但并未证明所提出的信息流机制是其原因。

探针对训练数据量的读出敏感性取决于深度，而且这种依赖关系因架构而异。这正是导致我们后来放弃深度剖面这一论断的结果。

在解读这一结果之前，必须先说明一个混杂因素。 两种数据量都训练相同的 10 个训练轮次，因此，包含 33,617 个话语的条件不仅获得了约 5.4× 倍的优化器更新次数，也包含约 5.4× 倍的独立语句。**因此，该实验识别的是固定训练轮数方案下对训练数据量的敏感性，而不是在优化相匹配条件下对独立语句数量的敏感性**。 更深的表征可能需要更多优化器步数来支持这一 CTC读出，而不是更多独立数据；较小数据量的条件可能是训练不足，而非数据匮乏。要区分这两者，需要进行步数匹配对照——将包含 6,200 个话语的数据集重采样至较大的步数，并以较小的步数训练大数据集——且应在几个具有代表性的深度上开展。我们尚未进行这一实验，因此全文将这一结果描述为**数据量敏感性** ，并避免将“数据效率”作为机制性论断（§9）。

表7. 按深度统计的探针读出对训练数据量的敏感性。增益是在固定10个训练轮次、同一说话人不重叠划分上，当探测训练数据增加5.4× （6,200 → 33,617条话语）时声调敏感词错误率的改善。

编码器	均值增益，跨层，6,200 → 33,617	最深层的增益	相对于该堆叠层的比率 mean
Whisper-small	−4.88 个百分点	第 12 层： **−9.15 个百分点**	**1.87×**
Qwen3-ASR	−3.03 个百分点	第 24 层： **−20.25 个百分点**	**0.74×**

Whisper-small 的最深层表征是其堆叠层中对数据量**最** 敏感的一层；Qwen3- 自动语音识别的则属于**最不**敏感的层之一。**图3** 直接展示了这一后果。

替代文本（图3）。与图2相同的两条深度剖面分别绘制了两次，每种探针训练量绘制一次。在每种架构内部，两种数据量描绘出的曲线形状相同，因此，当探针获得更多数据时，音素身份不再提升的深度并不会移动。标签给出了每条曲线的上升幅度。

这一后果提醒我们应保持谨慎。**请报告多个训练数据量下的探针曲线，否则不要比较不同架构的堆栈后段形状**。

两个最小值的形状不同，而这点是经过测量的，并非主观断言。两个探针均在匹配数据量下使用三个随机种子（0/1/2）重新运行，并保留每条话语的错误计数，以便将用于适配性对比的配对说话人 × 随机种子自助法（B =10,000；在每个随机种子内构造对比）应用于这些曲线。

在 Whisper-small 上，$\mathrm{L10}-\mathrm{L6}=-7.67$ $[-11.92,-3.98]$, $\mathrm{L11}-\mathrm{L10}=+1.83$ $[+0.49,+3.54]$以及 $\mathrm{L12}-\mathrm{L10}=+1.55$ $[+0.07,+3.11]$ ——三项结果均已解决。在 Qwen3-ASR 上，$\mathrm{L22}-\mathrm{L16}=-6.20$ $[-9.34,-2.75]$ 和 $\mathrm{L22}-\mathrm{L20}=-0.64$ $[-1.29,-0.02]$ 已解决，但末端对比 $\mathrm{L24}-\mathrm{L22}=-0.15$ $[-2.63,+2.43]$跨越零点。**Whisper-small 在越过最低点后显著上升；Qwen3-ASR 则没有** ——它的后段堆叠层走势平缓，与 Whisper-small 不同。

若以原始逐层随机种子标准差（Whisper-small 上的均值为 1.43 个百分点）为参照， +1.83 个百分点的变化看起来会像噪声。但事实并非如此：在同一随机种子内进行配对，可以抵消该种子整体难度的影响，这就是配对区间排除零点、而边际比较无法做到这一点的原因。每项结论也都取自其曲线上**估计最精确的点**（Whisper-small 第 10 层的随机种子标准差为 0.128 个百分点，而第 2 层为 2.75 个百分点），而说话人自助法区间

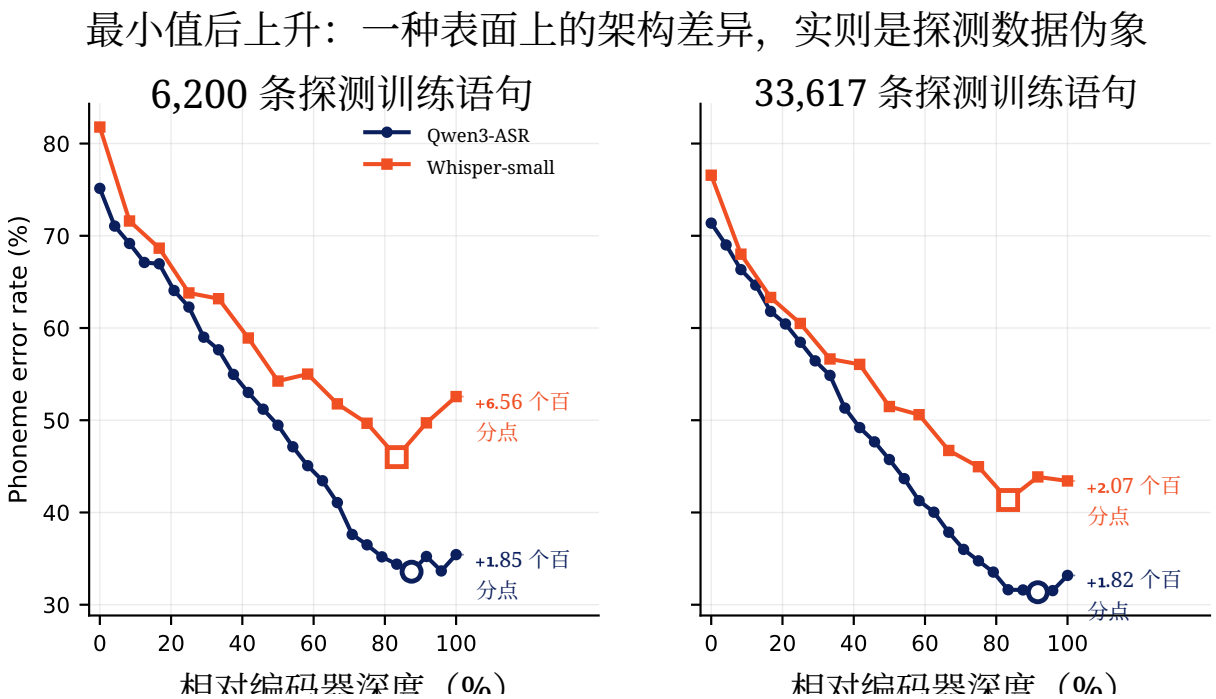


图3： 在两种探测训练数据量下，同样的两条深度剖面。标签给出了每条曲线从自身最小值到最后一层的上升幅度。在 6,200 条话语时，两种架构的后段堆栈行为似乎差异显著（+6.56 对 +1.85 个百分点）；而在 33,617 条话语时则并非如此（+2.07 对 +1.82 个百分点）。Whisper-small 的尾部趋于平坦，是因为其最深层模块是该堆叠层中对数据量最敏感的层；而 Qwen3-ASR 的最深层模块基本不受数据量影响（1.85 → 1.82）。在训练数据变化 5.4× 后，第10层最小值依然存在；这项稳健性检验针对的是一种特定操作——**而非**替代这些曲线仍然缺少的置信区间（§9，局限性17）。

单个层上的结果很宽泛——换言之，probecurves重述了§6.5在适配扫描中发现的结论：**说话人采样，而非重新训练，是不确定性的关键来源**。

5.6 在第二个语料库上重新测量的两个深度剖面

§5.1–§5.5中的所有内容都是CDSD测量结果，而同一对固定的探针曲线支撑着CDSD上的位置差异以及UASpeech上的位置一致性。要么音素身份曲线本质上无法预测该频带，要么它是**依赖语料库的**——在这种情况下，仍然可能从在当前语料库上测量的探针中读出位置。第二种解释会使本文的否定性结果明显变弱，而§5.1–§5.5无法区分这两种情况。**因此，我们在UASpeech上运行了探针B——两个模型、每个抽取点、三个随机种子**。

与CDSD探针协议相比，有两点不同限定了本小节的适用范围。

1. **目标是规范词典发音，而不是强制对齐**。存放UASpeech的机器上未安装对齐器，而UASpeech也不可再分发，因此无法将其转移到具备对齐器的计算环境中。探针B是CTC读出，需要目标序列，而不是时间对齐边界；同时，UASpeech封闭的449词词汇表使词典查找能够做到精确匹配。因此，目标采用CMU发音词典发音 [3]——0.7b版本，共135,166个条目——去除重音后，覆盖39个音素的ARPAbet音素集。**这意味着UASpeech探针的目标是由词标签所暗示的规范音素序列，而不是说话人实际产生的序列**。在一项病理学研究中，这一区分十分重要，因为替换、删除和扭曲都是所测现象的一部分：在UASpeech上，探针测量的是规范音系序列的可解码性，因此其绝对PER无法与CDSD基于对齐所得的目标相比较。这里只比较随深度变化的形状。**这也解释了UASpeech的两种测试计数**：适配扫描对全部3,798条留出语句进行评分，而探针为3,738条——相差60条，是因为这些话语中的词没有CMUdict发音，因而没有音素目标；探针必须将其丢弃，而自动语音识别扫描不会。覆盖率为**11,222/11,402条话语（98.4%）**——其中10,942条为直接匹配，280条通过复合词拆分恢复（退格 → 退 + 格）——另有180条话语、涉及10种词类型，因没有发音而被丢弃。

2. **不存在完整性门控**。 CDSD 探针沿用第4.1节的对齐门控，该门控剔除了14.6%的话语，因此也移除了极短和实际上为空的录音。这里没有任何等效机制，因为不存在可能失败的对齐。

因此，两个语料库的绝对音素错误率不可比较：两者的音位清单不同，一个语料库经过门控，另一个没有，而且一个使用孤立词，另一个使用朗读句子。只有**随深度变化的形状**可比较，以下也仅使用这种形状。数据划分与第6.5节使用的说话人互斥的8份训练集／2份验证集／5份测试集划分相同，共包含5,979／1,505／3,738条话语；探针架构、优化器、模型选择准则和逐层独立性均与第4.3节一致；两个模型的每个层级都使用随机种子0／1／2分别训练了三次。

表8. UASpeech 上按层统计的音素身份 PER（%），采用三个随机种子（0/1/2），均值 ±标准差。第 0 层为编码器输入；第 1–24 层（Qwen3-ASR）和第 1–12 层（Whisper-small）为区块输出。**层索引并非两个模型之间可比较的轴**——相对深度才是，且每条曲线的最小值以粗体标示。频带均值采用与……相同的比例定义

适配扫描。绝对值**不可**与表 3 或 §5.5 中的 CDSD 曲线进行比较。

层	Qwen3-ASR（24 个模块）	Whisper-small（12 个模块）
0	85.43 ± 0.55	93.12 ± 0.18
1	83.42 ± 0.81	90.44 ± 1.49
2	80.77 ± 1.04	86.02 ± 0.56
3	77.07 ± 0.57	81.02 ± 0.50
4	74.85 ± 0.37	73.01 ± 0.29
5	71.26 ± 0.56	71.25 ± 3.91
6	68.34 ± 1.00	69.53 ± 4.96
7	65.06 ± 0.46	63.51 ± 4.81
8	62.05 ± 0.48	55.64 ± 0.62
9	59.66 ± 0.61	53.90 ± 0.45
10	56.77 ± 0.16	54.46 ± 1.38
11	54.33 ± 0.71	**53.86 ± 0.81**
12	52.12 ± 0.44	69.79 ± 0.70
13	48.52 ± 0.50	—
14	47.14 ± 0.48	—
15	45.58 ± 0.11	—
16	44.93 ± 0.57	—
17	43.49 ± 0.59	—
18	42.06 ± 0.61	—
19	41.58 ± 0.38	—
20	**41.04 ± 1.13**	—
21	41.15 ± 0.66	—
22	41.55 ± 0.42	—
23	41.42 ± 0.67	—
24	42.08 ± 0.50	—
频带均值 — `lower`	72.85（模块 1–8）	82.62（模块 1–4）
频带均值 — `middle`	51.13（模块 9–16）	64.98（模块 5–8）
频带均值 — `upper`	**41.80（模块 17–24）**	**58.00（模块 9–12）**

两条曲线的最小值都出现在较深层，而且两者的频带排序都将 `upper` 排在最低。 Qwen3-ASR 的最小值位于**第 20 层（共 24 层）——相对深度的 83%**；Whisper-small 的最小值位于**第 11 层（共 12 层）——92%**。两者都没有显示出中层平台。两种模型上的频带均值排序均为 `upper` < `middle` <`lower`（41.80 < 51.13 < 72.85 和 58.00 < 64.98 < 82.62），这也是两种模型在 CDSD 上给出的排序。

表9. 在模型 ×语料库设计的全部四个单元格中，音素身份 PER 最小值的深度。每个条目都是依据本文自己的流程测得的探针曲线的最小值位置；CDSD 条目是 §5.5 中匹配数据量的三随机种子重新运行结果，UASpeech 条目则来自表8。百分比表示最小值所在的块占该堆叠层深度的比例。两列的 PER 值 不可相互比较——只有各自曲线转折时所处的深度可比。

模型	CDSD（普通话，朗读句子）	UASpeech（英语，孤立词）
Qwen3-ASR（24 个模块）	第22层 — **92%**	第20层 —**83%**
Whisper-small（12 个模块）	第10层 —**83%**	第11层 —**92%**

四个探针最小值均落在 83–92% 的区间内，这是测量所得，而非预设。 每个模型在第二个语料库上都保持了其后期成熟的形态：曲线沿堆叠层向下，至最后六分之一处转折，且无论在哪个语料库或哪种架构中，都未出现中层平台。最小值位置确实会随语料库而变——Qwen3-ASR 从 92% 变为 83%，Whisper-small 从 83% 变为 92%——但分别只移动了两个和一个块位置，**没有超出同一语料库不同随机种子之间的变动范围**。保持不变的是最小值所在的区间、没有中层平台这一点，以及三个比例区间的排序，而不是取得最小值的层。

这能说明什么。 在第6.9节的网格中，位置各不相同，但探针曲线在任何可能与之对应的方面都没有变化；而**在整个设计中近乎恒定的量，无法解释其中变化的结果**——这就排除了最后一种可能：语料库匹配探针或许能预测出继承探针未能预测的频带。这些曲线也明确了最有力候选规则（第8节）的适用范围：在区段平均 *PER* 最低处进行适配在 UASpeech 上为两个模型都选中了 `upper` ，两者都偏好 `all`（表14和表22）。

有两点需要注意，不能过度细致地解读这些曲线。最小值位置在不同随机种子之间并不稳定，而平坦的后期平台正是造成这一现象的原因：仅随机种子0就将Qwen3-ASR的最小值置于第23层、将Whisper-small的最小值置于第9层，而三随机种子均值分别为第20层和第11层——这两个位置的变化幅度都超过了它们在不同语料库之间的变化幅度，这也是论文始终报告**区间而非最小值层**的原因。此外，*Whisper-small*的后期曲线明显没有*Qwen3-ASR*那么平坦：其第12层从第11层的53.86上升至69.79，**+15单个块内就相差5.93个百分点**，而Qwen3-ASR的整个上层带仅跨越2.45个百分点。这是两条曲线在形状上真正存在差异的唯一方面，但并不能据此挽救任何位置规则。记录中还应保留一个不规则现象：Whisper-small在第5至第7层的随机种子标准差偏高（**3.91、4.96和4.81个百分点**），而其他位置的中位数为0.62个百分点，因此该频带的均值（64.98）应结合这一点解读。

---

6. 适配结果

6.1 四区间扫描

表 10. 来自 6 位未见说话人的 5,388 条留出话语上的字符错误率（%），**均值 ± 三个训练随机种子（42/43/44）的标准差**；未适配基线不涉及训练，仅进行一次确定性评估，并在三次运行中以完全相同的方式复现。三个频带的可训练参数预算均为 4.19 M； `all` 为 12.58 M。**粗体**表示最佳结果。严重程度标签源自字符错误率，而非临床标准（§9，局限性 13）。

变体	模块	参数	微平均字符错误率	中等难度	高-难度
基线（不进行适配）	—	0	32.06	23.91	74.55
`lower`	1–8	4.19 M	35.52 ± 1.15	28.79 ± 1.38	70.63 ± 0.35
`middle`	9–16	4.19 M	30.71 ± 0.31	23.72 ± 0.43	67.14 ± 0.36

变体	模块	参数	微平均字符错误率	中等难度	高- 难度
upper	**17–24**	**4.19 M**	**29.88 ± 0.36**	**22.76 ± 0.40**	66.96 ± 0.14
all	1–24	12.58 M	31.19 ± 0.37	24.34 ± 0.41	**66.87 ± 0.25**

排序： upper < middle < all < 基线 < lower，以及 upper 是每个随机种子中的最佳频带。§5.1中预先指定的预测部分成立。 upper 击败了在训练参数数量仅为其三分之一的情况下对整个编码器进行适配的方法，其区间不包含零（表12）。它在三个随机种子中也都击败了 middle ——即先前研究的探针对应指向的频带的比例对应项——但一**旦纳入优化随机性，这一差距便无法得到确认**，因此我们在§6.1.1中将其报告为方向性结果，而非已确立的结果。**这个upper-击败-all 句子本身并不能单独隔离位置因素**，因为 upper 和 all在适配器所处位置以及所部署的容量方面存在差异；§6.7通过位置 × 容量 2×2将二者区分开来。

表11. 每位说话人的字符错误率（%），同三个随机种子下的均值 ±标准差。每位说话人的最佳变体以**粗体**标示。

说话人	base	lower	middle	upper	all
CDSD_01	13.21	10.55 ± 0.32	10.56 ± 0.31	**9.02 ± 0.11**	11.46 ± 0.35
CDSD_03	53.29	49.74 ± 0.54	44.42 ± 0.46	44.32 ± 0.53	**43.20 ± 0.63**
CDSD_06	91.78	87.56 ± 0.46	85.55 ± 0.79	**85.30 ± 0.26**	86.04 ± 0.64
CDSD_08	**39.42**	59.06 ± 2.24	48.90 ± 1.73	51.42 ± 1.85	52.47 ± 1.94
CDSD_29	27.38	23.33 ± 0.19	21.98 ± 0.35	**20.03 ± 0.22**	22.23 ± 0.14
CDSD_38	16.87	25.26 ± 3.42	15.94 ± 0.34	**13.50 ± 0.10**	14.06 ± 0.47
微观	32.06	35.52 ± 1.15	30.71 ± 0.31	**29.88 ± 0.36**	31.19 ± 0.37

表12。 配对的**说话人 × 随机种子**自助法，B = 10,000：对说话人进行有放回抽样，并为每次重复抽样抽取一个训练随机种子；每个对比均在同一随机种子内形成，从而抵消某个随机种子的总体难度。这与表格15和23使用的是同一个估计量，并取代了本表早期版本中使用的仅按说话人自助法——有关更改内容及原因，请参见§6.1.1。“种子标准差”是三个随机种子之间对比值的标准差。

比较	Δ 字符错误率（百分点）	种子标准差	95%置信区间	判定
upper − lower	**−5.64**	1.27	[−10.87, −2.43]	不包含 0
upper − all	**−1.31**	0.16	[−2.16, −0.25]	不包含 0
middle − all	−0.48	0.65	[−3.13, +1.10]	**包含 0**
upper − middle	−0.83	0.59	[−2.06, **+1.86**]	**包含 0**
upper − 基线	−2.18	0.36	[−6.78, +4.60]	**包含 0**
middle − 基线	−1.35	0.31	[−5.51, +3.80]	包含 0
all − 基线	−0.87	0.37	[−5.63, +5.93]	包含 0
lower − 基线	+3.46	1.15	[−3.74, +12.26]	包含 0

这两部分表达了不同的结论，而论文的主张比点估计所暗示的更为有限。 在适配策略中， upper 在每个随机种子下都是最佳频带，且它的三个对比中有两个得到明确结果： upper − all = −1.31，置信区间为 [−2.16, −0.25], 和 upper −lower = −5.64，置信区间为 [−10.87, −2.43]。 **upper − middle 结果不明确**： −0.83，置信区间为 [−2.06,+1.86]。它在三个随机种子下均为负值（−0.26 / −1.43 / −0.81），因此方向一致，但幅度相差 5.5 倍；在对说话人和随机种子进行重采样后，区间包含零。 middle − all 的结果也不明确，且其符号随随机种子而反转。与完全不进行适配相比，**这里没有任何显著结果**：upper − 基线为 −2.18 个百分点，置信区间跨越 [−6.78, +4.60]。原因在于 CDSD_08：它在每个变体下都出现退化，因此在适配方法之间的对比中相互抵消，却主导了适配方法与基线之间的对比。对 upper < middle 进行符号检验，并未达到一致：随机种子 42 下为 6 位说话人中的 4 位，随机种子 43 和 44 下为 5/6，三随机种子均值下也为 5/6；每种情况下，CDSD_08 都是不同意见者（表11）。双侧检验中，符号检验的 p 值为 = 0.219，精确配对符号翻转检验的 p 值为 = 0.438（表26）。对于 upper < lower ，则确实达到一致——每个随机种子下均为 6/6，两种检验的 p 值均为 = 0.031。不同训练随机种子之间的一致，并不意味着不同说话人之间也一致。

因此，仅凭 CDSD 能站得住脚的结论范围有限：既然要在这个编码器中放置低秩适配器，上层带就比下层带和全部编码器模块更适合作为目标。在我们运行的每个随机种子下，上层带也优于中层带，但在当前样本量下，这一差距尚未得到确认。这个测试集并不能证明这种适配是否值得进行——不过，§6.5 的第二个语料库上的结果确实证明了这一点。

6.1.1 这些是三个随机种子的结果，先前的一项结论并未经受住检验

本节的早期版本报告了每个频带各一次训练运行的结果，并给出了通过仅按说话人自助法计算的区间。这个估计量量化的是恰好被留出的六名说话人所带来的影响，却没有体现适配器采用了哪种随机初始化和批次顺序。鉴于两个备受关注的 CDSD 对比约为 1.5 个百分点，这项遗漏至关重要，审稿人的指出是正确的。

因此，我们在其他实验方案完全相同的情况下，使用随机种子42、43和44分别对每个频带进行了重新训练；本文中的每个CDSD区间现在都来自说话人 × 随机种子估计量。**有一项结论发生了变化。** upper − middle 此前报告为 −1.52，置信区间为 [−2.27、 −0.54]，不包括零；现在为 **−0.83，置信区间为 [−2.06、 +1。86], 其中包括零**。我们不再声称 upper在CDSD上胜过 middle ，只能说在我们运行的三个随机种子中，它在全部三种情况下都胜出。

论文第一项贡献所依据的主张不受影响。 upper − all 从 −1.47变为 −1.31，并且仍已解决。§6.7中参数匹配的 2×2给出了与 −1.26相同的同秩配对，在符号和幅度上都一致；需要注意的是， 2×2自身的预算匹配对比并未解决（§6.7），因此它能够佐证这一对比，但不能独立确立位置效应。两个对比的随机种子敏感度相差超过三倍—— upper − all 的种子标准差为 0.16，而 upper − middle 为0.59——因此，在论文其他部分有独立佐证的对比，也是重新训练后保持稳定的对比；而此前没有独立佐证的对比，则是消解的那个。我们将此作为对单次运行频带扫描的一般性警示，包括我们自己的扫描：两个频带之间1.5个百分点的差异，处于随机种子可以造成的变动范围之内。

upper < all 结果与 Lin 等人 [14], 的结论相矛盾：他们认为，在自监督语音编码器中，向全部层均匀插入适配器的效果无法被超越。而在这里，仅在三分之一的模块中使用适配器，效果就优于向所有模块插入适配器，且置信区间不包含零；留一法分析也显示结果稳定。这是本文论点的又一例证：选择性位置是否有益，本身取决于编码器。**这一对比仅基于 CDSD**；UASpeech 无法支持任何一方的结论，因此，基于它的 2×2 以及第6.6节的跨模型比较，都是单语料库分析。

6.2 留一说话人敏感性

表 13. 在其余五位说话人上重新计算的微平均字符错误率（%），为随机种子 42/43/44 下的均值 ± 标准差，包含两项主要对比。基准列不涉及训练，仅进行一次评估。

排除	base	upper	middle	all	upper −基线	upper −middle
none	32.06	29.88 ± 0.36	30.71 ± 0.31	31.19 ± 0.37	**−2.18 ± 0.36**	−0.83 ± 0.59
CDSD_01	37.02	35.37 ± 0.48	36.01 ± 0.33	36.38 ± 0.38	−1.65 ± 0.48	−0.65 ± 0.69
CDSD_03	30.41	28.76 ± 0.35	29.65 ± 0.32	30.25 ± 0.36	−1.66 ± 0.35	−0.89 ± 0.56
CDSD_06	26.23	24.47 ± 0.41	25.36 ± 0.41	25.83 ± 0.41	−1.76 ± 0.41	−0.89 ± 0.74
CDSD_08	30.32	24.79 ± 0.04	26.42 ± 0.03	26.16 ± 0.03	**−5.53 ± 0.04**	−1.62 ± 0.02
CDSD_29	33.41	32.73 ± 0.43	33.24 ± 0.33	33.78 ± 0.51	−0.69 ± 0.43	−0.51 ± 0.69
CDSD_38	36.23	34.37 ± 0.48	34.76 ± 0.49	35.89 ± 0.57	−1.86 ± 0.48	−0.39 ± 0.85

upper − middle **在每个留一法子集中都保持相同的方向**（−0.39 至 −1.62 个百分点），因此没有哪个说话人单独主导了这一结果；但在每个子集中，其随机种子标准差都与其效应量相当。这与完整自助法得出的情况相同（§6.1.1），也正是我们不将其结论视为已解决的原因。 upper − 基线则在 −0.69 至 −5.53 个百分点之间波动，完全受测试集构成左右。这也是保留 CDSD_08 而非将其排除的原因：剔除它会使适配的表观收益从 −2.18 增至 −5.53 个百分点，造成 2.5× 倍的膨胀，而这是通过移除唯一持不同意见的说话人得出的。另请注意，排除 CDSD_08会使随机种子标准差降至 ±0.04——该测试集几乎所有随机种子间变异都来自这位说话人。

6.3 行为层面佐证： middle几乎不会改变模型

留出话语的比 例，其假设与未适配模型**逐字节完全相同** 's:

变体	与基线相同	%
middle	1,217 / 5,388	**22.6%**
all	579 / 5,388	10.7%
upper	129 / 5,388	**2.4%**

适配模块 9–16 后，未受影响的输出数量是适配模块 17–24 后的四倍多——这提供了独立于指标的佐证，表明该编码器的上三分之一区间才是产生重要影响的关键所在。

6.4 难度层级并不均匀

对于**中度**说话人，在三随机种子均值下， upper (22.76)和 middle (23.72)均低于基线（23.91），而 all (24.34)则不低于基线；在六名说话人上，这些差距均未得到解决，因此这只是数值上的排序，而非已证实的效应。对于**重度**说话人，三个深层变体均较基线有显著改善（66.87–67.14，相比74.55），且 all 略胜一筹。（本段较早版本引用了单次运行数据——22.96、24.63、24.77——这些数据已被§6.1.1的三随机种子重新运行结果

取代。）因此， upper的总体优势由中度说话人所贡献——在六名带有基于字符错误率的严重程度标签的测试说话人中，这是一项观察，而非一项结果。

6.5 第二个语料库：在英语孤立词上重复抽样的扫描

以上全部仅涉及一种语言、一个语料库、一种内容类型，以及基于字符错误率的严重程度标签。我们在**UASpeech** [12], 上重复了完全相同的四区间扫描，它在上述每一个维度上都与 CDSD 不同：是英语而非普通话，是**449 孤立词** 而非新闻朗读句子，并且采用**临床** 严重程度标签。实验方案、频带定义、参数预算、LoRA 配置、完整性门控和解码路径均保持不变。

语料库核算，以便重建数据划分。 原始 UASpeech 发布版本描述了 19 名构音障碍说话者和 455 个不同的词；当前发行包含 16 名。**我们的副本包含 15 名**：F02、F03、F04、F05、M01、M04、M05、M07、M08、M09、M10、M11、M12、M14、M16——即当前的 16 名减去**M06**。我们的清单覆盖**449 个不同词类型**，差额源于缺失的说话人以及被音频完整性检查剔除的话语。

数据划分按说话人互斥：8 名训练说话人 / 2 名验证说话人 / **5 名测试** 说话人（3,798 条话语），并按严重程度**和**性别分层；最初仅按严重程度分层，导致测试集全为男性，因此在启动前进行了修正。由于 UASpeech 的每位说话人都朗读同一份 449 词清单，因此从设计上就无法进行文本互斥划分——这是一个按设计构建的封闭词汇任务。评分时不区分大小写。每种适配条件均使用随机种子 42/43/44**训练三次**，并报告随机种子均值 ± 标准差。

表 14。UASpeech：来自 5 名未见说话人的 3,798 条留出语句。适配行报告三个随机种子的均值 ± 标准差。最佳结果以**粗体**显示。

变体	模块	参数	字符错误率（均值 ±标准差）	随机种子范围
基线（不进行适配）	—	0	41.95（单次运行）	—
`lower`	1–8	4.19 M	51.85 ± 0.38	51.60 – 52.29
`middle`	9–16	4.19 M	36.62 ± 0.12	36.47 – 36.70
`upper`	17–24	4.19 M	34.41 ± 0.32	34.10 – 34.75
`all`	1–24	12.58 M	**32.75 ± 0.30**	32.50 – 33.08

随机种子方差很小：百分点标准差为 0.12–0.38，相比之下，频带差异为 2–17 个百分点。重新训练使某个频带的字符错误率变化零点几；改变所适配的频带则会使其变化几个百分点。

表15。 UASpeech同时对比两种不确定性来源。“随机种子范围”是固定全部5位说话人后，针对每个随机种子重新计算的配对对比；“说话人 ×随机种子”则同时对说话人和随机种子进行重采样（$B = 10,000$）。对比在合并之前先在每个随机种子内形成，因此某个随机种子的总体难度会相互抵消。

对比	均值（百分点）	随机种子范围	说话人 × 随机种子 95%置信区间	跨随机种子的符号
upper − middle	**−2.21**	[−2.58, −1.73]	[**−4.37, −0.39**] 不包含 0	稳定
upper − lower	**−17.44**	[−17.91, −16.85]	[**−24.44, −10.05**] 不包含 0	稳定
lower − middle	**+15.23**	[+14.99, +15.59]	[**+8.79, +21.33**] 不包括 0	稳定
all − middle	**−3.87**	[−4.20, −3.59]	[**−6.72, −1.24**] 不包含 0	稳定
upper − all	+1.66	[+1.01, +2.09]	[−0.51, +4.46] 跨度 0	稳定

加入随机种子方差后，主要对比仍然成立： upper以2.21个百分点领先于 middle，说话人 × 随机种子区间不跨越零点，且三个随机种子中的符号均相同。 **upper − all 在方向上保持一致，但尚未得到解决** ——三个随机种子全部有利于all，但说话人 × 随机种子区间跨越零点；鉴于该效应量，这正是一个包含5位说话人的测试集应当预期产生的结果。

说话人而非随机种子，是决定不确定性的来源，这颠覆了通常的担忧。 在 upper − middle 中，固定全部五位说话人时，对比在三个随机种子间变化 0.85 个百分点；同时对说话人进行重采样，则会将区间扩大到 3.98 个百分点。（这两者是不同的统计量，因此我们不将它们的商称为不确定性比率。）增加随机种子对本研究帮助不大；增加测试说话人则会大有裨益。

表 16。 在两个语料库上进行配对说话人自助法，B = 10,000，按有放回抽样方式重采样说话人。每个语料库使用一组共享的重采样样本为所有变体评分，因此在每次重采样中，说话人难度会相互抵消。负值 Δ 有利于第一个变体。

比较	CDSD Δ（95%置信区间）	UASpeech Δ(95%置信区间)
upper − middle	−0.83 [−2.06, **+1.86**]不显著	−2.21 [−4.37, −0.39]
upper − lower	−5.64 [−10.87, −2.43]	−17.44 [−24.44, −10.05]
lower − middle	+4.81 [+1.10, +10.23]	+15.23 [+8.79, +21.33]
upper − all	−1.31 [−2.16, −0.25]	+1.66 [−0.51, +4.46]不显著
all − middle	+0.48 [−1.10, +3.13] 不显著	−3.87 [−6.72, −1.24]
upper − 基线	−2.18 [−6.78, +4.60] 不显著	−7.54 [−11.53, −3.83]
all − 基线	−0.87 [−5.63, +5.93] 不显著	−9.20 [−14.93, −4.37]
middle − 基线	−1.35 [−5.51, +3.80] 不显著	−5.33 [−8.25, −2.88]
lower − 基线	+3.46 [−3.74, +12.26]不显著	+9.90 [+5.17, +14.97]

标为不显著 的区间包含零。**在 UASpeech 上，除 upper− all 之外的所有对比都已得到明确结论；**而这正是本文最希望解读、也仍未能在该语料中得出结论的对比。CDSD 列现在在全文采用说话人 × 随机种子区间（§6.1.1）。有意省略各变体边际置信区间：在 5–6 位说话人的情况下，这些区间跨度为 33–54 个百分点，毫无参考价值。

关于什么是主要结果，我们在此直言不讳，因为早期版本曾对此处理错误。 该版本将 upper − middle 单一主要对比——即在训练任何适配器之前由探针预测的对比——列为主要对比，却将标题结论建立在 upper − all上，而后者并未事先声明。现在，我们基于**跨模型交互作用**的作用，将其指定为主要结果（§6.9.1）：这是贡献1所提出的主张，也是本文试图回答的问题

。它恰好也是这里唯一一个在微平均字符错误率、说话人等权加权和精确检验下都能得到明确结论的量；这是我们报告的一项性质，而非指定它为主要结果的理由——因为结果好看才选择主要对比，属于基于结果进行选择。 `upper − middle` 和 `upper − all` 作为两个预先指定的主效应报告。表 16 中所有其他条目均为**次要和探索性结果**，报告时未对每个语料库的九项对比进行多重性调整；对比有九项时，部分区间因偶然而不包含零是意料之中的。

有两项结果在两个语料库中的区间均不包含零，且经重复抽样后仍成立： `upper`优于`lower`，而 `lower` 不如 `middle`。第三项—— `upper` 优于 `middle` ——在 UASpeech 上已解决（-2.21，置信区间 $[-4.37, -0.39]$），但在 CDSD 上，重新抽样训练随机种子后**并未**解决（-0.83，置信区间$[-2.06, +1.86]$; §6.1.1），尽管 CDSD 的点估计在全部三个随机种子中都处于同一方向。因此，我们报告称，`upper` $>$ `middle` 在两个语料库中的方向一致，并且在其中一个语料库中已解决。**两个语料库中没有任何对比呈现相反方向的已解决结果。但二者也并非一致**： `upper − all` 点估计的符号反转（CDSD 上为 -1.31 个百分点，UASpeech 上为 $+1.66$ 个百分点），因此我们将参数效率结果视为单语料库结果——在仅有五名测试说话人的情况下，缺乏证据，不足以支持反向结论。

UASpeech 为 §6.1 中尚未解决的问题给出了定论。在 CDSD 上，所有适配与基线对比的置信区间都包含零；而在 UASpeech 上，四项对比均得出明确结果：适配上层带、全部编码器模块，甚至中层带，都优于不进行适配；而**适配下层带的表现则稳定地差于完全不适配**（字符错误率高出 $+9.90$ 个百分点，词错误率高出 $+16.22$ 个百分点，轻度说话人的字符错误率更是翻倍有余：$8.84 \rightarrow 19.68$）。还有两点值得注意： `all` 是 UASpeech 上表现最佳、CDSD 上表现第二佳的变体，因此即使针对性位置的效果更好，均匀插入仍是稳健可靠的默认选择；而 `lower` 对**轻度**说话人造成的相对损害最大，因为其基线字符错误率为 8.84，几乎没有理由用适配器去干扰这一表现。

6.6 第二种架构：在 Whisper-small 上进行相同的扫描

§6.5 在固定架构的同时改变语料库；本节则反其道而行之，这项比较正是论文第一项贡献的基础。在 §6.9 得出结论之前，应将本节结论视为仅适用于 CDSD。扫描使用 Whisper-small，并采用与 §6.1 相同的 CDSD 清单、说话人划分、完整性门控、LoRA 配置方案、随机种子规范和评估代码；区间按其 12 个模块的比例划分（`lower` 1–4、 `middle` 5–8、 `upper` 9–12、 `all` 1–12）。

表17。 来自同一批六名未见说话人的5,388条留出语句上的字符错误率（%），**均值 ± 三个训练随机种子（42/43/44）的标准差**；未适配基线无需训练，三次运行的结果均为77.567。

变体	CER	相较于未适配
未适配基线	77.57	—
`all` (1–12)	**53.56 ± 0.01**	**−24.01**
`middle` (5–8)	55.16 ± 0.13	−22.41
`lower` (1–4)	56.28 ± 0.07	−21.29
`upper` (9–12)	56.95 ± 0.08	−20.61

表18。 Whisper-small 在 CDSD 上的频带对比，使用三个随机种子。每个对比均报告三个估计量：贯穿论文使用的说话人 × 随机种子自助法（**微观** 字符错误率，对重采样说话人的编辑计数进行汇总）、针对**宏观**字符错误率的相同自助法（说话人等权重），以及针对全部 $2^6 = 64$ 说话人符号分配的**精确配对符号翻转**p值。§6.10 解释了为何要展示这三种估计量。负值 Δ有利于前面所述的变体。

对比	微观 Δ	95%置信区间	宏观 Δ	95%置信区间	精确 p 值
all − upper	**−3.40**	[−5.03, −1.84]	**−3.63**	[−4.94, −2.14]	**0.031**
all − lower	**−2.72**	[−4.42, −1.29]	**−3.05**	[−4.71, −1.56]	**0.031**
middle − upper	−1.80	[−4.38, **+0.27**]	**−2.01**	[−4.07, −0.04]	0.156
all − middle	−1.60	[−3.24, **+0.46**]	−1.61	[−3.05, **+0.11**]	0.125

all − upper **在全部三种估计量下均成立**，其精确 p 值达到六名说话人情况下的可达到下限——每名说话人都更倾向于 all。这并非它独有：同一表格中的 all − lower 也在全部三种估计量下成立，Qwen3-ASR × CDSD（表26）上的 upper − lower 亦然。唯独 all − upper 的特点是，它是该单元格结论所依据的位置 对比，并且在每种估计量下都成立，而其 Qwen3-ASR 对应项则不成立。该单元格的随机种子标准差比 Qwen3-ASR 扫描小一个数量级：all为0.013，其余为0.06–0.13个百分点，而后者为0.31–1.15个百分点；Whisper-small 在 CDSD 上的适配结果对随机种子的敏感度低得多。

图4 将两次扫描并列展示，每次均按照各自的未适配基线进行归一化，并叠加每位说话人的数据点，同时标出了不进行适配的基准线。

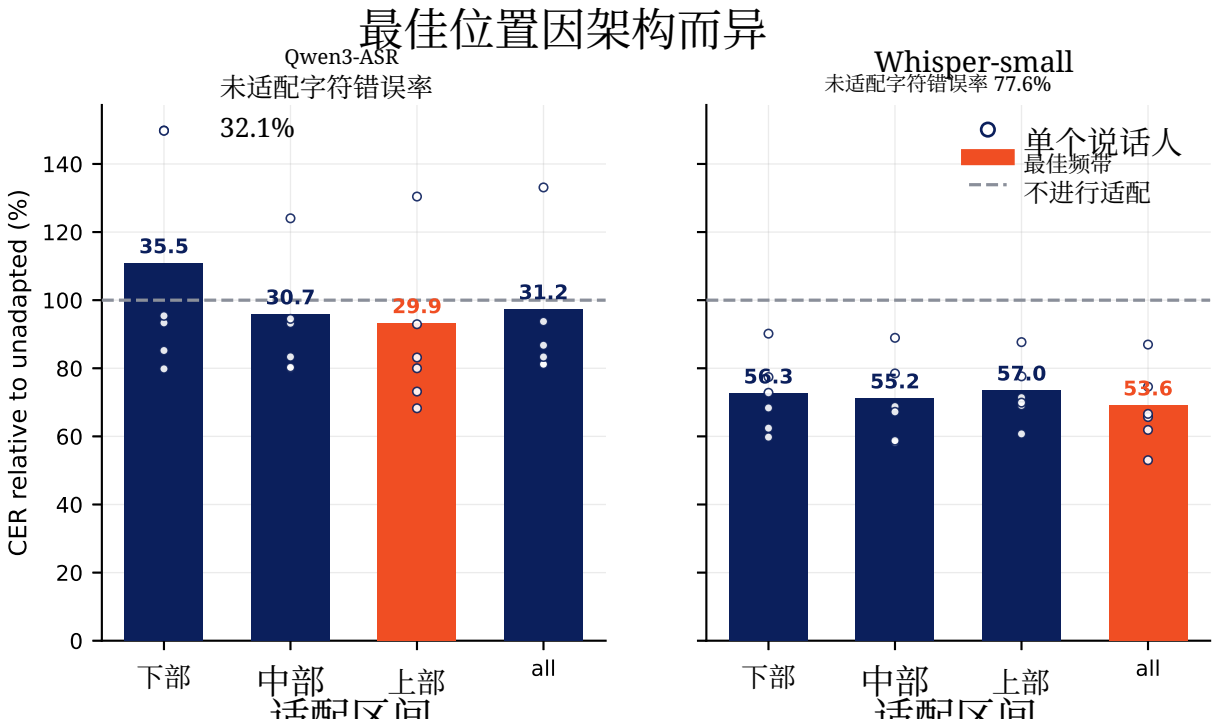


图4： 在CDSD上进行的四频带LoRA扫描，使用相同的清单、数据划分、门控、实验方案、参数预算和随机种子规范，仅替换编码器。由于使用共享的绝对坐标轴将难以读取，因此字符错误率（CER）显示为相对于各自架构未适配基线的数值（相对于32.06%和77.57%）；每个柱形图上标出了绝对CER。空心圆表示各位测试说话人的结果。虚线表示不进行适配。最佳频带以橙色标示： upper 在Qwen3-ASR上， all 在Whisper-small上，其中 upper 的秩在四个频带中最差。在Qwen3-ASR上， lower高于基准线——比完全不进行适配还差。

替代文本（图4）。两个面板并排排列，每个架构对应一个面板；每个面板均将四个 LoRA 放置区段——下层带、中部、上层带和全部——绘制为相对于该

架构自身未适配基线的 CER 变化，并带有误差线。在 Qwen3-ASR 上，上层带带来的改善最大，下层带最小。在 Whisper-small 上，排序有所不同：全部块适配带来的改善最大，而上层带最小。因此，两个面板对这些区段的排序不同，这正是论文直接检验的跨模型交互。

根据这些区间所允许的范围，谨慎地表述本研究的结论。

1. **Qwen3-ASR 上的结论无法推广，而且两个模型都无法判定这组对比** **.** 在 Qwen3-ASR 上，$\texttt{upper} - \texttt{middle} = -0.83$，置信区间为 $[-2.06, \mathbf{+1.86}]$ ——在全部三个随机种子上，方向均有利于 upper，但**尚未得到判定**（§6.1.1）。在 Whisper-small 上，同一方向的同一对比为 $\mathbf{+1.80}$，**置信区间为** $[\mathbf{-0.27}, \mathbf{+4.38}]$ ——符号相反，同样无法通过微平均字符错误率判定，但通过说话人等权加权后则可以判定（表18将其列为 $\texttt{middle} - \texttt{upper} = -1.80$）。因此，这组对比在两个模型上方向相反，但按主要指标均无法判定：这提示可能存在符号翻转，但并不能证明这一点。

2. **all 在 Whisper-small 上表现最佳；相较于 upper 的优势已得到判定，但相较于 middle则尚未得到判定**。$\texttt{all} - \texttt{upper} = -3.40$，置信区间为 $[-5.03, -1.84]$；$\texttt{all} - \texttt{middle} = -1.60$，置信区间为 $[-3.24, +0.46]$,，该区间跨越零点。在 Qwen3-ASR 上，upper 优于 all，因此，两个架构对 upper/all 这组对比得出了方向相反的判定 —— 这是论文中最有力的单项证据，表明位置并非紊乱本身的固有属性；不过，这一结论仅依据该语料库特有的第四个单元格。更狭义的主张，即全编码器适配优于所有等预算频带，缺乏证据支持：此处无法区分 middle 与 all 。3. **upper——在 Qwen3-ASR 上胜出的方案——在此处四个频带中排名最差**。4. **先前研究关于充分性的主张在我们的数据划分上并不成立**。他们在 v2表格中报告，第5–8层频带与全编码器适配的相对差距在 2.89% 以内；而我们的测量结果显示，其表现相对 all 差middle 3.0%，并且无法区分两者（表18）。

一个我们无法消除、且对上述所有内容都构成限制的混杂因素。 Whisper- small 在该语料库上的未适配字符错误率为 77.57，而 Qwen3-ASR 为 32.06——其基线能力仍不到后者的一半，不过差距比最初看来窄得多。**这个数值取代了此前的 103.57；后者部分是我们自身解码造成的伪影**，校正过程记载于 §6.6.1，因为这会改变此处可以得出的结论。在重复控制解码下，六名测试说话人中只有一人的字符错误率仍超过 100%（135.1；此前为 308.19），因此，对于该模型在该语料库上的表现，“插入错误多于参考文本”这一语料库层面的幻觉特征**并不** 成立，我们也不再以此作为任何论证的依据。能力较弱的模型仍可能从不同方向获益， lower中可以看出这一点：它会损害Qwen3-ASR（$+3.46$ 个百分点），却有助于 Whisper-small（-21.29 个百分点）。因此，两种架构不仅在深度、帧率、预训练语料库、规模和解码器类型方面不同，其基线能力也存在差异（§9，局限16）。本实验确实表明，这种差异**是真实存在的，并非跨论文伪影** ——同时也证伪了膝点规则，因为 Whisper-small 在第 10 层达到的 PER 最低值处于其 upper 频带内，而该规则预测这个位置应当能带来改善。这是四种架构中表现最差的。

6.6.1 未适配的 Whisper 基线对解码设置敏感，因此我们报告两种设置

我们首次测量 Whisper-small 未经适配时的 CDSD 字符错误率，结果为 103.57；由于超过 100%，很容易将其解读为出现幻觉的证据。这个数值是在贪心搜索和固定 200 个令牌的生成预算下得出的，而话语的平均时长仅为 2.4 秒——这为重复循环留下了充足空间，循环中的每个字符都会被计作插入错误。本文中的 **每一种**Whisper 条件都使用重复控制（no_repeat_ngram_size 3，束宽为 5）和时长缩放令牌上限重新运行，并据此报告全文中的图。采用较小的固定上限并不能

解决问题：CDSD 的话语时长可达 18.3 秒，UASpeech 则可达 19.5 秒，因此，若根据较短话语设定固定预算，就会以删除错误取代插入错误。

校正具有明显的非均匀性，而且方向与诊断预测一致：未适配基线提升了 26.01 个百分点，而适配区间仅提升 1.85 至 4.63 个百分点，这是因为已适配模型已经能够输出格式良好的结果，几乎没有需要抑制的循环。**本文的位置结论没有任何改变**，在两个语料库上均是如此；发生变化的是所有涉及未适配或弱适配条件的差距大小。有两个判定发生了变化，并且我们均按校正后的数值报告： $\texttt{all} - \texttt{middle}$ 此处（此前为已解决，现在跨越零点）以及 UASpeech CER 中的 $\texttt{lower} -$ 基线（§6.8，此前跨越零点，现在已解决）。我们之所以报告这些变化，而不是默默地复述原数字，是因为相同的解码默认设置被广泛使用，而以这种方式抬高的基线会让人得出模型失效的结论，尽管问题有一部分其实出在测试框架上。

6.7 位置还是容量？参数匹配的 2×2

到目前为止报告的每一项 `upper` 与 `all` 对比都同时改变了两个因素： `upper` 将适配器放置在 24 个模块中的 8 个模块，而 `all` 将其放置在全部 24 个模块中，但两者都使用 $\mathrm{r} = 64$，因此 `all` 还具有三倍的可训练容量。这留下了一种可能的解读： `all` 之所以没有胜出，并不是因为其位置，而是因为对于 33,617 个话语而言，它的参数量过大。本小节在 CDSD ×Qwen3-ASR 上将这两个因素分离开来；在该单元格中， `upper` 胜过 `all`。

设计。 位置 $\times$ 容量采用完全交叉设计，并为每个单元格选择秩，使沿容量轴的可训练参数数量保持一致。每个适配模块贡献四个 LoRA 模块，每个模块的 $\mathrm{r} \times 2 \times 1024$ 参数，因此配置的预算与（模块数 ×秩）成正比：

- **高容量，精确匹配**。 8 个模块 $\times\ \mathrm{r} = 192$ 和 24 个模块 $\times\ \mathrm{r} = 64$ 均得到 1,536 个秩-模块单位和 **12,582,912**个可训练参数——这是精确匹配，而非近似匹配。
- **低容量，98.4% 匹配**。 8 个模块 $\times\ \mathrm{r} = 64$ 得到 4,194,304 个参数；24 个模块 $\times \mathrm{r} = 21$得到 4,128,768 个参数。由于秩必须为整数，24 个模块无法实现精确匹配。

四个单元格均使用随机种子 42、43 和 44 进行训练，并始终采用 $\alpha = 2$<style id='4}r，清单、数据划分、门控、优化器、调度计划、训练轮数、合并流程和解码路径均与 §6.1 以及彼此之间保持一致。两个设计要点至关重要，之所以在此说明，是因为本节的早期版本曾将第二点表述错误。

随机种子。 </style}早期版本只在随机种子 42 下运行了此实验 2×2 。由于其主要结果的差距约为 1.5 个百分点——处于该语料库中随机种子导致的波动范围内（§6.1.1）——因此我们在三个随机种子下重新训练了每个单元格，以下每个区间均通过说话人 $\times$ 随机种子自助法计算得出。

缩放。 早期版本对**所有** 秩都采用 LoRA α 默认值 128，因此有效缩放α/r 并未**保持** 一致：高容量组中， `upper` $\mathrm{r}{=}192$ 为 0.667， `all` $\mathrm{r}{=}64$ 为 2.000；低容量组中则分别为 2.000 和 6.095。因此，这些实验组同时存在三方面差异——容量所处的位置、容量在各块中的集中程度，以及每次更新的缩放强度。此处 $\alpha = 2$四个单元格中的 r 均相同，由此消除了缩放因素，仅保留位置和每块秩集中度这两项差异。

表19。 CDSD 上 Qwen3-ASR 的位置 $\times$ 容量对比；随机种子 42/43/44 下的均值 ± 标准差，每个单元格中的 r 均相同。相同 5,388 条留出语句（来自始终使用的相同 6 位未见说话人）的字符错误率（%）。两个 12.58 M 单元格的参数完全匹配；两个~4.2 M 单元格的参数匹配度为 98.4%。最佳结果以**粗体**标示。

位置	模块	Rank	α	可训练参数	CER
基线（不进行适配）	—	—	—	0	32.06
all	1–24	64	128	12,582,912	31.19 ± 0.37
all	1–24	21	42	4,128,768	30.66 ± 0.49
upper	17–24	64	128	4,194,304	29.88 ± 0.36
upper	**17–24**	**192**	**384**	**12,582,912**	**29.90 ± 0.05**

表20。 说话人 × 在 α匹配的 2×2上进行自助法，B = 10,000——这是本论文其他每个CDSD对比（表12、16和18）所使用的相同估计量。“种子标准差”是该对比在三个随机种子之间的标准差。负 Δ 值有利于名称排在前面的变体。

对比	它所隔离的内容	Δ (pp)	种子标准差	95%置信区间	判定
upper r=64 − all r=64	原始对比，容量未受控	**−1.31**	0.16	[−2.16, −0.25]	**已解决**
upper r=192 − all r=64	位置于一个**精确的** 12.58 M	−1.28	0.37	[−2.36, **+0.08**]	**跨越零点**
upper r=64 − all r=21	匹配位置于 a ~4.2 M	−0.78	0.29	[−1.55, **+0.39**]	**跨越零点**
upper r=192 − upper r=64	匹配容量，位置固定为 upper	+0.03	0.38	[−1.33, **+0.73**]	跨越零点
allr=64 − all r=21	容量，位置固定为 all	+0.53	0.14	[−0.17, **+1。14**]	跨越零点

表 21. 低容量、参数匹配对的每位说话人的字符错误率（%），三随机种子均值。 Δ 为 upper r=64 减去 all r=21；负值有利于 upper。

说话人	upper r=64	all r=21	Δ
CDSD_01	9.02	10.39	−1.37
CDSD_03	44.32	42.85	**+1.47**
CDSD_06	85.30	86.24	−0.94
CDSD_08	51.42	51.59	−0.17
CDSD_29	20.03	21.26	−1.23
CDSD_38	13.50	14.49	−0.99

2×2所确立的内容。

1. **两种预算匹配的放置对比均未得到解决**。在参数精确匹配为 12,582,912 时， upper r $=192$ 领先 all r$=64$1.28 个百分点，置信区间为 $[-2.36, +0.08]$ ——该区间仅高于零八百分之一个百分点，因此不能排除零。在 ~4.2 M 的匹配下， upper r$=64$ − all r$=21$为 -0.78 个百分点，置信区间为 $[-1.55, +0.39]$。两者指向相同方向；在六名说话人和三个随机种子下，均无法区分。**预算匹配对照在本论文的早期版本中被报告为已解决（−1.55 个百分点，置信区间 $[-2.41, -0.57]$），但事实并非如此**。表 21 展示了低容量对的逐说话人拆分：六名说话人中有五名偏向 upper，但 CDSD_03 反向相差 1.47 个百分点，而 CDSD _08 持平于 −0.17 个百分点；在六人的汇总样本中，这足以使区间扩大。

2. **造成问题的是随机种子，而非缩放**。 这一点值得单独说明，因为两项校正是同时进行的。匹配 α/r使每个受影响的单元格变化了约十分之一个百分点—— upper r$=192$ 在随机种子42下从30.00变为29.88， all r$=21$从31.17变为31.08——而在该随机种子下，匹配后的对比更大（-1.67 对比 -1.55），而非匹配的对比则较小。因此，缩放不匹配确实是一个方法上的缺陷，但在此处几乎没有实际影响。真正拉宽区间的是训练随机性：同一对比在三个随机种子下分别为 -1.67、 -0.94和 -1.24；对于1.28个百分点的效应，种子标准差为0.37。与§6.1.1中的情况一样，这种幅度的频带差异处于随机种子所能造成的波动范围内。

3. **无论处于哪种位置，增加容量都毫无收益，甚至可能带来代价**。 在位置保持为 upper 时，将秩从64增至192，结果为 $+0.03$个百分点，置信区间为 $[-1.33, +0.73]$；在位置保持为 all 时，将秩从21增至64，结果为 $+0.53$个百分点，置信区间为 $[-0.17, +1.14]$。两者都未能消除差异，也都不支持更大的预算。因此， all 并非因为容量不足而表现不佳，而 upper 也没有从三倍的秩中获益——这是论文中对参数效率解读最有力的支持，也是本小节唯一强化而非削弱的结论。还要注意， upper r$=192$ 是此处所有测量单元中随机种子稳定性最高的单元（标准差为0.046）。

4. **等秩对比是唯一得到解决的对比**。 upper r$=64$ − all r$=64$ 为 -1.31 个百分点，置信区间为 $[-2.16, -0.25]$ ——与 §6.1 中使用的是同一组对照、随机种子和估计量，也是论文首项贡献所依据的对比。本小节中的任何内容都不会影响它。

因此，本小节支持的论断范围比原先更窄：**可训练参数总量并不能解释 upper 的优势——参数量为 4.19 M 的 upper 并不逊于参数量为 12.58 M 的 all ，增加容量对这两种放置方式也都没有帮助——但在匹配预算下，放置方式本身的对比在当前样本量下无法得到明确结论**。已得到解决的放置论断形式，是 §6.1 中的等秩论断。两者都是 CDSD ×Qwen3-ASR 结果，在另外三个单元格中均未测试（§6.9）。

6.8 第四个单元格：Whisper-small 在英语孤立词上的表现

第6.5节在架构固定的情况下改变语料库；§6.6 则在语料库固定的情况下改变架构。缺失的单元格——UASpeech 上的 Whisper-small——将决定 §6.6 中的位置差异究竟是两个模型本身的特性，还是 CDSD 的特性。此次扫描与其他所有单元格采用**相同的** UASpeech清单、说话人划分、音频完整性门控、LoRA 配置方案、覆盖 12 个模块的频带定义、解码路径、评分方式和随机种子规范：每种适配条件使用三个随机种子（42/43/44），并使用来自同一 5 位未见说话人的相同 3,798 条测试语句。

指标的选择，以及为什么在此处它并不只是呈现问题。UASpeech 话语都是单词，因此 WER 在一个单位为词元的任务上统计整词错误，而 CER 则衡量平均只有几个字符的字符串上的子词编辑距离。因此我们

将**WER 作为主要指标，CER 作为次要指标**。在 6.6.1 的校正解码下，这两个指标对频带排序达成一致。在未校正解码下，它们将 `middle` 和`upper` 排成相反的顺序，而这种分歧在校正后不复存在——它是膨胀条件的特性，而不是指标本身的特性。现在，两个指标都将 `upper` 排在`middle`之前，并且对于这两个指标而言，该对均未得到解决（表23）。

表22。 Whisper-small 在 UASpeech 上的表现：来自 5 位未见说话人的 3,798 条留出语句。适配行的数值为三个随机种子（42/43/44）的均值 ± 标准差；未适配基线不涉及训练，仅进行一次评估。各行按词错误率排序，该指标是此语料库的主要指标。最佳结果以**粗体**显示。

变体	模块	词错误率（均值 ± 标准差）	字符错误率（均值 ± 标准差）
基线（不进行适配）	—	118.78（单次运行）	59.09（单次运行）
`all`	1–12	**59.71 ± 0.53**	**40.41 ± 0.26**
`upper`	9–12	65.58 ± 0.20	44.42 ± 0.08
`middle`	5–8	72.37 ± 2.29	45.95 ± 0.86
`lower`	1–4	88.92 ± 1.45	51.34 ± 0.39

表23。 Whisper-small × UASpeech 对比。配对说话人 × 随机种子自助法，B = 10,000，对 5 名测试说话人和 3 个随机种子进行联合重采样；在合并之前，先在每个随机种子内形成对比，因此某个随机种子的总体难度会相互抵消。这与表15使用的是同一个估计量。负值 Δ 有利于第一个变体。在该语料库中，WER 是主要指标；字符错误率区块为次要指标。

指标	对比	Δ（百分点）	95%置信区间	判定
WER	`upper − all`	**+5.87**	[+3.17, +8.76]	已解决，倾向于 `all`
WER	`all − middle`	**−12.66**	[−22.49, −3.85]	已解决，倾向于 `all`
WER	`upper − middle`	−6.79	[−16.59, **+1。92**]	**跨越零点**
WER	`upper − lower`	**−23.34**	[−33.99, −11.64]	已解决
WER	`all − 基线`	**−59.07**	[−84.80, −33.01]	已解决
CER	`upper − all`	**+4.02**	[+2.36, +6.30]	已解决，倾向于 `all`
CER	`all − middle`	**−5.54**	[−9.97, −1.66]	已解决，更倾向于 `all`
CER	`upper − middle`	−1.53	[−5.54, **+1。80**]	**跨越零点**
CER	`lower −基线`	**−7.75**	[−14.82, −0.60]	已解决

该单元格表明了什么。

1. `all` **是 Whisper-small 在 UASpeech 上的最佳频带，且在两项指标上的优势均已确定 —** `upper − all` = +5.87 个百分点的词错误率，置信区间为 [+3.17, +8.76]，；以及 +4.02 个百分点的字符错误率，置信区间为 [+2.36, +6.30] ——全编码器适配的表现也优于 `middle` ，且两项指标的区间均不包含零。这在第二个语料库上重现了§6.6针对该架构在 CDSD 上得出的结论。

2. **适配值得进行，而浅层适配的收益较低**。 `all`相较于未适配模型的词错误率提高了59.07个百分点，置信区间为 [−84.80, −33.01], 而 `lower` 使

模型的词错误率保持在88.92，仅恢复了可获得改进的一半。其相较于基线的字符错误率对比为 -7.75，置信区间为 $[-14.82, -0.60]$, 不包含零： `lower` 确实对该模型有所帮助，但幅度不大。**这推翻了本节早期版本中的解读**；早期版本发现相同对比跨越零点，并据此得出结论： `lower` 在次要指标上完全没有得到有助于改进的证据；这一变化源于6.6.1中的解码校正，该校正同时消除了被夸大的基线及其伴随的大方差。尽管如此， `lower`在性质上仍与Qwen3-ASR不同，后者会受到明显的损害（在UASpeech上，相较于基线的字符错误率为$+9.90$个百分点，已解决）。在四个单元格中的三个单元格里， `lower` 都是四个频带中表现最差的；例外是CDSD上的Whisper-small，其中 `upper` 以56.95的数值劣于 `lower`的56.28（表17）。浅层适配是否比不进行适配更糟，取决于模型。

3. **`middle` 与 `upper` 在两个指标上都尚未解决，尽管如今二者对方向的判断已一致**。按 WER 计算， `upper` 领先 6.79 个百分点（CI $[-16.59$、 $+1.92]$）；按 CER 计算则领先 1.53 个百分点（CI $[-5.54$、 $+1.80]$）。两个区间都跨越零，因此在这个样本量下，无论从哪个方向看，这一对仍无法区分——但两个指标不再指向相反方向，不像 6.6.1 的解码校正之前那样。

4. **随机种子方差取决于频带，在排序尚未解决的频带中最大**。`all` 和 `upper` 的 WER 随机种子标准差分别为 0.53 和 0.20 个百分点，而 `middle`为 2.29 个百分点——这说明单次运行频带扫描普遍存在问题，因为单次运行频带扫描无法估计高方差频带中的某个随机种子可能偏离该频带均值多远。§6.1 和 §6.6 的 CDSD扫描实验本身也已不再是单次运行（§6.1.1）。这种离散程度远小于未校正解码下的情况（`lower` 10.60、 `middle` 6.62）： 6.6.1 中的伪影同时夸大了随机种子方差和水平，因此这些标准差高估了真实的随机种子敏感性。

运行此单元格是为了进行的比较。在 CDSD 上， `upper` $-$ `all` 在两种架构上得出的方向<甚至相反>（§6.6）。在 UASpeech 上则不是这样：Whisper-small 的结果倾向于 `all`，Qwen3-ASR 的点估计也倾向于 `all` ，但其区间跨越零点（§6.5，表 16）。**在此语料库上，两种架构对位置的判断并无分歧**。

6.9 模型 × 语料库设计汇总

四个单元格全部运行后， `upper` $-$ `all` 对比——跨模型位置主张所依据的对比——便可从同一网格中直接读出。为确保可比性，全文均使用字符错误率；Whisper × UASpeech 单元格在词错误率下也得出相同结果（表23）。负值表示倾向于 `upper`，正值表示倾向于 `all`。

表17。 `upper` $-$ `all` 在模型 × 语料库设计下的字符错误率对比（百分点），以及配对自助法 95% 置信区间。所有单元格均采用相同的 LoRA 配置方案、按频带比例定义、完整性门控和配对自助法估计量。 **全部四个单元格均为三随机种子均值，采用说话人 ×随机种子区间**。UASpeech 单元格均为三随机种子均值。

语料库	Qwen3-ASR（24 个模块）	Whisper-small（12 个模块）
CDSD（普通话，朗读句子）	-1.31 $[-2.16, -0.25]$ — 仅在微平均字符错误率下已解决（§6.10）	**$+3.40$ $[+1.84, +5.03]$— 在所有 估计量中均已解决， `all` 胜出**
UASpeech （英语，孤立词）	$+1.66$ $[-0.51, +4.46]$— **未解决**，点估计偏好 `all`	$+4.02$ $[+2.36$， $+6.30]$— 已解决， **`all` 获胜**

这种反转仅出现在 CDSD 上。 贡献1所依据的跨模型位置差异——一种架构的结果支持 upper，另一种支持 all，且语料库和实验方案相同——出现在 CDSD 行中，而在 UASpeech 行中不存在；在 UASpeech 上，没有任何单元格支持 upper。**第四个单元格并未将这一主张扩展到第二个语料库；它表明两个模型在该语料库上得出了相同答案。** Whisper-small 对all 的偏好在两个语料库中都很稳定；Qwen3-ASR 的偏好则依赖语料库：在 CDSD 上已得到明确结果，在 UASpeech 上则未能确定，且其点估计的符号发生反转（表24）。三个等预算区间之间的深度排序—— upper 最好， lower 最差——在四个单元格中的三个成立；在 Qwen3-ASR × CDSD 上，其中的 upper/middle 这一步与零无法区分（§6.1.1）；而在 Whisper-small × CDSD 上，排序发生反转， upper 在四者中排最后（表17）。无法成立的是将 §6.6 解读为两种架构通常偏好不同区间。

这个混杂因素并未被该单元格消除，反而更加凸显。Whisper-small 未进行适配的 UASpeech基线为 **118.78 词错误率** ——由于它预期的是连续语音，而参考文本是单个词语，因此其输出的词数仍多于参考文本中的词数——相比之下，Qwen3- 自动语音识别在相同话语上的字符错误率为 41.95。因此，在 UASpeech 上的能力差距比在 CDSD 上更大，所以仅从 CDSD 的图来看，第9节局限16 甚至有所低估。解读词错误率数值时需注意两点：首先，它取代了在 6.6.1 的未校正解码下得到的早期数值 250.80；其次，对于单词参考文本，只需多输出一个无关词，词错误率就可能超过 100%，因此，与连续语音中的同一数值相比，它是较弱的退化信号。相应的字符错误率为 59.09。

6.9.1 两个模型实际上存在差异吗？一项直接的交互作用检验

§6.6 报告了 upper − all 在 Qwen3-ASR 上得出负向结果、在 Whisper-small 上得出正向结果，而论文的首要贡献是指出这两个模型存在差异。**分别得出方向相反且已解决的两个区间，并不能检验这种差异**，因此我们直接进行检验。两个模型都在同一组六位留出说话人上进行了评估，因此可以在单次自助法重复样本中计算差中之差：联合重采样说话人，为每个模型抽取一个随机种子，且说话人难度在差值两侧均相互抵消

$$\mathrm{DoD} = (\texttt{upper} - \texttt{all})_{\text{Qwen3-ASR}} - (\texttt{upper} - \texttt{all})_{\text{Whisper-small}}$$

在单个自助法重复样本内形成：说话人被联合重复抽样，每个模型抽取一个随机种子，而说话人难度在两侧相互抵消。

表25。 CDSD 上的交互作用，每个模型使用三个随机种子，$\mathrm{B} = 10{,}000$。

估计量	Qwen upper−all	Whisper upper−all	交互作用	95%置信区间	判定
微平均字符错误率	−1.31	+3.40	**−4.71**	[−6.17, −2.89]	**已解决**
宏平均字符错误率	−0.98	+3.63	**−4.60**	[−5.94, −3.08]	**已解决**
精确符号翻转	—	—	—	p值 = **0.031**	**已解决**

交互作用是本文中最稳健的量。 在微平均字符错误率、说话人等权加权以及无分布精确检验下，它都能得到确认；在六名说话人的情况下，其未调整的 p 值处于可达到下限——每位说话人的交互项均为负值（−6.74、 −3.71、 −5.20、 −1.70、 −4.12、 −6.16）。**因此，我们将其视为主要对比。**这是一次有意的校正：早期版本将 upper − middle 提名为主要对比，随后却以 upper − all作为标题结论，这一顺序是错误的。交互作用才是贡献1实际提出的主张；不同于单独考察任一主效应，它不取决于说话人的加权方式（第6.10节）。

两个主效应仍然值得报告，而且它们以不同方式较弱：Whisper 的主效应在所有情况下都能得到确认，而 Qwen 的主效应仅在微平均字符错误率下得到确认。 **在所有估计量下都成立的主张是比较性的——在该语料库中，这两个编码器偏好的频带不同——而不是孤立地衡量任一模型的偏好**。

6.10 这些结果在多大程度上取决于估计量？微观、宏观与精确检验

迄今报告的每个区间都来自对六名 CDSD 或五名 UASpeech 说话人进行的百分位自助法。在如此少的聚类数下，百分位尾部覆盖率无法得到保证，因此论文的结论不应取决于某个估计量在其最不可靠的适用情形下的表现。因此，我们对每项主要对比都进行了两项检验。

第一项检验改变了**被估计量**。我们的自助法会汇总重采样说话人的编辑计数（**微观** 字符错误率），因此说话人所贡献的参考文本越多，权重就越大。另一种做法是让每位说话人权重相同（**宏观** 字符错误率）。两种做法都有其合理性：微观指标是该领域的标准自动语音识别指标，而宏观指标则对应于关于跨说话人泛化的主张所隐含指涉的量。

第二项检验不再假设分布具有特定形式，但并非不作任何假设。对于 $\mathrm{n} \leq 6$ ，配对随机化分布可以**精确枚举**——CDSD 有 $2^6 = 64$ 种符号赋值，UASpeech 有 $2^5 = 32$ 种——因此，除了区间，我们还报告精确的双侧符号翻转 p 值。它以**零假设下的符号对称性**取代参数族假设：每位说话人的配对差值与其相反数可交换；这对应于该说话人不存在位置效应的零假设。这种检验不依赖分布，但并非没有假设。此处报告的所有精确 p 值均**未经调整**；经过三假设族的 Holm 校正后，没有任何对比仍显著（最低值为 0.094）。这一点会在每处报告中说明。

表26。 论文主要对比的三种估计量。

Cell	对比	微观	宏观	精确 p 值
Whisper × CDSD	`all − upper`	$-\mathbf{3.40}$ 已解决	$-\mathbf{3.63}$ 已解决	$\mathbf{0.031}$
Qwen3-ASR × CDSD	`upper − all`	$-\mathbf{1.31}$ 已解决	-0.98 **跨度 zero**	0.156
Qwen3-ASR × CDSD	`upper − lower`	$-\mathbf{5.64}$ 已解决	$-\mathbf{5.32}$ 已解决	$\mathbf{0.031}$
Qwen3-ASR × CDSD	`upper − middle`	-0.83跨越零点	-0.63跨越零点	0.438

这四项中有三项在所有估计量下结论一致。有一项不一致，而它正是论文标题所强调的放置对比。 `upper − all` 在 Qwen3-ASR × CDSD 在微平均字符错误率下已解决，但在说话人等权加权下未解决；精确检验与宏观结果一致。原因可从逐说话人差异中看出：六名说话人中有五名更倾向于 `upper`，而 CDSD_03 以 1.12 个百分点的差异更倾向于 `all` ； 此外，CDSD_03 提供的参考文本太少，因此微平均会降低其权重。

因此，我们在陈述该结论时附带说明其依赖条件。在标准微观平均指标下，Qwen3-ASR 在 CDSD 上更偏好`upper` 而非 `all` ； 但这种偏好**在对说话人进行等权重处理时并不稳健**，鉴于六名说话人中有一人持不同意见，我们不将其称为在说话人普遍意义上已经确立的结论。不受影响的是跨模型比较的另一半：在同一语料库上，Whisper-small 对 `all` 的偏好在每一种估计量下都得到了解决，并且所有说话人的结果一致。

关于多重性的一点说明：显而易见的校正方法在这里并无参考价值。指定三个假设作为主要假设族，并对精确 p 值应用 Holm 校正

结果没有任何显著发现——包括一致的 5.6 个百分点 upper − lower 对比。这是结构性假象，而非一项发现：由于只有六名说话人，精确 p 值的最小可达值为 $2/2^6 = 0.031$，因此对三个假设进行 Holm 校正时，p值下限为 0.094；在看到任何数据之前，**就无法**达到 0.05。我们报告精确 p 值及其下限，而不报告一个由研究设计而非测量结果决定的校正后判定。坦诚地说，限制本研究的是六名说话人这一约束（§9，局限性 2），而不是效应不存在。

---

7. 浅层适配的一种失效模式

7.1 该模式

两名说话人造成了 lower的总体损害，而六人中有四人在其作用下实际上略有改善：

- **CDSD_08**：39.42 → 51.42（upper）、48.90（middle） → **59.06** （lower），三随机种子均值（表11）。每个变体都会使该说话人的表现变差，而 lower 是迄今为止最差的，但恶化程度并不随深度单调变化： middle 在此处低于 upper。
- **CDSD_38**：16.87 → 13.50（upper）、15.94（middle）、14.06（all） → **25.26**（lower），三随机种子均值（表11）。一个新的 失效模式：浅层适配会破坏一名说话人的表现，而深层适配却能很好地处理该说话人。

这表明，扰动最接近信号的模块，恰恰会破坏声学特征异常的说话人所依赖的低层表示；相比之下，上部模块可以重新调整目标，而不会破坏声学前端。以数据所支持的力度表述：**不同说话人对浅层适配的易感性存在异质性，而这里两个显著的失效案例正是声学特征异常的说话人**。由于有两名说话人出现失效，其中最严重的一名还是其人口统计学单元格中的唯一成员，因此这只是一个假设生成结果——该队列规模太小，无法确定哪些声学或人口统计学属性能够预测易感性，我们也不声称已经孤立出其中某一项属性。

7.2 CDSD_08 为何失败：语言先验接管

CDSD_08 是唯一一位来自训练分布中最稀疏单元格的测试说话人。**表 27**。 按（性别 × 严重程度）划分的训练集构成，共有来自 31 位说话人的 33,617 个话语。

Cell	话语	训练集占比	说话人
女性 + 重度	15,770	46.9%	14
女性 + 中度	8,976	26.7%	8
男性 + 重度	4,754	14.1%	5
男性 + 中度 2,088		**6.2%**	**2**
男性 + 轻度	1,022	3.0%	1
女性 + 轻度	1,007	3.0%	1

按话语计，训练集女性占76.6%。CDSD_01/29/38为女性+中度（占训练集的26.7%），且全部有所改善；CDSD_06 为女性+重度（46.9%），有所改善；CDSD_03 为男性+重度（14.1%），有所改善。CDSD_08 是6.2%单元格中唯一的测试说话人，也是唯一出现回归的说话人。

这种回归是广泛存在的，而不是由少数离群值造成的：在 upper下，**CDSD_08的话语中有43%变差，只有15%有所改善**（均值为 Δ +0.109 CER/话语）。代表性充分的CDSD_29 则呈现相反情况——36%有所改善，10%变差，均值为 Δ −0.089。

在最严重的案例中，这种失效模式一目了然：未适配模型转录正确，而已适配模型却替换成流畅、符合语料库典型特征的文本。

表28。 CDSD_08 under upper的代表性失误，**以描述而非引文呈现**。语料库许可禁止再分发，且声明指出不会发布转录文本或逐话语预测，因此这里仅对参考文本和假设字符串加以描述，而不予复现。正确性判断和长度均为我们自行测量所得。

参考文本（描述）	base	upper
5个字符的短语：“and real estate”	转录正确 ✓	替换为一个无关的四字公司名称
关于……的15字句子 基础设施投资	正确转录 ✓	替换为一个17字的一串互不相关的听起来像官方头衔的称谓
12字符句子 , an 在一月逮捕一名所述的 year	正确，但插入了一个数字 ≈✓	替代一个16字符通讯社电头，名称为一份全国性报纸

最后一个假设是典型的新闻通讯社开篇套语。CDSD 使用新闻文本进行读取，因此六个训练轮次的适配会在新闻措辞上建立强烈的先验；当已适配的声学前端无法很好地适应某位说话人时，解码器会退回到该先验，并输出与音频无关但看似合理的文本。这种退化在性质上不同于错误率升高：输出自信、流畅，却是错误的。这一机制并非该模型所特有——Glazer 等人 [4] 定位到上下文偏置产生于自动语音识别编码器内部并覆盖声学证据，而 Kolluri 等人 [13] 发现，LoRA 通过依赖语言模型先验来补偿编码器容量的降低。

这是替换错误，而不是过度生成错误，两者可以测量加以区分。由于“幻觉”通常通过输出长度来诊断，我们进行了检查：对于**每一位**说话人、**每个**频带，长度比|hyp|/|ref|均保持在 0.92–1.03 以内，包括 lower 以及 CDSD_08。（本句的早期版本还报告了 0.000 的语料级插入率。该数值无法根据已发布的汇总数据重建——这些数据提供的是编辑总数，而非插入、删除和替换计数——并且它与高于 1.0 的长度比不一致，因此现予撤回。长度比界限才是数据所支持的论断。）已适配模型输出的文本量大致正确；它输出的是错误的文本。这清楚地区分了这种错误与第6.6.1节中的 Whisper 解码伪影，后者属于过度生成，也排除了这样一种过于简单的解读，即浅层适配只是让解码器无休止地运行。**我们能够陈述的是，改变编码器早期表示会使冻结解码器的输出趋向于由其先验主导；解码器自身的权重从未改变，因此我们不对语言模型本身作出任何主张**。诊断指标：scripts/language_prior_diagnostics.py。

---

8. 讨论

测试的三种探测器摘要无法预测适配器的位置。这是该论文的核心负面结果，而限定语至关重要：失效的是从这些深度统计量映射到某个频带，而不是探测本身（§9，局限性 14）。一种自然且被广泛假定的推断是——测量语音信息在编码器中成熟的位置，然后将

适配器置于该处——但这一推断有三种合理的解读。在同一流程下测量两种架构后，三**种解读全部失效**。

解读*1*——在 *PER* 改进幅度最大的地方进行适配。这最接近先前研究本身的表述（“音素识别在中部层的改进幅度最大……因此这些层应是最有效的目标”）。将这一原则应用于 Qwen3-ASR 时，结果很糟：改进幅度最大的是模块 1–8，PER 下降了 16.97 个百分点（表4），而该频带是最差的目标——它是唯一一个适配后模型表现还不如完全不适配的目标。 upper频带的斜率最平缓，效果却最好。

解读*2*——在区段平均 *PER* 最低的地方进行适配。这是三种方法中效果最好的一种，且**作为从四个可用目标中选出最佳目标的规则，它在四种情况下有一种成立**。在 CDSD × Qwen3-ASR 上，该规则完全有效：区段平均 PER 的排序为 upper 35.20 < middle48.17 < lower 64.66，适配后字符错误率的排序为 upper 29.88 < middle 30.71 < lower 35.52，而且upper 也优于 all。在 CDSD × Whisper-small 上，区段平均 PER 给出的排序相同——upper 49.49 < middle 54.99 < lower 66.81——但适配后字符错误率的排序为 all 53.56 < middle55.16 < lower 56.28 < upper 56.95，因此 PER 最低的频带反而是最不适合适配的地方。在**UASpeech 上，该规则为两个模型都选出 upper 作为目标**——区段平均 PER 分别为 41.80 和 58.00（§5.6，表8）——而**两个模型都更倾向于 all** （表14和表22）。

一个较弱的结论仍然成立。仅考察三个等预算区间并忽略 all 时，区段平均 PER 与适配 CER 之间的秩对应关系对 Qwen3-ASR 的**两个** 语料库均成立；对 Whisper-small 的 CDSD 不成立；对 Whisper-small 的 UASpeech 则尚无定论（§6.8）。因此，探针曲线对于同一模型内等预算区间的相对 优劣仍有一定参考价值，但对于实践者实际面临的问题——包括将全块 LoRA 纳入选项——则毫无帮助。在两种架构中，频带排序均不随探针数量变化，因此无论结论仍成立还是不成立，都不是探针数量造成的假象。

解读 *3*——在拐点所在位置进行适配。有效区间是包含 PER 曲线拐点的区间。**我们用计算方式明确规定这一规则，使其能够在不依赖主观判断的情况下应用**：对于层 $\ell = 1, \ldots, L$ 上的曲线 $p(\ell)$，令 $\ell^* = \arg\min\, p(\ell)$，并将拐点定义为$\ell$满足 $p(\ell) - p(\ell^*) \le 0.05\,[p(1) - p(\ell^*)]$ 的最小 $\ell \le \ell^*$——即首次达到总可实现改善幅度 5% 以内的深度。包含该层的区间就是预测结果。

应用于已发布的曲线后，该规则在三个单元格中给出 upper ，在第四个单元格中给出 middle ：Qwen3-ASR × CDSD 在第20层出现拐点，Whisper-small ×CDSD 在第10层出现拐点，Qwen3-ASR × UASpeech 在第18层出现拐点，以及**Whisper-small × UASpeech 在第8层出现拐点，该层位于其编码器的 middle 第5–8频带**（表8）。因此，在 CDSD 上，Whisper-small 的拐点位于第10层，而不是其已发表曲线所暗示的第6层（§5.5）；这使其落在 upper 第9–12频带内——因此，该规则预测 upper 应在 Whisper-small 上胜出。它在四者中排名最差，而 all以已解决的优势胜出（§6.6）。

阈值是唯一的自由参数，其中四个预测有两个对该参数并不稳定。两个 CDSD 单元格在 2–10% 的每个阈值下都给出 upper 。两个 UASpeech 单元格在这一范围内发生变化：Whisper-small × UASpeech 在 2% 时给出 upper （第9层），在 5% 和 10% 时给出 middle（第8层）；Qwen3-ASR × UASpeech 在 2% 和 5% 时给出 upper （分别为第19层和第18层），在 10% 时给出 middle （第16层）。本段的早期版本曾声称预测在 2–10% 范围内保持不变；但根据已发布曲线重新计算四个预测后发现事实并非如此，这进一步说明，不应从探针读取频带，而应对其进行扫描。该规则由我们提出，并根据我们的 Qwen3-ASR 结果与先前研究已发表的 Whisper-small 曲线进行拟合；在唯一一个能够进行前瞻性测试而非拟合的架构上，该规则被证伪，因此我们撤回这一规则。

每条规则错得有多离谱？用选择遗憾，而非胜负标签来评估。如果仅根据启发式方法是否选中经验最小值来评分，就会把两种截然不同的错误视为相同：一种是在前两名统计上难以区分的单元格中，只比最佳频带差一点点；另一种则是选中了差出许多百分点的频带。因此，我们报告**遗憾**，即$R(h)$ = 字符错误率[所选频带的字符错误率与 h] − 字符错误率[最佳频带], 之间的差值，按单元格计算。

表29. 各启发式方法的选择遗憾（字符错误率百分点）。CDSD 单元格为三随机种子均值；UASpeech 单元格亦然。遗憾为 0.00 表示该启发式方法选出的频带就是经验最佳频带。

Cell	最佳频带	最大改进幅度 最大改进幅度	最低频带均值 *PER*	*knee*
Qwen3-ASR × CDSD	`upper`	`lower` **+5.64**	`upper` **0.00**	`upper` **0.00**
Qwen3-ASR × UASpeech	`all`	`lower` **+19.10**	`upper` +1.66	`upper` +1.66
Whisper-small × CDSD	`all`	`lower` +2.72	`upper` +3.40	`upper` +3.40
Whisper-small × UASpeech	`all`	`lower` **+10.94**	`upper` +4.02	`middle` +5.54
平均遗憾	—	**+9.60**	**+2.27**	**+2.65**

这让负面结果更加明确，而非有所弱化，也区分了论文此前评分相同的两种失败。最大改进幅度不仅是错了，而且**错得极其离谱**——它从未选中最佳频带，代价最高达19.10个百分点；这是因为PER曲线最陡峭的部分出现在前期上升阶段，而在四个单元格中，最早的频带有三个是最差的适配目标，在第四个中也排倒数第二（Whisper-small × CDSD，其中 `upper` 排最后，表17）。最低频带平均 *PER*在 `upper` 所有情况下都选中同一频带，因此只有在 `upper` 恰好胜出时才是正确的，其他情况均会出错，平均代价为2.27个百分点；拐点在三个单元格中与之结果一致，而在Whisper-small × UASpeech上选中`middle`，平均代价为2.65个百分点。**实践者若采用三种启发式方法中表现最佳的一种，平均而言，字符错误率仍会增加约2个百分点，最高可达4个百分点**；相比之下，扫描四次只需耗费几个GPU 小时，就能直接找到最佳频带。

三种规则均失效的更深层原因在于，这些深度剖面实际上并无差异。三种规则都试图用成熟深度差异来解释位置差异。然而，四个探针最小值都落在相对深度的 83% 至 92% 之间，没有一条曲线呈现已发表的比较所暗示的中层平台，而且四者对三个比例区间的排序完全相同（§5.6，表9）。**在 CDSD 上，两个探针最小值都位于编码器后段的模型，却偏好不同的适应带**，因此，无论是什么因素决定了适应带，都不是音素身份成熟的深度。完整的网格进一步凸显了这一点：位置会随模型 × 语料库设计而变化，探针曲线却不会随之变化，因此**这些特定的深度汇总指标无法证明其能够解释测试中表现最佳的频带为何不同**。我们有意止步于此。三种映射均告失败，并不能证明这种情况不可能：曲线在后段层上升、曲率、绝对 PER 和方差方面仍有差异，而且我们尚未测试针对这些特征学习得到的映射或非线性映射。我们的主张是，经过测试的成熟深度启发式方法无法迁移，而不是曲线的任何统计量都做不到这一点。一个显而易见的反驳是——与语料库匹配的探针或许能预测 UASpeech 上的位置——但测量结果已排除了这条路径。对于音素身份 PER，这条路径已被排除；对于本文未测试的诊断指标，则尚未排除（§9，局限性 14）。

这是探测研究文献多年来一直区分的一种语音适配情形，我们将其视为一个实例，而非一项发现。Hewitt 和 Liang [7] 表明，探针准确率在一定程度上反映的是探针本身的容量，而非表征的结构，并引入控制任务来区分二者；Ravichander 等人 [24] 直接表明，某种属性可以从表征中解码出来，却未必会被模型自身的计算过程所使用。因此，可解码性与功能相关性之间的概念鸿沟并非新鲜事。这里的新意在于，这一鸿沟在构音障碍自动语音识别的层选择性参数高效微调中有了一个具体且量化的实例：一种看似理应能够确定适配位置的诊断指标——它是在正确的模型、正确的语料库和正确的单元上测量的——却

仍未能在四个单元格中的三个里选出测试过的最佳频带。探测研究文献指出，这种失效是可能发生的；本文则报告了基于此作出放置决策所付出的代价。

两项对照设计上的失误导致了那项后来被撤回的主张，而且都很容易避免。首先，跨论文比较取代了受控比较：原始对比中 Whisper-small 的一半数据来自已发表图表，而非实际测量；当我们使用自己的流程进行测量后，结果并未复现（§5.5）。跨架构的深度比较，其可靠性取决于两项测量中较弱的一项。其次，研究将单一且规模不大的探针训练量视为充分条件：在 6,200 条训练语句时，两种架构的堆栈后段行为似乎差异显著；而在 33,617 条时，这一差异仅为 0.25 个百分点。**应在多个训练数据量下报告探针曲线，否则就不要比较不同架构的堆栈后段形状**。

哪些结论站得住脚，以及第四个单元格所限定的范围。在 CDSD 上，固定语料库和协议，仅更换模型就会改变胜出的频带（第6.6节）；在 Qwen3-ASR 一侧，预算匹配对照发现，无论采用哪种位置，增加额外容量都没有带来已得到验证的收益——不过，在匹配预算下，位置本身的对比在当前样本量下也未能得出明确结论，因此这并不表明容量无关紧要（§6.7）。若依据文献沿用频带，就会在 Qwen3-ASR 上选中 `middle` ——在四个选项中，该频带在三个随机种子下排名第二，且落后`upper` 0.83 个百分点——而最接近先前研究自身论证的选择，即最大改进规则，则会以明显差距选中最差的频带（§8，表29）。站不住脚的是从这一个语料库推广得出的结论。这否定了另一条途径，而非恢复第一条：实践者既不能从探针上读出频带，也不能从另一个语料库推知频带，因此必须针对手头的模型和语料库运行扫描来确定频带。

既然深度无法解释这一点，那么位置差异为何会出现。有两种候选解释与数据一致，但本文均未进行检验。第一种是**基线能力**：Whisper-small 在 CDSD 上的表现远弱于 Qwen3-ASR（字符错误率分别为 77.57 和 32.06），一个需要修复的问题多得多的模型，可能需要在所有模块中分散进行校正，而能力更强的模型只需在顶部进行调整—— `lower` 是最明确的证据：它损害了 Qwen3-ASR，却帮助了 Whisper-small。我们对这一解释的表述比未经校正的数据所允许的更为谨慎：经过 6.6.1 节的解码校正后，Whisper 基线在语料库层面已不再产生幻觉，因此其机制是“需要修复的内容更多”，而不是“必须先阻止其捏造”。第四个单元格使这一点更难检验：在两个模型对位置达成一致的语料库中，能力差距更宽，而不是在二者意见不一致的语料库中。第二种候选解释是**下游的因素**：Qwen3-ASR 编码器的上三分之一区间同时改善了音素身份并消除了跨条件差异，这正如投射到语言模型解码器中的层所表现的那样，Yegorova 等人 [32] 和 Pluth 等人 [22] 均观察到了这一点。仅凭两个模型实例，无法区分深度、帧率、预训练语料库、规模、解码器类型和基线能力。

与 DAMA 的关系。我们未能复现 Xiao等人[30] 针对多语言低资源适应报告的 U形适应性模式：在本研究中，早期频带并非第二个峰值，而是以明显差距成为最不适合进行适配的位置。这是在我们的研究设定中存在的矛盾，而非对该结论的否定，因为 DAMA 研究的是**语言**适配，而我们研究的是**领域**适配。一种合理的调和解释是，语言身份部分编码在早期声学语音层中，因此适配这些层有利于适应新语言；而构音障碍并不属于这些层能够吸收的变异类型（§7.1）。可以迁移的结论是：U形模式并非适用于所有适配目标。

实际意义，按置信度从高到低排列。

1. 对该编码器进行低秩适应以适配构音障碍语音时，应优先选择上三分之一区间，而非下部和中部。这个排序在两个语料库中都成立。至于单独使用 `upper` 还是 `all` 更好，结论则不能迁移：CDSD上 `upper` 表现更佳，而在UASpeech上，点估计倾向于 `all` ，但区间跨越零点（§6.9，表24）——因此

应将其表述为依赖语料库，而不是针对仅使用 upper的适配建议。该结论已解决的形式是等秩结论（-1.31个百分点，置信区间 $[-2.16,\ -0.25]$）。在匹配的参数预算下，两种放置对比都未能得出确定结论——在参数精确匹配为12,582,912时为 -1.28个百分点，置信区间 $[-2.36,\ +0.08]$；在 ~4.2 M时为 -0.78个百分点，置信区间 $[-1.55,\ +0.39]$ ——尽管两个点估计都倾向于同一方向，而且增加容量对任何一种位置都没有帮助（第6.7节）。

2. 不要以为浅层适配无害。在 Qwen3-ASR × UASpeech 上， lower的表现明显不如完全不适配（字符错误率高出 +9.90 个百分点，已解决）。这种影响**取决于模型，并非普遍存在**：在 Whisper-small 上，同一频带能带来小幅改善（UASpeech 字符错误率为 -7.75，置信区间为 $[-14.82, -0.60]$）。在四个单元格中的三个里， lower 是四个区间中表现最差的；例外是 Whisper-small × CDSD，在这种情况下， upper 的表现更差（56.95，而lower为 56.28，见表 17）。 lower 是否不如完全不适配，取决于模型。

3. 按架构**和语料库**确定频带，并**通过扫描而非探测**来确定。前半部分并非新观点——Pasad 等人 [21] 和 Beheshti 等人 [2] 提出了这一观点的一般形式，但该观点也存在争议 [14]。在固定参数预算下进行四区间扫描，成本约相当于四次适配运行（此处每次为 ~3 小时），并能直接回答这个问题；探测阶段成本更低，但无法回答这个问题，而且在现有语料库上进行探测也无法弥补这一不足（§5.6）。

4. 任何已发表的层推荐都应视为同时受限于**其对应的架构**和进行测量的语料库。在 Qwen3-ASR 上，采用已发表的频带推荐，在参数预算相同的情况下，字符错误率会增加 0.83 个百分点——在当前样本量下，这一差异尚无法确定（§6.1.1）；而在此处测试的语料库中，四个单元格里没有任何基于探测的规则能在一个以上的单元格中选出最佳频带（§8，表29）。

这并不意味着什么。这并不意味着逐层探测没有信息价值——§5.2、§5.3 和 §5.5 均报告了我们认可的探测结果，而数据效率不对称性本身也是一项探测结果。否定性结论的范围很明确：探针曲线无法确定自适应层段。这也不意味着 Qwen3-ASR 是比 Whisper 更好或更差的基础模型；本文并未进行这种比较。也不意味着对该编码器进行低秩适应能够提升普通话构音障碍语音识别：在这个包含六名说话人的测试集中，没有任何已适配变体与未适配模型存在显著差异。

---

9. 局限性

以下是具有约束作用的因素之所以在结论之前提出，是因为其中有几项限制了相关主张，否则这些主张读起来会显得比证据所支持的更强。

1. **随机种子的覆盖范围，以及目前仍属于单次运行的部分**。CDSD适配扫描不再仅使用单个随机种子：全部四个频带均已在随机种子42/43/44下重新训练，而本文中的每个CDSD区间现在都来自说话人 × 随机种子自助法（§6.1.1）。这一变更牺牲了一个结论—— upper − middle 不再能够解析——但第一项贡献所依托的对比仍然完好无损。仍属于单次运行的是

(a) **第5.1节的 CDSD 探针曲线**，因此无法断言任何探针曲线的最大值层，只能断言频带；模型 × 语料库设计中的每个适配单元现在都包含三个随机种子：Whisper-small × CDSD 扫描实验和第6.7 2×2 在早期版本中都只是单次运行，而现在不再是。加入这些随机种子后，两个结论发生了变化，第三个结论得到确认—— 2×2的预算匹配对比不再显示出已解决， upper− middle 在 Qwen3-ASR 上不再显示出已解决，而 Whisper-small 对 all

在 CDSD 上被证明是论文中最稳健的对比（第6.6节）。**我们现在会这样表述：单次运行频带扫描无论在哪个方向上，都不是支持低于2个百分点的论断的可靠依据：它可能夸大某种效应，也可能无法揭示真实效应的稳定程度**。在我们进行测量的所有情况下，随机种子敏感性都强烈依赖于频带——在 CDSD 上，随机种子标准差对于 `upper` 为 0.36 个百分点，对于 `middle` 为 0.31 个百分点，但对于 `lower` 为 1.15 个百分点；在 Whisper-small × UASpeech 上，对于`upper` 的词错误率为 0.20 个百分点，对于 `all` 为 0.53 个百分点，但对于 `middle` 为 2.29 个百分点——因此，高方差频带中的单次运行结果可能会显著偏离该频带的均值。**我们从自身经验中得出的总体谨慎原则是：约 1.5 个百分点的频带差异处于训练随机种子可能造成的变动范围内，不应根据单次运行报告为已解决**。

2. **CDSD上有六名测试说话人，UASpeech上有五名**。这仍然是对每一项适应性论断的唯一硬性约束。每次自助法重采样都从五人或六人的池中抽取样本，因此这些区间应被解读为提示性而非决定性结果；在这个样本量下，即使结果完美，也没有任何符号检验能够经受多重性校正（§6.10）。在两个语料库中进行重复实验可以缓解这一问题，但无法消除它。两个语料库中各变体边际置信区间的宽度均为30–54个百分点，我们不报告其中任何一个；在当前样本量下，只有配对对比具有可解释性，因为在一次重复抽样中，说话人难度会相互抵消。

3. `upper` − **基线在CDSD上不显著**，但在UASpeech上显著（−7.54个百分点，置信区间不包含零），其余三个适配与基线的对比也是如此。CDSD的结果主要由CDSD_08主导：该说话人在每种变体下的表现都出现回归，使其与未适配基线的每项对比差异被显著拉大，同时在频带对频带的对比中相互抵消。因此，“适配该编码器有帮助”这一点在UASpeech上得到确立，但在CDSD上并未得到确立；我们不合并两个语料库，因为它们在内容类型上的差异过大，无法支持合并估计。

4. **始终使用两个语料库和两种架构，既用于探针，也用于扫描——但两个语料库终究还是两个**。频带排序在模型 × 语料库设计的全部四个单元格中进行检验，探针结果也涵盖相同的四个单元格（§5.6）。UASpeech 探针使用词典目标且没有完整性门控，因此其绝对 PER 无法与 CDSD 曲线进行比较；两个语料库在关键单元上的规模都很小（分别有 6 名和 5 名测试说话人）；并且二者都不包含混合病因队列——UASpeech 以脑瘫为主，而 CDSD 未按说话人记录病因构成。

5. **跨模型的位置差异仅基于 CDSD；在 UASpeech 上，两种模型的结论一致**。这是论文第一项贡献所面临的约束。该`upper` − `all` 对比在 CDSD 的两种架构上呈相反方向（Qwen3-ASR 为 -1.31，置信区间为 $[-2.16、-0.25]$；Whisper-small 为 $+3.40$，置信区间为 $[+1.84、+5.03]$）——不过，只有 Whisper 部分经说话人等权加权和精确检验后仍然成立（§6.10），因此这种反转依赖于对 Qwen 部分的微平均解读。在 UASpeech 上，两种模型都没有偏向 `upper`：Whisper-small 的结论为 `all` （$+4.02$，置信区间为 $[+2.36、+6.30]$），而 Qwen3-ASR 完全未能对此作出结论（$+1.66$，置信区间为 $[-0.51、+4.46]$），但点估计位于同一侧（表24）。从推断角度看，两个语料库并不相互矛盾——但建立在反转之上的主张需要确实存在反转，而该现象在这里并不存在。

6. **跨模型比较在内部受到控制，但存在严重的混杂因素**。这已不再是跨论文比较——探针（§5.5）和频带扫描（§6.6、§6.8）均已在此处使用我们自己的语料库、数据划分、门控、实验方案和评分方式重新运行。剩下的差异在于，Whisper-small 和 Qwen3-ASR 在深度（12 对 24 个模块）和帧率（20 对 77 毫秒）和解码器类型和预训练数据和规模以及最重要的**未适配能力**方面均不相同。因此，我们声称位置在**这两个**

同一语料库上的两个模型之间存在差异，而不是说编码器架构导致了这种差异；能够确定目前主要解释的实验，是加入一个能力匹配的第三个模型（§8）。

7. **跨条件设计采用的是分布匹配，而非话语匹配**。 先前研究对每个话语匹配转录文本，从而在设计上消除了词汇混杂因素；我们只是在期望意义上消除这些因素，并通过一项 411行严格配对一致性检查加以验证。d–u 对比还混合了说话人集合差异与条件差异，因此论点所依据的是 d–r 对比。散度曲线未计算置信区间；论点在于曲线形状，而非数值大小。

8. **探针 C 的难度标签来自被探测模型所属的模型系列**。 先前研究使用更强的外部模型进行标注；我们没有更强且未受污染的 普通话解码器，因此难度被定义为“对通用自动语音识别模型而言较难”，而非独立的真实值。层级边界也并非简单地按三分位数划分，原因见 §4.3。

9. **人口统计偏差机制仅针对一位说话人得到验证，尚未确立为普遍规律**。 §7.2 展示了 CDSD_08 中的这一机制，并将失败量化为替换而非过度生成，但单个说话人不足以证明训练集中的代表性不足可以预测回归。本文的早期版本曾根据一项十种语言的合成分析提出这一观点；该分析所依据的数据，其来源和权利依据并未在本文中得到确立，因此我们删除了这项分析，而没有为其辩护。这个问题仍未解决。

10. **微调检查点的次级探针曲线受到污染**。该模型接触过81.9%的探针话语以及全部六名测试说话人。报告了其层的定位；其幅度不能被解释为泛化能力，且没有任何核心结论依赖于此。

11. **适配集比先前研究的小14.6%**，因为它继承了使探针与适配清单具有可比性所需的对齐完整性门槛。

12. **探针训练量保持匹配，但仅适用于Whisper-small和Qwen3-ASR，且每种架构各采用一种设置**。核心探针曲线最初使用6,200条话语进行训练，而先前研究使用 ≈39,000条；§5.5将两种架构都在33,617条话语上重新运行。在6,200条话语下的绝对音素错误率无法在不同研究之间进行比较；在低训练数据量下，曲线形状本身也对数据量敏感，并且这种敏感性会因深度和架构而异。UASpeech探针仅在单一数据量下运行（5,979条训练语句），因此带有完全相同的限制：在CDSD上改变训练数据量时，每条曲线发生转折的深度不受数据量影响（图3）；但我们尚未在UASpeech上改变数据量，因此其稳定性是推断得出的。

13. **该语料库中的严重程度标签是根据字符错误率得出的，而非临床评定**。 该语料库不提供临床严重程度评级，也不包含健康对照组，因此，表10和表13中的难度细分是按识别难度代理指标进行分层的，应据此解读。

14. **这一负面结果针对的是三条具体规则和一个信号，而不是所有可能的基于探针的预测器**。我们没有拟合或验证一个将探针曲线映射到最优频带的通用模型，也没有声称这样的模型不可能存在。我们所确立的是，文献中实际提出的三种解读，在两种内部控制架构中的至少一种上均告失败，而支撑这三种解读的成熟深度直觉在此处并不适用。该信号是**音素身份 PER**，除此之外别无其他；基于梯度统计、残差流结构或注意力熵构建的规则，原则上可能在这些规则失效之处取得成功。

15. **我们在测量第二种架构之前提出的膝点规则现已撤回，取而代之的负面结果建立在两种架构之上**。该规则是事后提出的，基于两个编码器拟合而成，并且其两个输入之一来自已发表图表，而非实际测量。在我们自行测量该输入后，发现它是错误的；此外，该规则对 Whisper-small 的前瞻性预测也被证伪（§6.6）。替代性主张建立在两项内部控制测量之上，但仍然是一个基于两种架构的结果：第三个编码器可能会重新支持某种能够适用于这三者的基于探针的规则拟合，但无法挽救此处已被反驳的具体规则。

16. **跨架构比较将基线能力与架构混为一谈，而且这种混杂因素在 UASpeech 上比在 CDSD 上更严重**。Whisper-small 在 CDSD 上明显弱于 Qwen3-ASR（字符错误率分别为 77.57 和 32.06）；在 UASpeech 上差距更大（词错误率 118.78，对比字符错误率 41.95）。因此，第6.6节中的位置差异已确认为真实差异，而非跨论文伪影，但我们并未将其归因于任何单一因素；基线能力是我们认为最可能、同时也是控制最不足的候选因素（§8）。由于差距最大的是两个模型一致的单元格，因此仅凭这四个单元格也无法得出将能力与位置联系起来的简单单调关系。

17. **并非每条探针曲线都包含三个随机种子**。 第5.5节中匹配数据量的 CDSD 重跑结果以及第5.6节中的两条 UASpeech 曲线都包含三个随机种子；第5.1节中原始的 6,200 条话语 CDSD 曲线仅运行了一次，因此不应将其中任何探针最小值点解读为可在层级分辨率上主张的结果。在存在三个随机种子的情况下，它们带有配对区间，但这些随机种子并不独立于解码过程。即使在固定 随机种子下重复运行，结果仍会相差 0.1–0.4 个百分点，因为 GPU 上的 CTC 训练并非位级可复现（涉及 cuDNN 算法选择和原子操作），因此这里的“随机种子”总体上界定的是优化方差，而不仅仅是初始化方差。即使有区间，单个最小值层仍无法作为可主张的结果：相邻的后期层之间存在大量重叠，这也是第5.1节和第5.5节报告区间和对比，而不是精确的最小值所在层的原因。

18. **未能复现先前研究中的中层平台，其原因尚未得到解释**。我们已经分别检查并排除了探针配置、填充处理和训练数据量（§5.5）。**基线能力尚未被排除**：与先前研究的 v2 图表相比，我们修正后的基线高出 12.4 个百分点，因此它仍是一个有待验证的候选解释（§5.5，第 2 项）。两条流水线之间还存在其他差异，而我们尚未找出这一差异。读者应相应谨慎解读我们的 Whisper-small 曲线：它在内部受控，并且在四种配置之间保持一致，但由于我们无法说明的原因，它与一项已发表的结果不一致。

19. **在两种预算下，位置/容量混杂因素都尚未得到解决，而且 2×2仅适用于 CDSD $\times$ Qwen3-ASR**。 §6.7 通过秩匹配参数量，并在四个单元格及三个随机种子之间保持$\alpha = 2r$不变。两种预算匹配的放置对比都无法排除零：在精确匹配 12.58 M 时为 -1.28 个百分点，置信区间为 $[-2.36, +0.08]$；在 $\sim$4.2 M 时为 -0.78 个百分点，置信区间为 $[-1.55, +0.39]$。六名测试说话人和三个随机种子无法在匹配预算下区分位置与每块秩集中度。 2×2 所确立的是一个具有否定性但很有用的结论：**增加容量并不会改善任一位置**（$+0.03$ 和 $+0.53$ 个百分点，二者均未得到解决），因此 `all` 并非因为参数不足而处于劣势。已经得到解决的位置主张是 §6.1 的等秩对比（-1.31 个百分点，置信区间为 $[-2.16, -0.25]$），而不是这一项。2×2 也仅在模型 $\times$语料库设计的一个单元格中运行。

10. 可复现性

所有实验均在一台工作站 GPU 上运行。探测阶段使用 8,800 条话语——包括特征提取、25 层 $\times$ 2 模型，以及用于两个探针的条件相似度分支和

难度解码——耗费的 GPU 时间不到两小时。§5.5 中的匹配规模重跑规模更大：对 47,687条语句清单进行 Whisper-small 特征提取，13 层共产生 107 GB 数据，扫描耗时 ~1.7 小时；Qwen3-ASR 分支的 25 层共产生 73 GB 数据（以 19.5 条话语/秒的速度运行 41 分钟），扫描耗时 ~55 分钟。UASpeech 探针的特征提取耗时分别为 Qwen3-ASR 484 秒、Whisper-small 551 秒，而探针本身约需**2.5 小时**；因此，在新语料库上复现表9的四单元格探针网格，用一台 GPU 花一个下午即可完成。Qwen3-ASR 的每个适配变体在 CDSD 上耗时 3.1–3.5 小时，在 UASpeech 上耗时 0.53–0.60 小时，因此推荐的四区间扫描在 CDSD 上约需 12–14 小时，在 UASpeech 上不到 2.5 小时；Whisper-small 的 ×UASpeech 单元格还要更省时，每个变体仅需 0.36–0.55 小时。参数匹配的 2×2 还需在 CDSD 上额外运行四次，每次耗时 3.2–5.1 小时。

发布的产物包含什么。 由于原始临床音频无法再分发，因此这里通过标识符、配置和衍生统计数据来实现可复现性，而非通过数据。仓库发布的内容包括：各语料库自己的语句键及其数据划分归属和清单指纹；每个随机种子；支撑本文**每个** 区间的说话人级和随机种子级分数；自助法、精确符号翻转和宏观估计器代码；以 CSV 而非仅以图表形式提供的探针曲线；LoRA 参数计数脚本；以及汇总结果表。

冻结的解码规范，因为解码并非此处的实现细节。 第6.6.1节表明，生成设置可使 Whisper 字符错误率变动数十个百分点，因此发布的产物以机器可读形式固定了以下内容：模型和检查点标识符、`transformers` / `peft` / `torch` / CUDA 版本、提示语言和任务标记、束宽度、`no_repeat_ngram_size`、按时长缩放的最大输出长度规则、停止条件、采样标志、文本规范化、CER/WER分词、检查点键重映射兼容层，以及小型回归测试清单的预期分数。最后一项最为重要：它是唯一能够发现环境悄然变化的手段，而本项目已经遇到过一种加载路径，它生成的模型看似正常运行，但检查点键重叠为**零**。

有两个特有的操作陷阱， 2×2，在弄清原因之前，它们都导致一次运行失败。首先，**必须强制在中央处理器上进行适配器合并**。 `peft`默认会先将适配器加载到 CUDA，再将其合并到基础权重中；在统一内存硬件上，将秩为192的适配器合并到一个含有1.7 B个参数的模型时，会耗尽共享内存池，导致进程被终止，而且没有任何错误信息指出合并是原因。先在中央处理器上加载基线和适配器，在那里完成合并并保存，然后再移至其他设备，即可避免此问题，但合并速度会较慢。其次，**训练和评估需要不同的环境**：训练器需要一个 `transformers` 版本，而封装推理路径无法接受该版本；评估器则需要一个能够加载此检查点、且不会悄悄以随机值初始化模型的版本（见下方警告）。因此，发布的脚本会明确指定解释器，而不是沿用当前激活的环境；在训练器的环境中运行评估器，就会产生看似合理、实则毫无意义的字符错误率。

关于统一内存硬件上复现实验的重要内存说明。 在本文使用的 GB10 上，主机和设备分配共享同一个内存池，而每层几十 GB 的特征加载会导致 CUDA 在模型创建时无法分配内存，即使它报告仍有可用内存。因此，发布的扫描脚本会**为每个层组启动一个进程** ，并在特征加载前预留一个 CUDA 内存块；如果使用单个进程扫描全部层，则会在第一次层转换时被终止，因为某一层的分配尚未完全释放，下一次加载便已开始。

图。 发布的 `scripts/make_figures.py` 可根据汇总 CSV 文件重新生成全部四幅图；既不需要 GPU，也不需要访问语料库。

代码和汇总结果发布于

`https://github.com/SMF-Labs/encoder-depth-dysarthric-asr`。该代码仓库按运行顺序收录了完整流程、支撑本文每张表格的结果文件，以及一份 README，介绍运行环境、评分规范和下文总结的实现陷阱

。**预印本发布时，代码仓库也会立即公开**——无需等待期刊接收——并同时提供冻结解码规范及其回归测试清单。公共代码仓库以单次提交的形式发布，提交为 `4ea79e2`，其文件树与经独立审计的开发分支头完全一致；开发历史则予以私下保留，因为早期的一些提交记录了本地文件系统路径。两者均不包含任何语料库内容。

需要特别提醒一个方法学问题，因为它决定了这些数字是否值得信赖。 此处所有解码均通过模型自带的封装推理路径完成。在一个包含369条话语、且记录了参考预测结果（字符错误率为41.87%）的重叠集上，通用的聊天模板加上`generate` 路径得分为**124.57%** ，而封装路径得分为**32.38%** ——这种巨大的差异足以使所有难度标签失真，却不会显示在任何错误消息中。使用任何新的推理路径前，都应先用已知结果进行验证。仓库中还记录了另一种类似的陷阱：在当前`transformers` 版本下加载此检查点，会悄无声息地得到一个随机初始化的模型。

数据可用性。 CDSD由其保管方根据2026年2月19日签署的学术许可提供，该许可仅允许用于研究，禁止向第二方提供音频，并要求致谢部分所列的署名和引用。UASpeech是根据语料库首席研究员于2026年2月6日出具的书面许可获取的；双方未另行签署文件。我们在访问申请中承诺仅将该语料库用于学术、非商业研究且不予再分发，并在获准访问时接受了这一条件。因此，音频、转录文本、强制对齐、提取的激活值、逐话语预测和探针检查点均保留在由基金会控制的计算环境中，不予发布；两个语料库均须向其保管方申请获取。

存储库确实包含的内容，严格按照第 §10 节所述：数据划分分配，以及带有清单指纹的**每个语料库自身的语句键**——对于 CDSD，格式为“说话人/子集/话语”；对于 UASpeech，格式为“说话人/块_词”。这些信息是复现者重建数据划分所需的内容；同时，它们在我们的机器上不包含音频、转录文本或路径，因此不会重新分发任何语料库内容——此外还包括每个随机种子、构成各区间基础的逐说话人和逐随机种子得分、自助法、精确符号翻转和宏观估计器代码、以 CSV 格式提供的探针曲线、参数计数脚本、带有回归测试清单的冻结解码规范，以及汇总结果表。不会发布适配后的模型权重。

11. 结论

针对构音障碍语音进行自动语音识别基础模型参数高效适配的层放置建议，既无法在不同编码器系列之间可靠迁移，也无法在不同语料库之间可靠迁移；而通常用于解释此类差异的深度剖面，也无法解释这些差异。

正半部分是局限于单个语料库的受控放置结果。在保持语料库、说话人划分、完整性门控、LoRA 配置方案和随机种子规范不变并更换模型后，CDSD 上的最佳频带发生了变化：在一个解码器为语言模型的 24层基础模型上，上层带表现最佳，并且仅使用三分之一的参数就优于适配整个编码器；不过，在匹配的参数预算下，这一对比已无法得到统计确认（§6.7）；在 12 层 Whisper-small 上，全块 LoRA 表现最佳，而上层带在四种方案中排名最末。**在模型的全部四个单元格中运行相同的对比 ×语料库设计表明，这一反转是 CDSD 所特有的**：在 UASpeech 上，两种模型均不偏好上层带（`upper − all` $=+4.02$ 个百分点 CER，CI 为 $[+2.36,\ +6.30]$ 对于 Whisper-small；对于 Qwen3-ASR，结果未得到确认，但点估计位于同一侧）。Whisper-small 对 `all` 的偏好在两个语料库上都成立；Qwen3-ASR 的`upper`与`all` 之间的结果在其观测到的点估计中依赖语料库：在 CDSD 上得到确认，而在 UASpeech 上符号反转且未得到确认（表24）；此外，语料库之间观测到的变化并不等同于统计上已确立的跨语料库差异，而本设计并未检验后者。在两个语料库上都成立的是 24层模型的三层带深度排序方向——在我们运行的每个单元格中，均为上层优于中部优于下部——以及

适配下部带确实不如不调整编码器这一发现 **在 Qwen3-ASR 上 × UASpeech** ——这是唯一一个该对比得到确认的单元格；CDSD 上的点估计同样有害，但未得到确认，而在 Whisper-small 上，下部带则有所帮助。在这一排序中， `upper` 高于 `lower` 在两个语料库上都得到确认； `upper`高于`middle` 这一步在 UASpeech 上得到确认，但在重新抽样训练随机种子后，在 CDSD 上未得到确认（ §6.1.1 ）。

负面结论涉及从探针推断位置。根据我们自己的流程测得（而非引用既有结果），Whisper-small 的音素错误最低点位于相对深度的 83%，而 Qwen3-ASR 的最低点位于 83–92%：**两种探针的最低点都出现在编码器后段，但在 CDSD 上，两个模型偏好的频带不同**。将探针曲线与频带联系起来的三条候选规则——在 PER 改善最显著处进行适配、在区段平均 PER 最低处进行适配、在拐点处进行适配——至少会在一种架构上失效；最后一条正是我们自己提出的调和方案。完整网格还表明，位置依赖语料库，**而探针曲线并非如此**：四个最低点都落在相对深度的 83% 至 92% 之间，频带排序完全相同，且任何位置都没有中层平台。近乎恒定的量无法选出变化的结果，因此，语料库匹配探针可能预测频带的这一路径，对于音素身份 PER 而言已被排除。

两项方法论结果解释了为何被否定的主张仍得以成立，而且二者的适用范围都超出本文。探针读出对训练数据量的敏感性取决于深度，且具有架构特异性——在固定训练轮次预算下，数据量与优化器步数无法区分（§5.5）——而且，在单一且适中的数据量下训练的探针，会夸大深层读出难度不同的架构之间的后段堆叠层差异——此处表面上的 6.56 对 1.85 个百分点差距，在 5.4× 处会缩小到 0.25 个百分点。跨论文深度比较的可靠性，则取决于两项测量中较弱的一项：在匹配配置、填充或训练数据量的条件下，我们无法复现针对 Whisper-small 报告的中层平台，对此我们报告为尚无解释；两个模型的基线能力并不匹配，也不能排除其作为影响因素（§5.5）。

有两项限制界定了结果的适用范围，我们选择如实报告，而不将其掩盖：编码器的帧周期使得在该架构上进行帧级边界探测缺乏适定性；浅层适配则会用流畅、符合语料库典型特征的文本取代转写，从而损害声学特征异常的说话者。

致谢

本研究使用中文构音障碍语音数据库（CDSD）。© **Su-Jing Wang**。 我们感谢CDSD的管理方根据其签署的许可协议提供访问权限；该协议要求注明此致谢语并引用Wan等人 [28]。我们感谢Mark Hasegawa-Johnson教授和伊利诺伊大学厄巴纳-香槟分校，他们于2026年2月6日以书面许可的形式授予我们访问UASpeech语料库的权限。我们感谢Leo Kaiya独立审核发布仓库，发现本地路径暴露问题，并验证无需访问受限语音语料库即可复现图和规范结果。

生成式人工智能使用声明

由于患有运动神经元病，第一作者完全依靠人工智能辅助眼动界面进行交流。在整个研究过程中，生成式人工智能（Claude，Anthropic）被用作辅助沟通工具，应用于实验设计、代码开发、数据分析、统计计算、图表生成和手稿撰写。所有科学决策、解释和结论均由作者作出。人工智能作为一种无障碍工具，使一位

重度肢体残障的研究人员能够开展计算研究。所有人工智能生成的内容均经作者审阅和核实后才予以呈现。

声明

利益冲突。 作者声明不存在利益冲突。

资助。 本研究由Scott-Morgan基金会资助。

伦理审批和二次数据使用。 本研究是对两个现有、已获许可的语音语料库进行的二次分析。研究未收集新的人体受试者数据，也未联系任何参与者。两个语料库均在其创建机构的伦理审批下收集——UASpeech获得了伊利诺伊大学机构审查委员会对项目可听与可见构音障碍语音学的视听描述与识别 （NIH NIDCD 5R21 DC008090）的审批——并且本文均按照各自记录的条款使用。对于CDSD，这些条款是数据可用性部分所述的已签署学术许可。对于UASpeech，这些条款包括语料库首席研究员的书面许可，以及数据保管方公布的分发条款：访问权限仅限于学术和政府研究实验室，不得用于商业用途，数据必须受到保护且不得再分发。

我们分别说明每个语料库所依据的基础，因为二者有所不同。 对于UASpeech，原始收集工作具有上述指定的机构审查委员会审批。对于CDSD，语料库论文记录的是**书面知情同意**，而非伦理委员会编号：Wan等人（2024年，第2.1节）报告称，每位说话人在录音前均签署了知情同意书；对于五名未满18岁的说话人，则取得了父母或监护人同意以及未成年人的同意；九名进行视频录制的说话人还签署了额外的同意书；研究人员告知说话人，他们可以随时退出；研究从未收集姓名，而是将每段录音与一个序列号对应。**该来源未引用任何机构审查委员会审批编号，我们也不声称存在此类编号**。 本文依据一份已签署的定制学术许可使用CDSD，该许可禁止再分发，并要求在致谢中注明© Su-Jing Wang。研究未进行新的参与者联系或数据收集，也未向作者所在机构另行申请二次使用认定。我们报告的是实际采取的做法，而非普遍规则：某一机构是否要求其自身进行二次使用或非人体受试者研究认定，取决于该机构的政策和司法管辖范围；读者不应据此推断，原始审批加上数据使用协议在任何情况下都足够。

本文中的两项发现不仅是统计结果，也是公平性结果；我们更愿意直截了当地说明这一点。首先，音频完整性门控在受损最严重层中以更高的比例排除话语（§4），因此，这里每个字符错误率所测量的总体都偏离了最需要这项技术的说话人；在此类评估上调优的系统，对他们而言可能看起来比实际表现更好。其次，浅层适配并不只是提高平均错误——它会将损害集中在声学特征异常的说话人身上（§7.2），而且具体表现为自信、流畅却错误的输出，而不是明显退化的转写。**对于无障碍技术而言，一种让均值有所改善、却灾难性地辜负最不典型说话人的适配选择，正是最需要关注的失效模式**，也是语料库级指标最难显现的失效模式。正因如此，我们始终报告逐说话人结果。

数据保护。 未尝试重新识别任何说话人。文中使用的说话人标识符均为语料库自身的化名代码。未公开任何音频、转录文本或逐话语预测（§10）。

CRediT作者贡献

Bernard Muller： 概念化、方法论、软件、形式分析、调查、数据整理、验证集、可视化、写作——初稿。

LaVonne Roberts： 监督、经费获取、项目管理、资源、撰写—审阅与编辑。

按首次出现顺序列出的引用文献：Muthu 等人 [18]；Hu 等人 [8]；Wan等人 [28]；Shi 等人 [25]；Huber 等人 [10]；Pasad 等人 [21]；Beheshti 等人 [2]；Lin 等人 [14]；Pasad 等人 [20]；Ashihara等人 [1]；Yue 等人 [33]；Mojarad 和 Tang [17]；Glazer 等人 [4]；Pluth 等人 [22]；Ma 等人 [15]；Xiao 等人 [30]；Ogawa 等人 [19]；Xu 等人 [31]；Kolluri 等人 [13]；Yegorova 等人[32]；Kim 等人 [12]；McAuliffe 等人 [16]；Graves 等人 [5]；Radford 等人 [23]。

参考文献

[1] Takanori Ashihara、Marc Delcroix、Takafumi Moriya、Kohei Matsuura、Taichi Asami 和 Yusuke Ijima。自监督语音和说话人模型学到了什么？来自跨模型逐层分析的新发现。载于*IEEE声学、语音与信号处理国际会议（ICASSP）*，第 10166–10170 页，2024年。doi：10.1109/ICASSP48485。2024。10446422。arXiv：2401.17632。

[2] Marjan Beheshti、Majid Rostami 和 Bo Chen。面向可泛化音频深度伪造检测的自监督语音模型探针引导层选择。arXiv：2606.30791，2026年。网址 `https://arxiv.org/abs/2606.30791`。v1，2026年6月29日。arXiv 评论指出，该手稿已投稿至 Computer Speech and Language。

[3] 卡内基梅隆大学语音研究组。CMU发音词典，0.7b版本。https://github.com/cmusphinx/cmudict，2014年。135,166条词目；带有词汇重音标记的ARPAbet转写，并已针对本文使用的音位清单进行删减。

[4] Neta Glazer、Yael Segal-Feldman、Hilit Segev、Aviv Shamsian、Asaf Buchnick、Gill Hetz、Ethan Fetaya、Joseph Keshet和Aviv Navon。《超越转录：自动语音识别中的机制可解释性》。载于*AAAI*人工智能会议论文集，第40卷，第37407–37416页，2026年。doi：10.1609/aaai.v40i44.41073。第40卷，第44期。AAAI-26人工智能对齐专题赛道，发表于2026年3月14日；ojs.aaai.org/index.php/AAAI/article/view/41073。预印本arXiv：2508.15882，其记录未包含期刊引用信息。

[5] Alex Graves、Santiago Fernández、Faustino Gomez 和 Jürgen Schmidhuber。连接主义时间分类：使用循环神经网络标注未分段序列数据。载于第*23*届国际机器学习会议论文集（*ICML*），第369–376页，2006年。doi：10.1145/1143844.1143891。

[6] Yi Han、Hang Chen、Jun Du、Chang-Qing Kong、Shifu Xiong 和 Jia Pan。面向低资源多语场景的大型自动语音识别模型的层自适应低秩适配。载于国际中文口语语言处理研讨会（*ISCSLP*），2024年。doi：10.1109/ISCSLP63861.2024.10800407。

[7] John Hewitt 和 Percy Liang。使用控制任务设计并解读探针。载于*2019*年自然语言处理实证方法会议和第*9*届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集，第2733–2743页，2019年。doi：10.18653/v1/D19-1275。

[8] Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。LoRA：大语言模型的低秩适应。载于国际学习表征会议（*ICLR*），2022年。arXiv：2106.09685v2。发表于 ICLR 2022；arXiv 记录未注明期刊参考文献。[9] Shujie Hu、Xurong Xie、Mengzhe Geng、Jiajun Deng、Huimeng Wang、Guinan Li、Chengxi Deng、Tianzi Wang、Mingyu Cui、Helen Meng 和 Xunying Liu。用于构音障碍语音识别的语音基础模型零样本混合专家说话人适配即时路由。载于*Interspeech 2025* 论文集，第4608–4612页，2025年。doi：10.21437/Interspeech.2025-721。

[10] Christian Huber、Laura Kernahan 和 Alexander Waibel。将基础自动语音识别模型适配于构音障碍语音：案例研究。arXiv:2606.31722，2026年。URL `https://arxiv.org/abs/2606.31722`。v1，2026年6月30日。arXiv 记录中未注明发表 venue。单说话人（$n = 1$）研究。

[11] Yicong Jiang、Tianzi Wang、Xurong Xie、Juan Liu、Wei Sun、Nan Yan、Hui Chen、Lan Wang、Xunying Liu 和 Feng Tian。Perceiver提示：用于中文障碍语音识别的 Whisper 灵活说话人自适应。载于*Interspeech 2024*论文集，第 2025–2029 页，2024年。doi：10.21437/Interspeech.2024-852。

[12] Heejin Kim、Mark Hasegawa-Johnson、Adrienne Perlman、Jon Gunderson、Thomas S. Huang、Kenneth Watkin 和 Simone Frame。用于通用访问研究的构音障碍语音数据库。载于*Interspeech 2008*论文集，第 1741–1744 页，2008年。doi：10.21437/Interspeech.2008-480。

[13] Ganesh Pavan Kartikeya Bharadwaj Kolluri、Michael Kampouridis 和 Ravi Shekhar。论编码器深度的作用：在 SLAM-ASR 中剪枝 Whisper 并进行 LoRA微调。载于《低资源场景下的语音语言模型：性能、评估与偏差分析（*SPEAKABLE*）@ *LREC 2026* 论文集》，2026年。网址 `https://aclanthology.org/2026.speakable-1.20/`。ACL Anthology 2026.speakable-1.20；预印本 arXiv:2603.27981，其评论栏注明“已被 SPEAKABLE Workshop，LREC 2026 接收”。

[14] Tzu-Han Lin、How-Shing Wang、Hao-Yung Weng、Kuang-Chen Peng、Zih-Ching Chen 和 Hung yi Lee。《语音领域的参数高效微调：揭示最佳位置、合并策略与集成技术》。载于*ICASSP* 音频、语音及其他领域自监督研讨会（*SASB*），第705–709页，2024年。doi：10.1109/ICASSPW62465.2024.10625957。arXiv：2401.02122v2。[15] Yujian Ma、Xikun Lu、Jinqiu Sang、Xianquan Jiang 和 Ruizhe Li。《幕后解析：对经 LoRA 适配的 Whisper 进行机制可解释性分析，用于语音情感识别》。载于*IEEE*声学、语音与信号处理国际会议（*ICASSP*），2026年。arXiv：2509.08454v3。会议信息已通过 arXiv 评论栏确认（“Accepted at ICASSP 2026”）。[16] Michael McAuliffe、Michaela Socolof、Sarah Mihuc、Michael Wagner 和 Morgan Son- deregger。《Montreal Forced Aligner：使用 Kaldi 进行可训练的文本-语音对齐》。载于*Interspeech*，第498–502页，2017年。doi：10.21437/Interspeech.2017-1386。ISCA档案库：mcauliffe17_interspeech。[17] Hamid Mojarad 和 Kevin Tang。《逐层探测 wav2vec 2.0 和 Whisper 对非裔美国人英语辅音连缀简化的处理》。载于*Interspeech*，第2251–2260页，2026年。doi：10.21437/Interspeech.2026-808。ISCA档案库：mojarad26_interspeech，长篇论文类别；预印本 arXiv：2606.23948。[18] Darwin Jelestin Muthu、Navya Gupta、Wei Lin Tay、Zhengchen Zhang、Daniel Wang Zhengkui 和 Rong Tong。《分析构音障碍自动语音识别中的语音条件效应：一项逐层探测研究》。arXiv：2608.01865，2026年。URL `https://arxiv.org/abs/2608.01865`。

版本2，2026年8月15日，全文均据此引用。v1报告的未适配构音障碍字符错误率为105.08，全块 LoRA 为68.85；v2将其修订为89.98和64.35，并更正摘要中的相对差值为6.67%和2.89%。arXiv记录未列出会议或期刊，稿件页眉则标有 APSIPAASC 2026 论文集。

[19] Keith Ando Ogawa、Bruno Lopes Yamamoto、Lucas Lauton de Alcantara、Lucas Pellicer、Rosimeire Pereira Costa、Edson Bollis、Anna Helena Reali Costa 和 Artur Jordao。《逐层 LoRA 微调：一种相似度度量方法》。arXiv:2602.05988，2026年。预印本 arXiv:2602.05988v2，其评论字段注明“已被 ICPR 2026 接收”。由于无法通过主要会议论文集记录核实该归属，因此将其作为预印本引用。代码见 github.com/c2d-usp/Layer-wise-LoRA-with-CKA。

[20] Ankita Pasad、Ju-Chieh Chou 和 Karen Livescu。《自监督语音表征模型的逐层分析》。载于*IEEE 自动语音识别与理解研讨会（ASRU）*，第 914–921 页，2021 年。doi：10.1109/ASRU51503.2021.9688093。arXiv:2107.04734v3。会议信息已根据 arXiv 评论字段（“已被 ASRU 2021 接收”）确认。

[21] Ankita Pasad、Bowen Shi 和 Karen Livescu。自监督语音模型的逐层比较分析。载于*IEEE声学、语音与信号处理国际会议（ICASSP）*，第 1–5 页，2023 年。doi：10.1109/ICASSP49357.2023.10096149。arXiv：2211.03929v3。会议信息已根据 arXiv 评论栏（“Accepted to ICASSP 2023”）确认。

[22] Dan Pluth、Zachary Nicholas Houghton、Yu Zhou 和 Vijay K. Gurbani。使用稀疏自编码器探究 Whisper 编码的可解释性。arXiv:2605.12225，2026年。URL `https://arxiv.org/abs/2605.12225`。v2，2026年7月17日。arXiv 记录中未注明发表场所。
[23] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。通过大规模弱监督实现鲁棒语音识别。载于国际机器学习会议（*ICML*），第28492–28518页，2023年。arXiv:2212.04356。发表于《ICML 2023 论文集》，PMLR 202:28492–28518；arXiv 记录未附期刊引用信息。[24] Abhilasha Ravichander、Yonatan Belinkov 和 Eduard Hovy。探测范式探究：探测准确率是否意味着任务相关性？载于欧洲计算语言学协会分会第*16*届会议论文集：主会场（*EACL*），第3363–3377页，2021年。doi：10.18653/v1/2021.eacl-main.295。[25] Xian Shi、Xiong Wang、Zhifang Guo、Yongqi Wang、Pei Zhang、Xinyu Zhang、Zishan Guo、Hongkun Hao、Yu Xi、Baosong Yang、Jin Xu、Jingren Zhou 和 Junyang Lin。Qwen3-ASR 技术报告。arXiv:2601.21337，2026年。URL `https://arxiv.org/abs/2601.21337`。v2，2026年1月30日。arXiv 记录中未注明发表场所。代码见 github.com/QwenLM/Qwen3-ASR，Apache-2.0。注：报告给出了 AuT 编码器的参数量和隐藏层大小，但未给出其层数；本文采用的24块这一数值取自发布的检查点配置。[26] Tianyi Tan、Xinan Chen、Xiaohuai Le、Wenzhi Fan、Xianjun Xia、Chuanzeng Huang 和 Jing Lu。CBA-Whisper：基于课程学习的 AdaLoRA 微调方法，用于低资源构音障碍语音识别中的 Whisper。在*Interspeech 2025* 论文集中，第3309–3313页，2025年。doi：10.21437/Interspeech.2025-1705。[27] Dominik Wagner、Ilja Baumann、Natalie Engert、Seanie Lee、Elmar Nöth、Korbinian Riedhammer 和 Tobias Bocklet。利用大语言模型生成的转录文本生成可控合成语音，对构音障碍语音识别进行个性化微调。在*Interspeech 2025* 论文集中，第3294–3298页，2025年。doi：10.21437/Interspeech.2025-2155。

[28] Yan Wan、Mengyi Sun、Xinchen Kang、Jingting Li、Pengfei Guo、Ming Gao 和 Su-Jing Wang。CDSD：中文构音障碍语音数据库。载于*Interspeech*，第4109–4113页，2024年。doi：10.21437/Interspeech.2024-1597。ISCA档案库：wan24b_interspeech。作者署名说明：arXiv 预印本（arXiv:2310.15930）将第一作者写作“Yan Wang”；ISCA 已发表的记录则写作“Yan Wan”，本文采用后者。Su-Jing Wang 是最后一位作者。语料库许可要求，任何使用该音频的出版物均须注明署名行“(c) Su-Jing Wang”。

[29] Zilai Wang、Natarajan Balaji Shankar、Mohan Shi、Kaiyuan Zhang 和 Abeer Alwan。Gumbel-BEARD：面向低资源领域 Whisper 自监督适应的自动层选择。载于 *Interspeech*，第 4318–4323 页，2026年。doi：10.21437/Interspeech.2026-825。ISCA档案库：wang26o_interspeech；预印本 arXiv：2606.11429。

[30] Yang Xiao、Eun-Jung Holden 和 Ting Dang。适配关键所在：面向低资源语言高效多语种语音识别的深度感知适应。载于 计算语言学协会 *ACL 2026*年论文集，第 16776–16788 页，2026年。doi：10.18653/v1/2026.findings-acl.827。预印本 arXiv：2602.01008 v1，2026年2月1日发布，不包含期刊引用信息；会议、页码和 DOI 信息来自 ACL Anthology 记录。

[31] Yichen Xu、Yuyang Liang、Shan Dai、Tianyang Hu、Tsz Nam Chan 和 Chenhao Ma。《理解并指导大语言模型参数高效微调中的层位置》。arXiv:2602.04019，2026年。URL `https://arxiv.org/abs/2602.04019`。v2，2026年2月8日。arXiv 记录中未注明发表 venue。

[32] Yekaterina Yegorova、Argyrios Gerogiannis、Haolong Zheng、Julia Hockenmaier、Chang D. Yoo 和 Mark A. Hasegawa-Johnson。《SALSA：通过学习式引导激活向量实现语音感知大语言模型适应》。arXiv:2606.00460，2026年。URL `https://arxiv.org/abs/2606`.00460。v1，2026年5月30日。arXiv 记录中未注明发表 venue。

[33] Zhengjun Yue、Devendra Kayande、Zoran Cvetkovic 和 Erfan Loweimi。探测 Whisper 在构音障碍语音检测与评估中的表现。载于*IEEE International Conference on Acous-tics, Speech and Signal Processing (ICASSP)*，第 18422–18426 页，2026年。arXiv:2510.04219。会议名称和页码范围取自 Muthu 等人的引文 [18]；未针对 IEEE 记录进行独立核实。