arXiv:2611.29985v1 [cs.SD] 23 Nov 2026

无需合成即可评估二语发音

冻结流匹配文本转语音中的反事实速度对比

S

Zehan Liu 北京交通大学

摘要

流匹配文本转语音（TTS）模型学习母语说话人如何实现文本内容，但这类知识仅用于合成。我们表明，冻结的文本转语音模型无需生成音频，即可评估非母语者（二语）发音。对于学习者的语句，我们在同一加噪输入上，比较模型在目标文本和反事实文本（删除或交换）条件下的速度预测误差；二者之差表现为一种生成式发音质量分数。该信号仅集中在一小段高噪声时间步范围内，原生语音标准可定位这一范围，而闭式规则则可将其迁移到不同骨干网络。在说话人不重叠的 speechocean762 数据集上，四个冻结式文本转语音模型与人工评分的相关系数达到 r=0.58–0.63，处于零样本自动语音识别和音素级基线的范围内；将这些模型通过无标签融合加入上述基线后，r 由 0.698 提升至 0.757。缓存的帧级分数还可在不增加文本转语音成本的情况下标记词级错误。

关键词：发音评估、流匹配、文本转语音、计算机辅助发音训练、零样本评估。

1 引言

自动发音评估（APA）用于评定非母语（二语）说话人朗读给定文本的表现，是计算机辅助发音训练（CAPT）的核心。GOPT [1]等有监督评分器虽然准确，但每种语言和每类学习者群体都需要专家评定的二语语音。零样本评分器则通过复用基于母语语音训练的模型来避免这项成本：发音优度（GOP）会比较声学模型 [2, 3], 给出的音素似然或后验概率；较新的方法会读取 CTC 后验概率或对数几率 [4, 5], ；Liu 等人 [6] 则统计自监督模型未能恢复的掩码单元。它们全都是判别式方法：关注音频对文本或其单元的预测效果。

文本转语音模型回答的是相反的问题：给定文本，母语者听起来会是什么样？迄今为止，CAPT 仅将这些模型用作生成器：要么生成带有误差的训练语音 [7]，要么合成与学习者音频对齐的母语者“标准”参考语音 [8]。后一种方法每条提示都需要完整合成一次，而且要比较两个不同的声音。

在视觉领域，文本到图像扩散模型已被转化为零样本分类器，而无需生成任何
事物：最能帮助模型对图像去噪的提示胜出 [9, 10], ，而用无文本条件下的误差进行归一化，可使不同输入的分数具有可比性 [11]。我们想知道

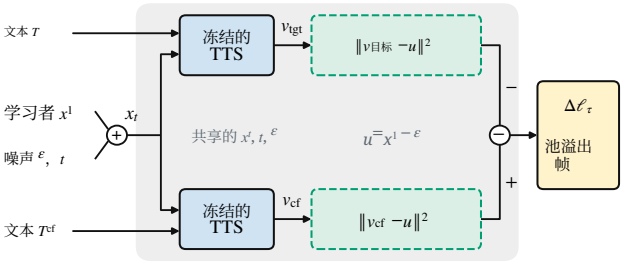


图1： 概览。学习者的加噪梅尔频谱图x_t 在目标文本T和反事实文本T_{cf}的条件下，分别通过同一个冻结式文本转语音模型。两种速度误差之间的逐帧差值 $\Delta\ell_\tau$ 被汇总为话语或词语评分。

同样的结论是否也适用于语音，其任务是等级评分，而非在多个类别中进行选择。一个流匹配文本转语音模型 [12, 13, 14, 15, 16] 在母语语音上训练，以预测文本条件速度场；它在学习者 语音上的预测误差能否反映发音质量?

直接使用时，它并不能发挥作用：原始误差与人工评分不相关（r<0.06），因为它主要受信号能量、说话人和通道影响。通过三种选择，可以将其转化为可用的分数：在相同的加噪输入上对目标文本和反事实文本进行对比，仅在语音帧上进行汇总，并评估高噪声时间步的一个窄频带。我们的贡献如下：

- **无合成评分器**。冻结式文本转语音模型的一次批量前向传播即可生成话语和词语评分；一个精确的恒等式展示了该对比所消除的内容。
- **可迁移的时间步规则**。针对母语语音的信噪比准则可定位有用的时间步，而等信噪比映射可将其迁移到特征尺度不同的骨干网络，无需重新调参。
- **互补性的证据**。在未见说话人上，该分数与零样本基线方法相当；在无标签融合中，它能进一步提升基线方法的效果，并且会响应文本编辑，但不会响应增益或静音。

2 方法

2.1 作为生成式GOP的速度对比

令$x_1 \in \mathbb{R}^{L\times D}$ 表示学习者朗读文本T时的梅尔频谱帧。采用最优传输路径的流匹配将数据与噪声 $\varepsilon \sim \mathcal{N}(0,\ I)$混合，

$$x_t = (1-t)\,\varepsilon + t\,x_1, \qquad u = x_1 - \varepsilon, \qquad t \in (0,1), \quad (1)$$

并训练$v_\theta(x_t,\ t,\ T)$来预测速度u。其直观原理（图 1）是：如果学习者像母语者一样发音T，那么已知T 有助于模型预测u；如果语音偏离T，文本的帮助就会减少。我们通过将这种帮助

与反事实文本T_{cf}进行对比来衡量。使用在相同$(x_t,\ t\ \varepsilon)$上计算的$v_{\text{tgt}} = v_\theta(x_t,\ t,\ T)$和$v_{\text{cf}} = v_\theta(x_t,\ t,\ T_{\text{cf}})$，第 τ 帧处的对比为

$$\Delta\ell_\tau(t,\varepsilon) = \tfrac{1}{D}\left(\|v_{\text{cf},\tau} - u_\tau\|_2^2 - \|v_{\text{tgt},\tau} - u_\tau\|_2^2\right). \quad (2)$$

共享噪声抽样遵循 [9]：误差本身在不同抽样之间变化很大，而差值的变化则小得多。

我们使用两种反事实。**无条件** 会移除文本，模型支持这种做法，因为训练时会随机丢弃文本以实现无分类器引导 [17]；这与 [11] 的无条件归一化相呼应。**交换** 则使用另一条语音的转录文本，其长度相差不超过两个字符，构成如 [9] 所述的竞争性假设。

式 (2) 具有 GOP 的形式 [2, 3], : 规范文本下的似然与另一种文本下的似然之对数比。对于理想的去噪器，对数似然等于所有噪声水平上的去噪误差积分的负值（相差一个常数） [18], , 而速度误差是经过重新缩放的去噪误差，$\|\hat{x}_1 - x_1\|_2^2 = (1-t)^2\|v_\theta - u\|_2^2$。因此，若以适当的权重对$t$ 积分，差值就能估计 $\log p(x_1|T) - \log p(x_1|T_{\text{cf}})$。我们使用少量时间步并赋予相同权重，因此我们的分数是似然比的替代指标，而非其估计值。

2.2 对比消除了什么

令 $\delta_v = v_{\text{tgt},\tau} - v_{\text{cf},\tau}$ 表示文本引起的预测速度变化。将式（2）展开可精确得到：

$$D\,\Delta\ell_\tau = 2\langle \delta_v,\, u_\tau - v_{\text{cf},\tau}\rangle - \|\delta_v\|_2^2. \quad (3)$$

分数是反事实残差在文本诱导方向上的投影，减去该方向的大小。首先，主导原始误差的能量项 $\|u_\tau\|_2^2$ 会完全抵消。其次，v_{cf} 已经根据x_t（音色、增益、通道）预测出的内容会从残差中消除；其余部分只有沿 δ_v 方向的分量才计入分数。这一点是经验性的，将在第 4.3 节进行检验。第三，没有任何因素能保护δ_v 本身：强噪声或语速变化会削弱它，而包含它的分数也会随之降低。

2.3 话语和词语评分

我们对M 个固定时间步t_m （使用成对的噪声采样 ε_m）以及由基于能量的语音活动检测器确定的语音帧 $\mathscr{A}$（均方根值在峰值 25 分贝范围内的帧）取平均：

$$S(x_1,T) = \frac{1}{M|\mathscr{A}|}\sum_{m=1}^{M}\sum_{\tau\in\mathscr{A}}\Delta\ell_\tau(t_m,\varepsilon_m). \quad (4)$$

所有M 次噪声采样的两条分支作为一个大小为 $2M$ 的批次运行，无需常微分方程求解器和声码器。在一个词的时间跨度内对缓存的逐帧值取平均，即可得到词分数，无需进一步进行文本转语音计算。

2.4 哪些时间步承载信号

所使用的时间步很重要，这一点与图像分类类似 [9, 10]。当 $t\to 1$ 时，x_t 几乎是干净的语音，两条分支都会复制局部声学特征，而 $\delta_v\to 0$。当$t\to 0$ 时，x_t 几乎是纯噪声，所携带的学习者音频信息很少。在 150 条带标签话语（第 3 节）上进行的初步实验将 F5-TTS 的有效窗口确定为$t \in [0.05, 0.14]$。

无需二语标签也能找到相同的窗口。在未标注的母语 LibriSpeech 音频片段 [19] 上，我们计算蒙特卡罗对比信噪比，

$$\text{MC-SNR}(t) = \frac{\mathbb{E}_{x,\varepsilon}\left[\Delta\ell(x,t,\varepsilon)\right]}{\sqrt{\mathbb{E}_x\left[\text{Var}_\varepsilon\left(\Delta\ell(x,t,\varepsilon)\right)\right]}}, \quad (5)$$

相对于噪声采样之间的离散程度，文本对母语语音的帮助。在t =0.08 时达到峰值，在 $[0.05, 0.15]$ 上保持在峰值的 80% 以上，并且在各时间步上与 L2 相关性曲线保持一致（无条件分支的皮尔逊相关系数为 0.883，交换分支为 0.874；图 2a、b）。

在一个骨干网络上找到的窗口不能直接照搬，因为不同骨干网络对特征的缩放方式不同：ZipVoice 会将梅尔特征乘以 0.1 [16]。当特征尺度为s时，x_t的信噪幅度比为$ts/(1-t)$。令其在骨干网络A和B 之间相等，可得

$$t_B = \frac{t_A/s}{1 - t_A + t_A/s}, \qquad s = s_B/s_A, \quad (6)$$

这会将 $[0.05, 0.14]$ 映射为 $[0.345, 0.619]$ ，对应s=0.1。我们将此窗口直接用于 ZipVoice，不进行调优。在一个包含 500 条话语的诊断子集（25 位说话人）上，照搬 F5-TTS 的窗口得到r=−0.154，而映射后的窗口得分为 0.481（同一子集上的 F5-TTS：0.384）；在测试集上，映射后的窗口得分达到 0.624（图 2c）。

3 实验设置

数据。我们使用 speechocean762 测试集 [20]：包含 125 位普通话母语者的 2,500 条话语，并附有专家给出的准确度、流利度、韵律和总分（0–10）。从 30 位说话人中分层抽取的 150 条话语子集（dev_150）用于所有试验方案的选择。结果报告基于其余 95 位说话人（未见过_说话人_1900；51 位成人、44 位儿童）。

文本转语音骨干网络。我们测试了四种冻结模型，使用 M=8和固定随机种子：F5-TTS v1 和 v0（336M，DiT，使用 100k 小时 Emilia [15, 21]训练）以及 E2-TTS（336M [14]）在 $t \in [0.05, 0.14]$ 上运行；ZipVoice 仅使用 585 小时 LibriTTS 训练（1.23亿 [16,22]），在映射窗口上运行。禁用音频提示。

基线方法。所有模型均为冻结模型，且不使用二语标签。(i) 我们重新实现了 [6] 的掩码单元评分器（HuBERT-Base，第 7 层，100 个聚类）。(ii) Whisper-large-v3-turbo [23] 在教师强制下的转录文本对数概率。(iii) Wav2Vec2-Base [24] 和 HuBERT-Large [25] 的字符级 CTC 分数。(iv) wav2vec 2.0 国际音标识别器 [26] 给出的音素级分数：logit 边际 GOP [5] 和无对齐 CTC 分数。(v) 黄金语音 [8]：通过动态时间规整将 F5-TTS 合成语音与学习者的音频对齐。

指标与融合。我们报告与人工评分的皮尔逊相关系数r以及通过对说话人进行自助法（2,000 次重采样，因此最小p值为 0.001 ）得出的 95% 置信区间。每位说话人贡献 20 个话语，因此语料库层面的r 混合了说话人排序与同一说话人话语排序；因此，我们还加入说话人内相关系数，计算方法是从分数和预测值中减去每位说话人的均值。为在没有标签的情况下融合评分器，我们将每个分数映射到其在 150 个开发集 _150 话语上的经验累积分布函数（仅使用分数，不使用评分），然后对映射值取平均（全部 600 个开发集说话人话语所得结果相同）。

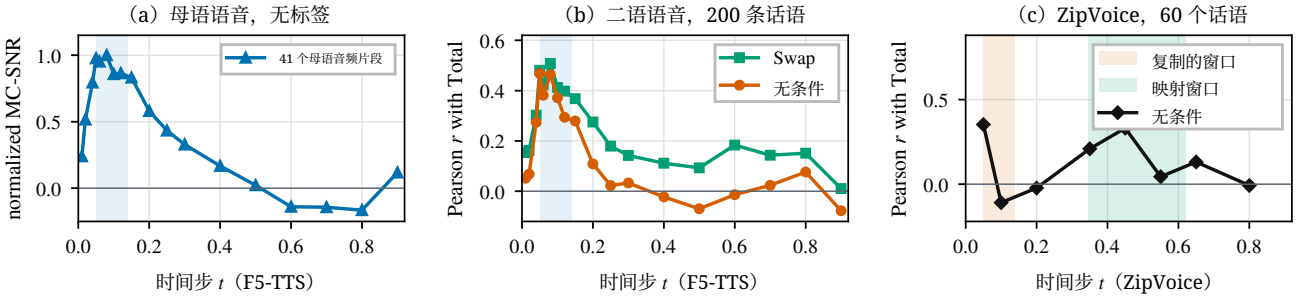


图 2： 在单个时间步上评估对比时，与人工评分的相关性。(a) 对比在 41 个母语者 LibriSpeech 音频片段上的 MC-SNR，并归一化至其峰值；未使用二语数据。(b) 在 200 条未见过的二语话语上，与总计的皮尔逊r；阴影标示 F5-TTS 窗口 $[0.05, 0.14]$。(c) ZipVoice 在 60 条话语上的相同诊断：复制的窗口未覆盖有用的时间步；根据公式 (6) 映射的窗口则覆盖了这些时间步。

表1： 说话人互斥的*speechocean762* 上的话语级结果（未见过的_说话人_1900：1,900条话语、95名说话人，其中51名成人和44名儿童）。r：与人工评分的皮尔逊相关系数；方括号：95%的说话人自助法置信区间；W-spk：总计上的说话人去均值r。除150标签对照组外，没有系统使用二语训练标签。粗体：每个区块中的最佳结果。

系统	参数量	准确率r	总计r [95% 置信区间]	说话人去均值r	成人 r	儿童 r
控制措施						
F5-TTS v1，原始误差（无对比）	336M	-.010	.003 [-.17，.18]	-.004	-.038	-.022
ZipVoice，原始误差（无对比）	123M	.044	.058 [-.13, .23]	.006	.027	-.030
ZipVoice，仅对非语音帧进行对比	123M	.201	.201 [.14, .25]	.020	.277	.079
长度回归：时长、字符数、速率、年龄（150 个标签）	—	.587	.627 [.53, .69]	.226	.714	.293
零样本判别和自监督学习基线						
HuBERT-Base 掩码单元错误数 [6]	94M	.591	.622 [.54，.68]	.236	.692	.304
Whisper-large-v3-turbo，转录文本对数概率。	809M	.542	.557 [.45，.64]	**.287**	.628	.361
Wav2Vec2-Base，字符CTC	95M	.535	.554 [.48，.62]	.215	.689	.342
HuBERT-Large，字符CTC	316M	.595	.623 [.55, .69]	.235	**.748**	**.385**
音素GOP，logit间隔 [5]	316M	.575	.599 [.50, .67]	.139	.702	.302
音素CTC，无对齐	316M	**.639**	**.665 [.57, .73]**	.191	.743	.340
我们的方法：冻结的流匹配文本转语音，速度对比（无自动语音识别）						
F5-TTS v0，交换，所有帧	336M	.572	.596 [.51, .66]	.164	.714	.307
E2-TTS，交换，所有帧	336M	.562	.590 [.51, .65]	.153	.710	.301
F5-TTS v1，无条件，语音活动检测	336M	.562	.575 [.49, .64]	.209	.701	.307
F5-TTS v1，交换，语音活动检测	336M	.602	**.626 [.55, .68]**	.187	**.742**	.334
ZipVoice，无条件，语音活动检测	123M	**.606**	.624 [.55, .68]	**.225**	.716	.357
ZipVoice，交换，语音活动检测	123M	.592	.613 [.54, .67]	.218	.715	**.382**
无标签融合（经验累积分布函数归一化分数的平均值）						
2-TTS：ZipVoice + F5-TTS v1	459M	.675	.701 [.62, .76]	.266	.804	.412
2-Disc：音素CTC +Whisper	1.1B	.670	.693 [.61，.75]	.294	.798	.420
4-Disc：2-Disc + HuBERT-Large + Wav2Vec2	1.5B	.671	.698 [.61，.76]	.303	.813	.425
2-Disc + F5-TTS v1	1.5B	.717	.743 [.67，.80]	.322	.840	.456
2-Disc + ZipVoice	1.2B	.720	.745 [.67，.80]	.340	.836	.479
2-Disc + ZipVoice + F5-TTS v1 (**4-Hybrid**)	1.6B	**.731**	**.757 [.69，.81]**	**.344**	**.847**	**.481**

4 结果

4.1 冻结式文本转语音模型能为发音打分吗？

这种对比对于获得可用的分数既必要又充分。没有这种对比，速度误差不包含任何信息（表1，$r \le 0.058$）。采用这种对比后，四种骨干模型的六种配置均达到总计r=0.58–0.63，处于零样本自动语音识别和音素基线方法的范围内（0.55–0.67）；无对齐音素CTC仍是表现最佳的单一模型（0.665）。ZipVoice 的参数量仅为 F5-TTS 的三分之一，训练数据不到其 1%，但通过映射窗口仍能取得同样出色的表现。将 ZipVoice 限定在语音帧上有助于交换（+0.021，p=0.003），但对无条件设置帮助不显著（+0.017，p=0.066）；仅在非语音帧上的对比效果较弱（0.201）。

语料库层面的 r 可轻易根据长度获得，因此我们将其作为控制变量。仅基于时长、字符数、语速和年龄进行回归，便可得到 r=0.627。掩码单元数量对于 [6] 达到 0.622，但它与时长相关，除以长度后降至 0.025。我们的分数并非仅由长度决定：在七个声学协变量之外，它们还能额外解释总计方差的 12.2%（ZipVoice）和 10.6%（F5-TTS），而掩码单元数量只能解释 0.8%。在说话人内部，VAD掩码TTS分数（0.19– 0.23）处于各基线方法（0.14–0.29）的范围内，其中 Whisper 最高。所有系统在儿童语音上的表现都弱得多（文本转语音：0.30–0.38；基线方法：0.30–0.39）。

4.2 它能为判别式评分器带来额外价值吗？

生成式分数能够在判别能力更强的模型无法发挥作用的地方提供帮助。从音素CTC加上Whisper（0.693）开始，再添加两个判别式模型几乎不会改变结果（0.698）。添加一个文本转语音模型后，得分达到0.743（F5- TTS）或0.745（ZipVoice）；添加两个模型后则达到0.757（+0.060 高于四模型判别式融合，95%置信区间为 $[0.038, 0.085]$，$p \le 0.001$）。说话人内r 从

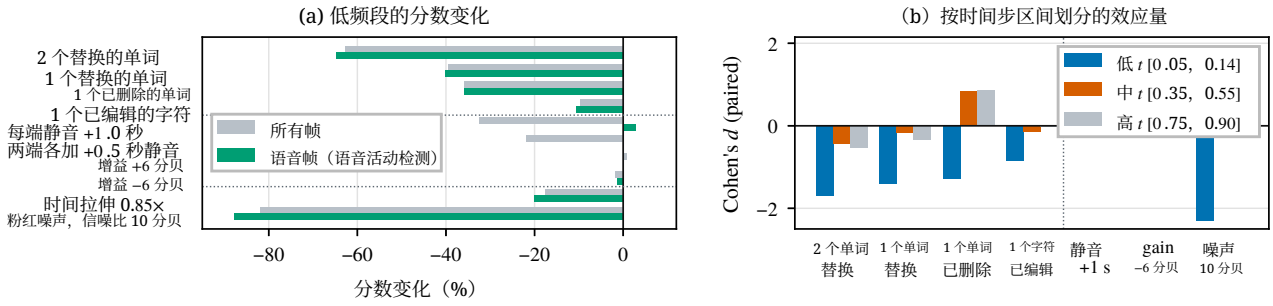


图 3： 对来自 15 位未见说话人的 100 条话语进行受控干预（F5-TTS v1，交换）。(a) 在低-t 时间步区间内，使用所有帧或仅使用语音帧时分数的平均变化。虚线将文本编辑、干扰因素变化以及模型范围之外的变化分隔开。(b) 三个时间步区间内的配对效应量（语音帧）。

0.303上升至0.344（p=0.003），且每个年龄组都出现了增益：12岁以下（0.424→0.476）、12–17岁（0.326→0.403）和成人（0.813→0.847）。仅使用两个文本转语音模型、不使用自动语音识别时，得分也达到0.701。排序融合（0.638→0.714）与150标签回归（0.713→0.744）的结果一致。

该贡献在控制协变量后仍然存在。 我们以总计为因变量，回归使用十三个协变量（七个声学协变量、五个自动语音识别和音素分数，以及掩码单元数量），并加入全部 95 位说话人的固定效应，同时使用按说话人聚类的标准误。两个文本转语音分数仍具有显著性（ZipVoice：t=4.24，p=5.3×10^{-5}；F5-TTS：t=3.16，p=0.002），而非语音帧对照项则不显著（t=−0.08）。

4.3 分数对什么因素有响应?

它响应的是文本，而不是录音电平。 我们对 100 段话语的文本或音频进行编辑（配对威尔科克森检验；图 3a）。将一个单词替换为另一个长度相同的单词，会使分数降低 40.1%（d=−1.40）；替换两个单词会使分数降低 64.8%，替换一个字符则会使分数降低 10.6%。增加 +6 分贝增益不会改变分数（+0.1%，p=0.78），而增加 −6 分贝会使分数变化 −1.3%。在两端各添加一秒静音，会使全帧分数降低 32.5%，但语音帧分数不受影响（+2.8%，p=0.13）。预测的局限也显现出来：时间拉伸 15% 会使分数降低 20–29%，而信噪比为 10 分贝的粉红噪声会使分数降低 88%。

只有选定的时间步会如预期般响应。 词语替换和字符编辑在低频t 频带中的效应量，是在其他频带中的 3 到 8 倍；而在中频和高频频带，删除一个单词甚至会提高分数（d ≈+0.85；图 3b）。

反事实的选择并不重要。在五种反事实文本中（F5-TTS，300 条话语，15 位未见说话人），即使不使用 F5-TTS 的词元长度掩码，空字符串的总计 r=0.543 仍为 0.399（控制转录文本长度和时长后）。随机交换是唯一一种其分数与转录文本长度相关的变体（−0.240）；匹配长度后，这种相关性便消失（−0.044），而对三个长度匹配的交换结果取平均，准确度最高（r=0.619）。

4.4 成本、黄金语音与词级使用

与黄金语音相比。 在表 2 中，将学习者语音与合成语音对齐，在梅尔特征上效果不佳（r=0.111），但在HuBERT特征上表现良好（0.737；在此子集上选择的层），优于这里所有单一对比得分（0.586– 0.655）。代价是需要合成：使用F5-TTS时，每条语句耗时160–375毫秒，而我们的分数仅需31.7毫秒（M=8），后者

表 2： 黄金语音（我们对 [8] 的改编）与对比：涵盖 15 位未见说话人的 300 条话语。单条话语的延迟，测试设备为一台 RTX 5090（批量大小为 1，音频平均时长为 3.99 秒）。

系统（参数量）	准确率 r	总计 r/ρ	ms	RTF
黄金语音：使用 F5-TTS v1 合成，然后进行动态时间规整				
16 步 +梅尔动态时间规整 (336M)	.087	.111/.138	138.3	.0347
16 步 +HuBERT-L12 动态时间规整 (651M)	.688	.737/.754	159.8	.0401
32 步 +HuBERT-L24 动态时间规整 (651M)	**.697**	**.739**/.718	375.3	.0941
16 步，参考音频已预先合成	.688	.737/**.754**	**18.4**	**.0046**
我们的方法：无需合成				
F5-TTS v1，交换，M=4 (336M)	.497	.525/.546	20.0	.0050
F5-TTS v1，交换，M=8 (336M)	.558	.586/.602	31.7	.0079
ZipVoice、无条件、M=8 (1.23亿)	.635	.655/.610	97.2	.0244

表3： 词级别、未见说话人。准确度为 ≤ 5（12,697 个单词中的 902 个；95%的说话人置信区间）的单词的 AUROC，以及 Top-1 最差单词准确度（515 个话语）。

系统	检测AUROC [95% CI]	Top-1
复制到单词的话语得分	.805 [.765，.832]	—
Wav2Vec2 字符CTC	.819 [.793，.844]	.493
我们的方法：缓存的F5-TTS v1帧对比		
无条件	.811 [.784, .831]	.497
Swap	.826 [.798, .847]	.454
无条件 + Wav2Vec2	.859 [.836，.879]	.540
交换 + Wav2Vec2	**.864 [.840，.883]**	**.542**

只需一个模型，且无需自动语音识别。对于带有预合成参考语音的固定提示，黄金语音是更好的选择（18.4毫秒）；对于开放文本，对比方法的速度快5–12倍。不同噪声种子下的分数稳定（在M=8处的组内相关系数为0.846，在M=32处为0.959）。

词错误。我们使用 CTC强制对齐器（表3）得到词跨度，并对 F5-TTS 缓存的帧对比分数进行汇总。在检测方面，交换的 AUROC 达到 0.826，高于将话语得分复制到每个词的基线（0.805，p=0.022），与 Wav2Vec2 CTC 相当（0.819）。在定位方面，无条件与 Wav2Vec2 相当（0.497 对比 0.493），而交换的得分较低（0.454）。在所有情况下，与 Wav2Vec2 的 z 分数取平均都有帮助（AUROC 0.864，+0.045，p≤0.001；Top-1 为 0.542）。

在没有人工评分的语料库中，该分数可以区分 LibriSpeech 中的母语语音与低分二语话语（AUROC 0.978），也可以区分 CommonAccent 中带口音的语音与母语语音 [27]（0.976；Wav2Vec2：0.904）。

5 局限性与结论

这个分数衡量的是与文本转语音模型先验的一致程度，而非语音学真值：在 F5-TTS 合成的 80 个音频片段中，它

会漏掉模型自身的字素到音素错误。它需要语速正常的干净音频，以及大小写正确的文本（全大写文本会使 F5-TTS 降至r=0.073）。该窗口最初是根据 150 条带标签的话语选定的，之后又在一个骨干网络上验证了无标签的信噪比标准。单凭这个分数，其表现并未超过最佳的音素级基线；所有人工评分结果均来自一个语料库，且母语背景也只有一种。

在这些限制下，冻结式文本转语音模型的速度对比是一种无需合成、类似 GOP 的发音分数，可作为基于自动语音识别的评分器的补充。

参考文献

[1] Y. Gong、Z. Chen、I.-H. Chu、P. Chang 和 J. Glass，“基于 Transformer 的多方面多粒度非英语母语说话人发音评估”，载于*IEEE* 国际声学、语音与信号处理会议论文集，2022年，第 7262–7266 页。

[2] S. M. Witt 和 S. J. Young，“面向交互式语言学习的音素级发音评分与评估”，*Speech Communication*，第 30 卷，第 2–3 期，第 95–108 页，2000 年。

[3] W. Hu、Y. Qian、F. K. Soong 和 Y. Wang，“使用深度神经网络训练的声学模型和基于迁移学习的逻辑回归分类器改进发音错误检测”，*SpeechCommunication*，第 67 卷，第 154–166 页，2015 年。

[4] A. K. Parikh、C. Tejedor-Garcia、C. Cucchiarini 和 H. Strik，“利用音系知识增强基于 CTC 的发音错误检测中的 GOP”，载于 *Interspeech* 会议论文集，2025年，第 5068–5072 页。

[5] ——，“评估用于发音错误检测的基于logit的GOP评分”，载于 *Interspeech* 会议论文集，2025年，第 2405–2409 页。

[6] H. Liu、M. Shi 和 Y. Wang，“零样本自动发音评估”，载于*Interspeech* 会议论文集，2023年，第 1009–1013 页。

[7] D. Korzekwa、J. Lorenzo-Trueba、T. Drugman 和 B. Kostek，“计算机辅助发音训练——语音合成几乎就是你所需要的一切”，语音通信，第 142 卷，第 22–33 页，2022年。

[8] T.-H. Lo、M.-T. Tsai、Y.-T. Sung 和 B. Chen，“作为黄金语音生成器的零样本文本转语音：一种系统化框架及其在自动发音评估中的适用性”，载于*ISCA SLaTE* 论文集，2025年，页码 26–30。

[9] A.C. Li、M. Prabhudesai、S. Duggal、E.Brown 和 D. Pathak，“你的扩散模型实际上是一个零样本分类器”，载于 *IEEE/CVF* 国际计算机视觉大会论文集，2023年，页码 2206–2217。

[10] K. Clark 和 P. Jaini，《文本到图像扩散模型是零样本分类器》，载于*NeurIPS* 论文集，2023年，第 58 921–58937 页。

[11] B. Krojer、E. Poole-Dayan、V. Voleti、C. Pal 和 S. Reddy，《扩散模型是视觉与语言推理器吗？》，载于 *NeurIPS* 论文集，2023年，第 8385–8405 页。[12] Y. Lipman、R. T. Q. Chen、H. Ben-Hamu、M. Nickel 和 M. Le，《用于生成建模的流匹配》，载于*ICLR* 论文集，2023年。[13] M. Le、A. Vyas、B. Shi、B. Karrer、L. Sari、R. Moritz、M. Williamson、V. Manohar、Y. Adi、J. Mahadeokar 和 W.-N. Hsu，《Voicebox：大规模文本引导的多语言通用语音生成》，载于*NeurIPS* 论文集，2023年，第 14 005–14034 页。

[14] S. E. Eskimez、X. Wang、M. Thakker、C. Li、C.-H. Tsai、Z. Xiao、H. Yang、Z. Zhu、M. Tang、X. Tan、Y. Liu、S. Zhao 和 N. Kanda，“E2 TTS：简单易用的全非自回归零样本文本转语音”，载于*IEEE* 口语语言技术研讨会论文集，2024年，页码 682–689。

[15] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen，“F5-TTS：一位以流匹配生成流畅且忠实语音的童话讲述者”，发表于 计算语言学协会年会论文集，2025年，第 6255–6271 页。

[16] H. Zhu, W. Kang, Z. Yao, L. Guo, F. Kuang, Z. Li, W. Zhuang, L. Lin, and D. Povey，“ZipVoice：基于流匹配的快速、高质量零样本文本转语音”，发表于 *IEEE* 自动语音识别与理解研讨会论文集，2025年。

[17] J. Ho and T. Salimans，“无分类器扩散引导”，发表于 *NeurIPS* 深度生成模型与下游应用研讨会，2021年。

[18]X. Kong, R. Brekelmans, and G. Ver Steeg，“信息论扩散”，发表于 *ICLR* 论文集，2023年。

[19] V. Panayotov、G. Chen、D. Povey 和 S. Khudanpur，《LibriSpeech：基于公有领域有声读物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论文集，2015年，第 5206–5210 页。

[20]J. Zhang、Z. Zhang、Y. Wang、Z. Yan、Q. Song、Y. Huang、K. Li、D. Povey 和 Y. Wang，《speechocean762：用于发音评估的开源非英语母语者英语语音语料库》，载于 *Interspeech* 会议论文集，2021年，第 3710–3714 页。

[21] H. He、Z. Shang、C. Wang、X. Li、Y. Gu、H. Hua、L. Liu、C. Yang、J. Li、P. Shi、Y. Wang、K. Chen、P. Zhang 和 Z. Wu，《Emilia：用于大规模语音生成的广泛、多语言且多样化的语音数据集》，载于 *IEEE* 口语语言技术研讨会论文集，2024年。

[22] H. Zen、V. Dang、R. Clark、Y. Zhang、R. J. Weiss、Y.Jia、Z. Chen 和 Y. Wu，“LibriTTS：一个从 LibriSpeech 派生、用于文本转语音的语料库”，载于 *Interspeech* 论文集，2019 年，第 1526–1530 页。

[23]A. Radford、J. W. Kim、T. Xu、G. Brockman、C. McLeavey 和 I. Sutskever，《通过大规模弱监督实现鲁棒语音识别》，载于 *ICML* 会议论文集，2023年，第 28 492–28518 页。

[24] A. Baevski、Y. Zhou、A. Mohamed 和 M. Auli，《wav2vec 2.0：用于语音表征自监督学习的框架》，载于 *NeurIPS* 会议论文集，2020年，第 12 449–12460 页。[25] W.-N. Hsu、B. Bolte、Y.-H. H. Tsai、K. Lakhotia、R. Salakhutdinov 和 A. Mohamed，《HuBERT：通过对隐藏单元进行掩码预测实现自监督语音表征学习》，发表于 *IEEE/ACM* 音频、语音与语言处理汇刊，第 29 卷，第 3451–3460 页，2021年。[26] Q. Xu、A. Baevski 和 M. Auli，《简单有效的零样本跨语言音素识别》，载于 *Interspeech* 会议论文集，2022年，第 2113–2117 页。[27] J. Zuluaga-Gomez、S. Ahmed、D. Visockas 和 C. Subakan，《CommonAccent：探索基于 Common Voice 的口音分类大型声学预训练模型》，载于 *Interspeech* 会议论文集，2023年，第 5291–5295 页。