合成孪生体：缩小土耳其语自动语音识别评估中文本转语音测试集与真实录音之间的差距

Nur Yağmur Tuncer数据科学与创新 *Ata TechnologyPlatforms*伊斯坦布尔，土耳其
yagmur.tuncer@atptech.com

Mustafa Sertaç Türkel数据科学与创新 *Ata Technology Platforms*伊斯坦布尔，土耳其
sertac.turkel@atptech.com

Fatma Nur Korkmaz数据科学与创新 *Ata TechnologyPlatforms*伊斯坦布尔，土耳其
fatmanur.korkmaz@atptech.com

Ahmet Tuğrul Bayrak数据科学与创新 *Ata TechnologyPlatforms*伊斯坦布尔，土耳其
tugrul.bayrak@atptech.com

arXiv:2611.30694v1 [cs.CL] 23 Nov 2026

摘要——**文本转语音引擎合成的语音是录音测试集的一种低成本替代方案，但对于估计识别器的真实错误率而言，现成的合成测试集的价值仅相当于几十份真实录音。本文考察真实录音的哪些属性在合成语音中有所缺失，以及这种差距能够缩小多少。针对 507 份土耳其语 Common Voice 录音，构建了**合成孪生体**：一组阶梯式合成副本，每次增加一个因素，包括说话人的音高、同一说话人的克隆、真实录音通道以及语速。在每一步中，都会测量自动语音识别在副本和真实录音上的行为一致性。在三种零样本克隆引擎以及由六个和 19 个识别器组成的识别器组中，说话人克隆是提升“哪些句子较难”这一判断一致性的关键步骤；对于 Chatterbox，它将斯皮尔曼相关系数从 .41 提高到 .55。当合成语音比真实语音更容易识别时，匹配信道或加入通用噪声可以校正错误水平。价值最高的孪生体相当于 70 至 122 份真实录音，具体取决于识别器组；相比之下，默认语音约相当于 9 份录音。在 VoxForge 上进行的复现实验得出了相同结论。VoxForge 是第二个语料库，其中说话人身份是已知的，而非推断得出：克隆将句子一致性从 .31 提高到 .50，并将价值从 8 份录音提高到约 50 份。对于“哪些句子较难”的一致性不同于对于“句子有多难”的一致性。本文最后给出了合成自动语音识别测试集的实用方案和报告检查清单。**

索引词—**语音识别、评测、文本转语音、合成测试数据、语音克隆、土耳其语**

I. 引言

评估自动语音识别（ASR）需要录制测试语音，而收集这些语音的成本高昂，取决于说话人的同意，并且对于大多数语言而言都很稀缺。文本转语音（TTS）提供了另一种选择：任何文本都可以以较低成本朗读出来，因此可以针对新领域、罕见现象或新语言生成测试集 [1], [2]。这类测试集的可靠性则不那么明确。合成语音会产生误报，即在真实语音中不会出现的错误，在21%至34%的情况下会出现 [2], ；而在合成音频上表现近乎完美的模型，在真实语音上可能表现不佳 [3]。

以往工作主要证明了差距的存在。本文将其量化为一个合成测试集所具有的真实录音数量，并探究为什么合成语音提供的信息少于真实语音，即它缺少真实录音的哪些属性，以及如何弥合这一差距。本文有三项贡献。首先，本文引入了合成孪生体，这是一组受控干预构成的阶梯，使每个合成话语逐步接近其真实对应项，在说话人音高、说话人身份、录音通道和时序方面逐渐实现匹配；该阶梯按照四个保真度层级进行评测，并通过合成数据集所具有的真实录音数量加以衡量。其次，利用三个克隆引擎、由 6 名和 19 名识别器组成的面板、一个通用噪声对照以及一个包含已知说话人的第二语料库，本文表明，匹配说话人能够复现部分句子级难度模式，而信道和噪声会改变错误水平。第三，本文针对合成语音上的自动语音识别评估提出了一套实用方案和报告检查清单。

II. 相关工作

CrossASR [1] 使用文本转语音生成用于差分测试的测试用例，Lau 等人 [2] 表明，以这种方式发现的许多故障都是误报，并训练了一个分类器对其进行过滤。2M-BELEBELE [4] 发现，当用 MMS-TTS 渲染结果替代人工 FLEURS 录音时，即使对于高资源语言，三个自动语音识别模型的排名也不可靠。WildASR [3] 观察到，从合成语音到真实儿童语音，错误数量增加了近六倍，因此避免使用文本转语音。

针对土耳其语，Tuncer 等人 [5] 表明，来自九种文本转语音引擎的合成数据集能够保持识别器排名，但其价值最多相当于 31 条真实录音。这些工作确立了这一差距，但没有对其进行分解。

Minixhofer等人 [6] 通过话语级说话人、韵律和环境统计量量化合成语音与真实语音之间的距离，并通过让文本转语音模型以真实值为条件，使该距离缩小了约 10%。Rossenbach 等人 [7] 考察了文本转语音模型的选择如何影响合成

发表于 ISAS 2026。© 2026 IEEE。允许个人使用本材料。对于所有其他用途，包括在任何当前或未来的媒体中重新印刷或重新发布本材料以用于广告或宣传目的、创建新的合集作品、转售或再分发至服务器或列表，或在其他作品中重复使用本作品的任何受版权保护的组成部分，均须获得 IEEE 的许可。

用于自动语音识别的训练数据。两项研究都没有探讨固定识别器在测试集的合成副本上的表现是否相同；本文将这一性质称为 评测 保真度，并按句子进行测量。与本文阶梯最接近的是，Lebryk等人 [8] 克隆了基准说话人的音色，发现模型在克隆音色上的行为比在通用声音上的行为更接近其在真实录音上的表现。他们的目标是检测基准特定优化，既没有匹配信道或时序，也没有测量句子级一致性。针对土耳其语临床自动语音识别，Tuncer 等人 [9] 使用合成语音进行训练，并在真实录音上确认了这些收益。

预测驱动推断 [10], [11] 将一小组金标准标签与大量低成本预测相结合，并以有效样本量量化这些预测的价值。本文采用同样的视角来表达合成测试集在真实录音中的价值。

III. 数据与度量

实验使用了来自 Common Voice 17 [12], 土耳其语测试划分的 743 条语句，这些语句按照固定种子抽取，作为平衡多源土耳其语基准 TurkASR- Bench [13]的一部分。贡献者使用自己的设备进行录音，因此不同语句之间的信道差异很大。第四节在 VoxForge [14], 土耳其语部分的 507 条朗读语句上重复了部分孪生实验，其中说话人身份已知。

每条话语的参考文本均由三个零样本引擎朗读：XTTS-v2 [15], 、Chatterbox [16] 和 Vox- CPM2 [17], ；每个引擎既使用内置音色，也使用真实说话人的音色克隆。每条合成话语都保留其对应真实录音的标识符，因此每次比较都按句子配对。合成测试集[1] 和合成孪生体[2] 均已公开。

所有音频均由 19 个开放式识别器转写：Whisper系列的八种规模 [18] ，最高至 large-v3-turbo；三种土耳其语 Whisper微调模型；五种土耳其语wav2vec2 和 XLS-R 模型；[19], MMS-1B-all [20], Qwen3-ASR-1.7B [21] 以及 SeamlessM4T v2 [22]。双生实验采用其中六个具有代表性的识别器组成评估组，分别为 Whisper tiny、medium 和 large-v3、在 Common Voice8 上微调的 XLS-R 300M、MMS-1B-all 和 Qwen3-ASR-1.7B；并使用全部 19 个识别器进行稳健性检验。评分前，参考文本和识别假设均规范化为 Unicode NFC，按照土耳其语规则转为小写并去除标点，数字则改写为文字。报告语料库级词错误率，数值越低越好。

A. 四个保真度层级

合成测试集可能适用于一种用途，却不适用于另一种用途。因此，我们从四个层级衡量其与真实语音的一致性，并在计算任何结果之前，为每个层级预先确定了通过阈值：

- 排名： 两个版本中识别器词错误率之间的肯德尔 τ [23], 以及顺序保持不变的显著不同识别器对所占的比例。阈值为 $\tau \geq .80$ ，至少有 90% 的此类识别器对保持原有顺序。

- 绝对误差：每个识别器的带符号差值 $\Delta = \text{WER}_{\text{合成}} - \text{WER}_{\text{真实}}$，并附有配对自助法区间。对于至少 80% 的识别器，阈值为 $|\Delta| \leq 2$ 个百分点。

- 句子难度：真实与合成版本的逐句 WER 之间的斯皮尔曼相关系数（在各识别器上取平均），以及虚警率，即仅合成版本出错的识别器—句子单元所占的比例。阈值为 $\rho \geq .60$ ，虚警率不超过 20%。

- 校准：以真实 WER 为因变量、合成 WER 为自变量进行线性回归，采用留一识别器法拟合，并以留出的识别器上的平均绝对误差进行评分。阈值为 3 个百分点。

区间基于对句子的 2,000 次配对自助法重采样；若要判定通过，经过针对各引擎和等级的 Holm–Bonferroni 校正后，自助法失败概率仍须低于 .05。

B. 合成测试集的价值是什么？

保真度也可以用一个数值来表示，即合成数据集 S的价值 。其定义为随机抽取的真实录音数量 n ，使得 $E|\hat{}W(n) - W| = |W(S) - W|$；其中， W 是识别器在所有真实录音上的词错误率，$\hat{}\ W(n)$ 是其在 n 抽取的录音上的词错误率，并对识别器取平均。价值为 50 的数据集所提供的关于真实词错误率的信息量，相当于 50 条随机录音，因此，从预测驱动推断的意义上说，价值就是该数据集的有效样本量 [10]。由于实验中的 507 条录音是根据说话人匹配选出的，而非随机抽取，其词错误率会偏离 W，所以即使是相同的真实录音，其价值也低于 507 条录音。

IV. 合成孪生体

每条合成话语都通过 XTTS-v2 一次改变一个因素，逐步构建成与其真实对应语音更加接近的孪生体。XTTS-v2 支持零样本语音克隆：

- *R0*（默认语音）：引擎内置的四种语音之一。
- *R1*（音高匹配语音）：其中位基频与真实说话人最接近，且与其位于 165 Hz 分界线的同一侧的内置语音。

- *R2*（说话人克隆）：以同一说话人的另一条话语为条件，对真实说话人进行克隆。
- *R3*（+信道）：将 R2 通过真实录音的信道处理。信道处理包括长期频谱均衡（三分之一倍频程平滑，限制在±20 dB）、取自真实录音中最安静的非静音帧并保持原始电平的背景噪声、响度匹配，以及以 64 kbit/s 进行 MP3 往返编码，与 Common Voice 原始录音的处理方式相同。

- *R4*（+ 语速）：R3，使用 XTTS-v2 的速度控制，将克隆语音重新合成为真实录音的语速，并采用

[1]https://huggingface.co/datasets/yagmurtuncer/turkasr-bench-tts [2] https://huggingface.co/datasets/yagmurtuncer/turkasr-bench-twins

语音时长比率限制在 0.6 到 1.8 的范围内。

- *R0 + 信道*：将 R3 的信道应用于默认语音，从而将说话人效应与信道效应分离。
- 预言机克隆：以目标话语本身作为条件生成的克隆，可作为克隆效果的上限。

由于此处使用的数据划分中，Common Voice 未公开说话人身份，因此使用 ECAPA-TDNN 说话人编码器为全部 9,650 条测试话语生成嵌入 [24], [25], ，并为每个目标选出与其最相似且句子不同的话语。XTTS 克隆与 ECAPA 匹配依赖不同的编码器，因此这种匹配并非自我验证。对于两条话语都带有性别标签的情况，全部 265 例的标签均一致。在 743 个目标中，513 个目标匹配到余弦相似度至少为 0.60 的话语；其中有音高估计的 507 个目标构成了实验，并由第三节中的六种代表性识别器进行转写。这些孪生语音与对应说话人相似：克隆语音与真实录音的平均 ECAPA 相似度为 .55，理想克隆为 .61；相比之下，默认语音为 .12，音高匹配语音为 .14，同一说话人的两段真实录音之间则为 .70。

表 I 和图 1 展示了 XTTS-v2 的阶梯。在表 I 中，Δ 表示合成语音与真实语音的词错误率平均有符号差值；当合成语音更容易识别时，该值为负。最后一行列出了相同句子的真实录音，它们设定了价值上限。图 1 的面板 (a) 以对数刻度展示价值，面板 (b) 展示句子一致性及其 95% 自助法区间；蓝色、橙色和绿色柱分别表示 XTTS 默认语音、XTTS 克隆和 Chatterbox (CB)。虚线表示相同真实录音的价值，以及六识别器难度度量的可靠性上限；点线表示通过阈值。表II 比较了各步骤之间的差异；其中的区间是对句子进行 2,000 次自助法重采样得到的 95% 区间，p 值则在每列和每个引擎内经过 Holm 校正。

匹配信道可以校正那些比真实语音更容易识别的音色所对应的错误水平。XTTS 的默认语音比真实语音更容易识别，差值为 $\Delta = -5.7$ 个百分点。匹配真实录音的信道后，差距从 5.8 个百分点缩小到 1.9 个百分点，显著减少了 3.9 个百分点；价值则从 9 段录音提高到 45 段录音，而句子一致性基本保持不变。匹配说话人比单独匹配音高更有效：说话人克隆也能显著缩小差距，减少 2.2 个百分点；而仅匹配说话人音高的音色反而使差距扩大了 2.1 个百分点。XTTS 克隆将句子一致性从 .38 提高到 .44，增幅方向符合预期，但经校正后未达到显著水平；以目标录音本身为条件的 Oracle 克隆则使其显著提高了 0.12。说话人和信道的影响并非简单相加。克隆已经继承了部分参考录音的信道特征，因此在 R3 中再应用目标信道，会使合成语音比真实语音进一步难识别 1.7 个百分点。

XTTS 双胞胎组中价值最高的是 32 个随机录音，而相同的 507 个真实录音的价值为 251，而 Oracle 克隆达到 $\rho = .51$，低于 .60 的阈值。在 R4 中匹配语速会使合成语音明显更难理解，并将 Δ 提高到 13.5 个百分点。由于未单独测量可懂度，这一增幅不能归因于时序本身，也可能反映了 XTTS-v2 速度控制的局限性，正如第五节所讨论的那样。

在相同参考录音的条件下，使用第二种克隆引擎 Chatterbox [16], 重复了四个核心步骤。表 III 报告了 Common Voice (CV) 和 VoxForge (VF) 上的结果，两者各包含 507 个句子，并使用六种代表性识别器；六种识别器的词错误率的 Kendall τ 在每种 CV 条件下均为 1.00，因此这里只报告 VF 的结果。Chatterbox 克隆比 XTTS 克隆更像原说话人，ECAPA 相似度为 .59，而 XTTS 克隆为 .55；这些结果进一步细化了 XTTS 的发现。现在，克隆显著提升了句子一致性，从 .41 提高到 .55，增幅为 0.14，达到 $p < .001$。相似度分别为 .59 和 .61 的 Chatterbox 克隆与 XTTS 理想克隆都显著提升了一致性，这表明，只要说话人克隆足够接近原说话人，就能承载句子难度的一部分。与 XTTS 不同，Chatterbox 的默认语音已经接近真实错误水平，价值为 56；加入信道后，其难度超过真实语音，价值降至 26。在两种引擎中，当合成语音比真实语音更容易时，加入信道都会改善匹配；第三种引擎也证实了这一规律。将比真实语音更容易的 Chatterbox 克隆与真实信道结合，得到本研究中价值最高的孪生语音：相当于 122 条录音，95% 置信区间为 49 至 239，差距为 1.4 个百分点。相同的真实录音价值达到 251，区间为 66 至 446，因此，基于六种识别器的价值估计应结合其区间一并解读。

作为第三种克隆引擎，VoxCPM2 [17],是一种近期推出的零样本引擎，其使用相同的参考录音及其转录文本作为条件。三种引擎中，它克隆出的音色与说话人最为接近，ECAPA 相似度为 .65，而其默认语音的相似度为 .12。VoxCPM2 印证了克隆效应，将句子一致性从 .50 显著提升至 .62，达到本研究中最高的句子一致性。不过，它生成的语音明显比真实语音更容易识别：在六个代表性识别器中，其克隆语音的词错误率比真实录音低 10.2 个百分点，因此这些克隆语音的价值仅相当于 3 段录音。与信道模式一致，真实信道将这一差距缩小了 3.3 个百分点，降至 6.9 个百分点，并将价值提高至 7。因此，引擎可以捕捉哪些 句子难以识别，却未必能捕捉它们有多难；所以应分别检查句子一致性和错误水平。

使用六个识别器测得的句子难度具有有限的可靠性。使用全部 19 个识别器重复分析后，真实语音的分半信度从 .78 提高到 .91。表 IV 比较了两个面板下的三个引擎。结论保持不变。对于 Chatterbox，克隆版在句子一致性方面的提升仍然显著，数值为 0.16；对于 VoxCPM2，数值为 0.09；二者均带有 $p < .001$；对于 XTTS

表 I 使用 XTTS-v2 的合成双胞胎阶梯

条件	句子 ρ (↑)	误报 (↓)	Δ 分	|Δ|分 (↓)	价值 (↑)
R0 原始语音	.38 [.30, .46]	24%	−5.7	5.8	9
R1 音高匹配语音	.38 [.30, .46]	23%	−7.9	7.9	5
R0 + 信道	.35 [.26, .44]	24%	−1.8	1.9	45
R2 说话人克隆	.44 [.35, .52]	25%	+0.5	3.6	32
R3 克隆 + 信道	.44 [.35, .52]	26%	+5.2	5.2	19
R4克隆 + 信道 + 速率	.41 [.33, .49]	29%	+13.5	13.5	3
Oracle 克隆（目标话语）	.51 [.43, .58]	24%	+0.5	3.5	32
相同的真实录音	1	—	0.0	0.0	251

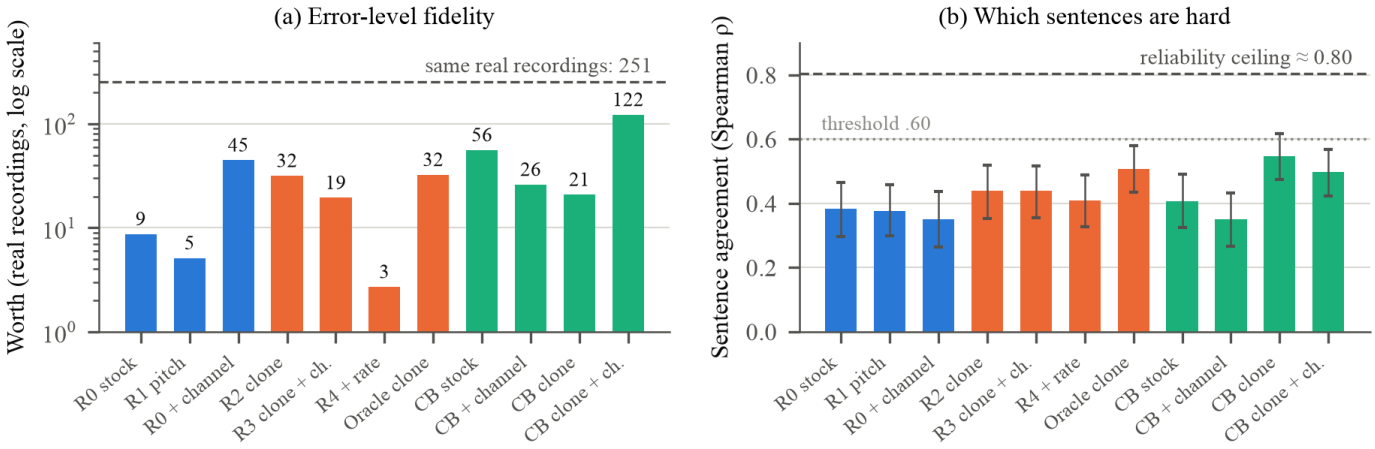

图 1. 双梯上价值与句子一致性的变化

表II 梯级之间的配对差异

Step	Δ 句子 ρ (↑)	p (Holm)	Δ |间隔|点 (↓)	p (Holm)
R0 原版 → R1 音高	−0.007 [−0.078, +0.061]	1.000	+2.1 [+1.0, +3.1]	0.005
R0 原版 → R0 + 信道	−0.033 [−0.078, +0.013]	0.625	−3.9 [−4.5, −2.8]	< .001
R0 原版 → R2 克隆	+0.055 [−0.022, +0.132]	0.625	−2.2 [−4.2, −0.7]	0.016
R1 音高 → R2 克隆	+0.062 [−0.015, +0.138]	0.625	−4.3 [−6.1, −2.5]	< .001
R2 克隆 → R3 克隆 + 信道	−0.000 [−0.050, +0.050]	1.000	+1.7 [+0.5, +3.0]	0.016
R2 克隆 → Oracle 克隆	+0.069 [+0.001, +0.137]	0.246	−0.1 [−1.0, +0.9]	0.867
R0 原版 → Oracle 克隆	+0.124 [+0.046, +0.208]	0.021	−2.3 [−4.3, −0.8]	0.016
Chatterbox: R0 原版 → R0 + 信道	−0.056 [−0.107, −0.006]	0.060	+2.5 [+0.2, +3.9]	0.111
Chatterbox: R0 原版 → R2 克隆	+0.141 [+0.055, +0.221]	< .001	+0.8 [−1.2, +2.6]	0.434
Chatterbox: R2 克隆 → R3 克隆 + 信道	−0.050 [−0.095, −0.006]	0.060	−2.0 [−3.6, +0.2]	0.188

则上升至 0.08。使用 19 个识别器时，价值最高的双胞胎是带信道的 Chatterbox 克隆版（70 条录音）和 XTTS 克隆（60 条），相比之下，原版 XTTS 语音为 10 条，真实录音为 309 条。这些双胞胎相对于原版 XTTS 语音的显著提升在两个面板中都保持稳定。

真实信道的一种通用替代方案，是加入来自标准集合的噪声或混响，即以固定信噪比加入 MUSAN 背景噪声 [26]，以及真实房间脉冲响应 [27]。应用于 Piper 引擎 [28],时，其语音比真实语音更容易识别 $\Delta = -8.6$，这种数据增强也会使错误水平接近真实水平，并将价值从 4 提高到 31 至 39 之间，但会使句子一致性从 .48 降至 .39 和 .41。在 XTTS 上，以 20 dB 信噪比加入噪声会将价值从 9 提高到 15，而真实信道的数值为 45。因此，通用退化可以调整错误水平，而说话人克隆则是改善句子级一致性的关键步骤。

Common Voice 中的说话人匹配是推断得出的，其信道包含 MP3 编码。因此，在 VoxForge [14]，语料库上重复 Chatterbox 的步骤。该语料库包含由志愿者在每次约朗读十个提示语的会话中录制的土耳其语朗读语音。一个会话可确定说话人，因此每个克隆音色都以同一会话 中的录音为条件；该录音使用了与目标话语最相似的另一个提示语，其 ECAPA 相似度中位数为 .80，因此无需设置相似度阈值。在为多源基准抽取的 743 条 VoxForge 语句中

表 III Chatterbox Twins 在 Common Voice 与 VoxForge 上

条件（Chatterbox）	Common Voice ρ(↑)	Common Voice $\lvert\Delta\rvert$ (↓)	Common Voice 价值（↑）	VoxForge ρ (↑)	VoxForge $\lvert\Delta\rvert$(↓)	VoxForge τ (↑)	VoxForge 价值 (↑)
R0 原始语音	.41	2.7	56	.31	4.5	.73	8
R0 + 信道	.35	5.1	26	.35	5.5	.73	5
R2 说话人克隆	.55	3.4	21	.50	2.9	.73	52
R3 克隆 + 信道	.50	1.4	122	.53	1.8	.87	54
相同的真实录音	1	0.0	251	1	0.0	1	156

表 IV：三种克隆引擎与 6 个和 19 个识别器

引擎 · 条件	ρ，6 个识别器（↑）	ρ，19 个识别器（↑）	$\lvert\Delta\rvert$，19（↓）	τ，19（↑）	价值，6（↑）	价值，19（↑）
XTTS · 默认语音	.38	.40	5.3	.80	9	10
XTTS · 股票 + 信道	.35	.37	2.2	.86	45	43
XTTS · 说话人克隆	.44	.48	2.5	.91	32	60
XTTS · 克隆 + 信道	.44	.47	4.9	.90	19	22
Chatterbox · 默认语音	.41	.47	2.5	.90	56	66
Chatterbox · 股票 + 信道	.35	.41	5.1	.90	26	21
Chatterbox · 说话人克隆	.55	.62	3.2	.95	21	23
Chatterbox · 克隆 + 信道	.50	.58	1.9	.91	122	70
VoxCPM2 · 默认语音	.50	.56	13.4	.78	2	2
VoxCPM2 · 默认 + 信道	.51	.55	11.5	.80	2	3
VoxCPM2 · 说话人克隆	.62	.65	9.3	.85	3	4
VoxCPM2 · 克隆 + 信道	.57	.61	6.5	.84	7	8
相同的真实录音	1	1	0.0	1	251	309

使用了来自 165 个会话的 507 条语句，其规模与 Common Voice 实验相当；其中 315 条按标识符顺序选取，192 条随机选取，随机子集的结果另行报告。音频为未压缩的 WAV，因此信道步骤省略了 MP3往返编码。克隆音色同样与说话人相近：其与真实录音的 ECAPA 相似度为 .67，而股票音色为 .12，两个真实录音之间则为 .79。

表 III 比较了两个语料库，结果表明说话人效应得到了复现：克隆使句子一致性从 .31 提高到 .50，在 $p < .001$ 处显著提升了 0.19，而此时说话人身份是已知的，而非推断得出的。仅看随机抽取的 192 个句子，增幅也相同，从 .34 提高到 .51，这证实部分有序的选择并不能解释这一结果。信道方面的模式也得到了复现。在 VoxForge 上，默认音色比真实语音更难 识别 4.5 个百分点，而信道使差距进一步扩大了 1.0 个百分点；相比之下，克隆比真实语音容易识别 2.3 个百分点，信道则使差距缩小了 1.1 个百分点。不使用信道时，克隆的价值相当于 52 段录音；使用信道时，相当于 54 段录音，95% 置信区间分别为 30 至 76 和 26 至 95；相比之下，默认音色的价值相当于 8 段录音，其区间为 3 至 18。在 VoxForge 上，τ从默认音色的 .73 升至使用信道时克隆的 .87；在六个识别器的情况下，一对相近系统的排名互换对应于 τ 中 .13 的差异。

使用不同的随机种子重新合成理想孪生，并对两个副本进行评分，可得到两个合成版本之间的句子一致性$\rho = .69$ 。真实录音与孪生之间的一致性介于 .51 和 .55 之间，因此达到孪生与其自身的第二次渲染结果之间一致性的 74% 至 80%；因此，剩余差距的一部分反映了零样本文本转语音的变异性，而非真实语音所缺失的某种属性。

句子难度本身也是带有误差的测量结果。当将六个识别器分成两半时，真实语音和理想孪生语音中每个句子的词错误率的 Spearman–Brown信度分别为 .78 和 .83，这将可观测的一致性限制在约 .80。校正这种衰减后，一致性从默认语音的 .48 提升至克隆的 .55，以及理想克隆的 .63。

逐句来看，真实录音与理想孪生语音之间的词错误率差异，与真实录音的语速、背景水平、停顿占比或克隆的说话人相似度均无关，相关系数分别为 .06、 –.07、–.03 和 –.01。只有句子长度显示出微弱的相关性，为 $\rho = .18$。明显的朗读偏差——在真实录音中，两个最强的识别器转写出了相同的非参考词，而在孪生语音中却转写出了参考文本——仅出现在 507 个句子中的 5 个。因此，这一差距并非集中在少数包含噪声、语速过快或误读的录音中，而是分布于众多句子之间，这与一个人对特定句子的发音方式存在细粒度差异相一致。鉴于表 II 中区间的宽度，此类阶梯研究需要数百个句子。

V. 讨论

合成语音支持对识别器进行粗略比较，但无法区分性能接近的系统、报告绝对准确率或定位故障。明确说明数据集的价值意味着

应明确其适用范围：与十段录音等效的数据集无法支持将词错误率报告到小数点后一位。

如果要用合成数据集取代真实录音，就应克隆一小份真实样本中的说话人，并根据该样本的价值度量来决定克隆引擎的选择和是否使用信道：只有当合成语音仍比真实语音更容易识别时，才通过真实录音通道处理音频。在本文的实验中，根据识别器组的不同，这一流程将合成数据集的价值从 Common Voice 上约 10 段录音提高到 70 至 122 段，并将其在 VoxForge 上的价值从 8 段提高到 54 段。一小份真实锚定集既用于作出这一决定，也用于报告合成数据集的价值。

报告基于合成语音的自动语音识别结果时，应同时提供以下信息：文本转语音引擎和语音音色，以及是否克隆了说话人；一小份域内真实录音数据集；带区间的排名一致性；每个系统的绝对词错误率差异；句子级一致性和虚警率；以及合成数据集以真实录音计的价值。

在 Common Voice 中，通过嵌入相似度找到参考录音对应的说话人；而在 VoxForge 中，录音会话会标明说话人，因此也能重现克隆效应，这表明该效应并非由这种匹配过程造成。语速步长受当前文本转语音速度控制所能清晰呈现的范围限制；若时长模型能够保留可懂度，就可以单独研究时序。XLS-R 识别器是在 Common Voice 上微调的，这可能使其在 Common Voice 录音上更具优势；但这一点不适用于 VoxForge。克隆语音仅用于衡量评估保真度，且只能应请求提供；其参考录音来自公共语料库，这些语料库以 CC0 和 GPL 许可发布。

该阶梯将以往研究未曾分别考察的因素区分开来。在匹配音色和信道后仍然存在的差距，部分源于文本转语音变异性，也可能部分反映个体朗读的特征，例如犹豫、强调和局部语速；第四节的录音层级指标无法捕捉这些特征，而当前的零样本文本转语音也无法根据参考录音重现它们。在三种引擎中，与说话人越相似的克隆语音，也越能重现句子难度。

VI. 结论

现成文本转语音音色的价值充其量相当于几十段真实录音。合成孪生体使测量并缩小这一差距成为可能。克隆说话人，并在合成语音比真实语音更容易的情况下匹配录音通道，可使合成数据集的价值提高七至十四倍：在六个识别器的情况下，从相当于 9 段录音提高到 122 段；在 19 个识别器的情况下，则从 10 段提高到 70 段；而且，与说话人高度相似的克隆音色能够复现每个句子的一部分难度。剩余差距的一部分源于文本转语音变异性，而其余部分可能反映了个人朗读句子的特有方式，而这部分差距可能反映了个人朗读句子的特有方式。

未来工作将把阶梯扩展到会话语音和更多语言，加入时长模型以单独研究时序，并确定真实锚定集的最小规模，同时确保其仍能选出正确的孪生配置。

参考文献

[1] M. H. Asyrofi, F. Thung, D. Lo, and L. Jiang, “CrossASR：通过文本转语音高效地对自动语音识别进行差分测试,” in Proc. IEEE ICSME, 2020, pp. 640–650. [2] J. K. Lau et al., “利用文本转语音合成语音测试用例? 自动语音识别测试中误报的实证研究,” in Proc. ACM ISSTA, 2023, pp. 1169–1181. [3] G. Tay, W. Ma, J. Lee, et al., “回归基础：语音代理时代重新审视自动语音识别,” arXiv:2603.25727, 2026. [4] M. R. Costa-jussà, B. Yu, P. Andrews, B. Alastruey, N. C. Camgoz, et al., “2M-BELEBELE：高度多语言的语音与美国手语理解数据集,” in Findings of ACL, 2025. [5] N. Y. Tuncer, F. N. Korkmaz, M. S. Türkel, and A. T. Bayrak, “土耳其语文本转语音引擎的朗读准确率：自动评测与人工评估,” Advances in Artificial Intelligence Research, 2026, 即将发表. [6] C. Minixhofer, O. Klejch, and P. Bell, “评估并缩小合成语音分布与真实语音分布之间的差距,” in Proc. Interspeech, 2023. [7] N. Rossenbach, R. Schlüter, and S. Sakti, “关于自动语音识别中用于合成数据生成的文本转语音模型选择问题,” in Proc. SynData4GenAI, 2024. [8] T. Lebryk, D. Ayllon, A. Baird, J. P. Cłapa, J. Madsen, and P. Tzi-rakis, “量化自动语音识别模型中的基准优化,” arXiv:2608.19936, 2026. [9] N. Y. Tuncer, M. Kumcu, A. B. Altınel Girgin, and M. Pınar, “TurkMed-STT：通过合成数据实现参数高效的土耳其语临床语音识别,” in Proc. 10th Int. Artificial Intelligence and Data Processing Symp. (IDAP), IEEE, 2026, pp. 1–7. [10] A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic, “预测驱动推断,” Science, vol. 382, no. 6671, pp. 669–674, 2023. [11] A. Fisch et al., “用于混合语言模型评测的分层预测驱动推断,” in Proc. NeurIPS, 2024. [12] R. Ardila et al., “Common Voice：大规模多语言语音语料库,” in Proc. LREC, 2020, pp. 4218–4222. [13] N. Y. Tuncer, F. N. Korkmaz, M. S. Türkel, and A. T. Bayrak, “TurkASR-Bench：多领域测试集及开放式土耳其语自动语音识别模型评测,” Advances in Artificial Intelligence Research, 2026, 即将发表. [14] VoxForge, “免费 GPL 语音语料库，土耳其语部分,” http://www.voxforge.org, 访问于 2026 年 10 月 25 日. [15] E. Casanova et al., “XTTS：大规模多语言零样本文本转语音模型,” in Proc. Interspeech, 2024, pp. 4978–4982. [16] Resemble AI, “Chatterbox Multilingual,” https://github.com/resemble-ai/chatterbox, 访问于 2026 年 10 月 8 日. [17] OpenBMB, “VoxCPM2,” https://huggingface.co/openbmb/VoxCPM2, 访问于 2026 年 10 月 12 日. [18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “通过大规模弱监督实现鲁棒语音识别,” in Proc. ICML, 2023, pp. 28492–28518. [19] A. Babu et al., “XLS-R：大规模自监督跨语言语音表征学习,” in Proc. Interspeech, 2022, pp. 2278–2282. [20] V. Pratap et al., “将语音技术扩展至 1,000+ 种语言,” J. Mach. Learn. Res., vol. 25, no. 97, pp. 1–52, 2024. [21] Qwen 团队, “Qwen3-ASR 技术报告,” arXiv:2601.21337, 2026. [22] Seamless Communication et al., “Seamless：多语言、富有表现力的流式语音翻译,” arXiv:2312.05187, 2023. [23] M. G. Kendall, “秩相关的新度量,” Biometrika, vol. 30, no. 1/2, pp. 81–93, 1938. [24] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN：基于时延神经网络的说话人验证中的强调通道注意力、传播与聚合,” in Proc. Interspeech, 2020, pp. 3830–3834. [25] M. Ravanelli et al., “SpeechBrain：通用语音工具包,” arXiv:2106.04624, 2021. [26] D. Snyder, G. Chen, and D. Povey, “MUSAN：音乐、语音与噪声语料库,” arXiv:1510.08484, 2015.

[27] T. Ko、V. Peddinti、D. Povey、M. L. Seltzer 和 S. Khudanpur，“面向鲁棒语音识别的混响语音数据增强研究”，载于 *IEEE ICASSP* 会议论文集，2017 年，第 5220–5224 页。[28] Rhasspy，“Piper”，https://github.com/rhasspy/piper，2026 年。