DAIEN-TTS：解耦音频插补用于环境感知的文本到语音合成

Ye-Xin Lu[1], Yu Gu, Kun Wei, Hui-Peng Du[1], Yang Ai[1], Zhen-Hua Ling[1]

[1]National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei, P. R. China

{yxlu0102, redmist}@mail.ustc.edu.cn, {yangai, zhling}@ustc.edu.cn

ABSTRACT

本文介绍了 DAIEN-TTS，一种零样本文本到语音（TTS）框架，通过解耦音频填充实现了环境感知合成。通过利用独立的说话人和环境提示，DAIEN-TTS 允许对合成语音的音色和背景环境进行独立控制。基于 F5-TTS 构建，提出的 DAIEN-TTS 首先整合了一个预训练的语音-环境分离（SES）模块，将环境语音分解为干净语音和环境音频的梅尔频谱图。然后在两个梅尔频谱图上应用长度不同的随机跨度掩码，这些掩码连同文本嵌入一起作为填充被遮罩的环境梅尔频谱图的条件，实现了个性化语音和个人化时间变化环境音频的同时延续。为了进一步提高推理过程中的可控性，我们采用双无类别引导（DCFG）方法处理语音和环境组件，并引入信噪比（SNR）自适应策略来将合成语音与环境提示对齐。实验结果表明，DAIEN-TTS 生成的具有高度自然度、强说话人相似性和高环境保真度的环境个性化语音。音频样本可访问 https://yxlu-0102.github.io/DAIEN-TTS。

Index Terms— 环境感知的零样本 TTS，解耦音频插补，流匹配，背景环境控制

1. 介绍

零样本文本转语音（TTS）是指根据说话者的简短语音提示生成未见过的说话者的语音。最近，先进的零样本 TTS 方法采用了上下文学习范式，例如基于神经编解码器的语言模型 [1, 2]、基于流匹配的语音填充 [3, 4] 及其混合方法 [5–7]，这些方法在语音自然度和说话人相似性方面取得了显著改进。通常，零样本 TTS 系统被期望生成高质量的个性化语音，而基于上下文学习的模型倾向于保留语音提示的声学环境特征，这推动了鲁棒性零样本 TTS 方法 [8–10] 的发展。然而，在有声读物和虚拟现实等场景中，通常需要合成具有背景环境的语音。尽管一些现有方法 [9, 11] 可以有效保留语音提示的背景环境，但它们仍然无法解耦环境以实现独立控制。

环境感知的 TTS 指的是使用独立的说话人和环境提示来分别控制合成语音的音色和声学背景。先前的方法 [12,13] 使用专门的说话人和环境编码器提取全局嵌入以进行解耦控制。然而，这样的方法受限于时间不变的声学环境如混响，并且难以处理时变的背景噪声。最近的文本到音频（TTA）方法已经展示了生成带有语音 [14,15] 的环境音频的能力。不过，VoiceLDM [14] 常常产生重复和含糊不清的语音，而 VoiceDiT [15] 缺乏零样本能力并且在语音与环境之间的对齐方面存在问题。UmbraTTS [16] 通过使用独立的干净语音和环境音频提示实现了环境个性化的语音合成，并且通过语音到环境比率（SER）嵌入额外控制了信噪比（SNR）。然而，它假设可以访问长度相等的纯净干净和纯净环境音频提示，在现实世界场景中这是不切实际的。

在本文中，我们提出了 DAIEN-TTS，一个基于独立音频填充的环境感知零样本 TTS 框架，通过说话人和环境提示实现音色和背景环境的独立控制。基于流匹配 F5-TTS 框架 [4]，DAIEN-TTS 集成了预训练的语音-环境分离（SES）模块，从环境语音中提取干净的语音和环境梅尔频谱图。在训练过程中，我们应用随机跨度掩码来模拟多样的提示持续时间。未被遮盖的环境梅尔频谱图通过交叉注意力与未被遮盖的语音梅尔频谱图和文本嵌入的连接一起整合到扩散变压器（DiT）块 [17] 中，实现语音和背景环境的同时填充。在推理过程中，我们采用双无分类器引导（DCFG）来增强对语音和环境组件的控制，并进一步引入信噪比自适应策略以使合成语音与环境提示相匹配。我们的主要贡献如下：

我们提出了 DAIEN-TTS，首个环境感知的零样本 TTS 框架，能够解耦并独立控制音色和时变背景环境。

2) 我们引入了一种基于交叉注意力的条件生成方案，以实现有效的环境控制，并进一步通过 DCFG 和 SNR 自适应增强推理过程中的生成可控性。

3) 大量实验表明，DAIEN-TTS 在客观和主观评估中均实现了高自然度、说话人相似性和环境保真度。

2. 方法论

所提出的 DAIEN-TTS 的整体结构如图 1 所示。基于 F5-TTS [4]，它将零样本 TTS 公式化为在流匹配框架 [3] 下的文本引导语音填充任务，DAIEN-TTS 提出了一种分离式的音频填充方法，以实现时间变化环境感知的语音合成。给定一个环境语音信号 $\boldsymbol{y} \in \mathbb{R}^T$，其中 T 是波形序列长度，引入了一个预训练的基于 Transformer 的 SES 模块来将 $\boldsymbol{y}$ 分解为清晰语音 mel 谱图 $\boldsymbol{c}_{spk} \in \mathbb{R}^{M\times L}$ 和环境音频 mel 谱图 $\boldsymbol{c}_{env} \in \mathbb{R}^{M\times L}$，其中 M 表示 mel 维度而 L 是谱图序列长度。为了模拟不同的提示条件，将不同长度的随机跨度掩码 $\boldsymbol{m}_{spk}$ 和 $\boldsymbol{m}_{env}$（取值为 0 或 1）分别应用于 $\boldsymbol{c}_{spk}$ 和 $\boldsymbol{c}_{env}$。最后，未被遮罩的部分 $(1-\boldsymbol{m}_{spk})\odot\boldsymbol{c}_{spk}$ 和 $(1-\boldsymbol{m}_{env})\odot\boldsymbol{c}_{env}$，连同扩展文本序列 $\boldsymbol{z}$（长度为 L），一起用作条件输入以重构被遮罩的环境语音成分 $\boldsymbol{m}_{spk}\odot\boldsymbol{x}_1$，其中 $\boldsymbol{x}_1$ 是环境语音 $\boldsymbol{y}$ 的梅尔谱图。模型结构的细节，以及训练和推理流程如下所述。

2.1. SES 模块

如图 2 所示，SES 模块在幅度谱图级别执行语音和背景环境的分离。具体来说，环境语音 $\boldsymbol{y}$ 首先通过短时傅里叶变换 (STFT) 转换为幅度谱图 $|\boldsymbol{Y}| \in \mathbb{R}^{F\times L}$，其中 F 表示频率 bin 的数量。随后，环境幅度谱图 $|\boldsymbol{Y}|$ 被输入到基于 Transformer 的掩码网络中，该网络由一个输入卷积层、K 个 Transformer 块和两个输出卷积层组成。掩码网络预测出两个掩码 $|\boldsymbol{M}^S|$ 和 $|\boldsymbol{M}^E| \in \mathbb{R}^{F\times L}$，分别对应语音和背景环境成分。这些掩码通过逐元素乘法应用于 $|\boldsymbol{Y}|$ 上，以抑制相应的不需要的成分，从而产生分离后的语音和环境幅度谱图 $|\boldsymbol{Y}^S| = |\boldsymbol{Y}| \odot |\boldsymbol{M}^E| \in \mathbb{R}^{F\times L}$ 和 $|\boldsymbol{Y}^E| = |\boldsymbol{Y}| \odot |\boldsymbol{M}^S| \in \mathbb{R}^{F\times L}$。最后，它们都经过 mel 滤波器组得到各自的 mel 谱图，这些 mel 谱图为后续的声音填充过程提供了说话人和环境条件。

2.2. 训练

所提出的 DAIEN-TTS 的训练过程如图 1 的左部分所示。与 F5-TTS 类似，我们在文本引导的音频填充任务上训练我

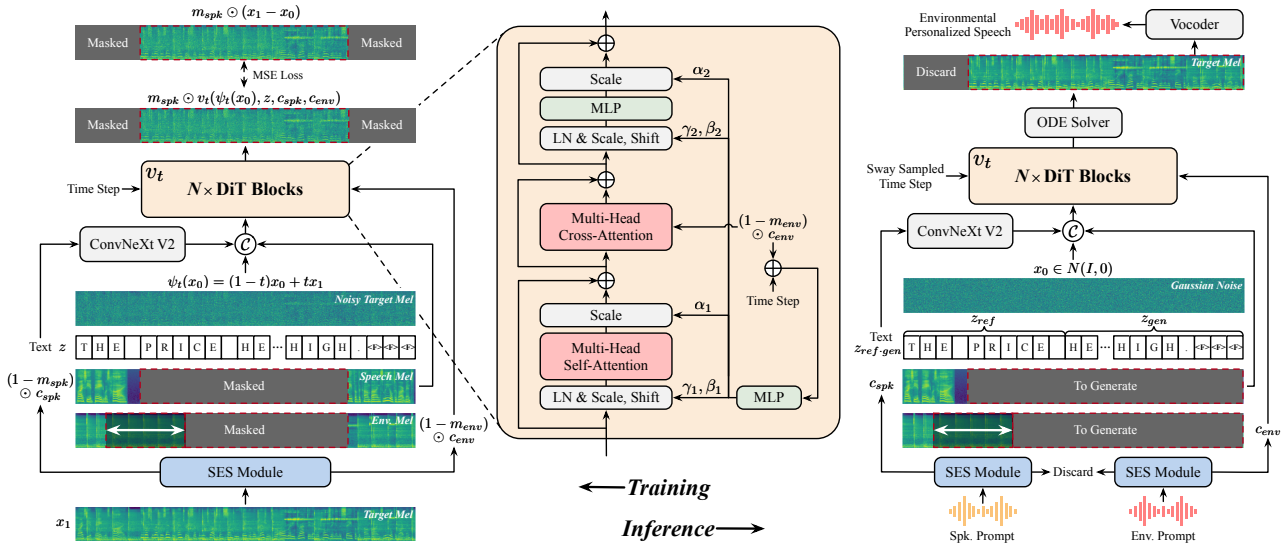

Fig. 1. 提出的 DAIEN-TTS 训练（左）和推理（右）过程概述。

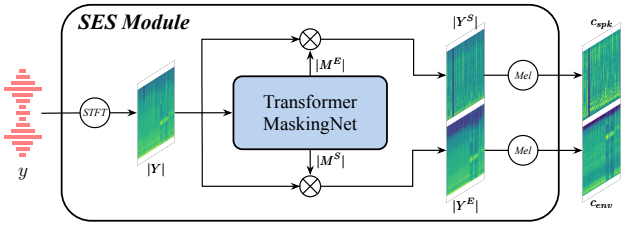

Fig. 2. SES 模块的模型结构。

们的模型，该任务旨在预测一段被遮蔽的环境语音片段，条件是完整的文本（包括周围部分和要生成部分的转录）以及其周围的清晰语音和通过 SES 模块分离出的背景环境音频。为了使模型在推理过程中能够处理不同长度的说话人和环境提示，我们将不同长度的随机跨度掩码应用于分离出的语音梅尔谱图 $\boldsymbol{c}_{spk}$ 和环境梅尔谱图 $\boldsymbol{c}_{env}$。基于条件流匹配（CFM）机制，我们将噪声环境语音 $\psi_t(\boldsymbol{x}_0) = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1$ 作为模型输入，其中 $\boldsymbol{x}_0$ 是采样的高斯噪声，t 是随机抽取的流程步骤。与 UmbraTTS [16] 直接将所有条件与噪声输入拼接到 DiT 模块不同，我们在每个 DiT 模块中插入一个多头交叉注意力层来引入 $(1-\boldsymbol{m}_{env})\odot\boldsymbol{c}_{env}$ 并注入环境信息 [7,15]。最后，模型被训练以在条件 $(1-\boldsymbol{m}_{spk})\odot\boldsymbol{c}_{spk}$，$(1-\boldsymbol{m}_{env})\odot\boldsymbol{c}_{env}$ 和扩展文本 $\boldsymbol{z}$ 的情况下重构 $\boldsymbol{m}_{spk}\odot\boldsymbol{x}_1$，使用以下训练目标：

$$\mathcal{L}_{\text{CFM}} = \Big\| \big(v_t(\psi_t(\boldsymbol{x}_0)|\boldsymbol{z}, (1-\boldsymbol{m}_{spk})\odot\boldsymbol{c}_{spk}, (1-\boldsymbol{m}_{env})\odot\boldsymbol{c}_{env}) - (\boldsymbol{x}_1-\boldsymbol{x}_0)\big)\odot\boldsymbol{m}_{spk}\Big\|_2^2, \tag{1}$$

其中 $v_t(\cdot)$ 是学习到的速度场。

2.3. 推理

所提出的 DAIEN-TTS 的推理过程如图 1 的右侧所示。给定一个演讲者提示语音 $\boldsymbol{y}_{spk}$ 和一个环境提示语音 $\boldsymbol{y}_{env}$，我们首先使用 SES 模块从 $\boldsymbol{y}_{spk}$ 中分离出语音成分，从 $\boldsymbol{y}_{env}$ 中分离出背景环境成分，分别作为演讲者条件 $\boldsymbol{c}_{spk}$ 和环境条件 $\boldsymbol{c}_{env}$。通过将说话人提示 $\boldsymbol{z}_{ref}$ 和文本提示 $\boldsymbol{z}_{gen}$ 连接在一起形成 $\boldsymbol{z}_{ref\cdot gen}$，我们使用常微分方程（ODE）求解器生成具有所需内容的目标梅尔频谱图。从采样噪声 $\psi_0(\boldsymbol{x}_0) = \boldsymbol{x}_0$ 开始，ODE 求解器根据微分方程：

$$d\psi_t(\boldsymbol{x}_0)/dt = v_t(\psi_t(\boldsymbol{x}_0), \boldsymbol{z}_{ref\cdot gen}, \boldsymbol{c}_{spk}, \boldsymbol{c}_{env}). \tag{2}$$

进行积分直至 $\psi_1(\boldsymbol{x}_0) = \boldsymbol{x}_1$。在获得生成的目标梅尔频谱图后，我们丢弃 $\boldsymbol{c}_{spk}$ 部分，并使用声码器重构环境个性化语音。

2.3.1. 双重无分类器引导

F5-TTS 使用 CFG [18] 来实现生成保真度和多样性的权衡。虽然 DAIEN-TTS 受三个项目的条件约束，但文本和说话人信息共同嵌入到语音信号中，并且独立于背景环境。因此，在推理过程中我们将它们视为一个统一的指导项。基于此，我们采用 DCFG 机制 [14] 来增强合成环境语音中的语音和背景环境成分的可控性：

$$\begin{aligned} v_{t,\text{DCFG}} =& v_t(\psi_t(\boldsymbol{x}_0), \boldsymbol{z}, \boldsymbol{c}_{spk}, \boldsymbol{c}_{env}) \\ &+\alpha_{speech}\big(v_t(\psi_t(\boldsymbol{x}_0), \boldsymbol{z}, \boldsymbol{c}_{spk}, \emptyset) - v_t(\psi_t(\boldsymbol{x}_0), \emptyset, \emptyset, \emptyset)\big) \\ &+\alpha_{env}\big(v_t(\psi_t(\boldsymbol{x}_0), \emptyset, \emptyset, \boldsymbol{c}_{env}) - v_t(\psi_t(\boldsymbol{x}_0), \emptyset, \emptyset, \emptyset)\big), \end{aligned} \tag{3}$$

其中 $\emptyset$ 表示空条件，α_{speech} 和 α_{env} 分别是语音和环境条件下 DCFG 的强度。

2.3.2. 信噪比自适应

对于环境提示，我们不仅旨在模仿其背景声学环境，还旨在保留合成语音中的信噪比。为此，我们在推理过程中引入了一个信噪比自适应策略。考虑到环境语音 $\boldsymbol{y}$ 由其语音成分 $\boldsymbol{y}^S$ 和背景环境成分 $\boldsymbol{y}^E$ 组成，使得 $\boldsymbol{y} = \boldsymbol{y}^S + \boldsymbol{y}^E$。根据信噪比的定义和离散帕塞瓦尔定理，我们有：

$$SNR = 10\log\left(\frac{\|\boldsymbol{y}^S\|_2^2}{\|\boldsymbol{y}^E\|_2^2}\right) = 10\log\left(\frac{\|\boldsymbol{Y}^S\|_2^2}{\|\boldsymbol{Y}^E\|_2^2}\right). \tag{4}$$

因此，我们可以对分离出的背景环境幅度频谱图 $\boldsymbol{Y}^E_{env}$ 应用一个缩放因子，以确保合成语音的信噪比与环境提示相匹配：

$$10\log\left(\frac{\|\boldsymbol{Y}^S_{spk}\|_2^2}{\|\boldsymbol{Y}^E_{env}\cdot Scale\|_2^2}\right) = 10\log\left(\frac{\|\boldsymbol{Y}^S_{env}\|_2^2}{\|\boldsymbol{Y}^E_{env}\|_2^2}\right). \tag{5}$$

相应地，我们可以推导出 $Scale = \sqrt{\|\boldsymbol{Y}^S_{env}\|_2^2 / \|\boldsymbol{Y}^S_{spk}\|_2^2}$。在推理过程中，$\boldsymbol{Y}^E_{env}$ 会根据计算出的因子进行缩放，然后转换为梅尔频谱图，这作为最终的环境条件用于生成环境语音。

3. 实验

3.1. 数据集和实验设置

我们使用了 LibriTTS 语料库 [19]，它包含 580 小时的语音数据，用于训练所提出的 DAIEN-TTS 模型。为了模拟广泛的时变背景环境，语音数据与 DNS-Challenge 数据集 [20] 中的环境音频混合使用，在 -5 分贝到 15 分贝之间均匀采样的信噪比。DNS-Challenge 数据集包含来自 AudioSet [21] 和 FreeSound[1] 的 68,000 段环境音频剪辑。为了评估，我们采用了 SeedTTS 测试-英集合 [5] 作为语音测试集，并使用在 0 分贝到 20 分贝之间均匀分布的信噪比将其与来自 SoundBible[2] 的环境音频剪辑增强。所有干净-环境语音配对均以 24 千赫采样率构建。

在 SES 模块中，我们将 Transformer 层的数量 K 设置为 8，具有 16 个注意力头，嵌入维度为 1024，前馈网络（FFN）的维度为 2048。在 TTS 模块中，每个 DiT 块被增加了一个交叉注意力层，也配置了 16 个注意力头。在训练过程中，我们采用了动态混合策略来构建干净与环境语音对。对于 SES 模块的预训练，所有语音数据都与环境音频进行了混合，以确保稳健的分离学习。对于 TTS 训练，我们应用了一种概率增强策略，其中每个语音样本有 50%的概率被与环境音频进行混合，其余情况下则与静音段落进行混合，以便促进文本和语音之间更好的对齐学习。SES 模块和 TTS 模块均在 24 个 NVIDIA V100 32G GPU 上使用 102,800 帧的批量大小训练了 600k 步，并遵循与 F5-TTS 相同的训练策略。对于推理，我们将 DCFG 强度 α_{speech} 和 α_{env} 设置为 2.0。

3.2. 基线和评估指标

由于环境感知 TTS 的核心在于背景环境的解耦和重建，我们对提出的 DAIEN-TTS 进行了两种场景下的评估：1) 首先使用静默作为环境提示来合成干净的个性化语音，以评估模型将环境成分与语音成分解耦的能力。对于基线系统，我们采用了 F5-TTS，并在 LibriTTS 数据集上训练了 600k 步。此外，我们实现了一个 DAIEN-TTS 的消融版本，称为 DAIEN-TTS（w/o CA），其中去掉了 DiT 块中的交叉注意力层，环境、说话人和文本条件直接拼接，类似于 UmbraTTS。2) 基于第一个场景，进一步使用背景环境提示来评估模型在合成环境语音中重建环境成分的能力。由于没有现成的适用于时变环境感知 TTS 的基线，我们仅将 DAIEN-TTS 与其消融版本 DAIEN-TTS（w/o CA）进行了比较。

我们使用客观和主观指标对所提出的 DAIEN-TTS 及基线进行了评估。对于客观评估，通过计算由 Whisper-large-v3 [22] 生成的转录本的词错误率（WER）来衡量语音可懂度。说话人相似度（SIM-o）是通过使用基于 WavLM-large 的说话人验证模型 [23] 计算合成语音和干净说话人提示之间的说话人嵌入的余弦相似性来量化的。对于主观评估，我们进行了三项平均意见评分（MOS）测试：自然度 MOS、说话人相似度 MOS（SSMOS）以及环境相似度 MOS（ESMOS），分别评估感知到的自然度、说话人相似度和环境相似度。我们在 Amazon Mechanical Turk 上招募了超过 30 名评价者，每位评价者对 20 个音频样本进行了评分，评分范围为 5 分制，并具有 0.5 分的分辨率。

[1] https://freesound.org.

[2] https://soundbible.com.

4. 结果与分析

4.1. 静默环境提示的评估

带有静音环境提示的环境感知 TTS 结果如表 1 上部分所示。我们首先将来自 Seed-TTS 测试-en 集的人类录音作为参考（标记为 “Human”），并通过声码器重新合成这些音频（标记为 “Vocoder”）。实验结果显示，声码器重新合成的参考音频在 SIM-o 和 WER 上略有下降，这可以视为清洁合成语音在客观指标上的上限。对于在清洁语音数据上训练的 F5-TTS 基线模型，在提供清洁说话人提示时，合成语音表现出高度的自然度、可懂度和说话人相似性。然而，当使用环境说话人提示进行条件设置时，无论是客观还是主观指标都观察到了显著下降，表明 F5-TTS 对背景环境敏感。

相比之下，所提出的 DAIEN-TTS 及其变体在所有指标下都显著优于 F5-TTS，特别是在环境说话人提示的情况下。这些结果表明 SES 模块有效地将背景环境与环境说话人提示分开，同时保持分离语音成分的高保真度。此外，在使用干净说话人提示的所有指标下，DAIEN-TTS 变体甚至略微超过了 F5-TTS，这可能是由于在配对的干净-环境语音数据上训练时引入的数据增强效果所致。最后，在 DAIEN-TTS 变体中观察到了可比的表现，表明当环境条件是静音时，删除交叉注意力层并没有显著影响合成质量。这进一步暗示了交叉注意力层主要贡献于建模背景环境。

4.2. 背景环境提示的评估

含环境感知的 TTS 与背景环境提示的结果呈现在表 1 的下半部分。为了构建真实数据，我们将人类录音与背景环境音频按照环境提示的信噪比混合（标记为 “Human + Environment”），并使用在环境语音数据上重新训练的声码器对其进行重合成，作为合成环境语音的上限。在这种情况下，DAIEN-TTS 变体在 WER 方面仍超过了声码器重合成的真实数据，并且达到了具有竞争力的 SIM-o 分数（比 “Vocoder” 低约 0.1）。此外，我们提出的 DAIEN-TTS 获得了与人类录音相当的 MOS、SSMOS 和 ESMOS 评分，突显了其在环境重建方面的有效性。

在这些变体中，所提出的 DAIEN-TTS 在说话人相似性方面略胜于 DAIEN-TTS（w/o CA），这一点通过 SIM-o 和 SSMOS 得以体现，同时展示了更优越的自然度和环境真实性，这从 MOS 和 ESMOS 可以看出。这些结果表明，通过交叉注意力引入环境信息比 DAIEN-TTS（w/o CA）和 UmbraTTS 所采用的直接拼接方法更能有效地建模背景环境。DAIEN-TTS（w/o CA）重构背景环境时产生的失真进一步影响了合成语音的自然度。总体而言，两种场景下的评估表明，我们提出的 DAIEN-TTS 成功地解耦并重构了说话人和背景环境组件，实现了高质量的环境感知 TTS。

5. 结论

在本文中，我们介绍了 DAIEN-TTS，这是一种环境感知的零样本 TTS 方法，它能够通过分离音频填充独立控制合成语音中的音色和时变背景环境。通过利用预训练的 SES 模块进行语音-环境分离，并采用基于流匹配的 TTS 框架进行环境 mel 谱图填充，我们的方法有效地解耦并重建了环境语音组件。在推理过程中，DCFG 机制被用来增强对语音和环境成分的控制能力，而信噪比适应方法则确保合成的环境语音与环境提示之间的一致性。实验结果表明，所提出的 DAIEN-TTS 在合成语音的自然度、音色控制和环境控制方面具有优越性。在未来的工作中，我们将探索语音与背景环境之间的相互影响。

Table 1. 环境感知 TTS 在 Seed-TTS 测试-en 数据集上的主观和客观评估结果。

Model	WER(%) ↓	SIM-o ↑	MOS ↑	SSMOS ↑	ESMOS ↑
静音环境提示					
Human	2.14	0.73	3.91 (± 0.09)	3.72 (± 0.09)	-
Vocoder	2.18	0.70	-	-	-
F5-TTS (Clean Spk. Prompt)	2.30	0.58	3.80 (± 0.09)	3.60 (± 0.09)	-
F5-TTS	2.87	0.49	3.09 (± 0.11)	2.92 (± 0.11)	-
DAIEN-TTS (w/o CA)	2.03	0.59	3.81 (± 0.08)	3.60 (± 0.09)	-
DAIEN-TTS	1.93	0.60	3.84 (± 0.09)	3.64 (± 0.09)	-
背景环境提示					
Human + Environment	2.80	0.70	3.86 (± 0.08)	3.81 (± 0.08)	3.72 (± 0.08)
Vocoder	3.03	0.65	-	-	-
DAIEN-TTS (w/o CA)	2.93	0.54	3.68 (± 0.10)	3.70 (± 0.09)	3.49 (± 0.10)
DAIEN-TTS	2.83	0.55	3.78 (± 0.08)	3.73 (± 0.08)	3.65 (± 0.08)

6. REFERENCES

[1] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al., "Neural codec language models are zero-shot text to speech synthesizers," arXiv preprint arXiv:2301.02111, 2023.

[2] Xiaofei Wang, Manthan Thakker, Zhuo Chen, Naoyuki Kanda, Sefik Emre Eskimez, Sanyuan Chen, Min Tang, Shujie Liu, Jinyu Li, and Takuya Yoshioka, "SpeechX: Neural codec language model as a versatile speech transformer," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 3355–3364, 2024.

[3] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al., "Voicebox: Text-guided multilingual universal speech generation at scale," Advances in neural information processing systems, vol. 36, pp. 14005–14034, 2023.

[4] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen, "F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching," in Proc. ACL, 2025.

[5] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al., "Seed-TTS: A family of high-quality versatile speech generation models," arXiv preprint arXiv:2406.02430, 2024.

[6] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al., "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," arXiv preprint arXiv:2407.05407, 2024.

[7] Keon Lee, Dong Won Kim, Jaehyeon Kim, Seungjun Chung, and Jaewoong Cho, "DiTTo-TTS: Diffusion transformers for scalable text-to-speech without domain-specific factors," in Proc. ICLR, 2025.

[8] Xiaofei Wang, Sefik Emre Eskimez, Manthan Thakker, Hemin Yang, Zirun Zhu, Min Tang, Yufei Xia, Jinzhu Li, Sheng Zhao, Jinyu Li, et al., "An investigation of noise robustness for flow-matching-based zero-shot TTS," in Proc. Interspeech, 2024, pp. 687–691.

[9] Leying Zhang, Wangyou Zhang, Zhengyang Chen, and Yanmin Qian, "Advanced zero-shot text-to-speech for background removal and preservation with controllable masked speech prediction," in Proc. ICASSP, 2025.

[10] Ye-Xin Lu, Hui-Peng Du, Fei Liu, Yang Ai, and Zhen-Hua Ling, "Improving noise robustness of LLM-based zero-shot TTS via discrete acoustic token denoising," in Proc. Interspeech, 2025, pp. 2465–2469.

[11] Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath, "VoiceCraft: Zero-shot speech editing and text-to-speech in the wild," in Prc. ACL, 2024, pp. 12442–12462.

[12] Daxin Tan, Guangyan Zhang, and Tan Lee, "Environment aware text-to-speech synthesis," in Proc. Interspeech, 2022, pp. 481–485.

[13] Ye-Xin Lu, Hui-Peng Du, Zheng-Yan Sheng, Yang Ai, and Zhen-Hua Ling, "Incremental disentanglement for environment-aware zero-shot text-to-speech synthesis," in Proc. ICASSP, 2025.

[14] Yeonghyeon Lee, Inmo Yeon, Juhan Nam, and Joon Son Chung, "VoiceLDM: Text-to-speech with environmental context," in Proc. ICASSP, 2024, pp. 12566–12571.

[15] Jaemin Jung, Junseok Ahn, Chaeyoung Jung, Tan Dat Nguyen, Youngjoon Jang, and Joon Son Chung, "VoiceDiT: Dual-condition diffusion transformer for environment-aware speech synthesis," in Proc. ICASSP, 2025.

[16] Neta Glazer, Aviv Navon, Yael Segal, Aviv Shamsian, Hilit Segev, Asaf Buchnick, Menachem Pirchi, Gil Hetz, and Joseph Keshet, "Umbratts: Adapting text-to-speech to environmental contexts with flow matching," arXiv preprint arXiv:2506.09874, 2025.

[17] William Peebles and Saining Xie, "Scalable diffusion models with transformers," in Proc. ICCV, 2023, pp. 4195–4205.

[18] Jonathan Ho and Tim Salimans, "Classifier-free diffusion guidance," arXiv preprint arXiv:2207.12598, 2022.

[19] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "LibriTTS: A corpus derived from librispeech for text-to-speech," in Proc. Interspeech, 2019, pp. 1526–1530.

[20] Harishchandra Dubey, Vishak Gopal, Ross Cutler, Ashkan Aazami, Sergiy Matusevych, Sebastian Braun, Sefik Emre Eskimez, Manthan Thakker, Takuya Yoshioka, Hannes Gamper, et al., "ICASSP 2022 deep noise suppression challenge," in Proc. ICASSP, 2022, pp. 9271–9275.

[21] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in Proc. ICASSP, 2017, pp. 776–780.

[22] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in Proc. ICLR, 2023, pp. 28492–28518.

[23] Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in Proc. ICASSP, 2022, pp. 6147–6151.