PAWP-TTS：面向紧凑型英尼语语音合成的音素感知权重编程

Firoj
Paudelfirojpaudel@gmail.com

代码：https://github.com/Firojpaudel/PAWP_TTS 音频：https://huggingface.co/spaces/Firoj112/pawp-tts-demo

arXiv:2611.29455v1 [cs.CL] 22 Nov 2026

摘要。本文介绍了**PAWP-TTS**，这是一种紧凑型非自回归文本转语音系统，支持英语、尼泊尔语以及英语–尼泊尔语混合语码语音，采样率为24千赫兹，推理时不使用语言模型。确定性的发音知识由基于规则和词典的双语前端处理，该前端保留尼泊尔语的本土对立特征并避免印地语音系，因此网络只需建模声学特征。统计所有参与推理的已学习张量（包括波形解码器）后，**PAWP-1.5M**（1,490,390个参数，5.98兆字节）在单一自动语音识别器下的英语和尼泊尔语字符错误率分别为1.50%和1.75%；在600条未参与训练的测试话语上，尼泊尔语和英语音素错误率（PER）分别为5.04%和4.02%。以解码器为核心的**281K参数**模型（1.1兆字节）仍清晰可懂（使用三个随机种子时平均PER为5.3%–6.6%）；在唯一参数数量相同的情况下，密集模型在每个随机种子上的表现都优于跨层基共享。修正前端与标签版本不匹配的问题后，尼泊尔语PER几乎减半，改善幅度超过任何架构或权重层面的改动。随后，本文研究权重空间编程：通过检查引导、无需训练的权重编辑。在严格控制条件下（仅使用验证集选择、测试集仅测试一次、配对自助法、匹配的随机基线和同参数量梯度基线），编辑展现出真实的选择性（使用96个权重时，词边界停顿的选择性达到27:1；在语言门控残差中没有跨语言附带影响），但相较于对照方法并未带来准确率提升。两个看似可行的目标最终被证实是标签或音系伪影。现已发布代码、ONNX图和权重。

1. 引言

近年来，神经文本转语音（TTS）系统越来越依赖数亿至数十亿参数的语言模型和神经音频编解码器，以实现自然韵律和多语言覆盖。这些架构虽在云端部署中表现出色，但会带来较高的内存占用、较长的首次音频生成时间（TTFA），且难以普及到尼泊尔语等低资源语言，尤其是在边缘设备、仅CPU或面向浏览器的场景中。

尼泊尔语存在一些特定的音系挑战，通用的印度语支或偏向印地语的多语言模型经常无法妥善处理这些挑战 [7]：标准加德满都尼泊尔语保留了四分的塞音喉部对立、齿音与卷舌塞音的区别（元音后会发生闪音化）、齿龈塞擦音（/ts, ts^h, dz, dz^h/）而非印地语后齿龈音、对立性的辅音 gemination、半开后不圆唇固有元音而非印地语中央央元音，以及跨正字法*sa/ˊsa/s.a*的齿龈咝音合流（/s/）。此外，日常尼泊尔语口语中经常会在句子中途切换至英语。

本研究探讨两个研究问题：

1. **紧凑型双语容量：** 当确定性语言知识是否清晰地拆分为规则和词典支持的前端?

2. **诊断性权重空间编程（WSP）：** 遵循将训练后的权重视为程序、可像程序一样读取和编辑的权重空间编程观点 [1]，，能否通过检查引导的零训练编辑（使用 ROME [10] 等秩一工具）对紧凑型文本转语音模型中的语音或时序缺陷进行精准修复，同时不损害另一种语言的表现？此类编辑能否经受住匹配的随机对照和参数量相同的梯度对照?

主要贡献如下：

- **参数量为 1.5M 和 281K 的开放式紧凑双语文本转语音模型：**对 PAWP-1.5M（1,490,390 个参数，FP32 下为 5.98 兆字节）和 PAWP-300K（281,078 个唯一参数／328,158 个实例化参数，FP32 下为 1.12 兆字节）进行了完整、经审计的参数核算（遵循 4 项指标标准），并采用语言门控低秩残差，可静态折叠为零开销的单语 ONNX图。

- **前端与对齐一致性：** 实证表明，消除前端/缓存版本不匹配（v2 → v3）后，验证集尼泊尔语词错率几乎减半（7.06% → 3.96%），并消除了 55.1% 的伪*va* → /w/ 痕迹。

- **对紧凑型文本转语音中零训练权重空间编程的受控审计：**仅使用验证集进行选择、单次测试

和配对自助法95%置信区间表明其确有选择性（96个权重中暂停选择性为27:1；语言残差内部没有跨语言附带影响），但与匹配对照相比，准确率并无提升；审计还找出了三个会伪装成模型缺陷的混杂因素：缓存标签漂移、母语音系不匹配和全局时长偏差。

2. 相关工作

非自回归文本转语音与傅里叶声码合成。声学骨干网络基于非自回归方差自适应合成 [12]，并结合傅里叶域 ConvNeXt1D + 逆短时傅里叶变换波形合成 [13]。传统流程通常不将参数量达数百万的声码器计入所报告的参数总数，而 PAWP-TTS 则统计从音素 ID 到 24千赫兹 PCM 波形推理所需的每个学习张量。发音特征条件化遵循以往低资源多语言文本转语音研究的做法 [9]。音素时长通常通过单调对齐搜索 [8] 或联合学习的对齐器 [2] 获得；在本文使用的混合多说话人双语数据上，联合学习的对齐器发生了坍塌，因此改用 CTC [4] 强制对齐（第4节）。

模型压缩与参数共享。 已有的压缩方法包括费舍尔加权奇异值分解 [5],自适应低秩分配 [15], 动态稀疏训练 [3], 以及跨层基共享 [14]。本研究在评估跨层共享基与唯一参数量相同的密集基线时，明确区分唯一存储参数 和实例化的计算参数 。

机制定位与权重编辑。采用协方差白化的秩一模型编辑（ROME） [10] 展示了在语言模型中进行有针对性的无梯度编辑的能力；权重空间编程 [1] 则将此类编辑视为像程序一样读取和改写网络权重，并指出编辑可以修复能力，但无法创造能力。本研究将“检查—定位—编辑—验证”工作流应用于紧凑型双语语音合成，并实施严格的评估控制：仅基于验证集的超参数选择、在未触碰的测试数据上进行一次性评估、配对自助法置信区间、范数匹配的随机编辑，以及针对相同参数的梯度微调对照。

3. 前端与紧凑架构

3.1 确定性双语前端

图 1概述了整体方案。PAWP-TTS 不将神经网络容量耗费在记忆正字法例外上，而是采用确定性的前端，将混合使用天城文和拉丁字母的文本映射到统一的 128 个词元国际音标清单，并显式标注语言 ID、词汇重音和词边界（<sp>）：

- **尼泊尔语音系：** 实现标准的加德满都尼泊尔语规则 [7]，涵盖尼泊尔语固有元音、齿龈塞擦音、喉音合流为 /s/（印地语后齿龈音 /sh/ 符号串入率为 0.0%）、元音后卷舌闪音化、辅音延长，以及一个包含 105,868 个词条的尼泊尔语词典（≈99% 由规则/范式生成，符号合规率为 100%， <unk> 占比为 0%），另含 76 个依赖语境的多音词（152 种读法）。
- **英语字素转音素：** 通过 ctypes 在进程内调用 libespeak-ng（<0.20 毫秒/句），与尼泊尔语词元无缝交错处理，以支持天城文–拉丁字母语码混合。

3.2 语言门控叠加与折叠

编码器和声学块中的每个逐点投影均参数化为:

$$y_t = W_{\text{shared}} x_t + m_{\text{en}}(t)\,\Delta W_{\text{en}} x_t + m_{\text{ne}}(t)\,\Delta W_{\text{ne}} x_t, \quad (1)$$

其中 $\Delta W_\ell = g_\ell U_\ell V_\ell$ 是紧凑的低秩语言残差，$m_\ell(t) \in [0,1]$ 是确定性的前端语言掩码。在代码混合合成期间，词元级和帧级掩码会跨文字边界平滑路由；对于单语部署， ΔW_ℓ 会在导出时静态折叠到W_{shared} 中，从而实现零运行时开销。

3.3 以解码器为中心的 1.5M 和 281K 配置

评估了两个主要参数配置（表 1）：

- **PAWP-1.5M：** 4 个编码器 ConvNeXt1D 块（d =128，k = 5，前瞻 1）、方差预测器（时长、$F0$、浊音、能量）、4 个声学块（d =128，k = 7）以及一个 3 层逆短时傅里叶变换解码器（d = 192、513 个短时傅里叶变换频点、24千赫、跳跃 = 240）。学习参数总数：**1,490,390**（5.98MB FP32）。

- **PAWP-300K（共享与密集）：** 解码器中心的分配方式（52.6% 的参数位于 24千赫 iSTFT 解码器中，$n$fft = 512，28 维发音动作条件）。共享基底变体（EXP-015b）在不同深度间汇集 K = 2 个基底（**281,078** 个独立参数 / 328,158 个物化参数，20.96 MMAC/s）；等预算稠密基线（EXP-015c）则取消基共享（**281,442** 个独立参数 =物化参数）。

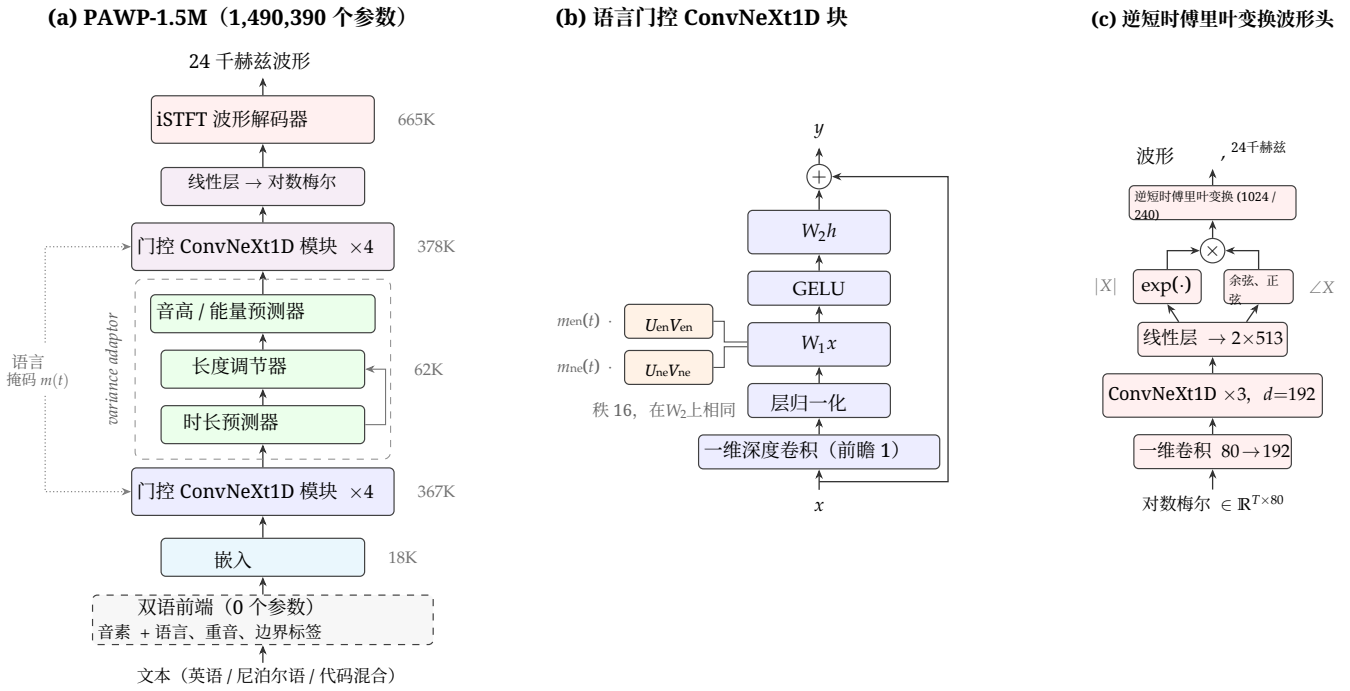


图 1：PAWP-TTS 架构。(a) 确定性前端（虚线所示，无参数）将混合的天城文/拉丁字母文本映射为带语言标签的音素；所有经过学习的推理张量均计入参数量，包括波形解码器（占参数的 44.6%）。时长将词元状态扩展为 10毫秒帧；预测的音高和能量为声学网络提供条件。(b) 每个编码器块和声学块的两个逐点层都添加秩为 16 的英语和尼泊尔语残差，由前端的语言掩码按词元选择，从而实现语码混合；对于单语导出，它们会折叠到W中。(c) Vocos 风格的头部预测 513 个频谱箱的对数幅度和相位，并通过逆短时傅里叶变换将其还原。

4. 实验设置与结果

数据与训练。模型使用约 225 小时的数据进行训练：77 小时英语朗读语音（LibriSpeech 的一个子集 [11]）、108 小时尼泊尔语（内部制作的录音棚录音和朗读语音，包括 OpenSLR 尼泊尔语），以及 34 小时混合语码语音。话语时长为 1–12 秒，验证集或测试集中的句子会从训练数据中移除。为获得单一跨语言目标音色，经过微调的 OmniVoice [6] 教师模型以同一音色合成英语和尼泊尔语语音（英语 22.7 小时，尼泊尔语 15.7 小时），仅保留由独立自动语音识别系统转录后字符错误率 $\leq$0为 .10（英语）或 $\leq$0.12（尼泊尔语）的样本。补充的合成尼泊尔语语音，其损失权重为人工录音的一半。训练采用多说话人设置（1,815 名说话人）；导出的模型保留一个说话人向量。训练在单张 RTX 4090 上以 BF16 精度运行，使用 L1 对数梅尔损失、多分辨率短时傅里叶变换损失、时长损失、F_0/浊音和能量损失，以及多周期/多分辨率对抗损失和特征匹配损失。在一半的训练步骤中，声学网络使用自身预测的韵律而非真实韵律作为条件，从而缩小了训练与推理之间的显著差距。时长由维特比强制对齐生成，使用的 CTC 音素识别器仅用于训练阶段 [4]。

4.1 Clean`v3`数据对齐（`EXP-014`）

早期训练（`v2`）早于最终版尼泊尔语词典，且 *va* → `/b/` 规则更新。使用当前前端重建包含 136,807 条话语的特征缓存（`v3`），重新训练冻结的 CTC 音素识别器（`ctc_对齐器_v3.pt`，CTC 损失 ≈1。0 → 0614），并重新提取维特比强制对齐时长，使单帧时长坍缩减少了 ≈3×，并使虚假的*va*词内 `/b/` → `/w/` 混淆在 300 条尼泊尔语测试集话语中从 55.1%（65/118）降至 **0.0%**（0/118）。随后，使用一致的标签进行微调，将验证集音素错误率（PER）从 7.06% 降至 3.96%（尼泊尔语），并从 5.29% 降至 4.33%（英语；图 2）。

4.2 双语合成质量与等预算 281K 消融实验

表1汇总了 `PAWP-1.5Mv3(EXP-015a`，在 `v3` 上进行 `EXP-012`的 10k 微调步数）以及预算相同的 281K 变体（`EXP-015b` 共享版与`EXP-015c` 密集版，从头开始训练 36k 步，使用`v3`，配方相同，随机种子=42）。留出集自动语音识别在 n = 30 英语、n = 30 尼泊尔语和 n = 9语码混用话语上使用单个识别器进行测量（同时报告人工录音室音频基准，作为可懂度代理指标，而非自然度比较）。

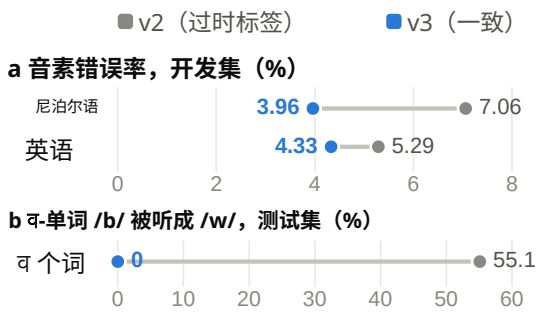


图 2：标签一致性比任何编辑都重要。 使用当前前端（v2→v3）重新构建特征缓存和音素识别器后，尼泊尔语词错率几乎减半（a），并消除了先前看起来像是模型缺陷的虚假 *va*-词 /b/→/w/ 错误（b）。

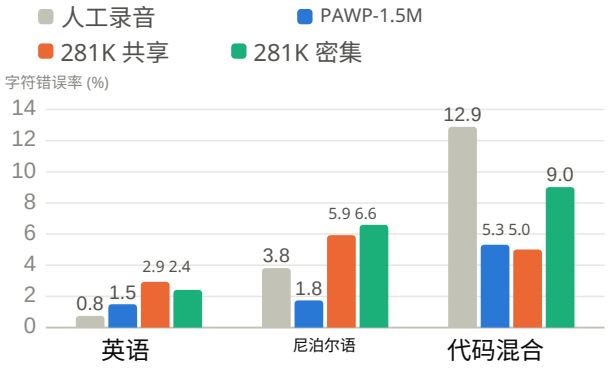


图 3：留出集自动语音识别字符错误率（一个识别器；$n = 30$ 英语、30 个尼泊尔语、9 个代码混合句子）。灰色柱表示同一识别器对这些句子的人工录音进行识别的结果；较低的合成字符错误率反映了干净、规范的朗读语音，是一种可懂度代理指标，而非质量比较。

性代理指标，而非自然度比较）。未修改测试集音素错误率是在$n = 300$ 尼泊尔语 $+ n = 300$ 英语话语上通过 ctc_aligner_v3.pt 测量的；最小对立对比则在一个内部 77 对尼泊尔语/双语数据集上进行声学评分。

多随机种子下共享与密集架构的权衡。在随机种子=42下，PAWP-300K Shared 的验证集_mel 较低（0.8768 对 0.8778），验证集_能量表现更好（1.50 对 1.60），且语码混合语音识别词错误率更低（12.02% 对 18.09%，n = 9）。然而，在规模更大的未触碰过的 300 + 300 test.jsonl 音素错误率基准测试中，参数量相同且唯一参数数量相等的 Dense-300K 在 3 个相互独立的随机种子（42、101、202）上均胜过 Shared，且两者没有重叠：尼泊尔语词错率平均为 **5.53**±**0.08%**，而 6.35±0.11%（−0.82 个百分点；随机种子=42 配对自助法 95%置信区间 $[+0.56, +1.41]$, p < 0.001）；英语词错率平均为 **5.34**±**0.28%**，而 6.57±0.47%（−1.23 个百分点；随机种子=42 95% 置信区间 $[+1.44, +2.35]$, p < 0.001），以更少3× 的参数量逼近 PAWP-1.5M v3 的 5.f（图 4）。

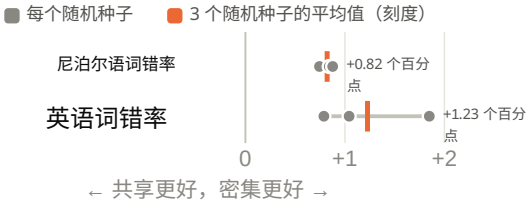


图4：相同唯一参数量（281K）下的共享基底与密集模型比较。三个训练随机种子（灰色）及其均值（橙色）在未触碰测试集上的音素错误率差异（共享 − 密集；$n = 300$尼泊尔语 + 300 英语）。密集模型在两种语言的每个随机种子上都表现更好。

这推翻了 $K = 2$ 跨层基础共享在 281K 规模下对单语音素保真度的作用，但验证了解码器中心的 281K 架构本身。

延迟与零样本压缩。 在一台 4 线程 x86_64 CPU 上，折叠后的 PAWP-1.5M ONNX 模型解码一个 60ms 音频块的耗时为 p50 1.11 ms（p99 1.52 ms），实时因子约为 0.0058；前端处理每个句子的 p50 耗时为 0.16ms。这些是语句编码完成后测得的每块计算时间；采用有状态流式处理的端到端首次音频生成时间留待未来研究。相比之下，事后进行的一次性 SVD（EXP-010）在没有恢复训练的情况下，参数量降至 500K 时模型便会崩溃，这表明在参数量低于 1M 时，端到端紧凑训练至关重要。

5. 诊断性权重空间编程：机制与混杂因素

在严格控制条件下，系统评估了零训练、激活引导的权重编辑PAWP-1.5M：仅使用 `val.jsonl` 进行超参数选择，并在未触碰的 300 + 300 `test.jsonl` 上进行单次验证；采用配对自助法95%置信区间、范数匹配的随机编辑，以及相同参数的梯度微调作为对照。三项实验揭示了真实的机制特性，也暴露了关键的方法学陷阱；图5总结了所有编辑及其对照。

5.1 无约束共享编辑与语言残差隔离

在 EXP-013a 中，对共享声学_blocks.3.pw2.weight 进行无约束秩-1编辑，降低了目标混淆，但使整体音素错误率恶化（尼泊尔语 +1.22%，英语+1.75%）。此外，字素审计发现，目标 /b/ → /w/ 错误中有 90.3%（65/72）是*va*词上的 v2缓存标签伪影。重建v3后，伪影被彻底消除，这凸显出必须先验证标签一致性，才能进行机制编辑。

表1： 端到端参数占用（计入所有推理张量）以及在干净的 v3 对齐上的评估。留出集自动语音识别使用单个识别器（n = 30/30/9，随机种子 =42）；未修改测试集音素错误率使用 ctc_aligner_v3.pt（n = 300NE + 300 EN），报告 EXP-015b/c 的三个随机种子的均值 ± 标准差（随机种子 42、101、202）。

模型（v3）	唯一值 / 总数参数	FP32 Size	英语自动语音识别 词错误率/字错误率	尼泊尔语 自动语音识别 词错误率/字错误率	混合自动语音识别 词错误率/字错误率	测试集 NE / EN 音素错误率（n = 600）区分 / 精确	CTC 77 配对
人工基准（自动语音识别）	—	—	2.63/ 0.75	11.60 / 3.82	25.91 / 12.88	—	—
PAWP-1.5M v3（EXP-015a）	1,490,390 / 1,490,390	5.98 兆字节	**4.02 / 1.50**	**12.35 / 1.75**	12.32 / 5.32	**5.04% / 4.02%**	**83.1% / 58.4%**
PAWP-300K 共享（EXP-015b）	281,078 / 328,158	1.12兆字节	6.88 / 2.93	**19.71 / 5.93**	**12.02/ 5.01**	6.35±0.11% / 6.57±0.47%	77.9% / 39.0%
Dense-300K (EXP-015c)	281,442 / 281,442	1.13 兆字节	**6.15 / 2.42**	20.21 / 6.59	18.09 / 9.01	**5.53±0.08% / 5.34±0.28%**	79.2% / 37.7%

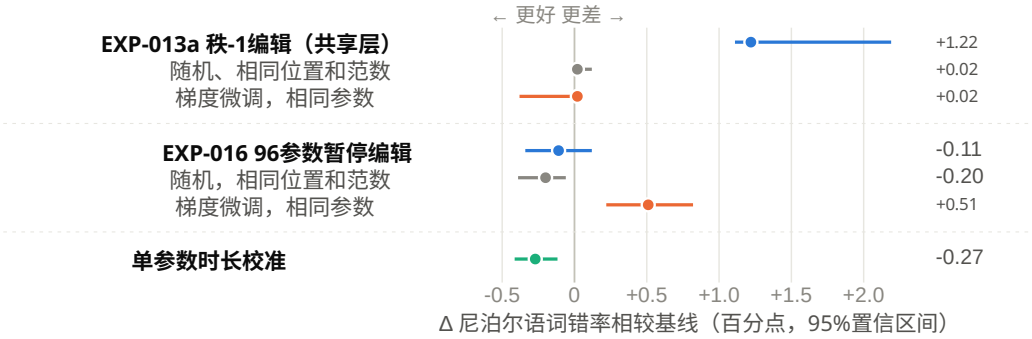


图5：与匹配对照组相比的权重编辑。在未触及的测试数据上，尼泊尔语词错率的变化（n = 300，配对自助法 95%置信区间）。编辑（蓝色）与相同位置和范数的随机编辑（灰色）以及对相同参数进行的梯度微调（橙色）进行比较。无约束秩-1编辑导致性能退步；停顿编辑并不优于其随机对照；在验证集上选出的单个全局时长偏置参数（绿色）带来了唯一显著的改进。

5.2 词边界停顿选择性与全局时长偏置（EXP-016）

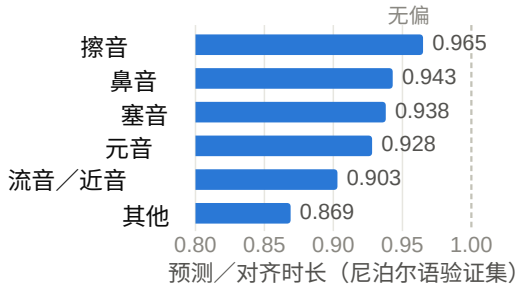
在干净的v3 验证数据上，词边界 <sp> 音素被低估（6.16 帧，而非 10.38 帧），且在时长_predictor.proj.weight的 96 维输入处，与常规音素线性可分（96 个参数， = 0占模型的 .0064%）。ROME 白化秩一更新在未触碰的测试数据上实现了**27:1 的精确选择性**（<sp> +1.001 帧，而音素为 .037 帧），减少了删除（+0 尼泊尔语，英语），并优于在开发集上调优的 96 参数梯度微调（尼泊尔语音素错误率为 4.93%，而非 5.55%；图6b）。

然而，这项修改在整体尼泊尔语词错率上并未胜过匹配的随机-96对照组（4.93%，p = 0.177 vs. 4.84%，p = 0.004）。按类别时长审计（169条尼泊尔语验证语句，13,450个音素）发现了原因：基线时长在每个音素类别中**都存在全局性偏短** −7.3%的情况（预测值为6.105帧，真实值为6.585帧；整体比率为0.927，范围为0.869至0.965），这与对数时长训练下的詹森不等式一致。仅根据val.jsonl（n = 150条尼泊尔语、 + 150条英语）选择一个参数的全局时长偏置偏移（duration_predictor.proj.bias）（选定值为Δb = +0.10，使验证集平均音素错误率从4.625%降至4.448%），并在未触碰的test.jsonl上**仅评估一次**（n = 300 条尼泊尔语、 + 300 条英语），将尼泊尔语音素错误率从4.744%（D = 401）降至**4.513%**（D = 373；$\Delta = -0.00271$，95%置信区间 $[-0.00414, -0.00118]$, $p < 0.001$），且没有造成英语性能退化（4.044% → 4.016%，p = 0.317）。因此，全局语速校准解释了随机-96带来的增益；恰当的修正方法是在训练时采用经过偏差校正的时长损失。

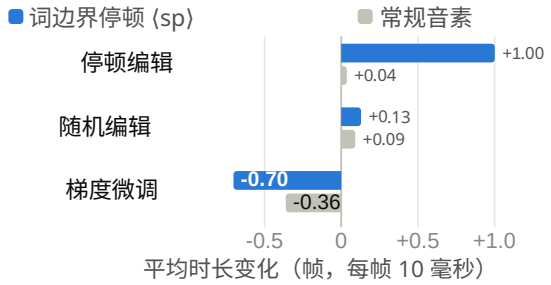
5.3 与母语者录音相比的目标有效性（EXP-017）

v3 重新对齐后，验证集中剩余最多的尼泊尔语替换错误是卷舌鼻音被替换为齿鼻音（国际音标：卷舌鼻音 → 齿 /n/）。在仅限尼泊尔语的低秩残差编码器_blocks.0.pw2.{u_ne, v_ne}中应用 ROME 零空间秩一编辑（6,144 个参数 =0.41%），按构造实现了**英语附带影响为 0.000%**，尼泊尔语 PER 附带影响为 0.00%（5.036% → 5.026%，p = 0.495）。

关键的是，直接在**真实的人类录音室尼泊尔语音频** 上运行 ctc_aligner_v3.pt，对 test.jsonl 中的语音进行分析（共 52 段语音，出现n = 62 次），发现**90.3%（56/62）**实现为齿音（平均裕度 +3.246），这与音系描述相符：口语加德满都尼泊尔语只有三个对立性鼻音（/m、n、ng/），且正字法中的*na*（卷舌音）除出现在卷舌塞音前之外，均实现为齿音 /n/。因此，EXP-017 的目标属于前端音系选择（拟议作为尚未实现的未来 v4 前端规则，



(a) 按音素类别划分的预测／对齐时长。



(b)96参数停顿编辑及对照项带来的时长变化。

图 6：时长诊断。 (a) 每种音素类别的预测时长都短了 3.5–13%，这是一种与对数域时长训练一致的全局偏差。(b) 停顿编辑会比常规音素更显著地延长词边界停顿 27×，但范数相同的随机编辑也会让所有时长略微增加，这解释了它为何能降低音素错误率。

因为 1,800 个词典条目仍会输出卷舌符号），而非神经网络模型缺陷。

6. 局限性与结论

局限性。 (1) 留出集自动语音识别评估使用的样本量较小（$n = 30/30/9$），且仅使用单一识别器；正式的多听者人工平均意见分评估仍有待开展。(2) 这份包含 105,868 个词的尼泊尔语词典有 ≈99% 是规则／范式生成的，且尚未与独立的母语者标注金标准集进行核验；文中提出的正字法卷舌鼻音 → 齿龈 /n/ 映射被记录为候选 v4 前端更新，但尚未在已发布的 v3 权重中实现。(3) 模型在推理时只合成单一目标说话人身份，不支持零样本语音克隆。(4) 目标语音语料库由教师模型生成；从合成语音蒸馏所产生的感知效应尚未测量。(5) 延迟以分块计算量报告；采用有状态流式处理时的端到端首音频输出时间尚未测量。

结论。 本研究表明，将基于语言学原理的英语–尼泊尔语前端与紧凑型、以解码器为核心的 ConvNeXt1D+ iSTFT 主干网络相结合，可在参数量为 1.49M（5.98MB）和 281K（1.12MB）的情况下，实现清晰易懂的 24千赫 双语及语码混合语音。更重要的是，受控的权重空间编程审计为语音合成中的权重编辑确立了实用的方法学准则：验证标签一致性，直接通过人工音频验证目标误差，将编辑范围限定在特定语言的低秩残差上，并与仅含 1 个参数的全局校准及预算匹配的梯度对照进行基准比较。

致谢。 这是作者开展的独立研究。作者感谢 WiseYak 提供图形处理器计算资源，并提供对以下内容的访问权限：(i) 内部制作的用于训练的尼泊尔语和混合语码语音录音；(ii) 内部微调的文本转语音模型，仅通过推理用于合成目标语音语料库；以及 (iii) 用于评估的内部语音识别服务。所有其他训练数据均来自公开语料库（第 4节）。权重空间编程实验（第 5节）是在人工智能研究代理（Gemini 4 Neon）的协助下设计并运行的，独立的人工智能代理对方法和结果进行了审查；报告中的每一个数值都经过与记录的运行结果核对。

参考文献

[1]权重空间编程。 https://rikiyat.github.io/blog/weight-space-programming/, 2026。博客文章。

[2] Rohan Badlani、Adrian Łańcucki、Kevin J. Shih、Rafael Valle、Wei Ping 和 Bryan Catanzaro。一种统一所有任务的文本转语音对齐方法。载于*IEEE* 国际声学、语音与信号处理会议（*ICASSP*），2022 年。

[3]Utku Evci、Trevor Gale、Jacob Menick、Pablo Samuel Castro 和 Erich Elsen。《操纵彩票：让所有彩票都成为赢家》。发表于国际机器学习会议 *(ICML)*，2020 年。

[4]Alex Graves、Santiago Fernández、Faustino Gomez 和 Jürgen Schmidhuber。《连接时序分类：使用循环神经网络标注未分段的序列数据》。发表于国际机器学习会议 *(ICML)*，2006 年。

[5] Yen-Chang Hsu、Ting Hua、Sungen Chang、Qian Lou、Yilin Shen 和 Hongxia Jin。《采用加权低秩分解的语言模型压缩》。

发表于*International Conference on Learning Representa- tions*（*ICLR*），2022 年。

[6]k2-fsa。OmniVoice：大规模多语言零样本文本转语音。https://github.com/k2-fsa/OmniVoice，2026。

[7] Rajesh Khatiwada。尼泊尔语。国际语音学协会期刊，39(3):373–380，2009。[8] Jaehyeon Kim、Sungwon Kim、Jungil Kong 和 Sungroh Yoon。Glow-TTS：通过单调对齐搜索实现的文本转语音生成流。载于神经信息处理系统进展（*NeurIPS*），2020。[9] Florian Lux 和 Ngoc Thang Vu。基于发音特征的低资源文本转语音语言无关元学习。载于计算语言学协会年会（*ACL*），2022。[10]Kevin Meng、David Bau、Alex Andonian 和 Yonatan Belinkov。定位并编辑 GPT 中的事实关联。载于神经信息处理系统进展（*NeurIPS*），2022。[11] Vassil Panayotov、Guoguo Chen、Daniel Povey 和 Sanjeev Khudanpur。LibriSpeech：基于公版有声书的自动语音识别语料库。载于*IEEE* 国际声学、语音与信号处理会议（*ICASSP*），2015。[12] Yi Ren、Chenxu Hu、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 和 Tie-Yan Liu。FastSpeech 2：快速、高质量的端到端文本转语音。载于国际学习表征会议（*ICLR*），2021。[13] Hubert Siuzdak。Vocos：弥合时域与基于傅里叶变换的神经声码器之间的差距，实现高质量音频合成。载于国际学习表征会议（*ICLR*），2024。[14]Jingcun Wang、Yu-Guang Chen、Ing-Chao Lin、Bing Li 和 Grace Li Zhang。基共享：用于大语言模型压缩的跨层参数共享。载于国际学习表征会议（*ICLR*），2025。

[15]Qingru Zhang、Minshuo Chen、Alexander Bukharin、Pengcheng He、Yu Cheng、Weizhu Chen 和 Tuo Zhao。面向参数高效微调的自适应预算分配。发表于国际学习表征会议（*ICLR*），2023 年。