arXiv:2611.30927v1 [eess.AS] 23 Nov 2026

SpeechSplice：通过替换编码器隐藏状态实现编码器–解码器大语言模型文本转语音中的发音控制

Kiyoshi Kurihara[1,2], Jumpei Ukita[1,2],YusukeIwasawa[1,2],YutakaMatsuo[1,2][1]Matsuo Institute, Inc.，日本，[2]东京大学，日本 kiyoshi.kurihara@matsuo-institute.com

摘要

大多数基于大语言模型（LLM）的文本转语音（TTS）方法都没有提供明确的发音纠正功能。因此，这些系统经常会误读许多日常词语，例如专有名词、异读词、数字和领域专用术语。虽然权重编辑或微调可以缓解这一问题，但这些方案可能导致质量下降，而且必须离线进行。我们提出 SpeechSplice，这是一种用于基于大语言模型的编码器–解码器文本转语音发音控制的免训练编码器输出替换方法。SpeechSplice 将某个词的编码器输出隐藏状态替换为其预期读音对应的隐藏状态。跨架构和跨语言实验涵盖了 10 个公开文本转语音系统，涉及 17 种模型—语言条件。在被替换的隐藏状态处，当生成器通过交叉注意力读取文本时，干预使发音更常转向目标读音；而当文本和语音共享同一自注意力序列时，这种效果较弱。在人工评估中，母语听众在 98.6–100.0% 的判断中认定，使用 SpeechSplice 生成的语音实现了目标读音；另一项独立研究则发现，与未经修改的合成语音相比，其自然度没有显著差异。通过阐明发音纠正在哪些条件下能够奏效，本研究为探究基于大语言模型的文本转语音架构中的发音控制提供了实证依据。

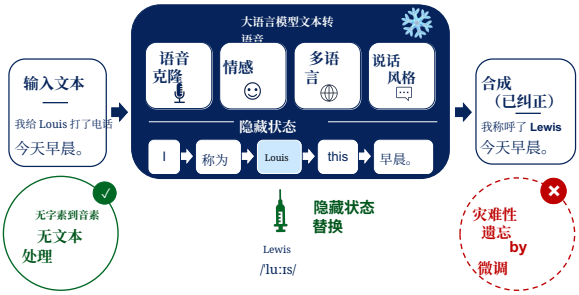


图 1：SpeechSplice 概览：无需字素到音素（G2P）前端即可控制发音。我们仅评估说话人相似度、自然度、音高和时长的保持情况。

1 引言

近期基于大语言模型（LLM）的文本转语音（TTS）方法——包括 VALL-E（Chen 等人，2025年）、CosyVoice 2（Du 等人，2024年）、Llasa（Ye 等人，2025年）、Qwen3-TTS（Hu 等人，2026年）和 T5Gemma-TTS（Arata 和 Kurihara，2026年）——将传统语言前端整合到单一序列模型中，直接把文本映射为语音。这种设计提升了自然度、多语言覆盖能力和零样本语音克隆能力，但在大多数此类系统中，它会移除一种可在推理时编辑的发音接口：在传统流程中（Shen 等人，2018年），发音歧义——异读词、多音字、同形词——由字素到音素（G2P）模块处理，并且可以逐条更新（Taylor，2009；Sproat，1996年；Zhang，2023年）。少数基于大语言模型的系统支持在音素或拼音层面控制发音（Cui，2025年；Gong，2026年；Deng，2025年）。微调可以恢复这种控制能力（Kato，2025年），但据报道，在基于大语言模型的文本转语音中，微调会降低质量，灾难性遗忘还会导致语音崩溃（Fujita，2024年；Kwon，2025年；Hsiao，2025年）。

我们提出**SpeechSplice**[1]，一种免训练的编码器输出替换方法，用于在基于编码器–解码器大语言模型的文本转语音中实现免 G2P 词语注入。通过在冻结文本编码器的最终输出处，将实体位置隐藏状态替换为编码目标读音的向量（图1），该方法使发音完全由模型内部处理，并能在不改动任何权重的情况下实现局部控制。

[1]代码、基准测试条目和听力测试材料：https://github.com/Kuri-san-jp/speechsplice

我们的贡献如下：

1. 我们提出了**SpeechSplice**。据我们所知，这是首个展示如何在基于编码器–解码器大语言模型的文本转语音中，于文本编码器输出处实现免 G2P 词语注入的方法；所有模型权重均保持冻结。

2. 我们在 17 种模型—语言条件下，对十种公开文本转语音系统进行研究，以确定干预能够影响哪些系统。区分干预有效与无效条件的关键要求是：在我们替换的张量处，生成器必须通过交叉注意力读取经编码的文本。另一项条件并非经过测试，而是预先假定的：分词器必须为源词和目标词分配不同的词元标识符；第 4.1 节将说明该条件不成立时的情况。这一发现为研究基于大语言模型的文本转语音架构中的发音控制提供了实证依据。

3. 我们邀请英语、日语和中文母语听者评估 SpeechSplice：使用带有“都不是”选项的配对比较测试评估目标词注入，并通过平均意见分（MOS）研究自然度。听者在 98.6–100.0% 的判断中选择 SpeechSplice 的输出，认为其实现了目标读音；MOS 显示，其自然度与未修改的合成结果没有显著差异。

2 相关工作

可控语音合成。 可控 TTS 领域的近期研究，既按照控制目标（如情感、零样本语音克隆、风格或发音）进行组织，也按照底层生成范式进行划分，包括自回归编解码器语言模型、扩散模型和混合系统（Xie 等人，2025年）。代表性方法利用全局风格词元控制风格（Wang 等人，2018年），利用文本提示控制声音和说话风格（Leng 等人，2024年），并支持零样本语音克隆（Chen 等人，2025年；Le 等人，2023年；Du 等人，2024年）。SpeechSplice 无需重新训练模型，即可加入发音控制。

另一条相关研究路线并非移除前端，而是通过神经字素到音素转换（Gorman 等人，2020年；Cheng 等人，2024年）和多音字消歧（Zhang 等人，2022年）。这些方法会在合成前确定读音，因此需要前端接收该读音。

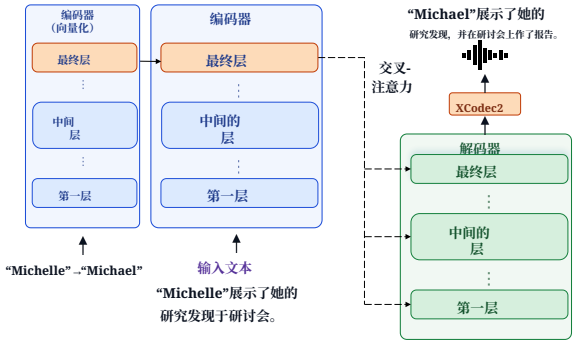


图 2：T5Gemma-TTS 上 SpeechSplice 的概览。源句子及其目标读音版本由同一个冻结的编码器进行编码。在最终编码器输出处，SpeechSplice 将源实体状态替换为对应的目标读音状态，解码器则通过交叉注意力读取修改后的序列。

无词典方法则在合成时进行纠正，借助非目标说话人参考语音（Fong 等人，2022年）、语音提示驱动的语音编辑（Kashyap 等人，2025年）或声学信息辅助拼写（Fong 等人，2023年）。FlowEdit 会针对每项纠正优化并存储一个文本嵌入扰动（Singh 等人，2026b）。SpeechSplice 无需目标词的录音，也无需优化，而是直接作用于最终编码器输出。

语音表示的逐层探查。对 wav2vec 2.0（Baevski 等人，2020年）和 WavLM（Chen 等人，2022年）等语音自监督学习（自监督学习）模型进行逐层分析发现，声学属性和语言属性并非均匀分布，而是在不同深度逐渐显现（Pasad 等人，2021年）；语音信息能在许多层中保留（Choi 等人，2024年）；中间层则能捕捉词级结构（Pasad 等人，2024年）。在问答、逻辑推理、语音识别和机器翻译模型中，让解码器跨注意力关注中间编码器层，而非最终编码器输出，结果表明，较低层和中间层已能完成较简单的子任务（Langedijk 等人，2024年）。以往的文本转语音研究使用自监督学习层来驱动从文本到语音信息的转换（Aihara 等人，2026年；Wang 等人，2023年；Sivaguru 等人，2023年；Lee 等人，2022年），而 SpeechSplice 采用了不同的方法。我们将相同的逐层视角应用于文本转语音模型的文本编码器。随后，在解码器读取这些状态之前，我们替换目标实体位置处的最终层状态

。

激活补丁与表征工程。激活补丁通过修改内部激活而非模型权重，在推理时引导模型行为（Vig 等人，2020年；Geiger 等人，2021年；Zou 等人，2023年；Li 等人，2023年）。在文本语言模型中，Li 等人（2023年）将学习到的方向添加到隐藏状态中。语言模型还可以用自然语言解释隐藏状态（Ghandeharioun 等人，2024年），而将一个前向传播中的隐藏状态移入另一个前向传播，可以改变翻译后的概念，但不会改变输出语言（Dumas 等人，2025年）。相关思路也已应用于文本转语音，包括反事实激活编辑（Lee 等人，2025年）、编解码器令牌编辑（Peng 等人，2024年；Wang 等人，2025年）、基于分数的语音编辑（Tae 等人，2022年）、频谱图级语音编辑（Lv 等人，2026年；Jiang 等人，2023年），以及通过激活引导进行口音编辑（Yang 和 Hansen，2026年；在语音识别领域，Sun 等人，2026年）。近期的文本转语音研究还利用激活补丁追踪情绪回路（Du 等人，2026年）、引导语速（Verdini 等人，2026年），并通过稀疏自编码器特征控制语音属性（Koriagin 等人，2026年）。SpeechSplice 通过将一个词的末端编码器状态替换为目标读音状态，改变该词的读音。我们的交叉注意力消融（第 5.6 节）会阻断注意力边；Geva 等人（2023年）也使用这种干预来定位信息传播至预测的位置。

基于微调的发音控制。另一种方法通过更新参数来控制发音，例如采用低秩适配（Hu 等人，2022年），如 UtterTune（Kato，2025年）和 SonoEdit（Singh 等人，2026a），对基于仅解码器大语言模型的文本转语音进行权重编辑。如第1节所述，据报道，对基于大语言模型的文本转语音进行微调会降低质量（Fu- jita 等人，2024年；Kwon 等人，2025年；Hsiao 等人，2025年）。而 SpeechSplice 会冻结所有参数。

3 方法

3.1 任务定义

我们的目标是在推理时仅修改目标实体的发音，同时保持模型权重和其余话语不变。给定一个冻结的编码器–解码器文本转语音模型、一条包含歧义实体 e 的输入句子 x，以及在推理时提供的期望读音 r_e 我们生成语音 $\tilde{a}$，其中e 读作 r_e，同时保持说话人特征和整体语音质量不变。

令 x' 为目标句，即将 e替换为带有读音 r_e的书写形式后的 x。令 $\mathcal{I}_{\text{src}}(e) = (s_1, \ldots, s_m)$ 和 $\mathcal{I}_{\text{tgt}}(e) = (u_1, \ldots, u_n)$ 分别表示 x 和x'中的实体令牌跨度。为找到 $\mathcal{I}_{\text{src}}(e)$，我们对整个句子进行分词并获取字符偏移量，然后选取偏移范围与实体重叠的每个令牌。我们使用同一个冻结的文本转语音编码器对 x 和x' 进行编码，得到 $\mathbf{H}^{\text{src}}$和 $\mathbf{H}^{\text{tgt}}$。我们不训练任何参数，因此 SpeechSplice 是免训练的。它将目标实体状态拼接在源文本的前缀和后缀之间（图2）：

$$\tilde{\mathbf{H}} = \left(\mathbf{h}^{\text{src}}_{1:s_1-1},\ \mathbf{h}^{\text{tgt}}_{u_1:u_n},\ \mathbf{h}^{\text{src}}_{s_m+1:T}\right), \qquad (1)$$

其中，$\mathbf{h}_{a:b}$ 是从索引 a到 b 的状态块，T 是 x 的编码器序列长度。拼接会将编码器长度从 T 变为 $T - m + n$，我们会重建注意力掩码和位置索引以使其匹配。对于 $m = n$，这是原位替换。当长度必须保持不变时，我们改用池化，使用 $A^e \in \mathbb{R}^{m\times n}$、$A^e_{jk} \geq 0$ 和 $\sum_{k=1}^{n} A^e_{jk} = 1$:

$$\bar{\mathbf{h}}^{\text{tgt}}_{s_j} = \sum_{k=1}^{n} A^e_{jk}\mathbf{h}^{\text{tgt}}_{u_k}, \qquad (2)$$

并且我们写入 $\bar{\mathbf{h}}^{\text{tgt}}_t$ 在每个 $t \in \mathcal{I}_{\text{src}}(e)$。第5.6 节的消融采用这种形式，其中 $A^e_{jk} = 1/n$。

3.2 面向解码器的编码器输出

在编码器–解码器文本转语音中，编码器将输入文本映射为 $\mathbf{H} = [\mathbf{h}_1, \ldots\ldots, \mathbf{h}_T] \in \mathbb{R}^{T\times d}$，其中 T是编码器序列长度，d是隐藏状态维度。解码器通过交叉注意力读取该序列，同时生成声学或编解码器标记。因此，最终编码器输出是文本编码与语音生成之间的文本侧接口，也是瓶颈。检索和注入都发生在最终编码器输出处。SpeechSplice在此读取目标实体状态，并将其拼接到源序列的实体标记位置，如公式（1）所示。附录 I规定了在第 5.3节比较的其他九个系统中，我们所替换的文本侧张量。

4 实验设置

4.1 数据集与模型

文本转语音模型。我们将十种公开的文本转语音系统作为冻结检查点进行评估（链接见附录I）。

其中五种系统中，解码器通过交叉注意力读取文本编码器输出（编码器–解码器）：T5Gemma-TTS、SpeechT5（Ao 等人，2022年）、LongCat-AudioDiT（Xin 等人，2026年）、Irodori-TTS（Arata，2026年）和 Kiseki-TTS（telecomadm1145，2026年）。另外五种系统中，文本与语音令牌共享一个自注意力序列（仅解码器）：Qwen3-TTS、XTTS-v2（Casanova 等人，2024年）、Llasa-3B（Ye 等人，2025年）、CSM-1B 和 Bark。T5Gemma-TTS 和 Qwen3-TTS 支持英语（EN）、日语（JA）和中文（ZH），其余系统各支持其中一到两种语言，共计 17 种模型—语言条件。SpeechT5 仅在英语（EN）下评估：其仅支持英语的分词器会为每组日语（JA）和中文（ZH）文本对中的源词和目标词分配相同的词元标识符，因此替换无法改变编码器输出；我们不将这两种条件计为失败，而是予以排除。因此，该方法的前提是分词器能够为源词和目标词分配不同的词元标识符。

基准测试。跨语言基准测试中，每种语言有 300 对样本，由 50 对手工编写的实体对和六个句子模板构成。声学交叉注意力消融实验和附录E同样使用的跨架构基准测试中，每种语言另有 300 对样本，由 30 对手工编写的实体对和十个模板构成，目标文本为日语姓氏，源文本为英语名字（英语）或中文姓氏（日语、中文）。源句和目标句之间恰好只有一个实体跨度不同。

传统系统比较。其他系统比较使用了跨语言基准测试的英语子集。我们将 SpeechSplice 与 VoiceCraft 零样本文本转语音（Peng 等人，2024年）和 EdiTTS 内容替换（Tae 等人，2022年）进行了比较。SpeechSplice 和 VoiceCraft 使用了相同的 3.25 秒 LibriTTS（Zen 等人，2019）测试集干净语音提示；EdiTTS 无法克隆任意提示说话人的声音，因此改为编辑其 Grad-TTS（Popov 等人，2021年）骨干模型合成的语音，该模型使用 LJSpeech 训练[2]。

人工聆听刺激材料。我们为两项研究构建了刺激材料。基线（BL）和注入（INJ）音频片段使用 T5Gemma-TTS 语音克隆生成，每种语言各使用一位母语参考说话人的语音，素材分别取自 LibriTTS（英语）、JSUT（Sonobe 等人，2017年）（日语）和 Databaker 中文标准普通话语料库[3]（中文），采用可变长度拼接实现。配对比较测试从跨语言基准中抽取了每种语言各 50 个姓名项目，但有一个日语项目在数据收集前被替换；此外，还使用了一个日语同表层集合：其中包含 50 个句子，冻结的 T5Gemma-TTS 基线会误读目标词，而书面句子保持不变。这些项目涵盖异读词、数字、送假名、地名和罕见姓氏；附录D列出了这些项目的来源和构成。自然度研究在英语、日语和中文中各使用另外 50 对基线与注入的名称替换配对，这些配对单独构建并合成，与 300 对及配对比较项目相互独立。BL 和 INJ 使用相同的文本、语音提示和解码配置，并在每次合成调用前将随机种子重置为相同的值，因此两种条件之间唯一的差异是对编码器输出进行拼接干预。

4.2 评估指标

自动化指标。为衡量注入语音是否包含目标读音，以及其余语音是否得到保留，我们采用以下指标。**Whisper-INJ** 表示包含目标读音的归一化自动语音识别（ASR）转录文本所占的注入样本百分比。转录文本由 Whisper large-v3（Radford 等人，2023年）生成；在跨语言基准和双实体测试中，日语改用 Furigana Whisper，中文则在拼音归一化后进行匹配。附录 F 报告了整句字符错误率（CER）。UTMOS（Saeki 等人，2022年）用于测量语音质量。说话人相似度是使用 WavLM-SV（Chen 等人，2022年）提取的说话人验证嵌入之间的余弦相似度，分别比较每个基线音频片段与其注入或编辑后的对应片段，以及每个注入音频片段与语音提示。比值 F_0 为相对于基线的基频平均绝对变化量，除以两个不同句子的基线话语之间的平均变化量；而 $|\Delta$时长$|$则是相对于基线的话语时长平均绝对变化量。

配对比较测试。为验证注入的目标词是否在生成语音中得到正确实现，我们按照 Speech Audio Corrector（Fong 等人，2022年）

[2] LJ 语音数据集

[3] Databaker 中文标准普通话语料库

和 Spell4TTS（Fong 等人，2023年）采用的发音评估协议开展了配对比较测试：母语听者以随机顺序听取一组 BL–INJ 对，并判断哪个音频片段读出了屏幕显示的目标词，也可以选择两个片段都不是。Fong 等人将预期发音作为参考音频或替代拼写呈现；而我们直接显示目标词，因为对于专有名称，若无参考，两个片段都无法显而易见地判定为正确。英语和中文显示正字形式，日语则只显示假名读法。每种语言使用 50 个项目，分为两个各含 25 个项目的区块；每种语言有 20 名听者（每个区块 10 名），每人判断分配给自己的 25 个项目及 3 个捕获试次，因此每个项目获得 10 个相互独立的判断，每种语言共 500 个判断。日语同表层集合采用相同的协议（附录 D）。界面会在两个音频片段均播放完毕前禁用作答功能；捕获试次中显示的词在两个片段中都不会被读出。我们报告无条件的 INJ选择率，将选择两个片段都不是计为失败，并提供项目级自助法置信区间。姓名集中的音频片段取自跨语言基准，因此对于每个项目，我们还会将其 10 个判断的多数结果与 Whisper 对同一组音频片段的判断进行比较；其中一个替换后的日语项目没有自动判断结果。没有听者符合预先设定的排除标准：捕获试次错答超过一次、直线作答，或完成的试次数少于 90%。

自然度MOS。 一项独立研究测量了注入是否会降低自然度。每种语言有20位母语听众，他们用五点评分量表对50段基线和50段INJ音频片段进行评分，平均每段音频片段获得10次评分；说明中明确指出，评分标准不包括发音正确性。我们对项目均值进行双侧配对 Wilcoxon 符号秩检验，以检验BL–INJ差异。附录G 记录了两项研究的招募、报酬、知情同意、说明和界面。

跨架构指标。为了比较不同系统中注入对语音的影响幅度，以及注入使实体发音向目标读音靠近的频率，我们报告整句的 BL–INJ 梅尔倒谱失真（MCD）（Kubichek，1993年）和音素注入改善率（PIB），并给出 Wilson 95% 置信区间（CI）。PIB 是指以下配对中，注入降低了相对于目标读音测得的实体音素错误率（PER）的比例： $N=300$ 配对测得的目标读音，

$$\mathrm{PIB}=\frac{1}{N}\sum_{p=1}^{N}\mathbf{1}\left[\mathrm{PER}_p^{\mathrm{INJ}}<\mathrm{PER}_p^{\mathrm{BL}}\right]. \quad (3)$$

我们用 L 表示文本在一个系统中经过的层数，用 l^* 表示替换位置之前的层数，因此，在编码器–解码器系统中，替换发生在 $l^*=L$ ，而在仅解码器系统中，替换发生在 $l^*=L-1$ ；第5.4 节在其他条件不变的情况下，对 l^* 进行遍历。

双实体测试。 为考察能否在同一句子中同时纠正两个实体，我们为每种语言另选了300个句子，每个句子都包含两个实体槽位。这些句子由50对同类专有名词和专门为其设计的六个模板构成。我们先纠正一个实体，再同时纠正两个实体，并按槽位对Whisper-INJ进行评分；我们汇总了来自300个句子的600个槽位判断。由于同一句子中的两个槽位并非相互独立，因此我们通过对句子进行重采样来比较各列（见附录J）。BL列对基线语音中要求其读出的词语进行评分。

交叉注意力消融。为了考察实体位置交叉注意力是否会将替换后的状态传递到生成语音中，我们还合成了阻断实体位置交叉注意力的基线和 INJ （ $\mathrm{BL}_{\mathrm{mask}}$，$\mathrm{INJ}_{\mathrm{mask}}$ ），并通过 MEL 相似度以及方差归一化 MCD，在实体跨度上将每种注入条件与其匹配的基线进行比较。MCD 的数值仅在表A1 （附录A）内具有可比性。

5 结果

5.1 跨语言注入与语音质量

听众直接确认了注入的读音（表1）：在姓名集中，没有听众选择基线，这与 Whisper-INJ 得分一致。整句字符错误率也呈现出相同的变化：在三种语言中，注入后的转录文本都接近目标句，且远离源句（附录F）。日语同表层集合并未替换词语，而是纠正自然发生的误读，结果也相同：只有少数判断选择了基线或两个片段都未选（表A4）。在每个集合中，选择率都显著高于随机机会；检验方法为广义估计方程（GEE）检验，中文则采用精确二项检验。对包含 160 个项目的日语朗读基准测试进行自动评分

语言	配对比较测试		自然度MOS			Whisper-INJ	UTMOS		说话人相似度			
	已选择INJ (%) ↑	z (p)	BL	INJ ↑	p	(%) ↑	BL	INJ ↑	BL–INJ ↑	INJ–提示 ↑	$\lvert\Delta F_0\rvert$ 比率 ↓	
EN	99.4 (+0.6/−0.8)	9.07 ($< 10^{-16}$)	3.56 ± 0.09	3.52 ± 0.08	0.45	96.33 (+1.61/−2.78)	3.86 ± 0.05	3.82 ± 0.05	0.83 ± 0.01	0.84 ± 0.01	0.98	
JA	98。6 (+1.4/−2.8)	4.24 (2.3×10^{-5})	3.68 ± 0.09	3.64 ± 0.09	0.74	92.33 (+2.50/−3.57)	3.39 ± 0.04	3.41 ± 0.05	0.94 ± 0.00	0.85 ± 0.01	0.71	
ZH	100。0 (+0.0/−0.0)	– (1.8×10^{-15})	3.67 ± 0.10	3.71 ± 0.10	0.96	96.00 (+1.70/−2.86)	3.33 ± 0.07	3.35 ± 0.07	0.91 ± 0.01	0.89 ± 0.01	0.88	

表 1：几乎每次判断中，听众都选择了注入音频片段，且平均意见分显示 BL–INJ 之间没有显著差异。我们未检验等效性。BL：T5Gemma-TTS 基线合成；INJ：SpeechSplice 注入。z (p)：针对随机机会的二项广义估计方程 (GEE) 检验 (Liang 和 Zeger，1986年)，或针对中文的精确二项检验（附录 G.1）。平均意见分 p：对项目均值进行双侧配对Wilcoxon符号秩检验。说话人相似度：两个被命名音频片段之间的 WavLM-SV 余弦相似度（提示：语音提示）。95%区间：对 INJ 选择率采用项目级自助法，对 Whisper-INJ 采用 Wilson 方法，如 (+上限/−下限)； $\pm 1.96\times$ 均值的标准误。指标：第4.2节。

系统	Whisper-INJ （%） ↑	UTMOS			说话人相似度		$\lvert\Delta F_0\rvert$ 比例 ↓	$\lvert\Delta$ 时长$\rvert$ (秒) ↓
		BL	注入/已编辑 ↑	p	BL–INJ ↑	INJ–提示 ↑		
SpeechSplice（本文方法）	96。33 (+1.61/−2.78)	3.86 ± 0.05	3.82 ± 0.05	0.07	0.83 ± 0.01	0.84 ± 0.01	0.98	0.00 ± 0.00
VoiceCraft（零样本）	90.00 (+2.91/−3.92)	3.62 ± 0.04	3.63 ± 0.04	0.54	0.87 ± 0.01	0.87 ± 0.01	0.98	0.20 ± 0.02
EdiTTS†	86.00 (+3.47/−4.38)	4.29 ± 0.02	4.28 ± 0.02	0.09	0.97 ± 0.00	–	0.74	0.13 ± 0.01

表 2：在跨语言基准的英语子集上与传统方法的比较。SpeechSplice 最常恢复目标读音，并且通过构造保持基线时长。SpeechSplice 行对应表 1 中的英语运行结果，其标题给出了区间；第 4.2 节定义了各列。UTMOS p：对每个片段的得分进行双侧配对Wilcoxon符号秩检验。VoiceCraft（零样本）：零样本语音克隆。基线：合成源句子（EdiTTS 使用 Grad-TTS）；注入/编辑（INJ）：SpeechSplice、VoiceCraft 的目标句子合成，或 EdiTTS 的编辑。 †EdiTTS 不基于大语言模型，且无法克隆声音，因此 INJ–提示不可用。

结果一致：注入纠正了全部 45 个严格基线误读（附录 C）。我们未发现其他方面存在代价。任何语言中的平均意见分（MOS）均未显示显著的 BL–INJ 差异，UTMOS 也未显示一致的变化；在对 300 个自动评估配对样本的 WavLM-SV 余弦相似度进行配对 Wilcoxon 检验后，各语言中与语音提示的说话人相似度均未发生变化，且 BL–INJ 相似度仍然很高；每个 $|\Delta F_0|$比率都低于 1，因此注入造成的变化 F_0 小于句子间通常的变化幅度。除八个姓名项目外，听众多数意见均与 Whisper 的判断一致；在这八处分歧中，有七处自动语音识别的判断更严格（附录 H），因此相较于人类感知，自动评分较为保守。

5.2 与传统方法的比较

我们在跨语言基准的英语子集上，将 SpeechSplice 与 VoiceCraft 零样本文本转语音以及 EdiTTS 进行了比较（表 2），以评估隐藏状态注入能否比现有的生成和编辑方法更频繁地恢复目标读音。

SpeechSplice 比 VoiceCraft（会对整段话语重新合成）和 EdiTTS（通过扩散替换内容）更常还原出目标读音。三个系统在不同的声学域中进行合成，因此我们只在各系统内部解读 UTMOS，而不跨系统比较；三个系统相较各自基线的变化都不显著，因此我们没有发现任何一个系统存在质量损失。对于说话人身份，我们直接测试了相对于语音提示的变化：在 300 组配对上进行配对 Wilcoxon 检验，SpeechSplice 的余弦相似度变化为 -0.001 ，95% 置信区间为 ± 0.009 和 $p = 0.98$ ；VoiceCraft 的变化为 -0.009 ，95% 置信区间为 ± 0.008 和 $p = 0.026$，这是唯一达到显著性的偏移。EdiTTS 无法克隆提示中的说话人，因此此项测试不适用于它。三个系统的音高变化都很小，其中 EdiTTS 的音高偏移最小。各系统在时长上的差异最大：按照设计，SpeechSplice 沿用基线生成时长（附录I），而 VoiceCraft 和 EdiTTS 会改变时长。因此，SpeechSplice 在基线时长下提高了准确率，且我们没有检测到其他方面的代价。

			在我们的替换点		层扫描		双实体句子（%）		
模型	语言	L	MCD（dB）	PIB（%）↑	l^*	PIB（%）最小值–最大值	BL	INJ（一个）↑	INJ（两者）↑
编码器–解码器（交叉注意力）									
T5Gemma-TTS	EN	26	9.95 ± 0.20	57.00 (+5.48/−5.66)	1/8/15/22/26	56.67–60.00	97.17	97.17	96.67
	JA	26	8.24 ± 0.21	52.00 (+5.59/−5.64)	1/8/15/22/26	48.00–52.00	88.00	88.83	89.17
	ZH	26	8.96 ± 0.25	55.00 (+5.53/−5.66)	1/8/15/22/26	52.33–55.67	78.83	81.83	81.00
SpeechT5	EN	12	3.53 ± 0.06	64.00 (+5.22/−5.58)	1/3/6/9/12	54.00–64.00	95.67	89.33	90.67
LongCat-AudioDiT	EN	12	10.00 ± 0.24	65.33 (+5.16/−5.55)	1/3/6/9/12	59.33–65.33	94.67	96.33	95.17
	ZH	12	8.65 ± 0.25	62.67 (+5.28/−5.60)	1/3/6/9/12	58.33–64.67	99.83	99.33	99.17
Irodori-TTS	JA	25	7.20 ± 0.24	49.33 (+5.63/−5.61)	1/7/13/19/25	49.33–59.33	96.33	94.33	94.67
Kiseki-TTS	JA	12	5.28 ± 0.17	41.67 (+5.65/−5.44)	1/3/6/9/12	41.67–44.67	2.67	7.83	7.00
仅解码器（自注意力）									
Qwen3-TTS	EN	28	2.30 ± 0.28	13.67 (+4.35/−3.43)	1/10/14/21/27	12.33–15.00	97.83	0.00	0.00
	JA	28	1.42 ± 0.27	12.00 (+4.17/−3.20)	1/10/14/21/27	10.00–13.67	91.83	0.00	0.00
	ZH	28	0.86 ± 0.15	12.33 (+4.20/−3.25)	1/10/14/21/27	9.00–12.67	99.83	0.00	0.00
XTTS-v2	EN	30	0.34 ± 0.08	2.67 (+2.51/−1.31)	1/9/16/24/29	2.67–54.67	96.33	0.00	0.00
Llasa-3B	EN	28	1.91 ± 0.30	13.33 (+4.31/−3.39)	1/8/15/22/27	13.33–51.33	96.00	0.00	0.00
	ZH	28	2.21 ± 0.28	20.33 (+4.92/−4.17)	1/8/15/22/27	20.33–49.00	97.17	0.00	0.00
CSM-1B	EN	16	4.74 ± 0.15	27.00 (+5.29/−4.71)	1/4/8/11/15	27.00–51.67	96.67	0.00	0.00
Bark	EN	24	0.58 ± 0.18	4.00 (+2.86/−1.70)	1/6/12/17/23	4.00–49.67	97.00	0.00	0.00
	JA	24	0.97 ± 0.26	7.67 (+3.57/−2.50)	1/6/12/17/23	7.67–41.67	57.33	0.00	0.00

表3：我们替换点处的 PIB（$l^*=L$；若为仅解码器则为 $L-1$）可区分编码器–解码器与仅解码器条件；大多数仅解码器模型在较浅的 l^* 处数值升高，而增加第二个实体不会造成可检测的准确率损失。INJ 会在 l^*（L：文本经过的层）处将一个实体的隐藏状态替换为交换后的句子的隐藏状态；BL 则不会。MCD：BL–INJ 的梅尔倒谱失真。PIB：与交换后的句子相比，INJ 能降低实体音素错误率的配对所占百分比。MCD 的 95% 置信区间为 ±；PIB 使用 Wilson 置信区间，按（+上限/−下限）的顺序给出。双实体句子：按预期读出的 600 个位置所占百分比（BL：原句；INJ：交换一个或两个实体）。

5.3 跨架构结果

本节考察这种干预在 T5Gemma-TTS 之外的适用范围，以及哪些特征能够区分干预有效与无效的系统。表3 汇报了全部 17 种模型—语言条件下的比较结果，以及第5.4 节的层扫描和第5.5节的双实体测试。总体而言，分组取决于生成器读取编码文本的路径。在我们进行替换的张量处，两组的 PIB 值没有重叠：编码器–解码器系统中最低的 Kiseki-TTS JA 高于仅解码器系统中最高的 CSM-1B EN。其余三个编码器–解码器系统采用了 T5Gemma-TTS 和 SpeechT5 未使用的两种解码器架构——LongCat-AudioDiT 和 Irodori-TTS 使用扩散变换器（Peebles 和 Xie，2023年），Kiseki-TTS 使用 Mamba-2 解码器（Dao 和 Gu，2024年）——因此，这一模式并非仅与这两个系统共有的自回归解码器有关。所有条件下的 MCD 均大于零，包括仅解码器系统；其中，即使是最小值 XTTS-v2 EN，其区间也不包含零。在这些系统中，替换会改变语音，但根据 PIB，语音向目标读音靠近的程度较低比任何编码器–解码器条件下都更频繁。

Bark 英语的 PIB 在表中排名第二低：在绝大多数配对中，注入的语义词元与基线完全相同，因此 PER 不会改变，而根据定义，该配对被计为失败。

5.4 替换深度的影响

深度对编码器–解码器系统的影响很小：在表3中，T5Gemma-TTS EN 的 PIB 在整个编码器范围内的变化幅度，小于任一单点置信区间的宽度；SpeechT5、LongCat-AudioDiT、Irodori-TTS 和 Kiseki-TTS 的表现也相同。这与解码器读取一个张量的情况一致：较早位置的替换仍会通过其余层传递影响。深度对仅解码器系统的影响则很大：CSM-1B EN、Llasa-3B EN 和 ZH、Bark EN 和 JA，以及 XTTS-v2 EN 的 PIB 在 $l^*=1$ 处均远高于我们的替换点。在这些系统中，语音令牌通过同一堆栈关注文本行，因此在深度l^* 处进行替换会重写其上方每一层的键和值。Qwen3-TTS 表明这种情况并非普遍存在：其 PIB 在所有扫描深度下都保持较低。

5.5 一次纠正两个实体

与只纠正一个实体相比，同时纠正两个实体并未产生可检测到的代价：表 3中的八种编码器–解码器条件下，两列的差异方向不一，且每种条件下配对句子差异的 95%区间均包含零（附录 J）。仅解码器条件下，模型在任何位置都不会读出已纠正的词；但其基线语音在绝大多数位置都会读出原词，Bark JA 中也超过半数。这表明问题出在传输，而非对刺激的读取。同时替换两个实体后，所有条件下的 UTMOS 仍与基线接近（附录 J）。

5.6 交叉注意力消融

第 5.3 节表明，解码器到编码器的交叉注意力会调节实体位置的编码器状态对语音生成的影响。我们按照第 4.2节所述，进行了包含四种条件的 T5Gemma-TTS 消融实验，并采用保持长度不变的注入方式：对源文本跨度内的目标实体状态进行均值池化，使编码器长度保持不变；同时仅屏蔽指向源文本实体列的交叉注意力。

表A1 （附录A）报告了结果：交叉注意力完整时，注入会使三种语言的声学特征都产生显著差异；阻断实体位置的交叉注意力后，输出几乎完全相同。令牌数量匹配的日语子集也得到相同结果（附录 A）。这些指标测量的是 BL–INJ 的声学差异，而非目标方向准确率。附录 B 还在同形词对上以音素层面重复了阻断实验。

6 讨论

人工评估。 两项听力研究分别回答了不同的问题：在三种语言以及同形集上，听者听出预期读音的比例都接近最高水平；MOS研究则发现自然度没有显著差异。对于多数听者的判断与自动语音识别判断不一致的情况，自动语音识别通常否定了听者认可的读音，因此自动评分低估了效果，而非高估。

跨架构比较。 在我们测试的系统中，编码器–解码器组与仅解码器组之间的划分（表 3）与一种解释一致：该解释建立在对生成器的一项要求之上，即生成器应通过被替换的状态读取文本侧内容。当满足这一条件时，替换会改变发音；而当大部分读音绕过被替换状态时，发音改变的情况则少得多。编码器–解码器系统在我们替换的张量处满足这一要求，因为该张量是其解码器通过交叉注意力读取的文本侧接口（第 3.2 节；附录 I）。五个仅解码器系统展示了当大部分读音绕过被替换状态时会发生什么：在我们的替换点处，它们会改变语音——在每种条件下，其 BL–INJ MCD 都非零——但与任何编码器–解码器条件相比，它们更少将该实体读成更接近目标读音的形式。因此，在这些系统中，在这一深度上，被替换的文本状态通过一次自注意力序列到达生成语音的可靠性低于通过交叉注意力路径到达的可靠性。该解释还预设了一个能将两种书写形式区分开的分词器：在任何隐藏状态能够区分它们之前，两者必须先获得不同的词元标识符，这也是 SpeechT5 仅以英语纳入的原因（第 4.1 节）。

替换深度。层扫描对这一对比作了进一步说明：除了 Qwen3-TTS 之外，在每个仅解码器系统中，结果都表明，我们替换点处的低值反映的是替换状态进入模型的深度，而非不存在任何能够读取替换状态的路径。在编码器–解码器系统中，即使较早进行替换，解码器读取的张量仍会受到影响（第5.4节）；在仅解码器系统中，每一层的音频标记都会读取文本行，因此当替换发生在 $l^{*}=L-1$ 时，只有最后一层会读取替换后的状态，而所有更早的层仍会读取原始词语。在我们找到的最佳深度下（局限性），XTTS-v2、Llasa-3B、CSM-1B 和 Bark 的 PIB 达到了与替换点处编码器–解码器条件相当的范围，如表3所示。有一项结果无法用上述解释说明：Qwen3-TTS，无论扫描哪个深度，其结果都无法进入该范围。

纠正两个实体。 双实体测试与同一解释一致。仅解码器条件下，修正后的词在表3的 INJ（一个）列和 INJ（两个）列中，在我们的替换点（附录I）均未出现在任何槽位中；这用另一项指标和另一个句子集重现了这种差异，并表明它并非 PIB 或单实体配对所特有。在编码器–解码器条件下，纠正第二个实体并未显著降低每槽位 Whisper-INJ（第5.5节），这与每个被替换的

片段都是通过交叉注意力读取其自身位置的这一解释一致；第5.6 节的消融为 T5Gemma-TTS 提供了支持。它们的 INJ（一个）列和 INJ（两个）列也都与 BL 列接近；BL 列的范围从 Kiseki-TTS JA（其基线语音很少读出要求其读出的词语）到 LongCat-AudioDiT ZH（其基线语音几乎总是能读出这些词语）。这表明，当生成器通过被替换的状态读取文本侧内容时，无论替换一个还是两个实体，纠正率都取决于各系统读取此类词语的可靠程度。

传输与传统方法的比较。消融实验表明，实体位置交叉注意力是主要的声学传输路径：阻断该路径会消除英语、日语和中文中的 BL–INJ差异。在受控的英语基准测试中，SpeechSplice 恢复目标读音的频率也高于 VoiceCraft 和 EdiTTS。这项比较涵盖了不同的接口——完整续接合成、基于扩散的内容替换，以及编码器状态拼接——然而，直接干预面向解码器的编码器张量，在保持基线时长的同时，取得了最高的发音准确率，且自然度没有可检测到的变化。

7 结论

我们提出了 SpeechSplice，这是一种免训练干预方法，可恢复基于编码器–解码器大语言模型的文本转语音中的发音纠正。在英语、日语和中文的几乎所有判断中，以及在一个基线会误读某个拼写不变单词的日语数据集上，母语听众都选择了注入音频片段（表1和A4）。另一项独立研究发现，与未经修改的合成结果相比，其自然度没有显著差异。在我们评估的公开语音合成与编辑系统中，SpeechSplice 恢复目标读音的频率也更高。

在我们进行替换的张量处，当生成器通过交叉注意力读取编码后的文本时，替换是成功的：在 17 个模型–语言条件下，五个编码器–解码器系统的 PIB 均高于五个仅解码器系统，且两组之间没有重叠（表 3）。免训练干预能够在基于编码器–解码器大语言模型的文本转语音中实现预期读音。

未来工作。 包括仅解码器 Transformer 在内的模型，可以从头开始训练，以支持注入，满足我们所确定的要求；而音素级条件控制则可以免去重新拼写目标词的需要。

局限性

- 我们对十种公开的文本转语音系统进行了机制比较，但解码器架构、训练数据、语言覆盖范围、分词器选择和模型规模仍同时存在差异，因此，该比较揭示的是哪些系统能够通过干预实现预期效果，而非究竟由哪个单一因素决定。

- 替换深度是通过经验确定的。在 Bark、Llasa-3B 和 Qwen3-TTS 英语中，我们发现效果最佳的深度是第一层，而不是表 3中报告的深度；在 XTTS-v2、CSM-1B 以及 Qwen3-TTS 日语和中文中，最佳深度则是中间层（l^*=9、4和21，依次对应）。目前没有可用于预先选定深度的标准。

- 该方法要求在推理时正确指定目标读音。

- 当源文本与目标文本的词元数量不同时，变长拼接能够保留所有目标实体状态，但需要一致地重建注意力掩码和位置索引；其对其他位置编码方案的可移植性仍有待测试。

- 平均意见分研究中，每种语言使用一名克隆参考说话人。

- 在听测之外，我们使用自动语音识别和音素识别器自动评估发音，而自动语音识别可能会漏掉听者能够听出的读音错误（Luoand Wan，2026年）。我们未评估音高重音。

- Kiseki-TTS在编码器—解码器系统中的PIB最低，而且在双实体测试中，其基线语音也很少被正确识别（表3）。

- 双实体测试评估发音。我们仅使用 UTMOS 估计其输出的语音质量（附录 J），未通过听者评估其自然度、可懂度或伪影。长文本、更多样化的说话人、低资源语言、间隔较近的实体，以及每句三个或更多实体的情况也尚未测试。

伦理声明

本研究编辑预训练文本转语音模型的内部表征，以正确读出专有名词。同一机制也可能带来风险，因此我们在下文列出主要的伦理考量。

滥用可能性。该方法可能被用于故意读错姓名。这种风险受到两个因素的限制：用户必须明确指定目标读音，且干预仅限于所选实体跨度。

说话人和人口统计学考量。在面向不同姓名来源和语言部署之前，还需要开展更多样化的评估。

人工评估。配对比较测试和自然度平均意见分研究的母语听众通过 Prolific 招募，英语和中文听众通过 Prolific 招募，日语听众通过 Lancers 招募；报酬高于 Prolific 的最低标准，或在 Lancers 上高于东京最低工资的两倍（附录G.2）。实验条件采用盲法，呈现顺序随机安排，并要求听众完整播放音频。配对比较测试包含捕获试次，没有听者符合预先设定的排除标准。我们仅使用参与者标识符进行分配、去重和质量控制，且未收集任何个人身份信息。

工件使用与许可。 除附录 E中的权重编辑基线外，本研究使用的所有第三方模型、数据集和评估工具均未经修改，并继续受其原始发布源的许可协议和使用条款约束。除以自身许可证发布的开源代码外，只有下文所述的样例页面以及附录 C 中的 ja-tts-g2p-bench 项目会根据相应条款允许再分发第三方材料：一段 JSUT 录音、日语语音提示，以及来自 ja-tts-g2p-bench 和维基百科的日语句子。我们发布的工件旨在用于研究和可复现性目的，不取代任何第三方资源原有的访问条件。发布的代码仓库（ h t tps://github.com/Kuri-san-jp/speechsplice）包含生成、评分和分析代码、容器定义以及需求规格、基准项目列表和预计算的评估输出，以及听力测试刺激材料清单、项目级聚合结果、英语、日语和中文版本的听者逐字说明，以及界面截图。该仓库还提供一个语音样本页面，其中的样本选自听力测试刺激材料（ https://kuri-san-jp.github.io/speechsplice/），并根据 LibriTTS、JSUT 和 Databaker语料库的条款发布；语音克隆提示源自这些语料库。此外，页面还包含来自 ja-tts-g2p-bench 和维基百科的日语同形项目。

数据筛选与文档编制。 我们手工整理了由广为人知的公众专有名词构成的基准实体列表，并将其用于研究和评估。由于专有名词可能包含公众人物的姓名，我们将其视为可能具有识别性的信息，并检查收集到的实体列表，删除包含除公开姓名之外的私人个人信息、冒犯性表达或明显不当内容的条目。人工评估记录与基准实体分开保存，仅用于配对比较测试和MOS研究的分析以及质量控制。

自动基准由三种语言的数据集构成——英语、日语和中文——每种语言包含两组各300对的实体替换对，以及一组包含300个双实体句子的集合（第4节）。其他系统的评估使用表1中的英语300对数据集；配对比较测试使用每种语言50个姓名项目，另加50个日语同形项目；MOS研究则分别使用英语、中文和日语各50对。本研究创建的伪影旨在用于研究并提高可复现性。

AI助手披露。 作者借助AI助手润色语言并准备清单。所有经AI辅助生成的文本均由作者审阅和编辑，作者对论文内容承担全部责任。AI助手未生成实验结果，也未替代科学分析。

参考文献

Ryo Aihara、Yoshiki Masuyama、François G. Germain、Gordon Wichern 和 Jonathan Le Roux。2026年。探索基于自监督表示的解耦神经语音编解码器

监督表示。载于*2026*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 21992–21996 页。

Junyi Ao、Rui Wang、Long Zhou、Chengyi Wang、Shuo Ren、Yu Wu、Shujie Liu、Tom Ko、Qing Li、Yu Zhang、Zhihua Wei、Yao Qian、Jinyu Li 和 Furu Wei。2022年。SpeechT5：面向口语语言处理的统一模态编码器-解码器预训练。载于第 *60* 届计算语言学协会年会论文集（第 *1* 卷：长篇论文），第 5723–5738 页，都柏林，爱尔兰。计算语言学协会。

Chihiro Arata。2026年。Irodori-TTS：一种由表情符号驱动风格控制的流匹配文本转语音模型。Hugging Face 代码仓库：https://huggingface.co/Aratako/Irodori-TTS-v4.1-Small。

Chihiro Arata 和 Kiyoshi Kurihara。2026年。T5Gemma-TTS 技术报告。*arXiv* 预印本，arXiv:2604.01760。

Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。2020年。wav2vec 2.0：一种用于语音表示自监督学习的框架。载于神经信息处理系统进展，第 33 卷，第 12449–12460 页。Curran Associates, Inc.

Edresson Casanova、Kelly Davis、Eren Gölge、Görkem Göknar、Iulian Gulea、Logan Hart、Aya Aljafari、Joshua Meyer、Reuben Morais、Samuel Olayemi 和 Julian Weber。2024年。XTTS：一种大规模多语言零样本文本转语音模型。*Interspeech2024*，第 4978–4982 页。Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao、Jian Wu、Long Zhou、Shuo Ren、Yanmin Qian、Yao Qian、Jian Wu、Michael Zeng、Xiangzhan Yu 和 Furu Wei。2022年。WavLM：用于全栈语音处理的大规模自监督预训练。*IEEE* 信号处理精选主题期刊，16(6)：1505–1518。Sanyuan Chen、Chengyi Wang、Yu Wu、Ziqiang Zhang、Long Zhou、Shujie Liu、Zhuo Chen、Yanqing Liu、Huaming Wang、Jinyu Li、Lei He、Sheng Zhao 和 Furu Wei。2025年。神经编解码器语言模型是零样本文本到语音合成器。*IEEE* 音频、语音与语言处理汇刊，33：705–718。Shiyang Cheng、Pengcheng Zhu、Jueting Liu 和 Zehua Wang。2024年。字素到音素转换方法综述。*Applied Sciences*，14(24)：11790。Kwanghee Choi、Ankita Pasad、Tomohiko Nakamura、Satoru Fukayama、Karen Livescu 和 Shinji Watanabe。2024年。自监督语音表征的语音学特征多于语义特征。发表于*Interspeech2024*，第 4578–4582 页。

Jiayan Cui、Zhihan Yang、Naihan Li、Jiankun Tian、Xingyu Ma、Yi Zhang、Guangyu Chen、Runxuan Yang、Yuqing Cheng、Yizhi Zhou、GuochenYu、Xiaotao Gu 和 Jie Tang。2025年。GLM-TTS 技术报告。*arXiv* 预印本，arXiv:2512.14291。

Tri Dao 和 Albert Gu。2024年。Transformer 即 SSM：通过结构化状态空间对偶实现的广义模型与高效算法。载于第 *41* 届国际机器学习会议论文集，第 235 卷，机器学习研究论文集，第 10041–10071 页。PMLR。

Steven B. Davis 和 Paul Mermelstein。1980年。连续语句中单音节词识别的参数化表示比较。*IEEE*《声学、语音与信号处理汇刊》，28(4)：357–366。

Wei Deng、Siyi Zhou、Jingchen Shu、Jinchao Wang 和 Lu Wang。2025年。IndexTTS：一种工业级、可控且高效的零样本文本转语音系统。*arXiv* 预印本，arXiv:2502.05512。

Hongfei Du、Jiacheng Shi、Yanfu Zhang 和 Ashley Gao。2026年。追踪基于大语言模型的文本转语音中的稀疏情感控制电路。收录于*2026*年自然语言处理实证方法会议论文集，第 8645–8658 页，匈牙利布达佩斯。计算语言学协会。

Zhihao Du、Yuxuan Wang、Qian Chen、Xian Shi、Xiang Lv、Tianyu Zhao、Zhifu Gao、Yexin Yang、Changfeng Gao、Hui Wang、Fan Yu、Huadai Liu、Zhengyan Sheng、Yue Gu、Chong Deng、Wen Wang、Shil- iang Zhang、Zhijie Yan 和 Jingren Zhou。2024年。CosyVoice 2：基于大语言模型的可扩展流式语音合成。*arXiv* 预印本，arXiv:2412.10117。Clément Dumas、Chris Wendler、Veniamin Veselovsky、Giovanni Monea 和 Robert West。2025年。将舌头与思想分离：激活补丁揭示 Transformer 中语言无关的概念表示。载于计算语言学协会第*63*届年会论文集（第*1*卷：长篇论文），第 31822–31841页，奥地利维也纳。计算语言学协会。Jason Fong、Daniel Lyth、Gustav Eje Henter、Hao Tang 和 Simon King。2022年。Speech Audio Corrector：利用非目标说话人的语音，对字素输入文本转语音中的发音错误进行一次性纠正。载于 *Interspeech2022*，第1213–1217页。Jason Fong、Hao Tang 和 Simon King。2023年。Spell4TTS：利用声学信息引导的拼写改善文本转语音发音。载于第*12*届 *ISCA* 语音合成研讨会（*SSW2023*），第8–13页。Kenichi Fujita、Hiroshi Sato、Takanori Ashihara、Hi- roki Kanagawa、Marc Delcroix、Takafumi Moriya 和 Yusuke Ijima。2024年。基于适配器、以自监督语音表示模型为条件的抗噪零样本文本转语音合成。载于*2024*年

IEEE 声学、语音与信号处理国际会议（*ICASSP*），第11471–11475页。

Atticus Geiger、Hanson Lu、Thomas Icard 和 Christo- pher Potts。2021年。神经网络的因果抽象。载于神经信息处理系统进展，第34卷，第9574–9586页。Curran Associates, Inc.

Mor Geva、Jasmijn Bastings、Katja Filippova 和 Amir Globerson。2023年。剖析自回归语言模型中的事实关联回忆。载于*2023*年自然语言处理实证方法会议论文集，第12216–12235页，新加坡。计算语言学协会。

Asma Ghandeharioun、Avi Caciularu、Adam Pearce、Lu- cas Dixon 和 Mor Geva。2024年。Patchscopes：用于检查语言模型隐藏表征的统一框架。载于第*41*届国际机器学习会议论文集，第235卷，机器学习研究论文集，第15466–15490页。PMLR。

YitianGong、Botian Jiang、Yiwei Zhao、Yucheng Yuan、Kuangwei Chen、Yaozhou Jiang、Cheng Chang、Dong Hong、Mingshu Chen、Ruixiao Li、Yiyang Zhang、Yang Gao、Hanfu Chen、Ke Chen、Songlin Wang、Xiaogui Yang、Yuqian Zhang、Kexin Huang、ZhengYuan Lin等人。2026年。MOSS-TTS 技术报告。*arXiv* 预印本，arXiv:2603.18090。

Kyle Gorman、Lucas F. E. Ashby、Aaron Goyzueta、Arya McCarthy、Shijie Wu 和 Daniel You。2020年。SIGMORPHON 2020 多语言字素到音素转换共享任务。载于第 *17* 届 *SIGMORPHON* 语音学、音系学与形态学计算研究研讨会论文集，第 40–50 页，线上。计算语言学协会。

Chi-Yuan Hsiao、Ke-Han Lu、Kai-Wei Chang、Chih- Kai Yang、Wei-Chih Chen 和 Hung-yi Lee。2025年。口语语言模型端到端训练中灾难性遗忘的缓解策略分析。收录于 *Interspeech2025*，第 3234–3238 页。

Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。2022年。LoRA：大语言模型的低秩适配。收录于 国际学习表征会议。

Hangrui Hu、Xinfa Zhu、Ting He、Dake Guo、Bin Zhang、Xiong Wang、Zhifang Guo、Ziyue Jiang、HongkunHao、Zishan Guo、Xinyu Zhang、Pei Zhang、Baosong Yang、Jin Xu、Jingren Zhou 和 Junyang Lin。2026年。Qwen3-TTS 技术报告。*arXiv* 预印本，arXiv:2601.15621。

Ziyue Jiang、Qian Yang、Jialong Zuo、Zhenhui Ye、Rongjie Huang、Yi Ren 和 Zhou Zhao。2023年。Flu-entSpeech：面向口吃的基于上下文感知扩散模型的自动语音编辑。载于 计算语言学协会的研究发现：*ACL 2023*，第11655–11671页，多伦多，加拿大。计算语言学协会。Ayan Kashyap、Neil Shah 和 Vineet Gandhi。2025年。Prompt-to-Correct：借助语音提示进行测试时自动发音纠正。载于 *2025*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），第1–5页。Shuhei Kato。2025年。UtterTune：基于LoRA的多语言文本转语音中的目标语言发音编辑与控制。*arXiv*预印本，arXiv:2508.09767。

NikitaKoriagin、Georgii Aparin、NikitaBalagansky 和 Daniil Gavrilov。2026年。 使用稀疏自编码器解释和引导文本转语音语言模型。 *arXiv* 预印本，arXiv:2606.10029。

Robert Kubichek。1993年。 用于客观语音质量评估的梅尔倒谱距离度量。载于 *IEEE* 太平洋地区通信、计算机与信号处理会议论文集，第 1 卷，第 125–128 页。

Ki-Joong Kwon、Jun-Ho So 和 Sang-Hoon Lee。2025年。通过跨语言持续学习对低资源文本转语音进行参数高效微调。载于*Interspeech 2025*，第 1613–1617 页。

Anna Langedijk、Hosein Mohebbi、Gabriele S arti、Willem Zuidema 和 Jaap Jumelet。2024年。Decoder-Lens：对编码器-解码器变换器进行逐层解释。载于 计算语言学协会研究发现：*NAACL 2024*，第 4764–4780 页，墨西哥城，墨西哥。计算语言学协会。

Matthew Le、Apoorv Vyas、Bowen Shi、Brian Karrer、Leda Sari、Rashel Moritz、Mary Williamson、Vimal Manohar、Yossi Adi、Jay Mahadeokar 和 Wei-Ning Hsu。2023年。 Voicebox：大规模文本引导的多语言通用语音生成。收录于 神经信息处理系统进展，第36卷，第14005–14034页。Curran Associates, Inc.

Kyowoon Lee、Artyom Stitsyuk、Gunu Jho、Inchul Hwang 和 Jaesik Choi。2025年。 用于文本转语音模型韵律事后调整及发音错误纠正的反事实激活编辑。收录于 *Interspeech 2025*，第434–438页。

Sang-Hoon Lee、Seung-Bin Kim、Ji-Hyun Lee、Eun-woo Song、Min-Jae Hwang 和 Seong-Whan Lee。2022年。 HierSpeech：通过使用自监督表示进行语音合成的分层变分推断，弥合文本与语音之间的差距。收录于神经信息处理系统进展，第35卷，第16624–16636页。Curran Associates, Inc.

Yichong Leng、Zhifang Guo、Kai Shen、Zeqian Ju、Xu Tan、EricLiu、Yufei Liu、Dongchao Yang、Leying Zhang、Kaitao Song、Lei He、Xiangyang Li、Sheng Zhao、Tao Qin 和 Jiang Bian。2024年。PromptTTS 2：

使用文本提示描述和生成语音。发表于国际学习表征会议。

Kenneth Li、Oam Patel、Fernanda Viégas、Hanspeter Pfister 和 Martin Wattenberg。2023年。推理时干预：从语言模型中 eliciting 真实回答。载于 神经信息处理系统进展，第 36 卷，第 41451–41530 页。Curran Associates, Inc.

Kung-Yee Liang 和 Scott L. Zeger。1986年。使用广义线性模型进行纵向数据分析。*Biometrika*，73(1)：13–22。

Shijun Luo 和 Lizhi Wan。2026年。ASR 往返评估可能掩盖中文新闻文本转语音中的上下文相关和依赖惯例的阅读错误。 *arXiv* 预印本，arXiv:2608.10606。

Sihan Lv、Yechen Jin、Zhen Li、Jintao Chen、Jinshan Zhang、Ying Li、Jianwei Yin 和 Meng Xi。2026年。AST：自适应、无缝且免训练的精准语音编辑。 *arXiv* 预印本，arXiv:2604.16056。

Kevin Meng、David Bau、Alex Andonian 和 Yonatan Belinkov。2022年。在 GPT 中定位和编辑事实关联。载于 神经信息处理系统进展，第 35 卷，第 17359–17372 页。Curran Associates, Inc.

Kevin Meng、Arnab Sen Sharma、Alex Andonian、Yonatan Belinkov 和 David Bau。2023年。在 Transformer 中大规模编辑记忆。载于 国际学习表征会议。Masanori Morise。2017年。Harvest：一种从语音信号中提取基频的高性能估计器。载于*Interspeech2017*，第 2321–2325 页。Ankita Pasad、Chung-Ming Chien、Shane Settle 和 Karen Livescu。2024年。自监督语音模型对词语了解多少？计算语言学协会会刊，12：372–391。Ankita Pasad、Ju-Chieh Chou 和 Karen Livescu。2021年。自监督语音表征模型的逐层分析。载于 *2021*年 *IEEE* 自动语音识别与理解研讨会（*ASRU*），第 914–921 页。William Peebles 和 Saining Xie。2023年。基于 Transformer 的可扩展扩散模型。载于 *2023*年 *IEEE/CVF* 国际计算机视觉会议（*ICCV*），第 4172–4182 页。Puyuan Peng、Po-Yao Huang、Shang-Wen Li、Abdelrahman Mohamed 和 David Harwath。2024年。Voice-Craft：面向真实场景的零样本语音编辑与文本转语音。载于计算语言学协会第 *62* 届年会论文集（第 *1* 卷：长篇论文），第 12442–12462 页，泰国曼谷。计算语言学协会。

Vadim Popov、Ivan Vovk、Vladimir Gogoryan、Tasnima Sadekova 和 Mikhail Kudinov。2021年。Grad-TTS：一种用于文本转语音的扩散概率模型。载于 第 *38* 届国际机器学习会议论文集，第 139 卷，收录于机器学习研究论文集，第 8599–8608 页。PMLR。

Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。2023年。通过大规模弱监督实现稳健的语音识别。载于 第 *40* 届国际机器学习会议论文集，第 202 卷，收录于机器学习研究论文集，第 28492–28518 页。PMLR。

Takaaki Saeki、Detai Xin、Wataru Nakata、Tomoki Koriyama、Shinnosuke Takamichi 和 Hiroshi Saruwatari。2022年。UTMOS：UTokyo-SaruLab 在 VoiceMOS 挑战赛 2022 中的系统。载于 *Interspeech2022*，第 4521–4525 页。

Hiroaki Sakoe 和 Seibi Chiba。1978年。 口语词识别的动态规划算法优化。 *IEEE*《声学、语音与信号处理汇刊》，26(1)：43–49。

Jonathan Shen、Ruoming Pang、Ron J. Weiss、Mike Schuster、Navdeep Jaitly、Zongheng Yang、Zhifeng Chen、Yu Zhang、Yuxuan Wang、R. J. Skerry-Ryan、Rif A. Saurous、Yannis Agiomyrgiannakis 和 Yonghui Wu。2018年。 通过以梅尔频谱图预测结果为条件的 WaveNet 实现自然文本到语音合成。载于 *2018*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 4779–4783 页。Ayush Pratap Singh、Harshit Singh、Nityanand Mathur、Akshat Mandloi 和 Sudarshan Kamath。2026a。SonoEdit：用于基于大语言模型的文本到语音发音纠正的零空间约束知识编辑。载于简约性与学习会议，第 328 卷 机器学习研究论文集，第 1090–1100 页。PMLR。Harshit Singh、Ayush Pratap Singh 和 Nityanand Mathur。2026b。 FlowEdit：流匹配文本到语音中的终身发音适应联想记忆。载于 *Interspeech 2026*，第 5050–5054 页。Ramanan Sivaguru、Vasista Sai Lodagala 和 S. Umesh。2023年。SALTTS：利用自监督语音表征改进文本转语音合成。载于 *Interspeech 2023*，第 3033–3037 页。Ryosuke Sonobe、Shinnosuke Takamichi 和 Hiroshi Saruwatari。2017年。 JSUT 语料库：用于端到端语音合成的免费大规模日语语音语料库。 *arXiv* 预印本，arXiv:1711.00354。Richard Sproat。1996年。 用于文本转语音合成的多语言文本分析。 《自然语言工程》，2(4)：369–380。Jinuo Sun、Yang Xiao、Sung Kyun Chung、Qiuchi Hu、Gongping Huang、Eun-Jung Holden 和 Ting Dang。

2026年。 大型音频语言模型中的口音适配激活引导。载于 *Interspeech 2026*，第 4634–4638 页。

Jaesung Tae、Hyeongju Kim 和 Taesu Kim。2022年。EdiTTS：用于可控文本转语音的基于分数的编辑。载于 *Interspeech 2022*，第 421–425 页。

Paul Taylor。2009年。文本转语音合成。剑桥大学出版社。

telecomadm1145。2026年。Kiseki-TTS：一款采用 Mamba2解码器的快速日语文本转语音模型。Hugging Face 代码库，https://huggingface.co/telecomadm1145/Kiseki-TTS。

Francesco Verdini、Antonis Asonitis、Aref Farhadipour、Marzieh Razavi、Pierre-Edouard Honnet、Vijeta Avijeet 和 Juan Pablo ZuluagaGomez。2026年。通过激活引导控制自回归文本转语音的语速。 *arXiv* 预印本，arXiv:2609.33810。

Jesse Vig、Sebastian Gehrmann、Yonatan Belinkov、Sharon Qian、Daniel Nevo、Yaron Singer 和 Stu- art Shieber。2020年。使用因果中介分析研究语言模型中的性别偏见。载于神经信息处理系统进展，第33卷，第12388–12401页。Curran Associates, Inc.

Helin Wang、Meng Yu、Jiarui Hai、Chen Chen、Yuchen Hu、Rilin Chen、Najim Dehak 和 Dong Yu。2025年。SSR-Speech：迈向稳定、安全且鲁棒的零样本文本语音编辑与合成。载于*2025*年 *IEEE*声学、语音与信号处理国际会议（*ICASSP*），第1–5页。

Siyang Wang、Gustav Eje Henter、Joakim Gustafson 和 Éva Székely。2023年。阅读语音和自发语音文本转语音中的自监督语音表征比较研究。载于*2023*年 *IEEE*声学、语音与信号处理国际会议研讨会（*ICASSPW*），第1–5页。

Yuxuan Wang、Daisy Stanton、Yu Zhang、R. J. Skerry- Ryan、Eric Battenberg、Joel Shor、Ying Xiao、Fei Ren、Ye Jia 和 Rif A. Saurous。2018年。风格标记：端到端语音合成中的无监督风格建模、控制与迁移。载于 第*35*届国际机器学习会议论文集，第80卷，机器学习研究论文集，第5180–5189页。PMLR。

Tianxin Xie、Yan Rong、Pengfei Zhang、Wenwu Wang 和 Li Liu。2025年。 大语言模型时代的可控语音合成：系统综述。载于 *2025*年自然语言处理实证方法会议论文集，第764–791页，中国苏州。计算语言学协会。

Detai Xin、Shujie Hu、Chengzuo Yang、Chen Huang、Guoqiao Yu、Guanglu Wan 和 Xunliang Cai。2026年。LongCat-AudioDiT：波形潜空间中的高保真扩散文本转语音。*arXiv* 预印本，arXiv:2603.29339。

Qiantong Xu、Alexei Baevski 和 Michael Auli。2022年。简单而有效的零样本跨语言音素识别。载于 *Interspeech 2022*，第 2113–2117 页。

Mu Yang 和 John H. L.Hansen。2026年。 用于口音中和零样本文本转语音的激活引导。*arXiv* 预印本，arXiv:2603.05977。

Zhen Ye、Xinfa Zhu、Chi-Min Chan、Xinsheng Wang、Xu Tan、Jiahe Lei、Yi Peng、Haohe Liu、Yizhu Jin、Zheqi Dai、Hongzhan Lin、Jianyi Chen、Xingjian Du、Liumeng Xue、Yunlin Chen、Zhifei Li、Lei Xie、Qiuqiang Kong、Yike Guo 和 Wei Xue。2025年。Llasa：扩展基于 Llama 的语音合成的训练时与推理时计算。*arXiv* 预印本，arXiv:2502.04128。

Heiga Zen、Viet Dang、Rob Clark、Yu Zhang、Ron J. Weiss、Ye Jia、ZhifengChen 和 Yonghui Wu。2019年。LibriTTS：一个源自 LibriSpeech 的文本转语音语料库。收录于 *Interspeech2019*，第 1526–1530 页。

Song Zhang、Ken Zheng、Xiaoxu Zhu 和 Baoxiang Li。2022年。一种用于汉语普通话多音字消歧的多音字 BERT。收录于 *Interspeech2022*，第 436–440 页。

Wen Zhang。2023年。日语汉字中的发音歧义。载于计算与书面语言研讨会论文集（*CAWL 2023*），第50–60页，多伦多，加拿大。计算语言学协会。

Andy Zou、Long Phan、Sarah Chen、James Campbell、Phillip Guo、Richard Ren、Alexander Pan、Xuwang Yin、Mantas Mazeika、Ann-Kathrin Dombrowski、Shashwat Goel、Nathaniel Li、Michael J. Byun、Zifan Wang、Alex Mallen、Steven Basart、Sanmi Koyejo、Dawn Song、Matt Fredrikson及另外2位作者。2023年。表征工程：一种自上而下的人工智能透明性方法。*arXiv*预印本，arXiv:2310.01405。

A 声学交叉注意力掩码消融

表 A1 给出了第 5.6 节总结的四种条件完整消融结果。实体跨度来自 Whisper 词级时间戳（Radford 等人，2023年）；对于没有空格的语言，则进一步细分到字符。MEL 相似度是对实体跨度进行时间平均后得到的对数梅尔余弦相似度。MCD 经过方差归一化，并通过动态时间规整（DTW）（Sakoe 和 Chiba，1978年）对梅尔频率倒谱系数（MFCC）帧（Davis 和 Mermelstein，1980年）进行计算，c_0 被剔除；归一化会提高数值尺度。将日语数据限制为词元数量匹配的配对（$\Delta_{\text{tok}}{=}0$），结果相同：基线 $_{\text{mask}}$和 INJ_{mask} 再次得到 MCD $= 0$，且 MEL 相似度几乎相同。

语言	指标		交叉注意力 完整	交叉注意力 阻断在实体处
EN	MCD（dB）	↑	$\mathbf{35.31 \pm 0.27}$	0.00 ± 0.00
	MEL 相似度	↓	$\mathbf{0.88 \pm 0.01}$	1.00 ± 0.00
JA	MCD（dB）	↑	$\mathbf{31.73 \pm 0.77}$	0.00 ± 0.00
	MEL 相似度	↓	$\mathbf{0.94 \pm 0.01}$	1.00 ± 0.00
ZH	MCD（dB）	↑	$\mathbf{35.37 \pm 0.39}$	0.00 ± 0.00
	MEL 相似度	↓	$\mathbf{0.86 \pm 0.02}$	1.00 ± 0.00

表 A1：在 T5Gemma-TTS 上采用保长度均值池化注入的声学交叉注意力掩码消融（第5.6节）。完整条件比较 BL–INJ，阻断条件比较 BL_{mask}–INJ_{mask}，即在英语、日语和中文中，针对 Whisper 词级时间戳确定的实体跨度，对交叉注意力屏蔽实体位置后合成的同一组数据。屏蔽对实体位置的交叉注意力后，三种语言中的 BL–INJ差异均消失。数值为均值及 95% 正态近似置信区间半宽。这些指标反映声学差异，而非目标方向准确率。

语言	完整 BL	完整 INJ	在实体处阻断 基线_{mask}	在实体处阻断 INJ_{mask}
目标匹配率（实体 *PER* < 0.3），% ↑				
JA	20.00 $(+4.89/-4.13)$	$\mathbf{38.00}$ $(\mathbf{+5.61/-5.31})$	0.00 $+1.26$	0.00 $+1.26$
EN	1.00 $(+1.90/-0.66)$	2.00 $(+2.29/-1.08)$	0.00 $+1.26$	0.00 $+1.26$
实体 *PER* 均值与目标读音参考对比 ↓				
JA	0.569 ± 0.297	$\mathbf{0.416}$ $\pm \mathbf{0.273}$	0.885 ± 0.097	0.885 ± 0.103
EN	0.898 ± 0.380	0.917 ± 0.416	1.222 ± 0.480	1.222 ± 0.480

表 A2：T5Gemma-TTS 上的音素级交叉注意力掩码消融。与表 A1 不同，注入采用公式（1）中的拼接方式：对所有英语配对，以及源跨度和目标跨度词元数相同的 73 组日语配对，均在原位进行替换。在日语中，交叉注意力完整时，INJ 会使实体发音趋向目标读音；而阻断交叉注意力时则不会。目标匹配率表示实体 PER< 0.3 相对于目标读音参考文本的数值。数值以百分比表示，并附有 Wilson 95% 置信区间，格式为 （+上限/−下限）（上方），以及均值± 标准差（下方）。对于比例为零的情况，仅显示 Wilson 区间非负的上限。

B 交叉注意力掩码消融的音素级验证（日语）

第 5.6 节的声学消融报告了表A1（附录A）中四种交叉注意力掩码条件下的 MCD 和 MEL 相似度。在交叉注意力阻断条件下，英语、日语和中文的 MCD $= 0$，而 MEL 相似度几乎相同，这表明阻断实体位置交叉注意力后，注入的声学效应消失。我们还使用 300 组日语同形配对进行音素级验证，以补充这项声学证据；这些配对的书面句子保持不变，只有预期读音不同。作为交叉检验，我们也使用了 300 组英语同形配对。

流程。对于每一组，我们生成五种语音信号：基线、INJ、基线_{mask}、INJ_{mask}，以及目标读音参考TGT，该参考通过直接合成目标读音句子获得，用于表A2中的分析。

结果。 表A2采用匹配基线比较逻辑：每项干预均与其匹配基线进行比较。在日语中，完整的交叉注意力在音素层面带来了适度的目标导向改进：与基线相比，INJ 在目标匹配率和相对于目标参考文本的 PER 两项指标上均有所提升。当实体位置的交叉注意力被阻断时，BL_{mask} 和 INJ_{mask} 的目标匹配率相同，PER 也几乎相同，这表明注入的编码器状态不再影响实体发音。在英语中，INJ 相较于 BL 并未带来可检测的提升，且两种屏蔽条件完全相同。

解读。结合表格中的声学 MCD 和梅尔频谱相似度结果，A1，表格A2支持传输路径解释。日语完整条件表明，传输的信号可以使输出在音素层面向目标文本靠拢，而阻断实体位置的交叉注意力则会消除声学和音素层面的差异。因此，表格A2应被视为交叉注意力路径的证据，而非独立的发音控制基准。

C 日语 160 项阅读基准

数据与合成。表A3 总结了日语朗读基准测试，其中包含八个类别的 160 个对抗性语境项目。每个项目提供一个汉字表记、语境预期读音、竞争读音，以及完整句子的预期读音/备选读音刺激。我们使用冻结的 T5Gemma- TTS 合成了基线、正确读音参考文本，以及

类别	项目	BL 正确	BL 错误	已纠正 / 错误
多音字	43	30	12	12/12
语境数字	34	25	8	8/8
戏剧特定读音	24	17	7	7/7
送假名	35	23	12	12/12
熟字训	6	4	1	1/1
动词变位	6	4	1	1/1
音训混合复合词	6	5	1	1/1
依领域而定的音读／训读	6	2	3	3/3
总计	160	110	45	45/45

表 A3：日语语境读音基准。BL 正确和 BL 错误分别统计基线合成中读作预期读音和竞争读音的项目，不包括有歧义的读音；已纠正／错误统计 BL 错误的项目中，经拼接注入后读作预期读音的数量。所有严格意义上的 BL 误读均通过拼接注入得到纠正，且所有纠正后的输出均符合直接包含判据，而非 CER 回退方案。

测量指标	值
听众	20
项目	50
判断	500
选择 INJ	493（98.6%）
选择基线	3（0.6%）
两者均未选中	4（0.8%）
INJ选择率（95% 置信区间）	98.6（97.4–99.6）
GEE z（p）	10.37（$< 10^{-16}$）
达到目标文本的项目级多数	50/50
捕捉试验通过率	100%
排除的听者	0

表 A4：对日语同表面纠正集的人工评估。听众几乎在每次判断中都选择了注入音频片段。INJ选择率为无条件比率，附带项目级自助法置信区间，并使用与表1相同的二项分布广义估计方程与随机机会进行检验。

错误读音参考文本，采用 JSUT 语音克隆、随机种子 42 和标准采样配置：top-k 采样，搭配k=30；top-p 采样，搭配 p=0.9；温度为 0。8。进行注入时，将预期读音的片假名形式编码，并将其完整实体跨度拼接到源编码器输出中。

读音判定。`Parakeet-Inc/furigana_whisper_small_jsut`[4] 直接将语音转录为假名。经过平假名归一化后，仅包含预期读音或仅包含竞争读音的转录文本会被标记为相应类别。其余情况则依据整句假名 CER 较低的一方进行判定；差异小于0.02 时标记为有歧义。五个有歧义的 BL 项目被排除在纠正分母之外。

该基准特意增加了低频读音的比例；BL 错误 $= 45/160$ 不应被视为自然文本错误率。为验证判定流程，我们还对 320 条极性已知的参考样本进行了评分（160 条合成结果采用预期读音，160 条采用竞争读音）。该流程对 317/320 条（ 99.06% ）返回了预期标签，没有出现极性反转，其余三条被标记为有歧义。

D 同表层纠正集的人工评估

配对比较测试的同形集包含 50 个日语项目，其中基线错误读出了目标词，而书面句子保持不变。其中 38 个项目来自 `ja-tts-g2p-bench`[5]，这是一个面向文本转语音的日语汉字读音基准，以知识共享署名-相同方式共享 4.0 发布，同时也提供附录C中的 160 个项目。这些项目在数据收集前经过筛选：首先从该基准的严格基线错误中出发，剔除问题无法得出唯一答案、音频退化或词语与其他项目重复的项目，然后从维基百科[6] 关于难读日语地名和姓氏的文章中选取困难地名和姓氏，将集合补充至 50 个；这些项目的误读和纠正均经过自动验证。刺激材料使用冻结的 T5Gemma-TTS 和 JSUT[7] 声音克隆提示词合成，遵循附录 C。最终构成为 9 个异读词、 8 个语境数字、 9 个送假名项目、 6 个特定戏剧读音、7 个地名、5 个罕见姓氏，以及来自较小类别的 6 个项目。

实验流程与第4.2节中的姓名集相同：20位母语听者、两个各含25个项目的区块、每个区块包含三个捕获试次，以及每个项目十次判断。表A4 报告了结果。听者几乎在每次判断中都选择了注入音频片段；按项目统计的多数选择在全部50个项目上都与目标一致，并且没有听者符合任何排除标准。

[4] Parakeet-Inc/假名标注_Whisper_小型_JSUT

[5] kawajiri-tellernovel/ja-tts-g2p-bench[6] 维基百科[7] JSUT 语料库

E 常规发音控制基线

表A5比较了 SpeechSplice 与作用于文本或模型权重的干预方法在跨架构基准测试的300组日语配对上的表现；各条件使用相同的实体跨度、说话人参考音频和自动语音识别流水线。基线合成原始句子；字素到音素基线将实体替换为自动预测的读音；文本提示保留实体，并在括号中附上以汉字书写的目标姓氏；ROME/MEMIT 则使用秩一模型编辑（ROME）（Meng等人，2022年）和 Transformer中的大规模记忆编辑（MEMIT）（Meng等人，2023年）编辑编码器权重。SpeechSplice 在每个样本中都注入了目标实体状态。五种条件下的音频均通过语音克隆合成，并依据自动语音识别转录文本评分，因此目标读音匹配（TRM，定义见下文）无法与表3中的 PIB 相比较；该表比较的是未使用语音提示合成的语音的实体音素错误率。

当多路径发音提取器应用于自动语音识别转录文本后，其至少一条路径中出现经过标准化的目标读音时，TRM 就将样本标记为正确。G2P-CER Match 也会接受某个样本：只要自动语音识别转录文本的 pykakasi 读音中存在一个窗口，且目标读音与该窗口的字符错误率不超过 0.30：

$$\mathrm{TRM} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\big[\,\mathrm{norm}(r_n^{*}) \in P(\hat{y}_n)\,\big] \tag{4}$$

$$\mathrm{G2P\text{-}CER} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\big[\,t_n \vee \min_{w} \mathrm{CER}(g_{n,w},\, \rho_n) \le 0.30\,\big] \tag{5}$$

这里，Eqs. (4) 和 (5) 中的 $\hat{}y_n$ 是样本 n 的自动语音识别转录文本，r_n^{*} 是其目标读音，N 是样本数量，t_n 是 Eq. (4) 中的指示函数，$P(\hat{}y)$ 是一个结合 pykakasi 读音与原始转录文本的多路径发音提取器，$\rho_n = \mathrm{norm}(r_n^{*})$，而 $g_{n,w}$ 是 norm(G2P($\hat{}y_n$)) 中长度与 ρ_n 相同的窗口，其中 G2P 是 pykakasi 读音。数值以百分比表示，附有威尔逊 95%区间，显示为 (+上限/−下限)。

G2P 结果为空是这一构造所特有的：源文本和目标文本是不同的实体，因此，保留输入内容的普通前端无法生成目标实体。文本提示部分揭示了目标读音，而上游权重编辑在任何样本中都无法恢复目标读音。

条件	TRM (%)↑	字素到音素字符错误率匹配 (%)↑
基线	0.00 +1.26	0.00 +1.26
字素到音素基线	0.00 +1.26	0.00 +1.26
文本提示	51.67 (+5.60/−5.64)	52.33 (+5.59/−5.65)
ROME/MEMIT	0.00 +1.26	0.00 +1.26
SpeechSplice	**98.67 (+0.81/−2.04)**	**99.00 (+0.66/−1.90)**

表 A5：在跨架构基准测试的日语配对上，与文本层面和权重层面的发音控制基线进行比较。在此划分中，直接干预编码器输出比文本改写、文本提示或编码器权重编辑更常恢复目标读音。对于比例为零的情况，仅显示 Wilson 区间的非负上界。

F 整句字符错误率

Whisper-INJ 检查转录文本中是否出现目标读音，因此会将每个样本简化为二元结果。作为连续性检验，我们计算自动语音识别转录文本相对于源句子和目标句子的字符错误率；对于英语，这些是附录 I 中 Whisper medium.en 生成的转录文本。表 A6 报告了 300 对基准测试的四种组合。

基线转录文本与源句子相近，与目标句子相距甚远；在三种语言中，注入都会使这两种距离反转。这一结论基于衡量整个句子的指标，而非单个子字符串，因此不会受到包含匹配的盲点影响。注入也保留了句子的其余部分：在日语和中文中，注入后的字符错误率（相对于目标句子）与基线字符错误率（相对于源句子）相近。英语中的同一比较则显示出更大的差距。

每种语言都使用各自的识别器和归一化方式，因此不同语言的字符错误率数值不可直接比较。英语统一转换为小写并保留字母数字字符；日语使用 pykakasi 转换参考文本后保留片假名；中文则保留汉字。置信区间为1.96× 300 个配对的标准误。

G 人工评估详情

本附录记录了第 4.2 节中的两项听力研究：招募和报酬、知情同意和数据处理、向参与者展示的完整说明，以及界面和质量控制。

转录文本与参考文本	EN	JA	ZH
基线与源文本 ↓	0.008	0.047	0.041
基线与目标文本 ↑	0.250	0.209	0.245
INJ 与目标文本 ↓	0.016	0.050	0.045
INJ 与源文本 ↑	0.255	0.215	0.246

表 A6：表 1 中名称替换对的整句字符错误率（95% 置信区间半宽为0。005 至 0。018）。在所有语言中，基线转录文本都接近源句子，注入转录文本则接近目标句。转录文本应与参考文本匹配时，数值越低越好；对于两行控制数据，数值越高越好。各语言使用的识别器和归一化方式不同，因此各列数值不可比较。

G.1 配对比较测试的统计检验

INJ 选择率采用二项广义估计方程（GEE）与随机机会（50%）进行比较（Liang 和 Zeger，1986年），将项目视为聚类单位，采用可交换工作相关结构，并使用聚类稳健标准误（双侧，经 Bonferroni 校正）$\alpha = 0.0125$。中文条件下，每项判断都选择了 INJ，因此 GEE 完全分离，其 Wald z 无法定义；我们报告预先指定的备用分析，即对 50 个项目级多数结果进行精确二项检验。

G.2 招募与报酬

英语和中文听众通过 Prolific[8] 招募，日语听众则通过 Lancers[9]，招募，后者是一个日本众包平台。Prolific 的筛选条件要求英语（英语条件下的配对比较测试和平均意见分研究）或普通话（中文条件）为第一语言，且获批率为 95–100%；英语研究仅面向英国和美国居民，中文研究则不限制居住地。日语招募说明要求参与者为日语母语听众。招募说明在参与者报名之前告知任务内容、预计时长和报酬，并要求参与者使用耳机。平台标识符和各研究排除名单用于避免参与者在同一研究中重复参与。

所有实际时薪均高于 Prolific 的每小时 8.00 美元最低标准，Lancers 的时薪则超过东京最低工资（每小时 1,226 日元）的两倍。所有纳入分析的提交均已获批准并支付报酬。十份英语配对比较测试提交因我方的研究配置错误而收到了错误的刺激材料区块；这些提交均已全额支付报酬，并已从分析中排除。

G.3 知情同意与数据处理

参与过程遵循两个平台的标准同意模式：招募文本说明了任务、时长、报酬和研究目的，只有接受这些条款的工作者才参与其中。参与完全自愿，且可随时退出。我们未使用单独的知情同意书。登录时，我们收集了平台参与者标识符、年龄段、性别类别和聆听设备。在 Prolific 上，标识符会从研究网址自动填入，且不会显示姓名字段，以遵守平台禁止收集真实姓名的规定。Lancers 表单还包含一个标记为姓名的字段。Lancers 上的委托方只能看到匿名工作者标识符，因此参与者在两个字段中都填写了该标识符，这两个字段仅用于匹配完成码和付款。我们没有存储真实姓名。我们没有收集联系信息、参与者的音频或自由文本回答。回答存储在只有研究人员可访问的研究服务器上，仅以汇总数据形式报告，不会公开参与者层级的原始回答。这些标识符仅用于分配、去重、质量控制和付款匹配。

G.4 伦理审查

作者所在机构没有机构审查委员会，也未获得正式批准或豁免认定。这些研究属于最低风险在线聆听任务：所有刺激材料均为作者生成的合成语音；任务不涉及欺骗或敏感内容；两个平台均仅允许成年人参与；参与完全自愿，且任务内容事先已明确告知；研究未收集任何个人身份信息；支付报酬高于 Prolific 的最低工资标准，或在 Lancers 上高于东京最低工资。

G.5 向参与者展示的说明

以下英文说明为逐字原文。两项研究的日语和中文说明也均按原文照录，并附有英文释义，见human_eval/instructions_verbatim.md（https://github.com/Kuri-san-jp/speechsplice）中的已发布代码库。配对比较测试的英语和中文说明，以及

[8]Prolific[9] Lancers

日语和中文平均意见分说明由作者翻译，未经专业译者独立验证。

配对比较测试（姓名集）。英语版本标题为“发音匹配测试”：

> 1. 您将依次听到两个音频片段。2. 第一个声音播放完后，→ 2 秒静音 →，第二个声音会自动播放。3. 屏幕顶部会显示一个姓名。请选择哪个音频片段读出了这个姓名。4. 请根据姓名的发音进行判断。请勿考虑声音的自然度或音频质量。5. 如果两个音频片段都没有读出显示的姓名，请选择“都没有读出显示的姓名”。6. 您将先完成 2 道练习题。练习题的回答不会计入结果。

每轮都会询问“哪个音频读出了显示的姓名？”，选项为第一个声音/ 都没有读出显示的姓名/ 第二个声音。练习页面还会显示“这是练习。您的回答不会计入结果。”

日语和中文版本（在我们的词汇表中，这两个版本的标题均为“姓名听辨测试”）呈现相同的六个项目、问题、选项和练习说明，但日语版的第 4 项写道：“仅根据假名序列（读音）进行判断。请勿考虑音高重音、声音的自然度或音频质量。”

配对比较测试（日语同表层集合）。同形研究采用了上述日语说明，并进行了四处更改（译注）：标题改为“单词读音听辨测试”；第3项改为“对于句子中的一个单词，请选择以屏幕所示读音读出该词的音频片段”；第5项指向中央按钮；每个试次询问“哪个音频片段以所显示的读音读出了该词？”

自然度MOS。英文说明：“您将评估语音样本的自然度和声学质量。首先，请完成这个简短的练习环节。”随后是以下步骤：“点击‘播放’按钮收听音频。/ 收听完整的音频片段。/ 使用五点评分量表评价自然度和声学质量。/ 点击‘下一步’继续查看下一个样本。”随后显示一条醒目的警告内容为：“注意：请勿评判发音是否正确。请不要评估专有名词的发音是否正确。仅评价语音的声音质量、流畅度和自然度。”评分量表定义如下： 5 = 完全自然、质量高的语音； 4 =略显不自然，但质量高； 3 = 有些不自然，但可以理解； 2 =明显不自然，听起来较困难； 1 =非常不自然，质量低；按钮标签为 1 差 / 2 较差 / 3 一般 / 4 良好 / 5 优秀。每次试验都会重复该警告，并要求在继续之前收听完整的音频片段。

日语和中文的平均意见分说明是作者对同一文本的翻译，其中包括警告和五项量表定义。

G.6 界面与质量控制

两项研究均在浏览器应用程序中进行，流程如下：填写附录G.3中的四个字段登录、显示说明页面、进行两次练习试次、进行正式试次，最后显示完成码屏幕。配对比较测试中，两段音频片段会自动依次播放，中间间隔两秒静音；在两段音频片段均播放完毕之前，答题按钮始终处于禁用状态；在平均意见分研究中，下一步按钮会在音频片段播放完毕之前保持禁用。每个会话的项目顺序由服务器端随机打乱；每个配对比较测试试次中的基线/INJ 位置由服务器端均匀随机抽取，并与回答一同记录。界面不显示条件标签；音频文件名包含条件信息，但只有检查页面时才能看到。

捕获试次重复使用真实的刺激语音，并显示一个两个音频片段都未说出的第三个词；正确答案是都不是 按钮。通过率分别为 100%（英语）、98.33%（日语姓名）、100%（中文）和 100%（日语同形），且没有听者在三个捕获试次中失败超过一次。MOS研究不包含捕获试次。在MOS研究中，每种语言的 100 个音频片段被分成两个包含 25 组配对的 50 个音频片段区块；日语和中文的每个区块均由 10 名听者评分，英语则分别由 9 名和 11 名听者评分，因此每种语言获得 1,000 次评分。

H 听力研究揭示的失败案例

多数听者与 Whisper 的判断在八个姓名项目上存在分歧。其中七个是识别错误：听者听到了目标读音，但转录文本中没有出现该读音，因为识别器

转写成了一个发音相似的词。这些项目在自动评分中被计为失败，但语音本身是正确的，因此我们将自动评分视为保守的。

第八项是反向情况，也更能说明问题。注入音频片段呈现了目标读音，但多出了一个拍；大多数听众选择了两者皆非，而不是接受它。相比之下，子串匹配会在较长形式中找到目标读音，并将该项判为成功。这正是自动评分会漏掉的失败模式，也是配对比较测试提供两者皆非选项的原因。

I 计算与模型详情

模型与参数量。所有文本转语音模型和辅助模型均使用公开发布的冻结检查点，但附录E 中的 ROME/MEMIT 基线会编辑编码器权重；不进行微调。表A7 列出了这些模型。T5Gemma-TTS 使用 XCodec2（Ye 等人，2025年）作为其音频分词器和声码器。我们根据`wav2vec2-xlsr-53-espeak-cv-ft`[10]（Xu 等人，2022年）的国际音标（IPA）转录文本计算所有 PER 和 PIB 值。

检查点与工具。文本转语音检查点包括 T5Gemma-TTS[11],SpeechT5[12], LongCat-AudioDiT[13],Irodori-TTS[14], Kiseki-TTS[15], Qwen3-TTS[16], XTTS-v2[17], Llasa-3B[18],CSM-1B[19], 和 Bark[20]。说话人相似度使用 WavLM-SV[21]。Whisper large-v3 使用贪心语言强制解码。未提供语音提示的转录文本时，T5Gemma-TTS 语音克隆会使用 Whisper large-v3-turbo 转录提示音频[22]。表1 中的 JA Whisper-INJ 结果以及表3中双实体列的结果使用 Furigana Whisper[23]，，附录F 中的英语字符错误率则使用另一组转录文本

[10]facebook/wav2vec2-xlsr-53-espeak-cv-ft[11] Aratako/T5Gemma-TTS-2b-2b[12] microsoft/speecht5_文本转语音[13] meituan-longcat/LongCat-AudioDiT-1B [14] Aratako/Irodori-TTS-v4.1-Small[15] telecomadm1145/Kiseki-TTS[16] Qwen/Qwen3-TTS-12Hz-1.7B-Base[17] coqui/XTTS-v2[18]HKUSTAudio/Llasa-3B[19] sesame/csm-1b[20]suno/bark[21] microsoft/wavlm-base-plus-sv[22] openai/whisper-large-v3-turbo[23] Parakeet-Inc/furigana_Whisper_小型_JSUT

模型	Role	#参数量
T5Gemma-TTS-2b-2b	主文本转语音模型	5.31B
SpeechT5（文本转语音）	（编码器–解码器）对比文本转语音	144M
LongCat-AudioDiT-1B	（编码器–解码器）对比文本转语音	1.26B
Irodori-TTS-v4.1-小型	（编码器–解码器）对比文本转语音	766M
Kiseki-TTS	(编码器–解码器)对比文本转语音	411M
Qwen3-TTS-12Hz-1.7B	（编码器–解码器）对比文本转语音	1.93B
XTTS-v2	（仅解码器）对比文本转语音	441M
Llasa-3B （Ye 等人，2025年）	（仅解码器）对比文本转语音	3.41B
CSM-1B	（仅解码器）对比文本转语音	1.55B
Bark	（仅解码器）对比文本转语音	1.09B
XCodec2（Ye 等人，2025年）	（仅解码器）音频分词器 / 声码器	830M
VoiceCraft	零样本文本转语音	840M
EdiTTS	基于分数的内容替换	1480万
Whisper large-v3	多语言自动语音识别	15.4亿
Whisper large-v3-turbo	语音提示评估	809M
Whisper medium.en	英语字符错误率转录文本	762M
wav2vec2-xlsr-53-espeak-cv-ft (Xu 等人，2022年)	国际音标音素转写	316M
UTMOS22	平均意见分预测器	103M

表 A7：本文使用的模型、其角色以及根据已发布权重计算的参数量。对于文本转语音系统，参数量不包括用于生成波形的音频编解码器、自编码器或声码器。

来自 Whisper medium.en[24]。我们使用 F_0 和 pyworld 提取[25] Harvest（Morise，2017年）处理有声帧。针对不同语言的文本处理，我们采用了适合相应任务的音素化和归一化工具，包括 g2p_en[26],pykakasi[27], 和 pypinyin[28]。

每个系统替换的张量。表3 每个系统替换一个张量，该张量选取为生成器读取的文本侧状态。在五个编码器–解码器系统中，这是到达交叉注意力的编码器输出：T5Gemma-TTS 和 SpeechT5 中编码器的最后一个隐藏_状态_；在 LongCat- AudioDiT 中，则是交叉注意力输入，它将层归一化嵌入加回层归一化后的

[24]openai/whisper-medium.en[25]pyworld 包[26]字素到音素_英语包 (Kyubyong/g2p)[27] pykakasi 包 (miurahr/pykakasi)[28] pypinyin 包 (mozillazg/python-pinyin)

最终编码器状态；在 Irodori-TTS 中，则是经过 768→512 投影及其归一化后的文本状态；在 Kiseki-TTS 中，则是经过最终归一化后分发给解码器交叉注意力的键和值的编码器输出。因此，最后三个并非层输出本身。在 Irodori-TTS 中，替换最后一层的输出（层扫描所采用的做法）会得到相同的音频，因为归一化和投影都是对每个词元分别进行的。在五个仅含解码器的系统中，被替换的张量是预填充期间自注意力序列中的一个隐藏状态，取自实体词对应的行：XTTS-v2 中第 30 个 GPT 块里的第 29 个，Llasa-3B 中 28 个解码器层里的第 27 层，CSM-1B 中 16 个主干层里的第 15 层，Bark 中 24 个语义 Transformer 层里的第 23 层，以及 Qwen3-TTS 中 28 个说话人层里的第 27 层，计数方式如第4.2节所述。第5.4节会改变这一选择。

生成长度。 T5Gemma-TTS 合成的音频片段长度在解码前即已固定：其时长估算器根据输入文本和语音提示设置以秒为单位的时长。对于 INJ，我们传入根据源句子估算的时长，即 BL 使用的时长，因此两个音频片段具有相同的时长，无需先合成 BL。表 1 和2 中的 T5Gemma-TTS 运行均采用了这种方式。

计算。 所有实验仅进行推理，并在单个 NVIDIA A100 40GB GPU 上运行（单实体 Irodori-TTS 运行使用 H100 80GB）。合成耗时约 275 个 GPU 小时，评分耗时约 40 个 GPU 小时。

J 双实体测试中的配对差异与语音质量

配对准确率差异。 为比较表 3 中的 INJ（一个）和 INJ（两个），同时避免将句子的两个槽位视为相互独立，我们对每个句子的两个槽位判定结果取平均值，并计算每个句子的 INJ（两个） − INJ（一个）之差。平均差值以百分点表示，并根据 300个句子（10,000次重采样）计算 95%百分位数自助法区间，结果为 -0.50 $[-2.17, +1.17]$，$+0.33$ $[-2.33, +3.00]$，以及 -0.83 $[-4.67, +3.00]$，分别对应 T5Gemma-TTS EN、JA 和 ZH；$+1.33$ $[-0.83, +3.50]$ 对应 SpeechT5 EN；-1.17 $[-2.67, +0.33]$ 以及 $-0.17[-0.83, +0.50]$ 对应 LongCat-AudioDiT EN 和 ZH；$+0.33$ $[-1.33, +1.83]$ 对应 Irodori-TTS JA；以及 -0.83 $[-2.17, +0.50]$ 对应 Kiseki-TTS JA。每

模型	语言。	UTMOS ↑ BL	INJ（一个）	INJ（两个）	INJ（两者）− BL
编码器–解码器（交叉注意力）					
T5Gemma-TTS	英语	3.92	3.93	3.90	-0.0145 ± 0.0572
	日语	3.35	3.32	3.31	-0.0408 ± 0.0475
	中文	3.05	3.07	3.08	$+0.0319 \pm 0.0960$
SpeechT5	英语	4.20	4.19	4.15	-0.0574 ± 0.0269
LongCat-AudioDiT	英语	3.45	3.46	3.47	$+0.0147 \pm 0.0545$
	中文	2.55	2.53	2.52	-0.0337 ± 0.0459
Irodori-TTS	日语	3.77	3.76	3.77	-0.0037 ± 0.0324
Kiseki-TTS	日语	2.42	2.42	2.42	$+0.0021 \pm 0.0496$
仅解码器（自注意力）					
Qwen3-TTS	英语	4.52	4.51	4.51	-0.0030 ± 0.0026
	日语	4.36	4.36	4.36	-0.0003 ± 0.0107
	中文	3.73	3.72	3.72	-0.0101 ± 0.0218
XTTS-v2	英语	4.23	4.23	4.23	-0.0007 ± 0.0088
Llasa-3B	英语	3.59	3.59	3.61	$+0.0222 \pm 0.0364$
	中文	2.80	2.83	2.84	$+0.0403 \pm 0.0415$
CSM-1B	英语	4.21	4.21	4.20	-0.0054 ± 0.0253
Bark	英语	3.67	3.67	3.67	$+0.0019 \pm 0.0092$
	日语	2.47	2.47	2.47	$+0.0012 \pm 0.0123$

表 A8：表 3 中双实体句子的 UTMOS，按模型和语言分别列出基线（BL）以及替换一个或两个实体（INJ）的结果。最后一列为 INJ（两个）与 BL 之间的配对差值；±：95%置信区间。替换两个实体后，各种条件下的 UTMOS 均接近基线；SpeechT5 EN 和 Qwen3-TTS EN 的降幅很小。

每个区间都包含零；我们没有检验等效性。在仅解码器条件下，每个句子的两列值都为零。

语音质量。 表 3 中两个实体的列只对发音评分。为检查在一个句子中替换两个实体是否会降低语音质量，我们使用表 1所用的 UTMOS 预测器，对双实体测试中的每个音频片段进行评分。INJ（一个）是每个句子中替换一个实体的两个音频片段的平均分。最后一列是每种条件下 300 个句子的 INJ（两个）与基线之间的平均配对差值。表A8显示，所有条件下这一差值都很小；只有 SpeechT5 英语和 Qwen3-TTS 英语的区间不包含零，且 UTMOS 略有下降。UTMOS 是一种自动估算指标，我们没有对这些音频片段进行听测。