arXiv:2601.05554v1 [cs.SD] 9 Jan 2026

SPAM：基于提示的文本转语音的风格提示遵循度指标

Chanhee Cho†、*Nayeon Kim*†、*Bugeun Kim*

中央大学，{cch991112、nayeonkim、bgnkim}@cau.ac.kr

摘要

基于提示的文本转语音（TTS）旨在生成遵循文本提示中提供的细粒度风格线索的语音。然而，大多数先前研究既不依赖合理性指标，也不依赖忠实度指标来评估提示遵循度。也就是说，它们无法确保评估是否以提示为依据，并且是否与人类判断相似。因此，我们提出了一种新的自动指标——风格提示遵循度指标，该指标明确满足合理性和忠实度这两个要求。受 CLAP 启发，我们的方法将语音分解为声学属性，并将其与风格提示对齐。此外，我们使用监督式对比损失训练评分器，从而能够更清晰地区分不同的语义。我们从两个角度开展了两项实验。合理性实验表明，SPAM 与平均意见分（MOS）具有很强的相关性。此外，忠实度实验表明，SPAM 能够成功以给定的风格提示为依据，因为它可以区分提示的不同语义。我们相信，SPAM 能够为评估合成语音的风格提示遵循度提供一种可行的自动化解决方案。

索引词— 评估指标、提示遵循度、基于提示的文本转语音、合理性、忠实度

1. 引言

随着文本转语音（TTS）技术的发展，自然语音合成已广泛应用于各种场景。因此，基于提示的文本转语音逐渐受到关注，它将更丰富的风格线索融入文本提示中 [1, 2, 3, 4, 5]。然而，能够定量评估合成语音对提示的忠实实现程度的自动化指标仍未得到充分研究。大多数研究依赖平均意见分（MOS），而这既耗费人力又耗时。因此，我们提出了一种用于评估提示遵循度的新型自动化指标。

研究人员采用了各种方法，自动评估合成语音对给定风格提示的遵循程度。早期研究人员尝试提取风格嵌入，并检查这些嵌入是否形成一个紧密的相似提示聚类 [1, 2, 3]。然而，这种评估是否恰当地反映了人类感知，仍值得商榷。由于该方法依赖对聚类布局的主观视觉检查，嵌入之间的距离未必能直接对应感知距离。因此，这种方法无法用于比较不同的文本转语音模型。

因此，其他研究者开始采用大语言模型评审范式，以获得提示遵循度的自动评分[6]。他们让大型多模态模型评估给定提示—语音对的提示遵循度。然而，这种方法无法确保评判是否真正基于风格提示的内容。由于大语言模型对细微扰动很敏感 [7],，结果可能并不忠实。在此，我们遵循 [8], ，将评估定义为：当语义相似的提示产生一致结果，而语义不同的提示产生不同结果时，该评估就是忠实的 。

为解决现有研究的这两项局限，我们提出了一种受 CLAP [9] 启发的方法。具体而言，我们认为，用于计算文本-音频相似度的对比学习框架非常适合我们的问题。鉴于其潜力，RA-CLAP [10]采用了 CLAP 风格的模型来处理与我们相似的任务，即情感语音风格检索。然而，与该任务不同，我们的任务需要对特定声学属性进行对齐。尽管如此，将 CLAP 方法应用于我们的任务仍面临两项挑战。首先，需要对声学属性进行因子分解。风格提示通常会规定某些声学属性。由于现有评估器并未明确地对属性进行因子分解，因此无法确定模型在评估过程中是否确实考虑了这些属性。其次，单个批次中可能频繁出现多个正例。受生日悖论影响，大批次中很可能包含多个正例。标准 CLAP 损失并未考虑这一悖论，因此我们需要另一种损失函数来利用这种多正例情形。

因此，我们提出了**风**格**提**示**遵**循**度**量（SPAM），这是一种旨在克服上述局限性的新指标。为实现定量自动评估，我们采用基于 CLAP 的评分器，用于衡量合成语音对给定提示的遵循程度。此外，为了通过因子分解区分各项属性，我们设计了 SPAM，能够显式分解音高、语速、能量、说话人和转录文本。最后，为了利用多个正例，我们采用监督式对比（SupCon）损失 [11] ，而非标准 CLAP 损失。此外，我们还测试了 SPAM 的合理性和忠实度。

† 共同贡献 本研究得到由韩国政府（MSIT）资助的韩国信息通信技术规划评估院（IITP）项目支持 [RS-2021-II211341，以及人工智能研究生院项目（中央大学）支持]。

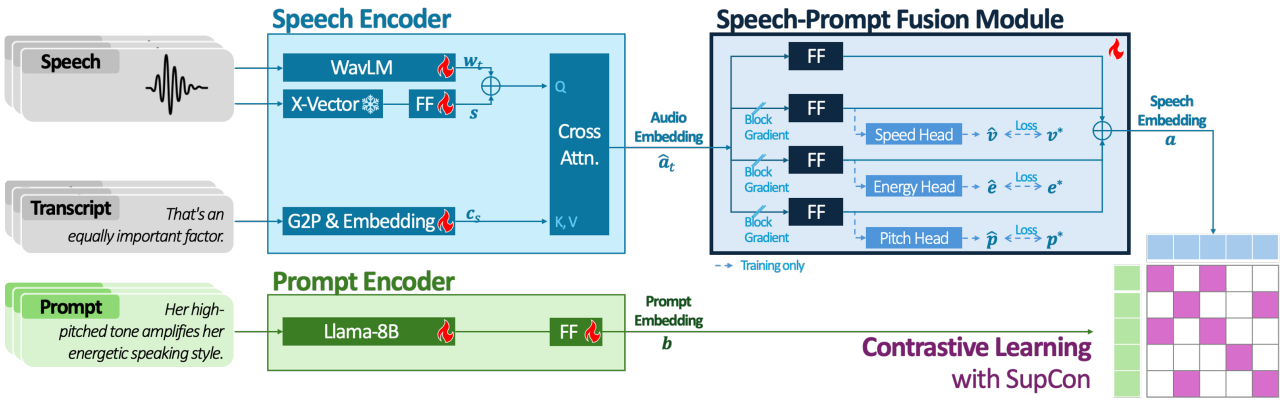

图 1。SPAM 的架构

2. SPAMMETRIC

SPAM 可以将音频和文本映射到共同空间，并计算两者之间的相似度。具体而言，如图 1 所示，我们设计 SPAM 来考虑语音特定因素，并通过各个模块输出相似度分数。首先，**语音编码器** 将波形与其他线索结合，生成信息丰富的音频嵌入。其次，**提示编码器** 将文本提示转换为提示嵌入。第三，**语音-提示融合模块** 利用声学信息强化音频嵌入，并计算其与提示的相似度。

2.1. 语音编码器

为了获得比波形本身更细粒度的语音表示，SPAM 将波形与说话人和转录文本结合起来。我们采用三个编码器分别对这三者进行编码。首先，对于波形，我们使用 WavLM [12]，它在包括 RA-CLAP [10, 13, 14] 在内的许多语音处理系统中都表现可靠。给定一个 16kHz 波形，编码器会生成帧级嵌入 $\mathbf{w}_t \in \mathbb{R}^h$。其次，对于说话人，我们采用 X-Vector [15], ，这是一种广泛使用的说话人特征嵌入。我们冻结了 X-Vector 模块，并采用了一个前馈适配器，将 x-vector 的输出映射到 $\mathbf{s} \in \mathbb{R}^h$。第三，对于转录文本，我们直接将给定转录文本中的音素嵌入 $\mathbb{R}^h$ 空间。具体而言，我们使用字素到音素（G2P）模块和嵌入查找表，获得转录文本嵌入 $\mathbf{c}_s$。

接下来，我们通过结合$\mathbf{w}_t$、$\mathbf{s}$ 和$\mathbf{c}_s$，获得给定语音的音频嵌入 $\hat{}\mathbf{a}_t$ 。我们使用交叉注意力层来捕捉波形与转录文本之间的关系。具体而言，SPAM 会计算$\mathbf{w}_t + \mathbf{s}$ 向量与$\mathbf{c}_s$ 向量之间的注意力。因此，我们获得每一帧对应的 $\hat{}\mathbf{a}_t$ 。

2.2. 提示编码器

为了区分不同的文本提示，SPAM 使用语言模型将文本提示转换为提示嵌入$\mathbf{b}$ 。

该过程与 CLAP 或 RA-CLAP 类似，不同之处在于所使用的语言模型。我们采用带有前馈适配器的 Llama-3.1 8B [16] 。由于文本提示中的细微差异（例如语调、情绪）可能会导致生成的语音出现显著差异，因此提示编码器应足够大，能够区分这些差异。

2.3. 语音-提示融合模块

为了让 SPAM 分别识别声学属性，我们将音频嵌入 $\hat{}\mathbf{a}_t$ 输入四个并行分支：全局波形、语速、能量和音高。首先，全局波形分支会针对整体波形输入生成一个全局表示。我们采用该分支来规范训练过程，防止模型对某一特定属性过拟合。

另外三个分支会生成逐属性嵌入。每个分支通过一个前馈层将 $\hat{}\mathbf{a}_t$ 转换为特定属性的嵌入。需要注意的是，为了引导每个声学分支对应相应的信号，SPAM 使用三个辅助预测头；其中，方差预测器 [17] 用于语速，多层感知机预测头用于能量和音高。这些预测头分别生成帧级估计值 $\hat{}v_t$、$\hat{e}_t$ 和 $\hat{}p_t$ 。

生成四个嵌入后，我们得到最终的语音嵌入 $\mathbf{a}$。为了获得声学增强的表示，我们将每一帧对应的四个嵌入相加，并对所有帧的相加结果取平均。随后，将两个嵌入$\mathbf{a}$ 和 $\mathbf{b}$ 传入后续计算。在训练过程中，我们计算对比损失和辅助损失。在推理过程中，我们计算二者之间的余弦相似度。

2.4. 训练损失

我们在训练过程中使用四种损失计算总损失 $\mathcal{L}$：对比损失 $\mathcal{L}_{con}$、语速损失、能量损失和音高损失，具体如下：

$$\begin{aligned} \mathcal{L} &= \lambda_c \mathcal{L}_{con} + \lambda_p \mathcal{L}_\delta(\hat{p}) + \lambda_v \mathcal{L}_\delta(\hat{v}) + \lambda_e \mathcal{L}_\delta(\hat{e}), \\ \mathcal{L}_{\text{con}} &= \frac{1}{2}\Big(\mathcal{L}^{\text{sup}}(\mathbf{a}, \mathbf{b}) + \mathcal{L}^{\text{sup}}(\mathbf{b}, \mathbf{a})\Big) \end{aligned}$$

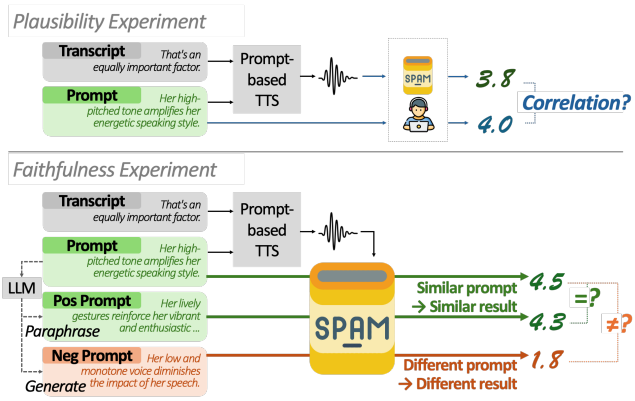

图 2。合理性与忠实度实验

$$\mathcal{L}^{\text{sup}}(X,Y) = \sum_{\mathbf{x}\in X}\left(\mathbb{E}_{\mathbf{p}\sim P(\mathbf{x})}\left[-\log\frac{e^{\mathbf{x}^\top\mathbf{p}}}{\sum_{\mathbf{y}\in Y}e^{\mathbf{x}^\top\mathbf{y}}}\right]\right),$$

其中， $\mathcal{L}_{con}$ 和 $\mathcal{L}_{\delta}$ 分别是SupCon损失 [11] 和Huber损失。$\hat{}p$、$\hat{}v$和$\hat{e}$ 是通过对各帧的$\hat{}p_t$、$\hat{}v_t$和$\hat{e}_t$ 取平均得到的话语级音高、语音和能量预测值。此外， P（$\mathbf{x}$）是锚点$\mathbf{x}$的正例集合。

为定义 P(a) 和 P(b)，SPAM 使用风格键。常见的基于提示的文本转语音数据集提供风格键 [18, 19, 20],，用于表示语音音频的声学特征；例如，具有高音高和正常语速的男性声音。因此，我们利用这类键值信息来识别正向样本对；当提示和音频映射到同一个风格键时，该提示-音频对即为正向样本对。请注意，根据风格键，提示与音频之间的匹配问题会变成多对多问题；因此，我们使用 SupCon 而非 InfoNCE，以在训练过程中获得更好的可靠性。

3. 实验

为确保 SPAM 指标的有效性，我们进行了两项实验（见图 2）。首先，合理性实验考察该指标能否像人类一样成功评估对提示的遵循程度。其次，忠实度实验考察该指标能否清晰地区分遵循提示的语音与未遵循提示的语音。

3.1. 合理性实验

为评估合理性，我们将各项指标与人类对提示遵循度的判断进行比较。具体而言，我们通过 CloudResearch 招募了 320 名标注员，请他们使用 5 点李克特量表对遵循度进行评分。对于每一组提示–语音配对，由 8 名标注员评分，并取其平均值作为平均意见分。

我们使用话语级相关性来评估指标评估的合理性。当平均意见分与某项指标之间的相关性足够高时，我们便可以得出结论：该指标能够很好地反映人类评估结果。我们使用了三种度量方法：线性相关系数（LCC）、斯皮尔曼等级相关系数（SRCC）和肯德尔τ系数（KTAU）。

3.2. 忠实度实验

此外，为评估忠实度，我们在提供原始提示的变体时，比较各项指标的响应。忠实的指标应满足两个条件：(1) 输入与语音配对的原始提示和语义等价的释义时，分数分布应相同；(2) 评估语义不等价的提示时，分数分布应相较于原始提示呈左移。在本文中，我们将前者称为正向提示，将后者称为负向提示。对于每一组提示–语音配对，我们生成了 10 个正向提示和 10 个负向提示，并比较各项指标对原始提示、正向提示和负向提示的不同反应。

我们通过两种方式评估忠实度：遵循率（AR）和配对 t 检验。首先，AR 衡量正向提示的得分是否高于负向提示。与自助法类似，我们将所有可能的正向提示与负向提示配对中出现此类情况的平均概率计算为 AR。其次，配对 t 检验用于衡量是否满足两项条件。我们检验了以下备择假设。

(1) H_1: $\mu_+ \neq \mu_0$ 表示正向提示（μ_+）的得分与原始提示（μ_0）不同。优秀的指标应拒绝 H_1。(2) H_2: $\mu_- < \mu_0$ 表示负向提示（μ_-）的得分低于原始提示。优秀的指标应接受 H_2。

3.3. 数据集与基线

训练 SPAM 时，我们使用了 TexTrolSpeech [18] 和 SpeechCraft [19]，的合并训练集，仅采用来自高质量语音的真实标注数据。

实验中，我们使用了两个测试集：TextrolSpeech 和 LibriTTS-P [20],，后者用于验证模型对未见风格提示的遵循情况。我们从每个数据集中随机抽取了 50 组提示—语音配对。

此外，我们使用五种基于提示的文本转语音系统，为每条真实语音生成五条合成语音：PromptTTS [1],PromptStyle [2], VoxInstruct [4], Parler-TTS-mini-v1 和 Parler-TTS-large-v1 [5]。因此，每项实验都包含 100 条真实语音和 500 条合成语音。此外，我们使用 600 组原始配对和 600 组负向配对测量平均意见分，以纳入提示遵循度较低的案例。

使用该数据集，我们将两个版本的 SPAM 与 RA-CLAP 教师模型 [10]进行比较。SPAM 的两个变体采用了不同的波形编码器：WavLM 和 CLAP 编码器。RA-CLAP 与我们的方法略有不同，因为它旨在进行情感语音风格检索。此外，RA-CLAP 区分声学属性，而 SPAM 可以。

	合理性			忠实度					
	LCC	斯皮尔曼等级相关系数	肯德尔τ系数	AR	原始	正向	(H_1 拒绝？)	负向	(H_2 接受？)
基于*TextrolSpeech* 数据集									
SPAM (WavLM)	0.584	0.584	0.405	**0.862**	0.361±0.153	0.357±0.143 **-2.025**	✓	0.050±0.221 **-20.145*****	✓
(CLAP)	0.554	0.560	0.389	0.841	0.039±0.026	0.035±0.025 -3.699***		-0.005±0.030 **-17.538*****	✓
RA-CLAP	0.520	0.514	0.357	0.852	0.400±0.324	0.380±0.312 -3.479**		-0.020±0.219 **-16.912*****	✓
在*LibriTTS-P*数据集上									
SPAM（WavLM）	0.580	0.568	0.400	**0.771**	0.279±0.171	0.294±0.162 3.200**		0.052±0.234 **-15.105*****	✓
(CLAP)	0.516	0.499	0.346	0.766	0.027±0.027	0.028±0.027 **0.239**	✓	-0.006±0.033 **-14.689*****	✓
RA-CLAP	0.429	0.435	0.304	0.750	0.249±0.285	0.284±0.297 2.645*		0.003±0.199 **-7.628*****	✓

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

表1。在整个测试集上的合理性与忠实度实验结果

	SPAM (WavLM)	SPAM (CLAP)	RA-CLAP
在*TexTrolSpeech* 数据集			
真实音频	0.721	0.673	**0.726**
ParlerTTS-mini	**0.749**	0.710	0.709
-large	0.648	**0.653**	0.615
PromptTTS	**0.665**	0.584	0.416
PromptStyle	0.515	**0.518**	0.503
VoxInstruct	0.462	0.429	**0.473**
基于 *LibriTTS-P*数据集			
真实音频	**0.718**	0.667	0.545
ParlerTTS-mini	**0.752**	0.712	0.536
-大型	**0.716**	0.610	0.531
PromptTTS	**0.560**	0.495	0.375
PromptStyle	**0.519**	0.503	0.417
VoxInstruct	**0.454**	0.403	0.371

表 2。每个模型和数据集的人工MOS与自动化指标之间的线性相关系数

4. 结果与讨论

合理性。 合理性实验的结果如表1（左）所示。总体而言，无论采用哪种相关系数，我们都观察到平均意见分与所有指标之间存在中等程度的相关性。具体来说，SPAM（WavLM）的 LCC 约为 0.58，高于 RA-CLAP（0.520 和 0.429）。此外，根据表2分析各模型的系数时，我们观察到 SPAM 与平均意见分之间的相关性更强且更稳定。例如，对于每个数据集中的真实音频，SPAM 都呈现出较强且稳定的相关性（约为 0.72）。与此同时，RA-CLAP 在不同数据集之间的差距较大：分别为 0.726 和 0.545。在不同的文本转语音模型（例如 ParlerTTS）中也持续观察到这一现象。SPAM 的这种一致性揭示了其对未见过的风格提示的泛化能力。我们推测这种一致性可能有两个原因：语音-提示融合模块能够区分声学属性，以及 Llama 3 提示编码器能够增强上下文理解。得益于这两点，SPAM 可以取代平均意见分来评估提示遵循度，因为无论提示风格如何，它都能提供稳定且合理的评分。

忠实度。 忠实度实验的结果如表1（右）所示。对于 AR，SPAM (WavLM) 的 AR 分数（0.862 和 0.771）高于 RA-CLAP（0.852 和 0.750）。较高的 AR 分数表明，SPAM 能够成功区分正例和负例，为正向提示赋予比负向提示更高的分数。配对 t检验的结果提供了更严谨的解释。检验结果显示，只有 SPAM 系列几乎能够成功地为正例赋予与原始样本相同的分数，因为备择假设 (H_1) 被拒绝。与此同时，RA-CLAP 未能拒绝假设，这表明存在不忠实的情况，即正例的分数可能与原始样本的分数不相等。对于对比损失旨在学习的负例，所有指标均通过接受备择假设 (H_2) 成功通过检验。我们推测，正例方面的这一差异源于采用了 SupCon损失，它比标准 CLAP损失更能利用正例之间的关系。因此，SPAM 展示了忠实的评估。

5. 结论

本文提出了用于评估合成语音是否遵循提示的**S**tyle **P**rompt **A**dherence **M**etric（SPAM）。我们对 CLAP 进行了改进，使其满足该任务的两项要求：对齐声学属性，以及考虑多正样本。通过合理性与忠实度实验，我们评估了 SPAM 的评估结果是否与人类相似，以及它能否区分正向提示和负向提示。结果表明，SPAM 能够对提示遵循度进行合理且忠实的评估。

6. 参考文献

[1] Zhifang Guo、Yichong Leng、Yihan Wu、Sheng Zhao 和 Xu Tan，“Prompttts：利用文本描述实现可控文本转语音”，载于*2023 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，2023 年，第 1–5 页。

[2] Guanghou Liu、Yongmao Zhang、Yi Lei、Yunlin Chen、Rui Wang、Lei Xie 等人，“Promptstyle：通过自然语言描述实现可控的文本转语音风格迁移”，载于*Interspeech 2023*，2023 年。

[3] Reo Shimizu、Ryuichi Yamamoto、Masaya Kawamura、Yuma Shirahata、Hironori Doi、Tatsuya Komatsu 等，“Prompttts++：使用自然语言描述控制基于提示的文本转语音中的说话人身份”，载于*2024*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2024年，第12672–12676页。[4] Yixuan Zhou、Xiaoyu Qin、Zeyu Jin、Shuoyi Zhou、Shun Lei、Songtao Zhou 等，“Voxinstruct：采用统一多语言编解码器语言建模的富有表现力的人类指令到语音生成”，载于第*32*届*ACM*国际多媒体会议论文集，美国纽约州纽约，2024年，第554–563页。[5] Dan Lyth 和 Simon King，“使用合成标注通过自然语言引导高保真文本转语音”，2024年。[6] Kexin Huang、Qian Tu、Liwei Fan、Chenchen Yang、Dong Zhang、Shimin Li 等，“Instructttseval：对文本转语音系统中复杂自然语言指令遵循能力的基准评测”，2025年。[7] Robin Jia 和 Percy Liang，“用于评估阅读理解系统的对抗样本”，载于*2017*年自然语言处理经验方法会议论文集，Martha Palmer、Rebecca Hwa 和 Sebastian Riedel 编，丹麦哥本哈根，2017年9月，第2021–2031页，计算语言学协会。[8] Chirag Agarwal、Sree Harsha Tanneru 和 Himabindu Lakkaraju，“忠实度与合理性：关于大语言模型解释的（不）可靠性”，2024年。[9] Yusong Wu、Ke Chen、Tianyu Zhang、Yuchen Hui、Taylor Berg-Kirkpatrick 和 Shlomo Dubnov，“采用特征融合和关键词到描述增强的大规模对比式语言-音频预训练”，载于*2023*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），2023年。

[10]Haoqin Sun、Jingguang Tian、Jiaming Zhou、Hui Wang、Jiabei He、Shiwan Zhao 等，“RA-CLAP：关系增强的情感说话风格对比式语言-音频预训练，用于语音检索”，载于*Interspeech2025*，2025年，第2995–2999页。[11] Prannay Khosla、Piotr Teterwak、Chen Wang、Aaron Sarna、Yonglong Tian、Phillip Isola 等，“监督式对比学习”，载于第*34*届神经信息处理系统国际会议论文集，美国纽约州雷德胡克，2020年，NIPS ’20。[12] Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen 等，“Wavlm：面向全栈语音处理的大规模自监督预训练”，载于*IEEE*信号处理选题期刊，第16卷第6期，第1505–1518页，2022年。[13] Hyungchan Song、Sanyuan Chen、Zhuo Chen、Yu Wu、Takuya Yoshioka、Min Tang 等，“探索WavLM在语音增强中的应用”，2022年。[14] Shujie Hu、Long Zhou、Shujie Liu、Sanyuan Chen、Ling-wei Meng、Hongkun Hao 等，“WavLLM：迈向稳健且自适应的语音大语言模型”，载于计算语言学协会论文集：*EMNLP 2024*增刊，2024年11月，第4552–4572页。[15] David Snyder、Daniel Garcia-Romero、Gregory Sell、Daniel Povey 和 Sanjeev Khudanpur，“X向量：用于说话人识别的稳健深度神经网络嵌入”，载于*2018*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），2018年，第5329–5333页。[16] Aaron Grattafiori、Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle 等，“Llama 3模型家族”，2024年。[17] Yi Ren、Chenxu Hu、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 等，“FastSpeech 2：快速且高质量的端到端文本转语音”，2022年。[18] Shengpeng Ji、Jialong Zuo、Minghui Fang、Ziyue Jiang、Feiyang Chen、Xinyu Duan 等，“TexTrolSpeech：采用编解码器语言文本转语音模型的文本风格控制语音语料库”，载于*2024*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2024年，第10301–10305页。[19] Zeyu Jin、Jia Jia、Qixin Wang、Kehan Li、Shuoyi Zhou、Songtao Zhou 等，“SpeechCraft：带有自然语言描述的细粒度富有表现力语音数据集”，载于第*32*届*ACM*国际多媒体会议论文集，2024年10月，第1255–1264页。[20] Masaya Kawamura、Ryuichi Yamamoto、Yuma Shirahata、Takuya Hasumi 和 Kentaro Tachibana，“LibriTTS-p：包含说话风格和说话人身份提示的文本转语音与风格描述语料库”，2024年。