arXiv:2601.02753v1 [eess.AS] 6 Jan 2026

VCLIP：基于人脸的　　通过人脸-语音关联学习实现说话人生成

石耀[1], 徐云飞[1], 索宏斌[1], 万玉龙[1], 刘海峰[2]

[1]数据与人工智能工程系统，OPPO，中国北京[2]中国科学技术大学，中国合肥 {shiyao1,xuyunfei,suohongbin,wanyulong,blade}@oppo.com

摘要

本文探讨基于人脸的语音合成任务。这是一种个性化语音合成，其合成语音在感知上受到约束，需与参考人脸图像相匹配。由于缺乏 TTS 质量的视听语料库，以往方法要么合成质量较低，要么因知识迁移方案而产生领域不匹配。本文提出一种名为 Vclip 的新方法，利用 CLIP 编码器在噪声视听数据中蕴含的面部语义知识，高效学习人脸与语音之间的关联，并在 Voxceleb 测试集上实现了 89.63% 的跨模态验证 AUC 分数。随后，所提方法采用基于检索的策略，并结合基于 GMM 的说话人生成模块，为下游 TTS 系统根据参考图像生成可能的目标说话人。实验结果表明，所提 Vclip 系统结合检索步骤，能够弥合基于人脸的语音合成中人脸与语音特征之间的差距。此外，利用从下游 TTS 中提取的反馈信息，有助于合成与参考面孔高度匹配的语音。演示请访问 sos1sos2sixteen.github.io/vclip。

索引词— 语音合成、说话人生成、人脸-语音关联学习

1. 引言

当前语音合成研究不仅追求准确、自然的内容传递，也关注个性化说话人语音或风格控制等更高级的应用场景。文献中通常将语音个性化定义为语音克隆任务 [1]。但将合成语音以其他模态作为条件，例如类别标签 [2]、人脸图像 [3, 4] 或描述性文本 [5]，正逐渐受到业界关注，尤其是在基于文本的图像生成方法取得成功之后。在用于生成个性化合成语音的各种线索中，人脸图像模态尤其受到关注，因为在日常体验中，语音与视觉本来就会自然共现，而且这类成对数据很容易从大量在线视频中获取 [6, 7, 8]。面部特征与语音特征之间的联系具有双重性。面部特征反映了性别、体重和年龄等基本生理状况，而这些因素会影响说话人发声器官对语音的生理生成 [9]。从心理学角度而言，人类会通过自身经历学会将某些人脸特征与特定的声音品质联系起来，这种现象被称为人脸-语音一致性 [10]。

语音合成中，人脸与语音条件控制的一大差异在于现有数据集的质量。合成任务需要高质量音频用于训练，但 LRS3 [6] 等主要视听数据集通常噪声较多，无法满足主流 TTS 的需求。由于这些限制，常见的基于人脸的语音合成方法采用迁移学习方案，尝试利用神经网络学习从人脸到语音特征的映射。转换后得到的图像特征在下游文本转语音推理过程中被用作说话人嵌入的替代项 [3, 11, 12] 。然而，基于域外 [13] 特征合成新说话人会跨越两个领域差距。而且，学习到的特征映射并未考虑预测基于人脸的说话人嵌入之后进一步累积的误差，导致上游映射模块与下游文本转语音模块之间缺乏协调。这些设计选择可能会导致文本转语音 [14] 的性能下降。此外，我们认为，通过对参考图像进行确定性映射来预测真实说话人嵌入本质上难以实现，因为仅凭人脸图像无法提供足够信息来还原一个人的声音。相反，可以说，单张参考人脸同样适合多种不同的声音。针对这一问题，VoiceMe [15] 提出采用人在回路采样策略，为一张人脸生成可能的说话人嵌入。该技术还考虑了下游文本转语音的端到端合成结果。然而，由于需要大量人工交互，其扩展性较差。

因此，我们将基于人脸的语音合成视为**条件式说话人生成任务 [2]，旨在为参考人脸生成可能的候选合成语音**。虽然 [15] 借助众包人工评估者将语音与人脸关联起来，但我们旨在从公开可用的配对音视数据中高效学习这种关联。我们提出*Vclip*，这是一种类 CLIP 对比学习 [16] 方案，利用低质量视听数据提炼并关联预训练 CLIP 图像编码器与说话人编码器的语义知识。随后，我们提出一种基于检索的策略，用于基于高斯混合模型的说话人生成模型 [17]，以便根据参考图像生成可能的说话人嵌入。该过程将 Vclip 中的跨模态关联知识迁移到一个使用高质量数据训练的现有零样本文本转语音系统中。此外，该方法还可以直接扩展，以考虑特定下游 TTS 系统造成的误差。实验结果表明，在 Vclip 中，检索步骤对于弥合对比学习方案 [13] 导致的人脸特征与语音特征之间的领域差距至关重要。并且，从下游文本转语音模块提炼的反馈信息有助于合成与参考人脸更匹配的新颖语音。

所提出的 Vclip 系统在 Voxceleb 测试集上进行跨模态验证，取得了 89.63% 的 AUC 分数 [18], , 超过了强基线方法 [19]。以该系统为指导，我们证明所提出的策略能够合成与参考人脸高度相关的声音，这一点得到了感知评估和自动评估研究的证实。

2. 背景

人脸—声音关联。人脸与语音为一系列个人属性提供一致且冗余的信息，例如男性化程度、年龄、体重等。这证明了在现实世界中，由人类评估者进行静态

面孔—声音匹配的可行性，如 [10] 所示。在机器学习领域，Learnable PINs [20] 提出了一项跨模态特征学习任务，将人脸图像与语音片段关联起来，这项任务称为人脸—声音关联（FVA）学习。近年来，这项任务引起了研究人员的关注 [21, 18], ，其中 Self-Lifting [19] 是最先进的方法之一。Self-Lifting 采用迭代学习过程，从未标注的共现视听数据中捕捉人脸与语音之间的关系，并在 VoxCeleb [18] 上取得了最先进的结果。

基于人脸的语音合成。 与传统的面部-声音关联学习不同，传统方法将问题表述为验证或检索任务，而 Speech2face [22]则将问题扩展到生成领域，旨在根据语音特征生成面部图像。要将这一过程反向用于文本转语音，主要障碍在于缺少用于训练的高质量音视频数据。因此，Face2Speech [3] 通过在噪声数据上使用 SGE2E损失，将人脸编码器的嵌入空间与现有的说话人编码器耦合起来。他们建议在合成过程中使用推断得到的面部特征替代说话人嵌入，从而有效地将从噪声数据中学到的配对知识迁移到文本转语音。后续研究 [12, 11] 沿用了这一路线，提出了不同的耦合模块设计。相比之下，[4] 使用成对的语音和人脸输入来训练文本转语音模型，所用数据集 LRS3 [6], 是一个用于唇读的音视频数据集。但这种设置的合成质量受语料库质量低劣的严重限制。鉴于目前没有公开可用的高质量大规模音视频语料库，我们认为，基于迁移的方案仍然是基于人脸的文本转语音所必需的。

将 CLIP 扩展到音频领域。 CLIP 是一种在互联网规模数据上预训练的语言-视觉匹配模型，在判别式和生成式领域均有广泛的视觉相关应用 [16, 23]。文献中已有将 CLIP 扩展至音频等其他模态的研究。AudioCLIP [24] 从视频中共现的视听数据进行学习，生成一个适用于音视频事件定位的三模态嵌入空间。Wav2CLIP [25] 采用 LiT 式方案 [26] ，以 CLIP 图像编码器为参照训练音频编码器，用于音频事件分类。然而，据我们所知，尚无先前研究将 CLIP特征与语音中的说话人特征联系起来。

3. 方法

我们提出了一种由两部分组成的基于人脸的说话人生成方法。一个该方法的概览如图 1 所示。第一部分提出了一种名为 Vclip 的匹配系统，用于从在线语音视频中的大量视听数据中学习面部-声音关联知识。第二部分提出了一种针对基于 GMM 的说话人生成模块所生成的新说话人的检索流程。该方法同时考虑经训练的 Vclip 模块提供的关联知识，以及下游文本转语音模块所造成的误差，为参考人脸生成一组高质量的目标说话人嵌入。

3.1. 用于人脸—声音关联的 Vclip

与 Wav2CLIP 类似， [25], Vclip 由两个独立模块组成，分别用于人脸（记为 i）和语音（记为 v）模态。每个模块都包含一个针对输入的预训练且冻结的特征提取网络 $f_{\cdot}(\cdot)$，以及一个投影网络 $\phi_{\cdot}(\cdot)$，用于从相应特征中进一步提取一致信息。我们提出使用 CLIP图像编码器作为人脸编码器 f_{i}，因为众所周知，CLIP 能够捕捉有意义的面部属性 [23] ，并且在人脸识别方面表现出色 [27]。我们使用一个简单的多层感知机作为面部特征投影 ϕ_{i}。至于语音编码器 f_{v}，则采用预训练的说话人验证网络。我们使用流模块 [28]作为语音投影 ϕ_{v} ，将预训练说话人表征的分布转换为投影后面部特征的分布。

给定一个包含 N 对面部-声音样本（$\mathbf{X}_{\mathrm{i}}$，$\mathbf{X}_{\mathrm{v}}$）的小批次，Vclip 提取的跨模态表征为 $\mathbf{E}_{\mathrm{i}}=\phi_{\mathrm{i}}\circ f_{\mathrm{i}}(\mathbf{X}_{\mathrm{i}})$

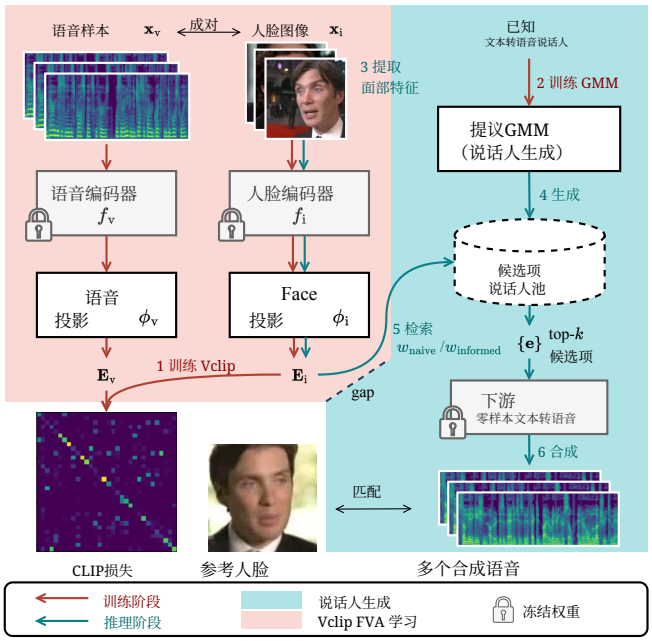

图 1。所提方法概述。首先训练 Vclip 和说话人生成模型（红色，步骤 1–2）。随后，使用训练好的 Vclip 作为零样本评分函数，从生成的候选语音池中检索与参考人脸图像提取的图像特征最匹配的前 k 个可能语音（蓝色，步骤 3–6）。

和$\mathbf{E}_{\mathrm{v}}=\phi_{\mathrm{v}}\circ f_{\mathrm{v}}(\mathbf{X}_{\mathrm{v}})$。这里， $\circ$ 表示函数复合。我们使用式 1 所示的 CLIP 对比损失来训练 Vclip，其中 $c=\|\mathbf{E}_{\mathrm{v}}\|_2\|\mathbf{E}_{\mathrm{i}}\|_2$ 表示余弦相似度归一化因子，tr(·)表示矩阵迹， $\sigma_{\max}$ 表示 Softmax函数。该公式只需要共同出现的样本对，而不需要说话人身份标签，这与以往面部-声音关联学习工作中的无监督设置一致 [19]。

$$\mathcal{L}=\frac{1}{2N}\left(\mathrm{tr}\left(\log\sigma_{\max}\left(\frac{\mathbf{E}_{\mathrm{v}}^{T}\mathbf{E}_{\mathrm{i}}}{c}\right)\right)+\mathrm{tr}\left(\log\sigma_{\max}\left(\frac{\mathbf{E}_{\mathrm{i}}^{T}\mathbf{E}_{\mathrm{v}}}{c}\right)\right)\right) \tag{1}$$

CLIP损失促使正样本对的投影特征在余弦距离上彼此相似，负样本对则相反。因此，学习到的投影会被引导去提取人脸和语音输入中相互一致的信息。但相似并不意味着可以互换。研究表明， [13] 多模态模型中存在模态差距：来自不同源模态的特征总是被嵌入到不同的子空间中。因此，我们推测，将人脸嵌入直接替代说话人嵌入，并不能为基于人脸的文本转语音带来最优结果。

3.2. 使用 Vclip 进行新颖的说话人生成

训练好的 Vclip 模型包含将语音与人脸关联起来所需的信息。但由于存在模态差距，Vclip 提取的人脸嵌入对于文本转语音而言属于分布外说话人嵌入，这可能会降低合成质量 [14]。为了解决这一问题，我们提出一种生成并检索策略，在说话人嵌入域内生成与参考人脸高度相关的样本。

首先，按照 [17] 的方法，使用高斯混合模型（GMM）对下游零样本文本转语音嵌入空间中已知说话人的说话人嵌入无条件分布 $p($**e**$)$ 进行建模。为了获得给定参考人脸

图像$\mathbf{x}_i$对应的可能说话人嵌入样本，我们针对从 $p(\mathbf{e})$ 中抽取的 N 个候选样本 $\{\mathbf{e}\}$，使用 Vclip 作为零样本评分函数 $w(\cdot;\mathbf{x}_i)$，以 $\mathbf{x}_i$ 为参照。选取得分最高的 k 个样本作为生成说话人。

朴素评分。¨假设 **Vclip 与下游文本转语音模型共享说话人表示（使用文本转语音说话人编码器作为），一种朴素方法是将候选样本** $\mathbf{e}\}$ 的 Vclip 投影结果 $\phi_v(\mathbf{e})$ 作为检索索引进行评分，如公式 2 所示。

$$w_{\text{naive}}(\mathbf{e};\mathbf{x}_i) := \cos(\phi_v(\mathbf{e}), \phi_i \circ f_i(\mathbf{x}_i)) \tag{2}$$

基于文本转语音特征签名的评分。 值得注意的是，朴素方法忽略了零样本文本转语音系统 [29]输入与输出之间的说话人不匹配。也就是说，$f_v \circ \text{tts}(\mathbf{e}) \neq \mathbf{e}$。此处，$\text{tts}(\cdot)$ 表示下游文本转语音系统，其输入为目标说话人嵌入；为简洁起见，此处忽略文本输入。实践中，我们可以对多个句子取平均。我们可以考虑这一偏差，并在评分过程中通过穷举，将 $f_v \circ \text{tts}(\mathbf{e})$ 作为公式 2 中$\mathbf{e}$ 的替代项。但在采样期间对每个候选$\mathbf{e}$ 执行合成的成本过高。注意到，对于给定的 Vclip 模型，组合 $f_v \circ \text{tts}(\cdot)$ 仅由文本转语音系统决定。该映射体现了文本转语音系统输入输出中的说话人不匹配现象。它捕捉了给定输入嵌入时不可避免的声音克隆错误，而这些错误是相关文本转语音系统固有的。事实上，完美的零样本语音克隆应在任何位置都产生恒等映射。鉴于此，我们选择将这一特征签名信息 的近似版本蒸馏到一个简单的前馈网络 $\text{s}(\cdot)$ 中，并使用余弦嵌入损失 $\mathcal{L}_{\text{signature}} = \sum_{\mathbf{e}\in\mathcal{E}} \cos(\text{s}(\mathbf{e}),\ f_v \circ \text{tts}(\mathbf{e}))$。随后，我们在公式 3 中定义基于文本转语音特征签名的评分函数 $w_{\text{informed}}(\cdot)$。

$$w_{\text{informed}}(\mathbf{e};\mathbf{x}_i) := \cos(\phi_v \circ \text{s}(\mathbf{e}), \phi_i \circ f_i(\mathbf{x}_i)) \tag{3}$$

3.3. 生成语音的自动评估

我们使用一个**独立的** Vclip 模型作为自动评估器，用于评估生成语音与其参考面孔之间的面孔—声音匹配质量。给定一对$\mathbf{x}_i, \mathbf{x}_v$人脸—语音样本，我们的方法会根据参考人脸$\mathbf{x}_i$生成一组$k$ 可能的说话人嵌入 $\mathcal{E}$。

语音重建。我们将 v2v（式 4）定义为生成语音与参考人脸对应的真实语音之间的余弦相似度。该指标衡量仅凭人脸图像重建其原始语音的能力。

$$\text{v2v} := \frac{1}{k} \sum_{\mathbf{e}\in\mathcal{E}} \cos(f_v \circ \text{tts}(\mathbf{e}), f_v(\mathbf{x}_v)) \tag{4}$$

面孔—声音匹配。我们将 f2v（式 5）定义为生成语音与其参考人脸之间的余弦相似度。该指标反映了根据评估器学习到的关联关系，$\mathcal{E}$ 中提出的说话人与输入人脸的匹配程度。

$$\text{f2v} := \frac{1}{k} \sum_{\mathbf{e}\in\mathcal{E}} \cos(\phi_v \circ f_v \circ \text{tts}(\mathbf{e}), \phi_i \circ f_i(\mathbf{x}_i)) \tag{5}$$

说话人生成质量。理想情况下，说话人生成系统 [2], 所生成说话人的统计特征应与真实说话人相同。因此，我们用一个包含 4 个分量的 GMM 对已知文本转语音说话人嵌入的统计特征进行建模。从合成话语中提取的说话人嵌入相对于该参考分布的对数似然，可作为衡量文本转语音领域失配程度的指标。

4. 实验

4.1. 实验设置

数据集。 对于FVAlearning，除非另有说明，我们使用来自 Voxceleb2-dev [8] 的视频作为所提出方法的训练数据和开发数据

表 1：Vox1-test 人脸-声音 AUC 结果

系统	曲线下面积(%) ↑ 封闭式 数据集	开放数据集
Learnable PINs [20]	81.88	84.70
Self-Lifting [19]	86.70	89.40
Vclip	**88.18**	**89.63**

表 2。Vclip 上的消融研究。

系统	曲线下面积(%) ↑
提出的方法	**88.78**
bcsz = 320	86.54
−CLIP损失	87.14
$-\phi_v(\cdot)$ 流	88.16

Vclip。从训练集中保留500名随机说话人，并将其留作开发集。对于每个视频片段，我们提取第一帧图像和完整音轨，作为成对的人脸-语音数据。测试时使用来自 [19]的Voxceleb1 [7]预处理版本。请注意，[19] 中使用的完全相同的测试试验也用于比较各FVA方法的性能。对于零样本文本转语音，我们使用LibriTTS-R [30], 中的全部2130名训练说话人进行训练；LibriTTS-R是LibriTTS的高质量修复版本。

实现。 对于面部-声音关联学习，我们使用 OpenAI 发布的 ViT-B/32 CLIP 作为预训练图像编码器。此外，我们使用在 VoxCeleb2 上训练的基于H/ASP的说话人验证模型 [31] 作为预训练语音编码器。两个编码器共用一个 512 维的嵌入空间。我们使用 ReLU激活的多层感知机 [25] 作为图像投影层（ϕ_i），并使用 RealNVP [28] 流作为语音投影层（ϕ_v）。所有 Vclip 模型均在单张 NVIDIA V100 GPU 上训练，批大小为 1024，因为我们观察到使用大型小批量训练时性能会提升 [26]。模型持续训练，直到开发集的 AUC分数开始下降。对于零样本文本转语音，本章所有实验均训练一个标准的多说话人VITS [32] 。我们遵循 [17] ，在主成分上实现一个包含 100 个分量的对角协方差高斯混合模型，该主成分保留已知文本转语音说话人嵌入数据方差的前 99%，并将其用作说话人生成系统。

4.2. Vclip 的 FVA 性能

我们使用跨模态验证中的 ROC 曲线下面积（AUC）来评估 Vclip 模型的 FVA 性能。表 1 报告了 Vox1 测试集的 AUC分数。在封闭 设置下，仅使用 Vox1 的训练分区。在开放 设置下，Pins 和 Self-Lifting 使用 AVSpeech [33] 进行预训练，并在 Vox1 上微调。由于我们无法直接使用 AVSpeech，Vclip 使用 Vox2 进行预训练；Vox2 的规模大致相同，但身份数量较少。评估结果表明，所提出的 Vclip 优于 Self-Lifting，并且在两种数据设置下均取得了最佳性能。与 [27], 一致，我们的初步实验表明，基线中使用的监督式人脸识别模型在人脸识别任务上的表现优于 CLIP。我们认为，面部-声音关联学习中观察到的性能提升归功于语义信息丰富的 CLIP特征，这表明它能提供与语音特征更加一致的信息。

我们进一步研究了 Vclip 中一些设计选择的影响。表 2 报告了最终版 Vclip 和三个变体的 AUC分数。最显著的偏差出现在批大小缩减至 320 时，这表明 CLIP损失需要较大的小批量才能提升性能 [26]。相反，即使批大小较大，用 SGE2E损失 [3] 替换 CLIP损失也会对系统产生负面影响。这些证据支持了我们为所提出的 Vclip 模型所做设计选择的合理性。

表 3。生成语音的自动评估结果。

系统	v2v（±标准差）	f2v（±标准差）↑	似然↑
参考值	.588 (±.08)	.307 (±.13)	557.40 (±22.42)
不使用检索	.131 (±.09)	.021 (±.21)	360.21（±16.33）
基线	.229（±.09）	.305（±.13）	328.82（±37.41）
含 w_{naive}	.249 (±.10)	.301 (±.13)	**518.78 (±13.84)**
带 w_{informed}	**.254 (±.10)**	**.313 (±.13)**	516.22 (±15.31)

4.3. 基于人脸的说话人生成的自动评估

我们使用第 3.3 节所述的指标，对所提出的基于人脸的文本转语音的生成质量进行自动评估。我们使用一个在 Vox1 上单独训练的 Vclip 来计算分数。

基线。 鉴于基于人脸的文本转语音的相关研究采用了截然不同的设置，因此无法对准确复现的实现进行公平比较。遵循 [3], 的主要思路，我们实现了一个采用 SGE2 2、3 E 损失的 监督式 Vclip 变体（表格 rd 项），作为强有力的特征映射基线。在生成过程中，投影后的人脸嵌入 $\phi_i \circ f_i(\mathbf{x}_i)$ 直接用作生成的说话人嵌入（$k=1$）。与 [3], 相比，该基线得益于 CLIP特征的使用以及更强大的零样本文本转语音系统。

设置。 我们从 Vox1 试验中抽取 $M = 500$ 个正脸语音对，并评估每个系统的 v2v、f2v 平均值（在 k, M 上），以及生成说话人的对数似然。检索时使用的超参数为 $k=10$ 和 $N=5000$。为便于理解，我们提供基于同一测试集计算的各项指标参考值。v2v 的参考值是文本到语音系统在零样本语音克隆任务上的余弦相似度；f2v 的参考值是真实脸语音对的余弦相似度。似然值的参考值则是已知文本转语音说话人的似然值。

结果。 结果如表 3 所示。我们注意到，与同一文本到语音系统上的真实语音克隆结果相比，所有 v2v 值都较低。这表明，无论如何，生成语音都与对人脸背后真实语音的预测相去甚远。然而，我们的系统生成的 f2v 分数接近甚至超过了参考值，这意味着生成的参考语音匹配具有实际意义。采用朴素评分（w/w_{naive}）的生成与检索流程，在 f2v 分数方面表现与我们的基线相当，但其似然值接近参考值。这证明，基于说话人生成的策略能够弥合基线映射范式所造成的领域差距，从而应用于基于人脸的文本转语音。此外，将下游文本到语音系统的知识与我们提出的基于文本转语音特征签名的评分结合到流程中（w/ w_{informed}），可以获得更高的 f2v 分数。这表明，从下游文本转语音模块中提炼出的反馈信息能够补充我们的方法，并帮助检索出与参考面孔更匹配的语音。

4.4 基于人脸的说话人生成主观评估

我们对生成语音的自然度 和匹配度 进行了主观评估。表 4 和表 5 中的每项至少由 12 名评估者评估 15 个样本。

自然度。 我们采用 1 至 5 分的量表，衡量不同设置下文本转语音的自然度。我们还纳入了直接使用 LRS3 训练的 Face-TTS [4] 样本，以比较质量。评估结果见表 4。我们发现，在迁移学习设置下，其表现与普通语音克隆没有显著差异。但 Face-TTS 的表现明显较差，这表明基于人脸的文本转语音仍然高度依赖高质量数据，才能生成听感悦耳的语音。

匹配。 FVA 匹配的评估人员需要为匹配 MOS（m-MOS）打分，以评估成对

表 4。主观自然度结果。

task	设置	平均意见分（±95%CI）↑
语音克隆	已知	4.01 (±0.12)
	零样本	4.15 (±0.11)
基于人脸	Face-TTS [4]	2.46 (±0.19)
	ours	4.14 (±0.10)

表 5。主观人脸/语音匹配结果。

系统	平均主观意见分 (±95%CI)↑ 匹配的 配对	随机 配对
真实值	3.46（±0.10）	1.96（±0.16）
语音克隆（预言机）	3.33 (±0.10)	2.31 (±0.17)
ours	3.26 (±0.11)	

面部和语音刺激之间的匹配程度。我们采用 1–4 分制，范围从“不匹配”到“匹配良好”，与 [3] 中使用的量表方向相反。结果见表 5。为便于解读，我们构建了多个参考系统。真实值 条目的音频刺激选自 Vox1 中的录音。语音克隆 包含由真实匹配音频克隆出的零样本语音与人脸图像配对。最后，我们的方法是一个采用 w_{informed} 系统的 Vclip，如第 4.3 节所述。我们观察到，$matched$ 配对的 m-MOS 始终高于随机配对基线，且所提方法取得了与具有预言机信息的语音克隆基线相当的结果。这表明，Vclip 捕捉到了人脸图像与生成的说话人嵌入之间真正具有感知意义的面部-声音关联知识。

5. 讨论

与文本模态的关系。 由于 Vclip 将 CLIP 扩展到了语音领域，因此将人物外貌描述提示迁移至合成语音在概念上是可行的。初步实验表明，基于说话人提示生成的语音仅对反映目标语音性别的简单概念敏感。我们发现，未受训练时约束的文本条件缺乏可控性，仍需进一步研究。

属性不匹配。 尽管所提出的策略有效缩小了生成的说话人嵌入的领域差距，但属性层面的不匹配始终存在，因为在野外采集的视听语料库包含语言、族裔和环境因素方面的差异，而这些差异在朗读式文本转语音数据中并不存在。因此，尽管所提出的 Vclip 捕捉到了这些因素的多样变化（见我们的在线演示），以说话人嵌入为中心的文本转语音范式只能捕捉到与 LibriTTS 相符的属性。

6. 结论

我们将问题建模为条件说话人生成任务，并提出了一种由两部分组成的基于人脸的文本转语音方法。第一部分提出了一种名为 Vclip 的面部-语音关联方法。Vclip 利用 CLIP 中语义丰富的人脸表征和对比学习方案，在公开的面部-语音关联学习基准上取得了 89.63% 的 AUC 分数。第二部分将 Vclip 的面部-声音关联知识迁移到基于高斯混合模型的说话人生成模型中，为现有的零样本文本转语音系统生成概率较高的新型说话人嵌入。该条件说话人生成过程结合从下游文本到语音模型中提取的反馈信息后，取得了优于特征映射对照方法的匹配结果。未来，我们计划研究表现力更强的下游文本到语音方法，以缓解我们在传统多说话人文本转语音系统中遇到的属性不匹配问题。

7. 参考文献

[1] Xu Tan、Tao Qin、Frank Soong 和 Tie-Yan Liu，《神经语音合成综述》，*arXiv:2106.15561*，2021 年。[2] Daisy Stanton、Matt Shannon、Soroosh Mariooryad、RJ Skerry- Ryan、Eric Battenberg、Tom Bagby 和 David Kao，《说话人生成》，载于 *ICASSP 2022*，第 7897–7901 页。[3] Shunsuke Goto、Kotaro Onishi、Yuki Saito、Kentaro Tachibana 和 Koichiro Mori，《Face2Speech：利用从人脸图像预测的嵌入向量实现多说话人文本转语音合成》，载于*INTERSPEECH 2020*，第 1321–1325 页。[4] Jiyoung Lee、Joon Son Chung 和 Soo-Whan Chung，《Imaginary voice：用于文本转语音的人脸风格扩散模型》，载于 *ICASSP 2023*，第 1–5 页。[5] Dongchao Yang、Songxiang Liu、Rongjie Huang、Guangzhi Lei、Chao Weng、Helen Meng 和 Dong Yu，《Instructtts：在离散潜在空间中借助自然语言风格提示建模富表现力的文本转语音》，*arXiv:2301.13662*，2023 年。[6] Triantafyllos Afouras、Joon Son Chung 和 Andrew Zisserman，《Lrs3-ted：用于视觉语音识别的大规模数据集》，*arXiv:1809.00496*，2018 年。[7]Arsha Nagrani、Joon Son Chung、Weidi Xie 和 Andrew Zisserman，《Voxceleb：野外大规模说话人验证》，*Comput. Speech Lang.*，第 60 卷，2020 年。

[8]Joon Son Chung、Arsha Nagrani 和 Andrew Zisserman，《Voxceleb2：深度说话人识别》，载于 *INTERSPEECH 2018*，B. Yegnanarayana 编，第 1086–1090 页。

[9] Adam C Lammert 和 Shrikanth S Narayanan，《根据共振峰频率短时估计声道长度》，*PloS one*，第 10 卷，第 7 期，文章编号 e0132193，2015 年。[10] Harriet MJ Smith、Andrew K Dunn、Thom Baguley 和 Paula C Stacey，《面孔与声音中的一致性线索：检验备份信号假说》，*Evolutionary Psychology*，第 14 卷，第 1 期，文章编号 1474704916630317，2016 年。[11] Jianrong Wang、Zixuan Wang、Xiaosheng Hu、Xuewei Li、Qiang Fang 和 Li Liu，《基于人脸图像的残差引导个性化语音合成》，载于*ICASSP 2022*，第 4743 页。[12] Hsiao-Han Lu、Shao-En Weng、Ya-Fan Yen、Hong-Han Shuai 和 Wen-Huang Cheng，《基于人脸的语音转换：学习面孔背后的声音》，载于*ACMMM 2021*，第 496–505 页。[13] Victor Weixin Liang、Yuhui Zhang、Yongchan Kwon、Serena Yeung 和 James Y Zou，《关注差距：理解多模态对比表示学习中的模态差距》，*Advances in NIPS 2022*，第 35 卷，第 17612–17625 页。[14] Wenbin Wang、Yang Song 和 Sanjay Jha，《基于解耦表示的泛化零样本说话人自适应语音合成》，载于 *INTERSPEECH 2023*，第 4454 页。[15] Pol van Rijn、Silvan Mertes、Dominik Schiller、Piotr Dura、Hubert Siuzdak、Peter M. C. Harrison、Elisabeth Andr´e 和 Nori Jacoby，《VoiceMe：文本转语音中的个性化语音生成》，载于*INTERSPEECH 2022*，第 2588–2592 页。[16] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等，《通过自然语言监督学习可迁移的视觉模型》，载于 *ICML 2021*，第 8748–8763 页。

[17] Henry Turner、Giulio Lovisotto 和 Ivan Martinovic，《使用混合模型生成身份以实现说话人匿名化》，*Computer Speech& Language*，第 72 卷，2022 年。[18] Chuyuan Xiong、Deyuan Zhang、Tao Liu 和 Xiaoyong Du，《语音-人脸跨模态匹配与检索：一项基准》，*arXiv:1911.09338*，2019 年。[19] Guangyu Chen、Deyuan Zhang、Tao Liu 和 Xiaoyong Du，《Self-Lifting：一种用于无监督语音-人脸关联学习的新型框架》，载于*ICMR 2022*，第 527–535 页。[20] Arsha Nagrani、Samuel Albanie 和 Andrew Zisserman，《可学习的 Pins：用于人物身份的跨模态嵌入》，载于 *ECCV 2018*，第 71–88 页。[21] Peisong Wen、Qianqian Xu、Yangbangyan Jiang、Zhiyong Yang、Yuan He 和 Qingming Huang，《探寻声音的形状：一种学习语音-人脸关联的自适应框架》，载于*CVPR 2021*，第 16347–16356 页。[22]Tae-Hyun Oh、Tali Dekel、Changil Kim、Inbar Mosseri、William T Freeman、Michael Rubinstein 和 Wojciech Matusik，《Speech2Face：学习声音背后的面孔》，载于*CVPR 2019*，第 7539–7548 页。[23] Rinon Gal、Or Patashnik、Haggai Maron、Amit H Bermano、Gal Chechik 和 Daniel Cohen-Or，《StyleGAN-NADA：CLIP引导的图像生成器域适应》，*ACM Transactions on Graphics (TOG)*，第 41 卷，第 4 期，第 1–13 页，2022 年。[24] Andrey Guzhov、Federico Raue、J¨orn Hees 和 Andreas Dengel，《AudioCLIP：将 CLIP 扩展至图像、文本和音频》，载于*ICASSP 2022*，第 976–980 页。[25] Ho-Hsiang Wu、Prem Seetharaman、Kundan Kumar 和 Juan Pablo Bello，《Wav2CLIP：从 CLIP 学习鲁棒的音频表征》，载于 *ICASSP 2022*，第 4563–4567 页。[26] Xiaohua Zhai、Xiao Wang、Basil Mustafa、Andreas Steiner、Daniel Keysers、Alexander Kolesnikov 和 Lucas Beyer，《LiT：通过锁定图像文本调优实现零样本迁移》，载于*CVPR 2022*，第 18123–18133 页。[27] Aaditya Bhat 和 Shrey Jain，《CLIP 与十亿级图像数据集时代的人脸识别》，*arXiv:2301.07315*，2023 年。[28] Laurent Dinh、Jascha Sohl-Dickstein 和 Samy Bengio，《使用 Real NVP 进行密度估计》，载于*ICLR 2017*。[29] Edresson Casanova、Julian Weber、Christopher D Shulby、Ar- naldo Candido Junior、Eren G¨olge 和 Moacir A Ponti，《YourTTS：面向所有人的零样本多说话人文本转语音与零样本语音转换》，载于*ICML 2022*，第 2709–2720 页。[30] Yuma Koizumi、Heiga Zen、Shigeki Karita、Yifan Ding、Kohei Yatabe、Nobuyuki Morioka、Michiel Bacchiani、Yu Zhang、Wei Han 和 Ankur Bapna，《LibriTTS-R：一个修复后的多说话人文本转语音语料库》，*arXiv:2305.18802*，2023 年。[31] Hee Soo Heo、Bong-Jin Lee、Jaesung Huh 和 Joon Son Chung，《用于 VoxCeleb 说话人识别挑战赛 2020 的 Clova 基线系统》，*arXiv:2009.14153*，2020 年。[32] Jaehyeon Kim、Jungil Kong 和 Juhee Son，《用于端到端文本转语音的对抗学习条件变分自编码器》，载于*ICML 2021*，第 5530–5540 页。[33] Ariel Ephrat、Inbar Mosseri、Oran Lang、Tali Dekel、Kevin Wilson、Avinatan Hassidim、William T Freeman 和 Michael Rubinstein，《看向鸡尾酒会，聆听说话者：一种用于语音分离的说话人无关音视频模型》，*arXiv:180403619*，2018 年。。